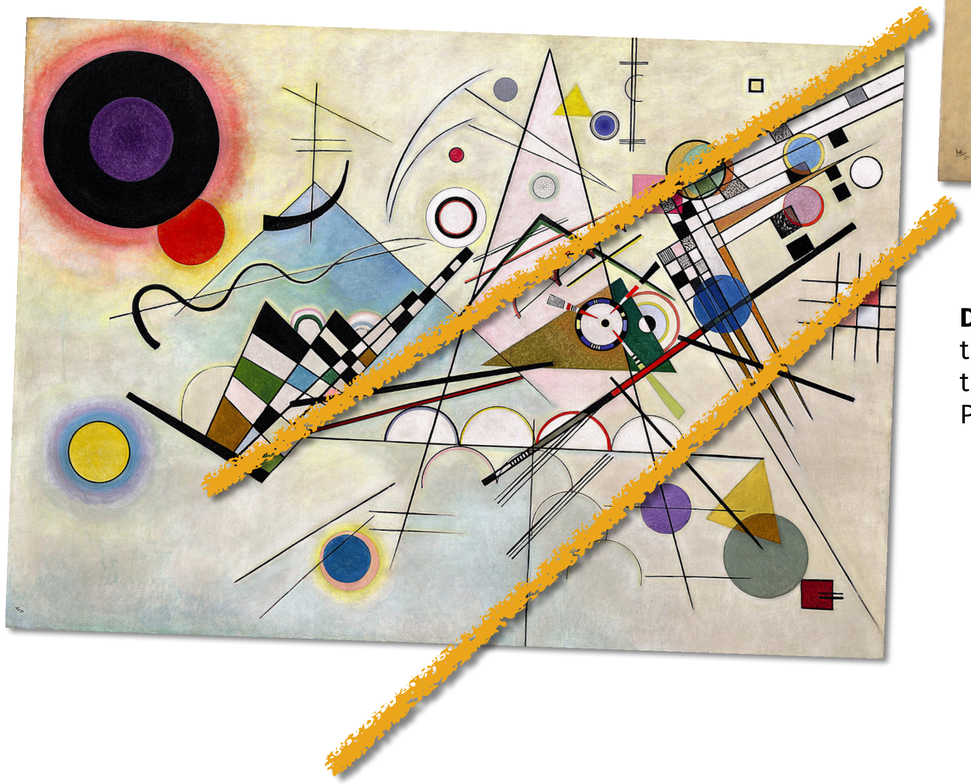
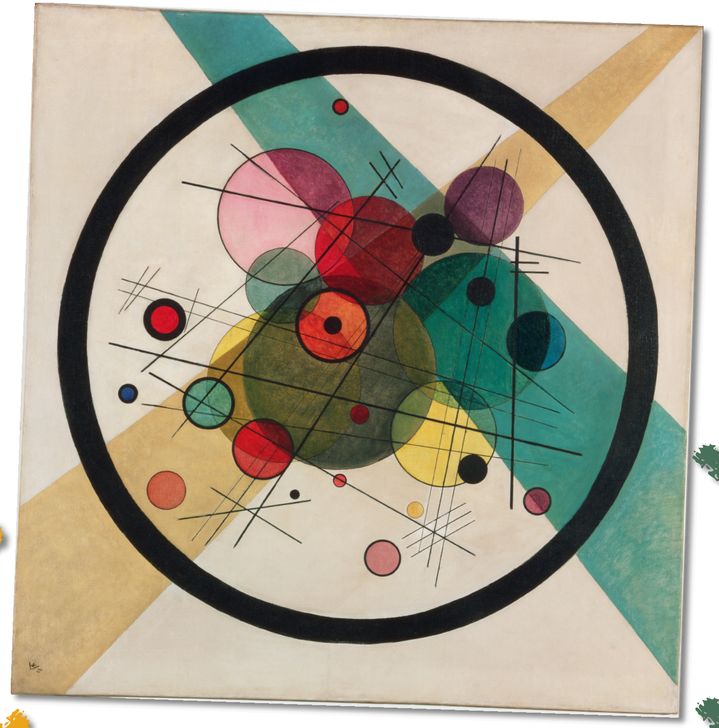
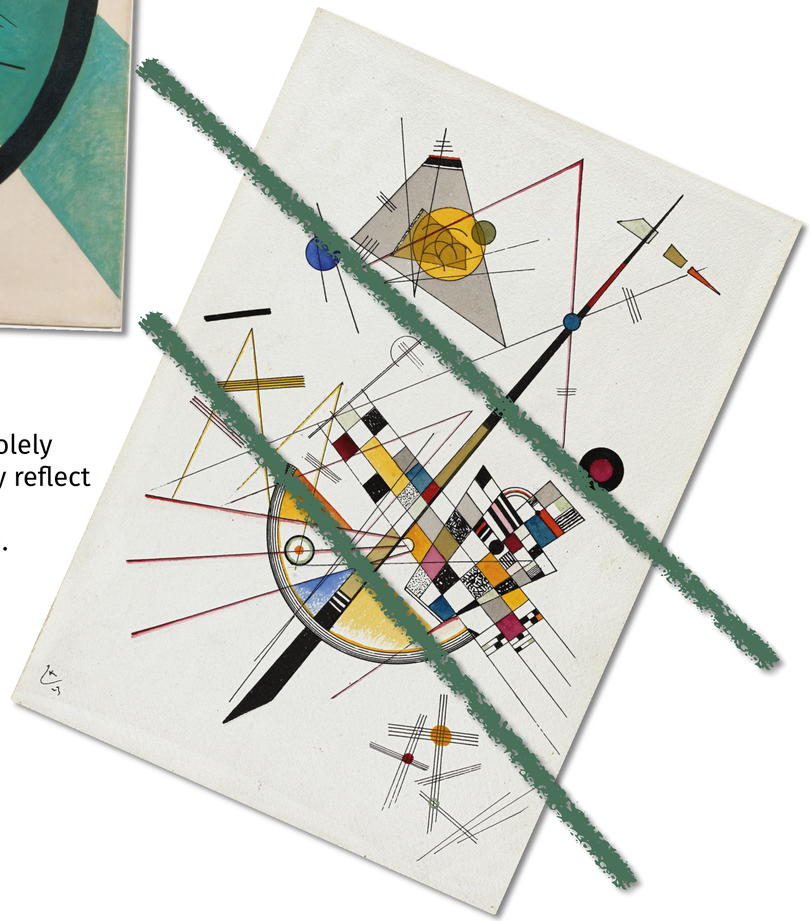
Diverse Table Scans to Cohesive Datasets with LLMs

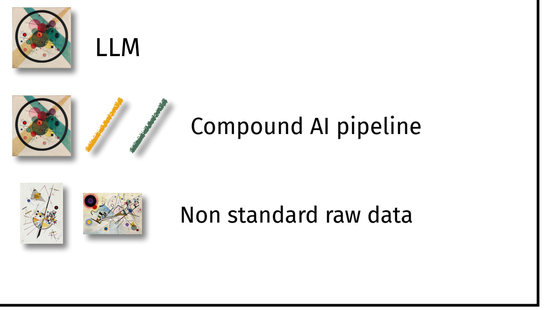
Vitaly Meursault (Philly Fed)
Chris Severen (Philly Fed)
Veronica Backer Peral (MIT)

From Rasters to Rows
Federal Reserve Bank of Philadelphia
August 16, 2024





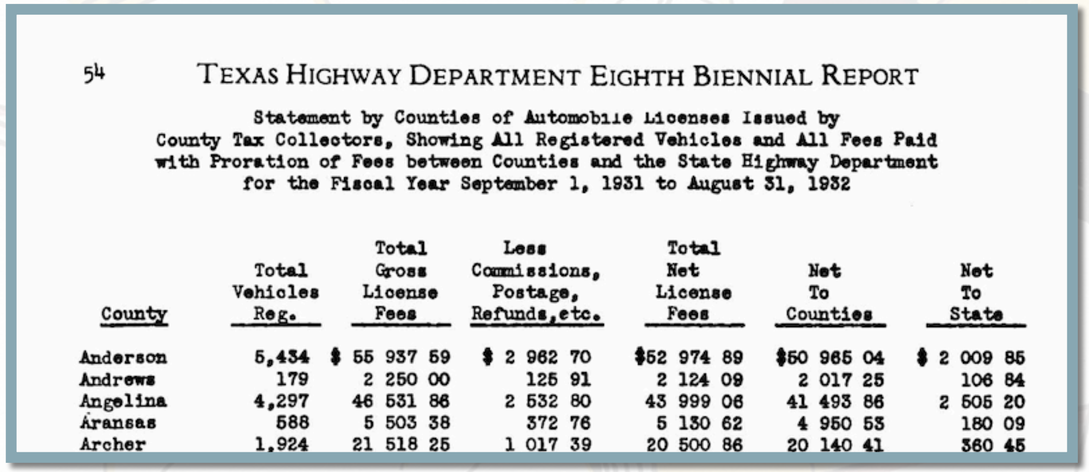




Disclaimer: The views expressed here are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

The data are scattered across diverse scans. What do you do?

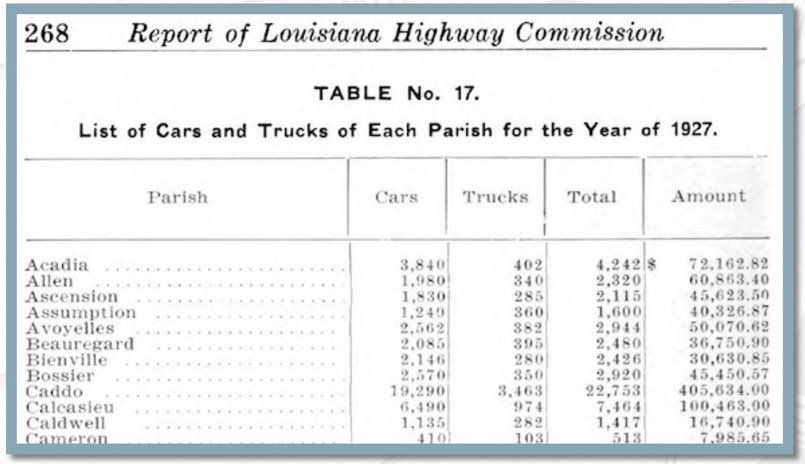
The data are scattered across diverse scans. What do you do?



54 TEXAS HIGHWAY DEPARTMENT EIGHTH BIENNIAL REPORT

Statement by Counties of Automobile Licenses Issued by County Tax Collectors, Showing All Registered Vehicles and All Fees Paid with Proration of Fees between Counties and the State Highway Department for the Fiscal Year September 1, 1931 to August 31, 1932

County	Total Vehicles Reg.	Total Gross License Fees	Less Commissions, Postage, Refunds, etc.	Total Net License Fees	Net To Counties	Net To State
Anderson	5,434	$ 55 937 59	$ 2 962 70	$52 974 89	$50 965 04	$ 2 009 85
Andrews	179	2 250 00	125 91	2 124 09	2 017 25	106 84
Angelina	4,297	46 531 86	2 532 80	43 999 06	41 493 86	2 505 20
Aransas	588	5 503 38	372 76	5 130 62	4 950 53	180 09
Archer	1,924	21 518 25	1 017 39	20 500 86	20 140 41	360 45

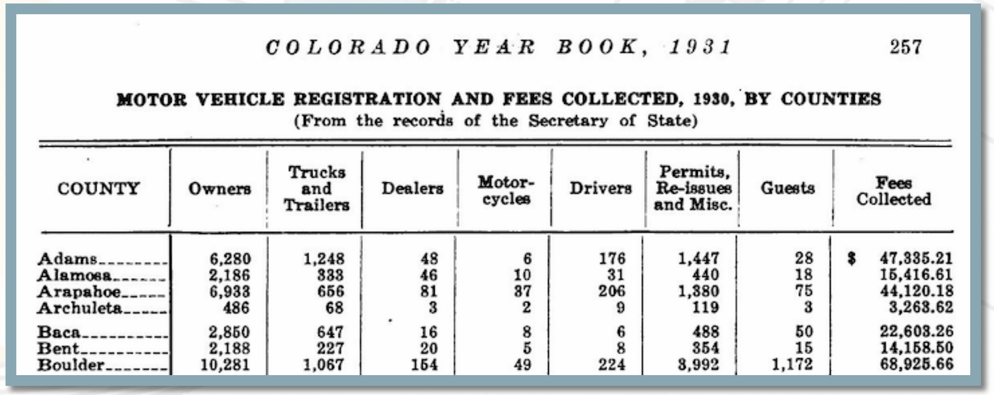


268 *Report of Louisiana Highway Commission*

TABLE No. 17.

List of Cars and Trucks of Each Parish for the Year of 1927.

Parish	Cars	Trucks	Total	Amount
Acadia	3,840	402	4,242	$ 72,162.82
Allen	1,980	340	2,320	60,863.40
Ascension	1,830	285	2,115	45,623.50
Assumption	1,240	360	1,600	40,326.87
Avoyelles	2,562	382	2,944	50,070.62
Beauregard	2,085	395	2,480	36,750.90
Bienville	2,146	280	2,426	30,630.85
Bossier	2,570	350	2,920	45,450.57
Caddo	19,290	3,463	22,753	405,634.00
Calcasieu	6,490	974	7,464	100,463.00
Caldwell	1,135	282	1,417	16,740.90
Cameron	410	103	513	7,985.65

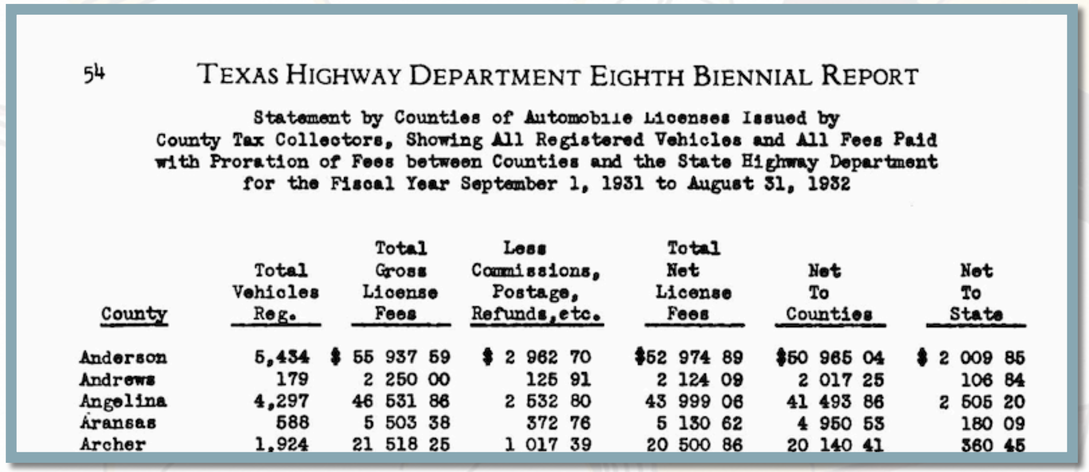


COLORADO YEAR BOOK, 1931 257

MOTOR VEHICLE REGISTRATION AND FEES COLLECTED, 1930, BY COUNTIES
(From the records of the Secretary of State)

COUNTY	Owners	Trucks and Trailers	Dealers	Motor-cycles	Drivers	Permits, Re-issues and Misc.	Guests	Fees Collected
Adams	6,280	1,248	48	6	176	1,447	28	$ 47,335.21
Alamosa	2,186	333	46	10	31	440	18	15,416.61
Arapahoe	6,933	656	81	37	206	1,380	75	44,120.18
Archuleta	486	68	3	2	9	119	3	3,263.62
Baca	2,850	647	16	8	6	488	50	22,603.26
Bent	2,188	227	20	5	8	354	15	14,158.50
Boulder	10,281	1,067	154	49	224	3,992	1,172	68,925.66

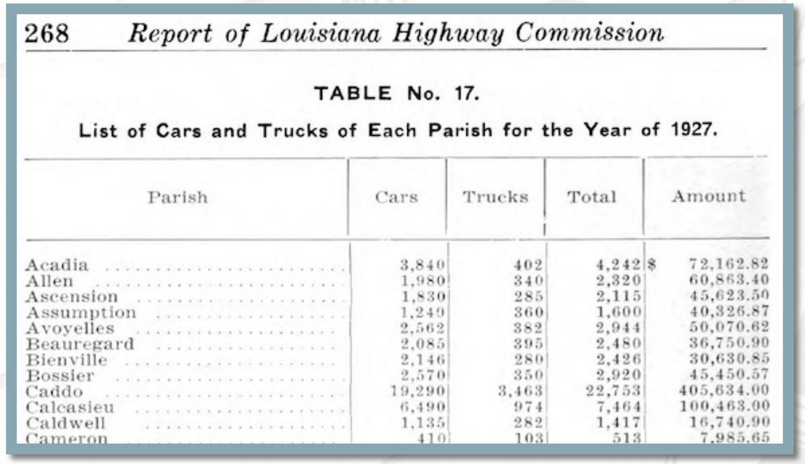
The data are scattered across diverse scans. What do you do?



54 TEXAS HIGHWAY DEPARTMENT EIGHTH BIENNIAL REPORT

Statement by Counties of Automobile Licenses Issued by County Tax Collectors, Showing All Registered Vehicles and All Fees Paid with Proration of Fees between Counties and the State Highway Department for the Fiscal Year September 1, 1931 to August 31, 1932

County	Total Vehicles Reg.	Total Gross License Fees	Less Commissions, Postage, Refunds, etc.	Total Net License Fees	Net To Counties	Net To State
Anderson	5,434	$ 55 937 59	$ 2 962 70	$52 974 89	$50 965 04	$ 2 009 85
Andrews	179	2 250 00	125 91	2 124 09	2 017 25	106 84
Angelina	4,297	46 531 86	2 532 80	43 999 06	41 493 86	2 505 20
Aransas	588	5 503 38	372 76	5 130 62	4 950 53	180 09
Archer	1,924	21 518 25	1 017 39	20 500 86	20 140 41	360 45

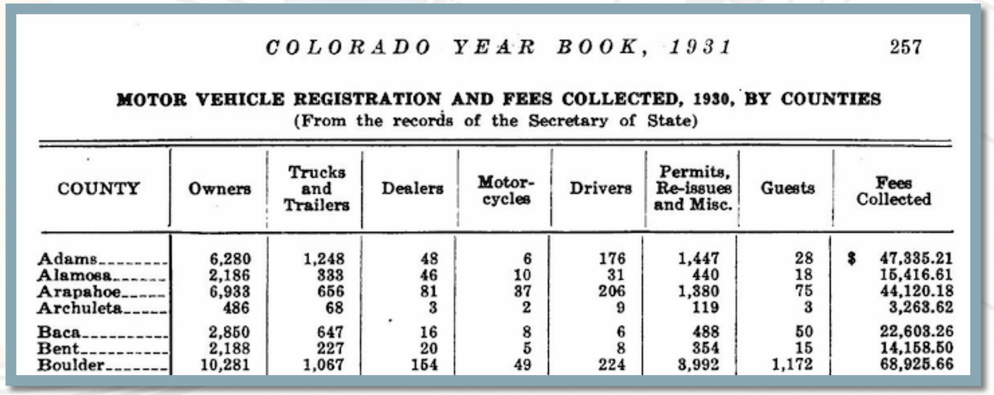


268 *Report of Louisiana Highway Commission*

TABLE No. 17.

List of Cars and Trucks of Each Parish for the Year of 1927.

Parish	Cars	Trucks	Total	Amount
Acadia	3,840	402	4,242	$ 72,162.82
Allen	1,980	340	2,320	60,863.40
Ascension	1,830	285	2,115	45,623.50
Assumption	1,240	360	1,600	40,326.87
Avoyelles	2,562	382	2,944	50,070.62
Beauregard	2,085	395	2,480	36,750.90
Bienville	2,146	280	2,426	30,630.85
Bossier	2,570	350	2,920	45,450.57
Caddo	19,290	3,463	22,753	405,634.00
Calcasieu	6,490	974	7,464	100,463.00
Caldwell	1,135	282	1,417	16,740.90
Cameron	410	103	513	7,985.65

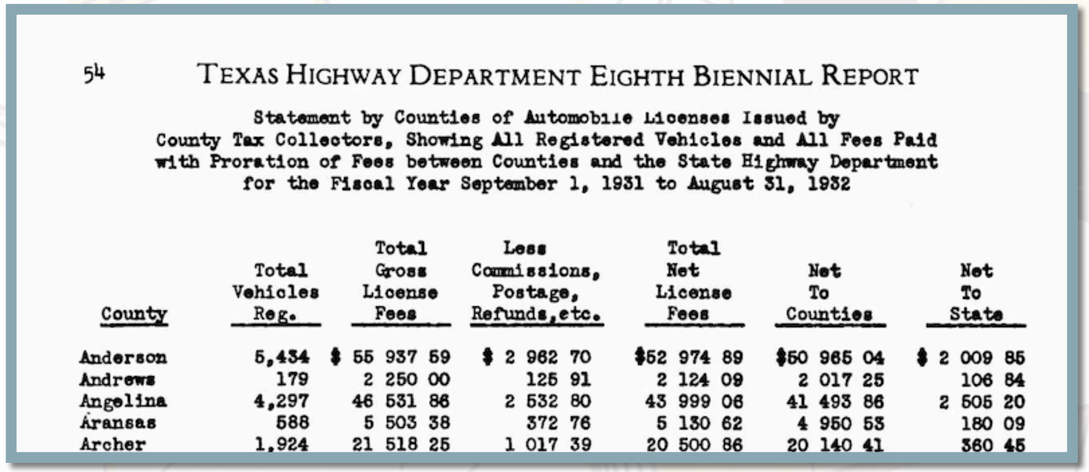


COLORADO YEAR BOOK, 1931 257

MOTOR VEHICLE REGISTRATION AND FEES COLLECTED, 1930, BY COUNTIES
(From the records of the Secretary of State)

COUNTY	Owners	Trucks and Trailers	Dealers	Motor-cycles	Drivers	Permits, Re-issues and Misc.	Guests	Fees Collected
Adams	6,280	1,248	48	6	176	1,447	28	$ 47,335.21
Alamosa	2,186	333	46	10	31	440	18	15,416.61
Arapahoe	6,933	656	81	37	206	1,380	75	44,120.18
Archuleta	486	68	3	2	9	119	3	3,263.62
Baca	2,850	647	16	8	6	488	50	22,603.26
Bent	2,188	227	20	5	8	354	15	14,158.50
Boulder	10,281	1,067	154	49	224	3,992	1,172	68,925.66

The data are scattered across diverse scans. What do you do?

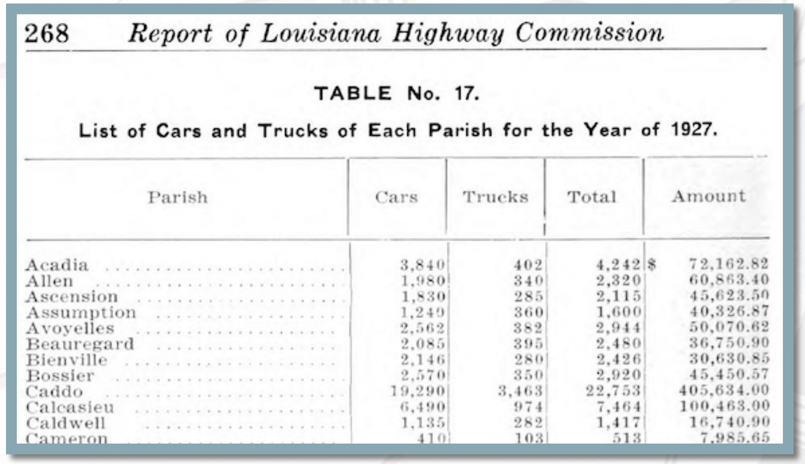


54 TEXAS HIGHWAY DEPARTMENT EIGHTH BIENNIAL REPORT

Statement by Counties of Automobile Licenses Issued by County Tax Collectors, Showing All Registered Vehicles and All Fees Paid with Proration of Fees between Counties and the State Highway Department for the Fiscal Year September 1, 1931 to August 31, 1932

County	Total Vehicles Reg.	Total Gross License Fees	Less Commissions, Postage, Refunds, etc.	Total Net License Fees	Net To Counties	Net To State
Anderson	5,434	$ 55 937 59	$ 2 962 70	$52 974 89	$50 965 04	$ 2 009 85
Andrews	179	2 250 00	125 91	2 124 09	2 017 25	106 84
Angelina	4,297	46 531 86	2 532 80	43 999 06	41 493 86	2 505 20
Aransas	588	5 503 38	372 76	5 130 62	4 950 53	180 09
Archer	1,924	21 518 25	1 017 39	20 500 86	20 140 41	360 45

State	County	Year	Cars	Trucks	Total
TX	Anderson	1932			5,434
LA	Acadia	1927	3,840	402	4,242
CO	Adams	1930		1,248	
...	...	...	...	...	...

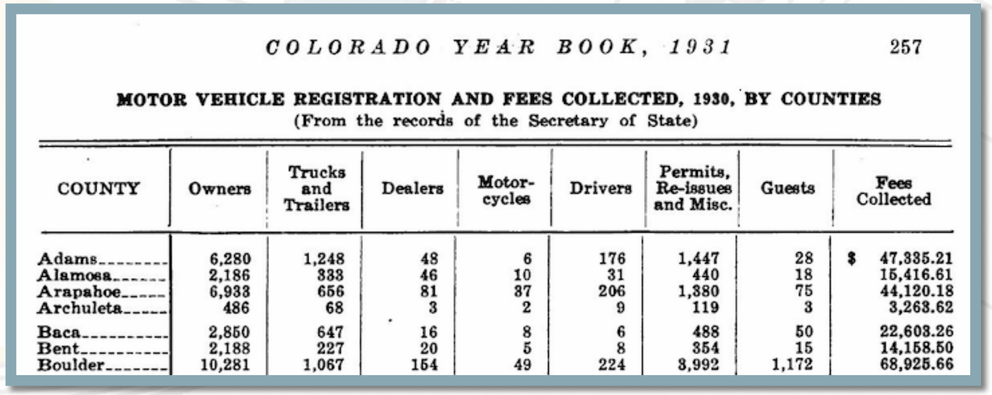


268 *Report of Louisiana Highway Commission*

TABLE No. 17.

List of Cars and Trucks of Each Parish for the Year of 1927.

Parish	Cars	Trucks	Total	Amount
Acadia	3,840	402	4,242	$ 72,162.82
Allen	1,980	340	2,320	60,863.40
Ascension	1,830	285	2,115	45,623.50
Assumption	1,240	360	1,600	40,326.87
Avoyelles	2,562	382	2,944	50,070.62
Beauregard	2,085	395	2,480	36,750.90
Bienville	2,146	280	2,426	30,630.85
Bossier	2,570	350	2,920	45,450.57
Caddo	19,290	3,463	22,753	405,634.00
Calcasieu	6,490	974	7,464	100,463.00
Caldwell	1,135	282	1,417	16,740.90
Cameron	410	103	513	7,985.65



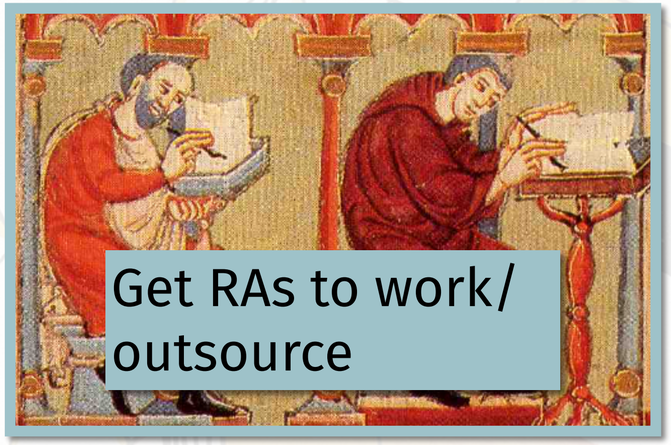
COLORADO YEAR BOOK, 1931 257

MOTOR VEHICLE REGISTRATION AND FEES COLLECTED, 1930, BY COUNTIES
(From the records of the Secretary of State)

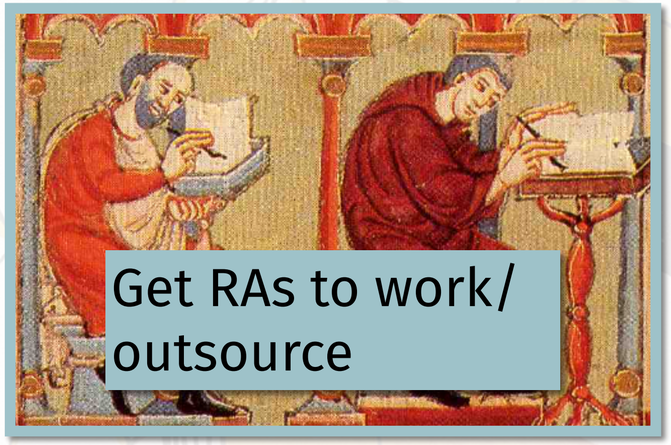
COUNTY	Owners	Trucks and Trailers	Dealers	Motor-cycles	Drivers	Permits, Re-issues and Misc.	Guests	Fees Collected
Adams	6,280	1,248	48	6	176	1,447	28	$ 47,335.21
Alamosa	2,186	333	46	10	31	440	18	15,416.61
Arapahoe	6,933	656	81	37	206	1,380	75	44,120.18
Archuleta	486	68	3	2	9	119	3	3,263.62
Baca	2,850	647	16	8	6	488	50	22,603.26
Bent	2,188	227	20	5	8	354	15	14,158.50
Boulder	10,281	1,067	154	49	224	3,992	1,172	68,925.66

The data are scattered across diverse scans. What do you do?

The data are scattered across diverse scans. What do you do?



The data are scattered across diverse scans. What do you do?



Layout Parser

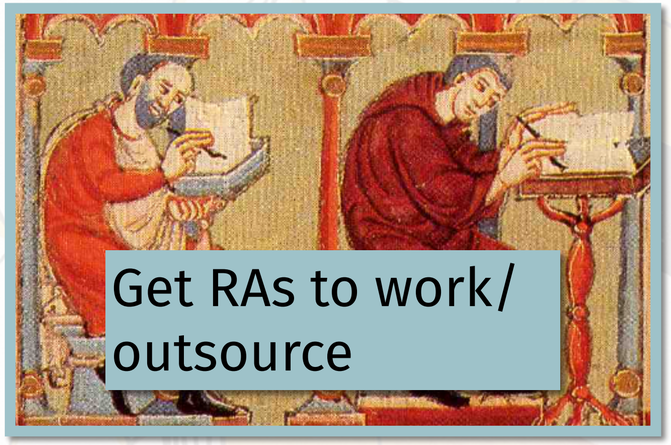
Tool up

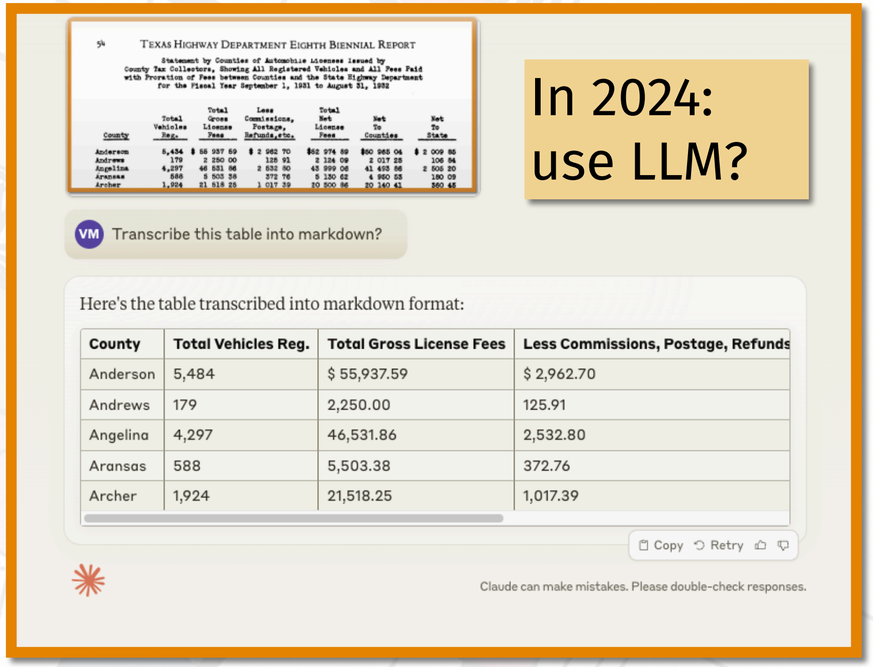
A unified toolkit for Deep Learning Based Document Image Analysis

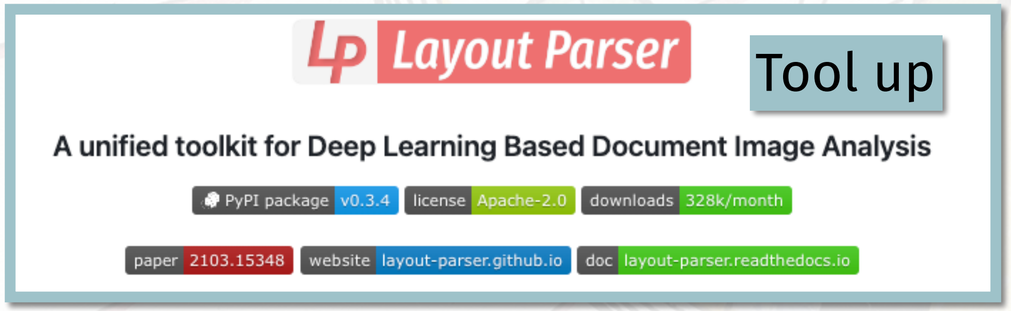
PyPI package v0.3.4 | license Apache-2.0 | downloads 328k/month

paper 2103.15348 | website layout-parser.github.io | doc layout-parser.readthedocs.io

The data are scattered across diverse scans. What do you do?





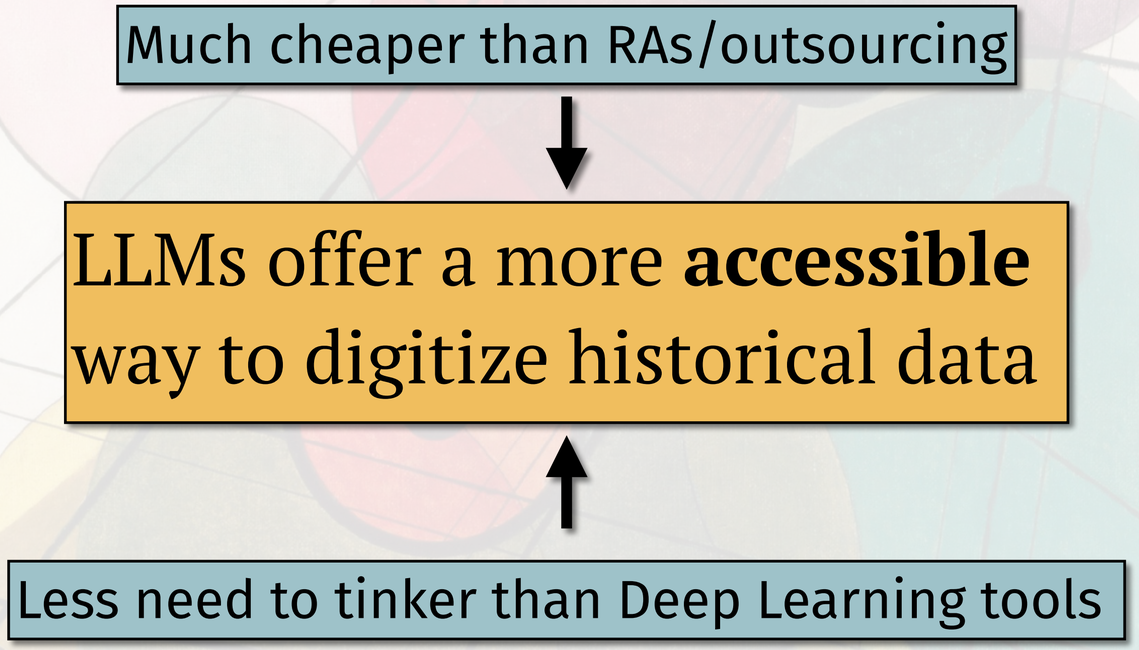


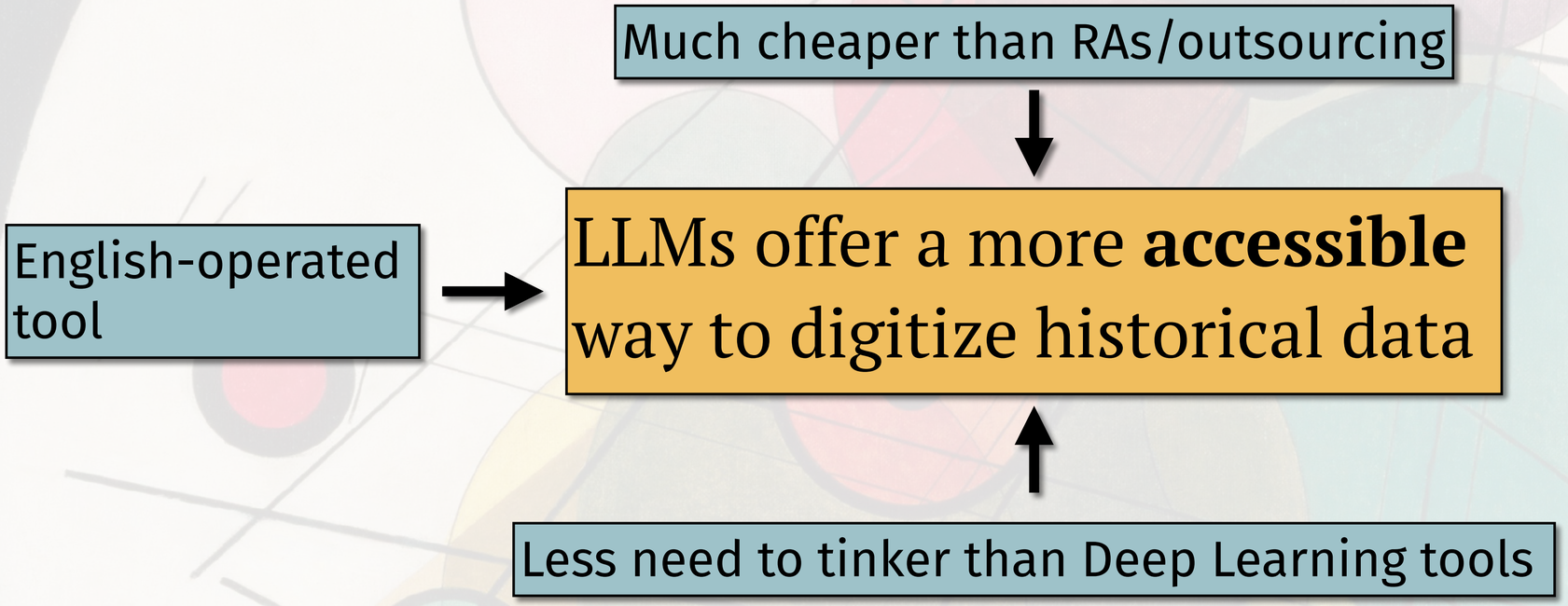
LLMs offer a more **accessible** way to digitize historical data

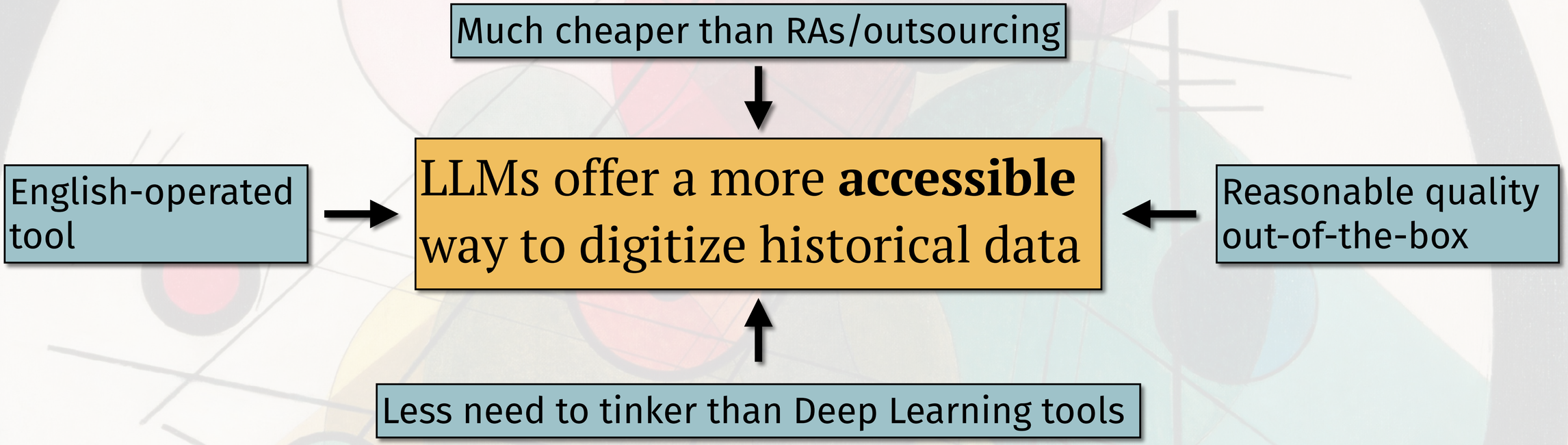
Much cheaper than RAs/outsourcing

↓

LLMs offer a more **accessible** way to digitize historical data


Much cheaper than RAs/outsourcing
LLMs offer a more **accessible** way to digitize historical data
Less need to tinker than Deep Learning tools

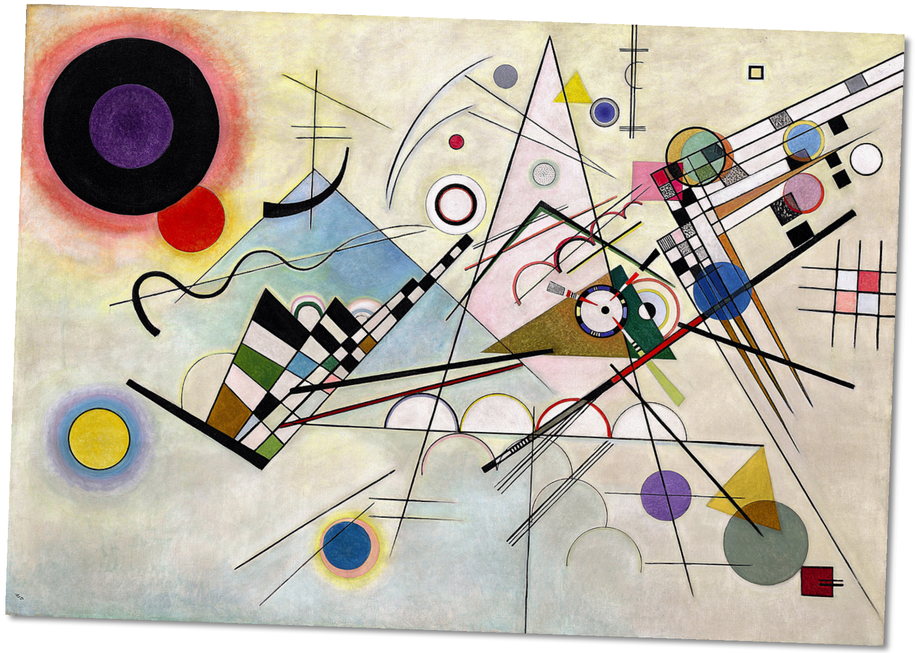

Much cheaper than RAs/outsourcing
English-operated tool
LLMs offer a more **accessible** way to digitize historical data
Less need to tinker than Deep Learning tools


Much cheaper than RAs/outsourcing
English-operated tool
LLMs offer a more **accessible** way to digitize historical data
Reasonable quality out-of-the-box
Less need to tinker than Deep Learning tools

Plan for this talk

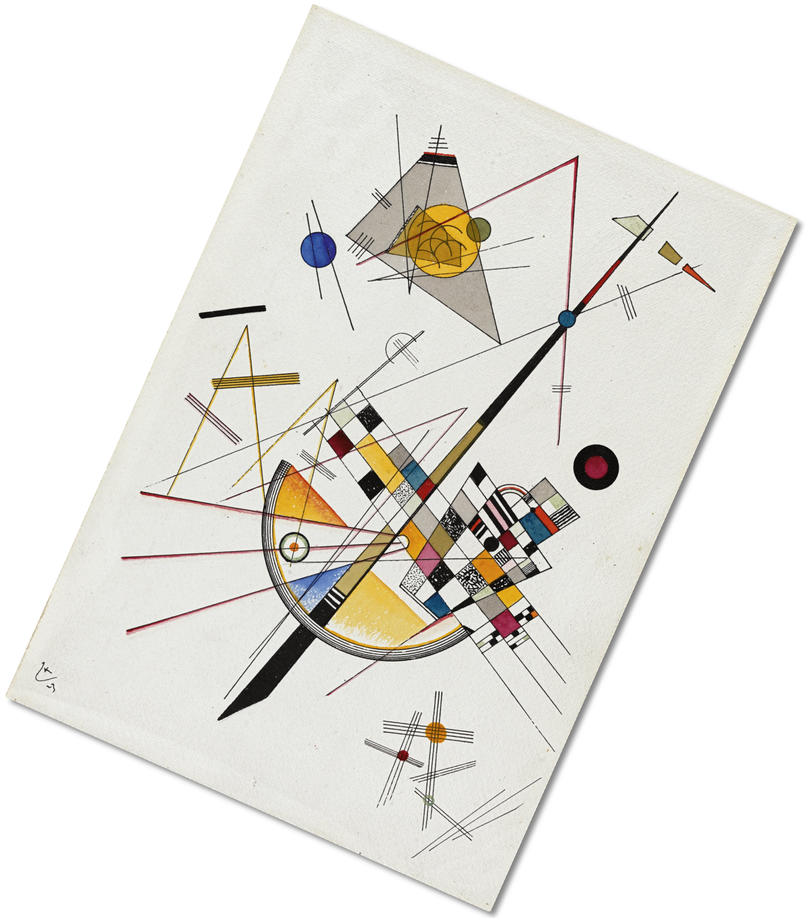
Plan for this talk

1. Example dataset: historical auto statistics





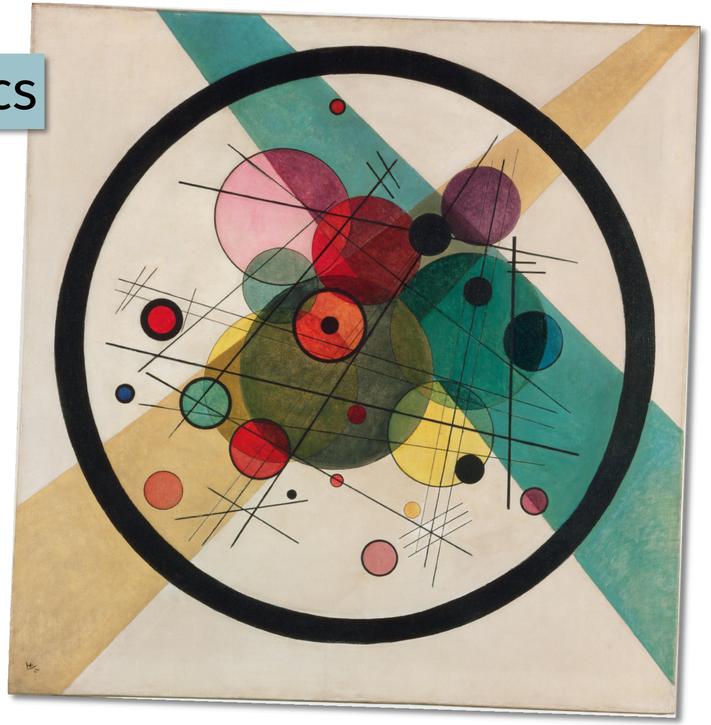
Non standard raw data

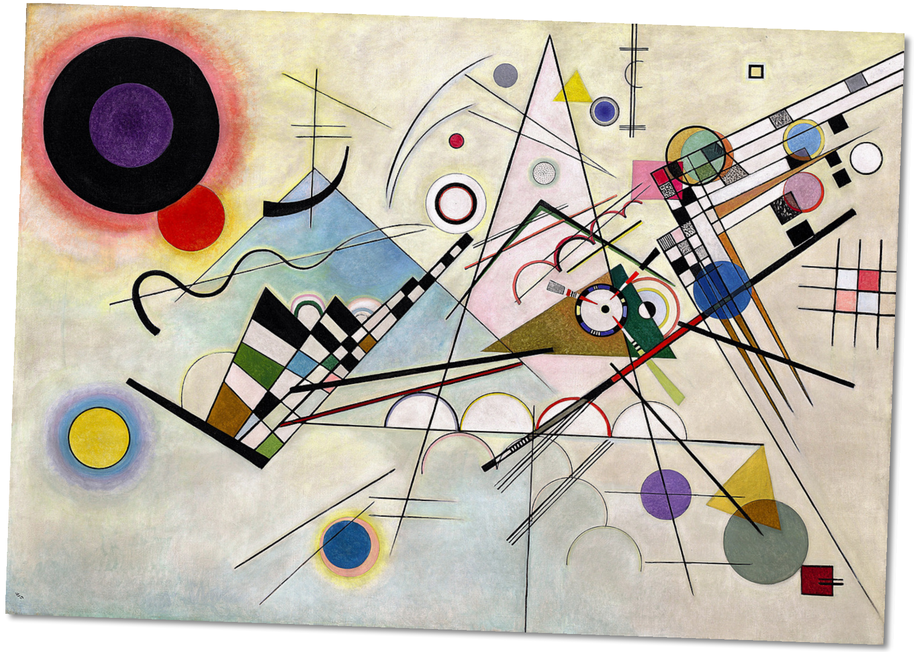


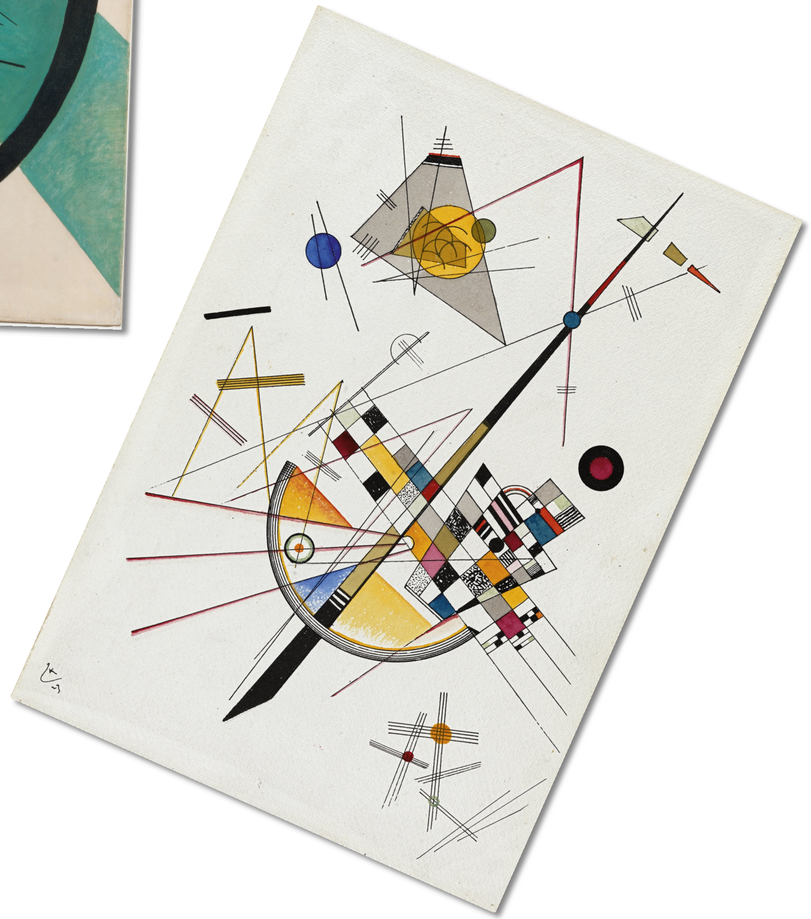
Plan for this talk

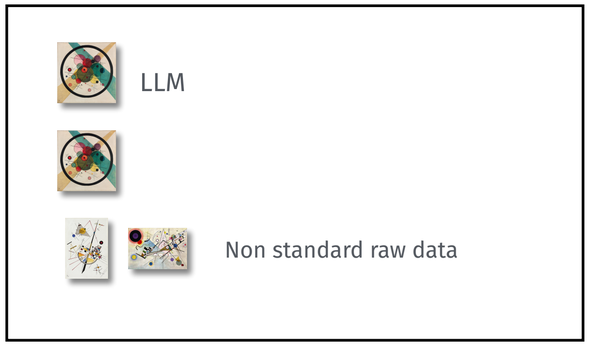
1. Example dataset: historical auto statistics

2. LLM



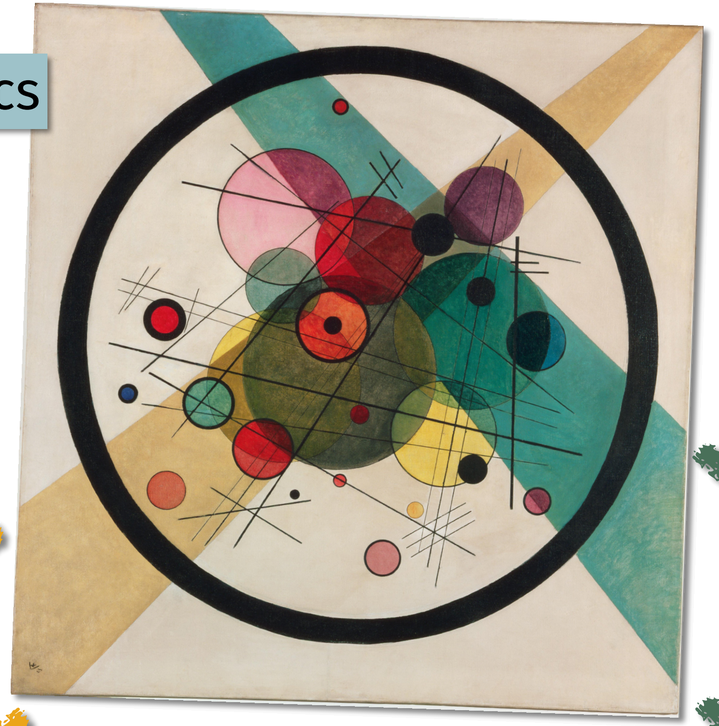


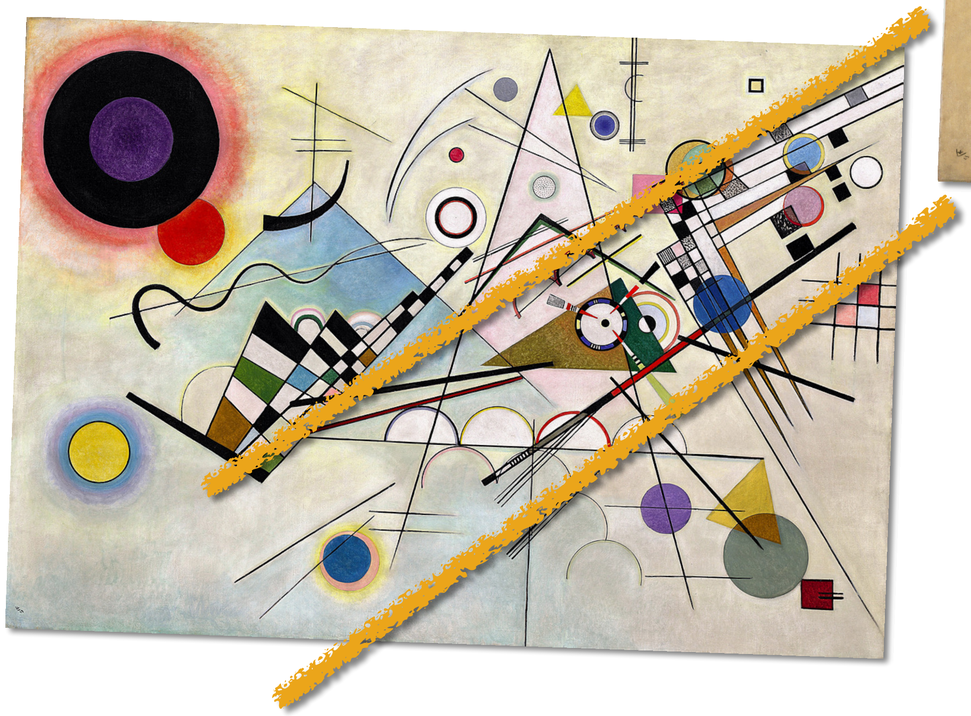


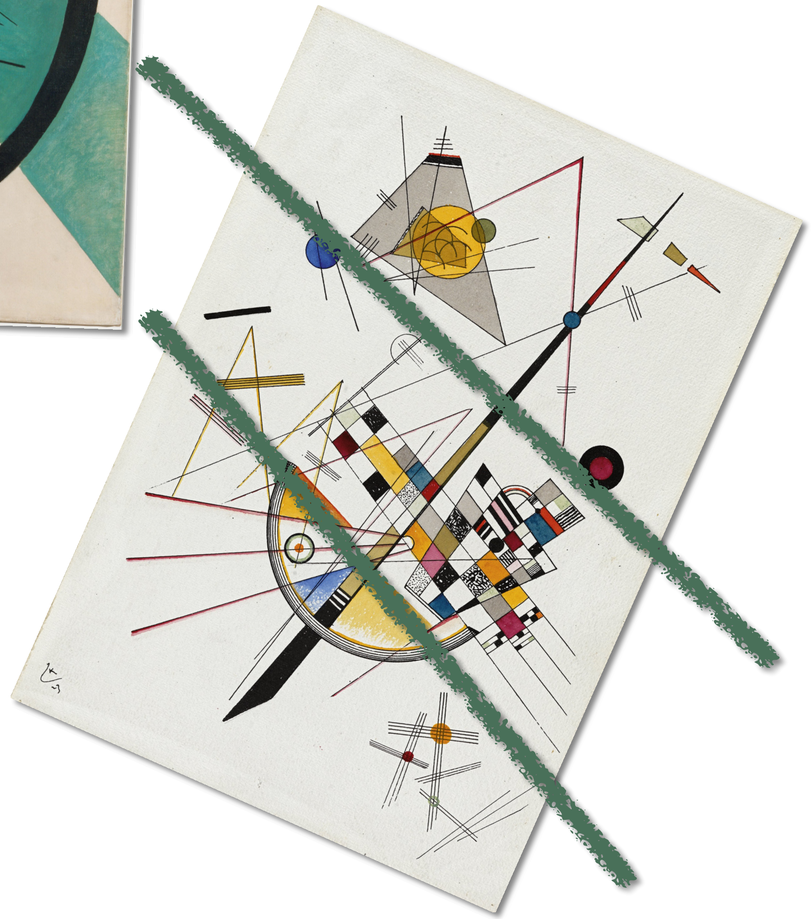


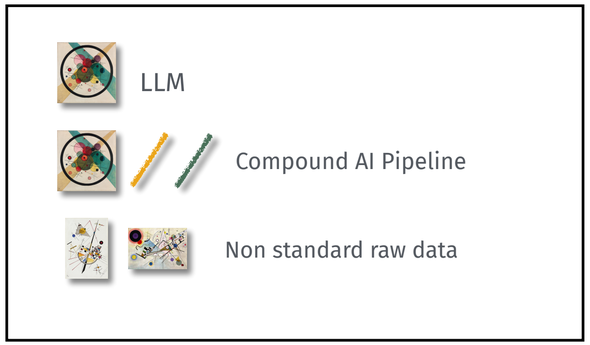
Plan for this talk

1. Example dataset: historical auto statistics
2. LLM
3. Compound AI pipeline



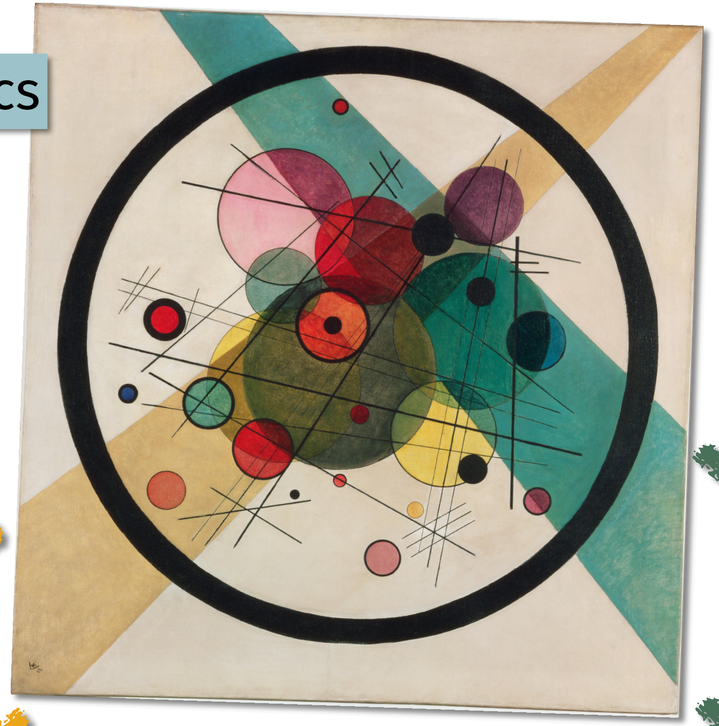


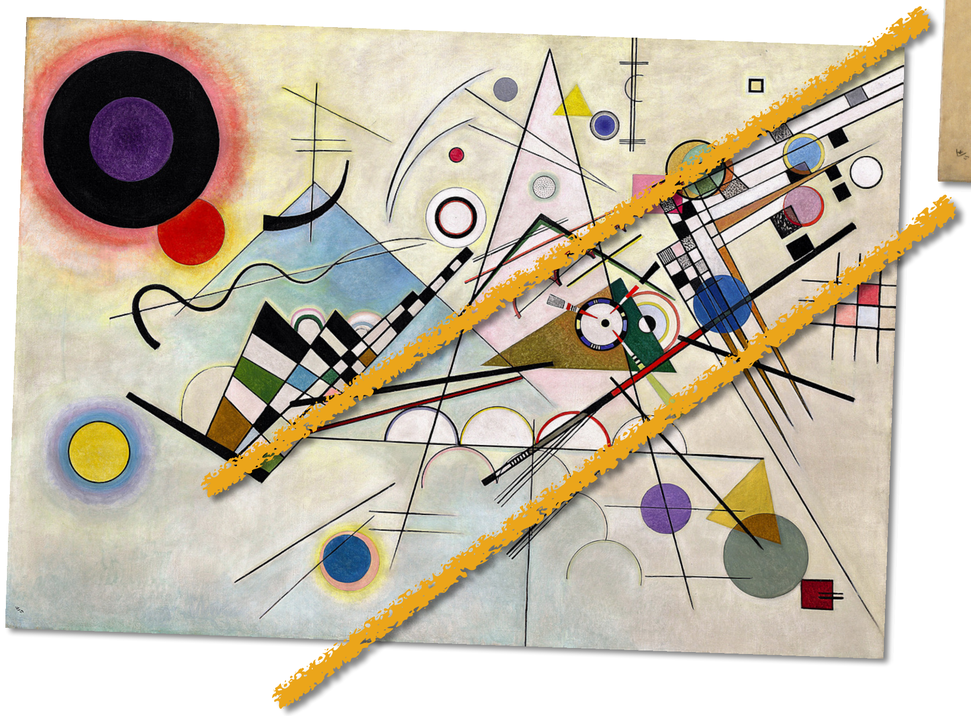


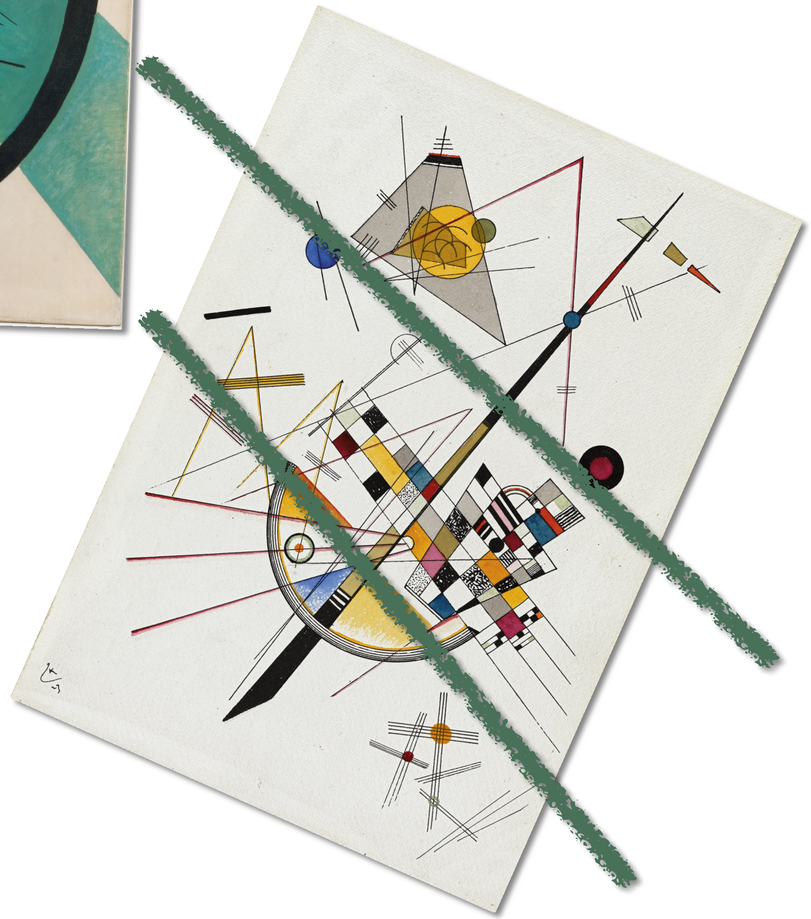


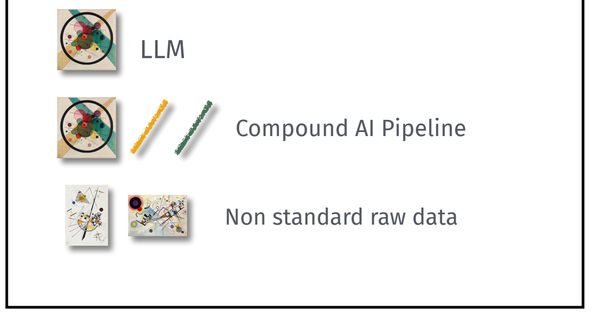
Plan for this talk

1. Example dataset: historical auto statistics
2. LLM
3. Compound AI pipeline
4. Evaluation



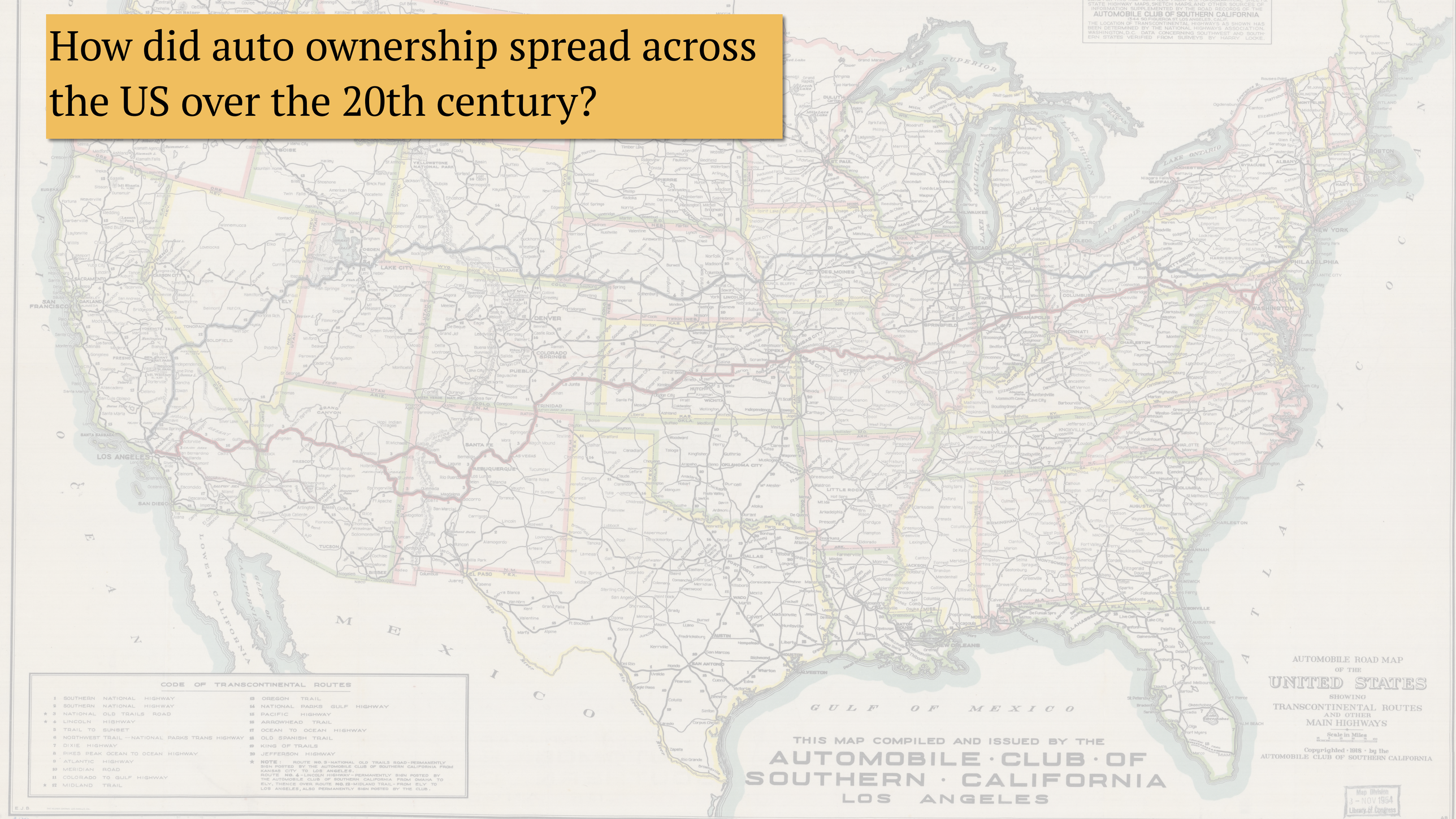







COMPILED FROM STATE HIGHWAY MAPS, SKETCH MADS, AND OTHER SOURCES OF INFORMATION SUPPLEMENTED BY THE ROAD RECORDS OF THE AUTOMOBILE CLUB OF SOUTHERN CALIFORNIA, 1344-50 FIGUEROA ST. LOS ANGELES, CALIF.
THE LOCATION OF TRANSCONTINENTAL HIGHWAYS AS SHOWN HAS BEEN DETERMINED BY THE NATIONAL HIGHWAYS ASSOCIATION, WASHINGTON, D.C. DATA CONCERNING SOUTHWEST AND SOUTHERN STATES VERIFIED FROM SURVEYS BY HARRY LOCKE.
CODE OF TRANSCONTINENTAL ROUTES
1 SOUTHERN NATIONAL HIGHWAY
2 SOUTHERN NATIONAL HIGHWAY
★ 3 NATIONAL OLD TRAILS ROAD
★ 4 LINCOLN HIGHWAY
5 TRAIL TO SUNSET
6 NORTHWEST TRAIL — NATIONAL PARKS TRANS HIGHWAY
7 DIXIE HIGHWAY
8 PIKES PEAK OCEAN TO OCEAN HIGHWAY
9 ATLANTIC HIGHWAY
10 MERIDIAN ROAD
11 COLORADO TO GULF HIGHWAY
★ 12 MIDLAND TRAIL
13 OREGON TRAIL
14 NATIONAL PARKS GULF HIGHWAY
15 PACIFIC HIGHWAY
16 ARROWHEAD TRAIL
17 OCEAN TO OCEAN HIGHWAY
18 OLD SPANISH TRAIL
19 KING OF TRAILS
20 JEFFERSON HIGHWAY
★ NOTE: ROUTE NO. 3 - NATIONAL OLD TRAILS ROAD - PERMANENTLY SIGN POSTED BY THE AUTOMOBILE CLUB OF SOUTHERN CALIFORNIA FROM KANSAS CITY TO LOS ANGELES.
ROUTE NO. 4 - LINCOLN HIGHWAY - PERMANENTLY SIGN POSTED BY THE AUTOMOBILE CLUB OF SOUTHERN CALIFORNIA FROM OMAHA TO ELY, THENCE OVER ROUTE NO. 12 - MIDLAND TRAIL - FROM ELY TO LOS ANGELES, ALSO PERMANENTLY SIGN POSTED BY THE CLUB.
E. J. B.
THIS MAP COMPILED AND ISSUED BY THE
AUTOMOBILE · CLUB · OF SOUTHERN · CALIFORNIA
LOS ANGELES
AUTOMOBILE ROAD MAP OF THE
UNITED STATES
SHOWING
TRANSCONTINENTAL ROUTES AND OTHER MAIN HIGHWAYS
Scale in Miles
Copyrighted · 1918 · by the AUTOMOBILE CLUB OF SOUTHERN CALIFORNIA
Map Division 3 - NOV 1954 Library of Congress

How did auto ownership spread across the US over the 20th century?



How did auto ownership spread across the US over the 20th century?

Historical auto statistics







How did auto ownership spread across the US over the 20th century?

Historical auto statistics

County-level, 1900—2000







How did auto ownership spread across the US over the 20th century?

Historical auto statistics

County-level, 1900—2000

Various state publications from Google Books







How did auto ownership spread across the US over the 20th century?

Historical auto statistics

County-level, 1900—2000

Various state publications from Google Books

This paper: ~700 state-year tables with human transcriptions **(enables evaluation!)**







CODE OF TRANSCONTINENTAL ROUTES

1	SOUTHERN NATIONAL HIGHWAY	13	OREGON TRAIL
2	SOUTHERN NATIONAL HIGHWAY	14	NATIONAL PARKS GULF HIGHWAY
★ 3	NATIONAL OLD TRAILS ROAD	15	PACIFIC HIGHWAY
★ 4	LINCOLN HIGHWAY	16	ARROWHEAD TRAIL
5	TRAIL TO SUNSET	17	OCEAN TO OCEAN HIGHWAY
6	NORTHWEST TRAIL — NATIONAL PARKS TRANS HIGHWAY	18	OLD SPANISH TRAIL
7	DIXIE HIGHWAY	19	KING OF TRAILS
8	PIKES PEAK OCEAN TO OCEAN HIGHWAY	20	JEFFERSON HIGHWAY
9	ATLANTIC HIGHWAY		
10	MERIDIAN ROAD		
11	COLORADO TO GULF HIGHWAY		
★ 12	MIDLAND TRAIL		

★ NOTE: ROUTE NO. 3—NATIONAL OLD TRAILS ROAD—PERMANENTLY SIGN POSTED BY THE AUTOMOBILE CLUB OF SOUTHERN CALIFORNIA FROM KANSAS CITY TO LOS ANGELES. ROUTE NO. 4—LINCOLN HIGHWAY—PERMANENTLY SIGN POSTED BY THE AUTOMOBILE CLUB OF SOUTHERN CALIFORNIA FROM OMAHA TO ELY, THENCE OVER ROUTE NO. 12—MIDLAND TRAIL—FROM ELY TO LOS ANGELES, ALSO PERMANENTLY SIGN POSTED BY THE CLUB.

THIS MAP COMPILED AND ISSUED BY THE
AUTOMOBILE · CLUB · OF
SOUTHERN · CALIFORNIA
LOS ANGELES

AUTOMOBILE ROAD MAP
OF THE
UNITED STATES
SHOWING
TRANSCONTINENTAL ROUTES
AND OTHER
MAIN HIGHWAYS

Copyrighted · 1918 · by the
AUTOMOBILE CLUB OF SOUTHERN CALIFORNIA

Map Division
3 - NOV 1964
Library of Congress

ID DOCS
T7000.01
1954

IDAHO STATE LIBRARY
3 9373 00450818 6

47th ANNUAL REPORT

BY THE
IDAHO BOARD OF
HIGHWAY DIRECTORS

FOR THE YEAR ENDING DECEMBER 31, 1954

IDAHO DEPARTMENT OF HIGHWAYS

ID DOCS
T7000.01
1954

IDAHO STATE LIBRARY
3 9373 00450818 6

47th ANNUAL REPORT

BY THE
IDAHO BOARD OF
HIGHWAY DIRECTORS

FOR THE YEAR ENDING DECEMBER 31, 1954

IDAHO DEPARTMENT OF HIGHWAYS

Table VI Motor Vehicle License Registration by Counties

1953 and 1954

County	1953 Autos	1953 Trucks	1953 Total	1954 Autos	1954 Trucks	1954 Total
Ada	29,877	6,490	36,367	32,000	7,179	39,179
Adams	1,017	509	1,526	993	553	1,546
Bannock	15,459	3,011	18,470	15,742	3,441	19,183
Bear Lake	2,563	886	3,449	2,598	978	3,576
Benewah	2,290	1,108	3,398	2,258	1,271	3,529
Bingham	7,451	2,436	9,887	7,601	3,301	10,902
Blaine	1,980	795	2,775	2,002	831	2,833
Boise	484	274	758	551	311	862
Bonner	5,868	2,306	8,174	5,666	2,432	8,098
Bonneville	12,570	3,340	15,910	13,110	3,804	16,914
Boundary	1,883	1,044	2,927	2,107	1,166	3,273
Butte	1,037	462	1,499	1,075	525	1,600
Camas	385	256	641	395	274	669
Canyon	19,345	5,367	24,712	20,041	6,673	26,714
Caribou	2,125	853	2,978	2,189	1,149	3,338
Cassia	4,732	1,759	6,491	4,755	2,261	7,016
Clark	437	202	639	496	225	721
Clearwater	2,894	1,173	4,067	3,059	1,355	4,414
Custer	1,000	622	1,622	1,010	618	1,628
Elmore	3,347	804	4,151	3,991	972	4,963
Franklin	3,707	1,089	4,796	3,638	1,407	5,045
Fremont	2,948	1,118	4,066	2,971	1,582	4,553
Gem	3,092	1,071	4,163	3,235	1,175	4,410
Gooding	3,490	1,425	4,915	3,433	1,562	4,995
Idaho	3,756	2,166	5,922	3,900	2,411	6,311
Jefferson	3,192	1,185	4,377	3,181	1,458	4,639
Jerome	3,902	1,726	5,628	3,966	1,919	5,885
Kootenai	11,347	3,027	14,374	11,619	3,414	15,033
Latah	6,922	2,486	9,408	6,875	2,891	9,766
Lemhi	1,982	995	2,977	2,060	1,022	3,082
Lewis	1,487	659	2,146	1,604	970	2,574
Lincoln	1,159	524	1,683	1,201	597	1,798
Madison	2,897	1,076	3,973	2,844	1,390	4,234
Minidoka	3,477	1,375	4,852	3,701	1,860	5,561
Nez Perce	9,784	2,917	12,701	10,038	3,408	13,446
Oneida	1,803	810	2,613	1,743	937	2,680
Owyhee	1,917	1,040	2,957	1,684	1,128	2,812
Payette	4,261	1,532	5,793	4,350	1,645	5,995
Power	1,301	597	1,898	1,376	1,023	2,399
Shoshone	8,345	1,875	10,220	7,964	1,961	9,925
Teton	971	428	1,399	988	517	1,505
Twin Falls	15,529	5,200	20,729	15,771	6,230	22,001
Valley	1,283	509	1,792	1,327	571	1,898
Washington	2,912	1,205	4,117	3,035	1,409	4,444
TOTALS	218,208	69,732	287,940	224,143	81,806	305,949

Non-homogeneity is an opportunity for LLMs

Non-homogeneity is an opportunity for LLMs

The number of registrations issued by Counties is as follows:

	Letter.	*Passenger Tags.*	*Commercial Registrations.*	*Omnibus Reg.*	*Total.*
Atlantic,	A	20,433	5,783	738	26,954
Bergen,	B	55,576	8,668	580	64,824
Burlington,	N	16,643	3,578	123	20,344
Camden,	C	33,620	6,886	562	41,068
Cape May,	F	7,159	1,673	92	8,924
Cumberland,	Y	13,291	4,222	134	17,647
Essex,	E 1-E 2-E	115,295	23,042	2,933	141,270
Gloucester,	Z	13,146	3,248	139	16,533
Hudson,	H	58,559	13,674	1,528	73,761
Hunterdon,	J	7,236	1,767	79	9,082
Mercer,	L	27,803	6,710	438	34,951
Middlesex,	K	28,937	5,116	293	34,346
Monmouth,	M	33,665	8,163	558	42,386
Morris,	V	21,052	3,088	272	24,412
Ocean,	O N	8.741	2,183	180	11,104
Passaic,	P	39,803	9.538	634	49,975
Salem,	S	7,185	2,022	30	9,237
Somerset,	I	11,049	2,201	135	13,385
Sussex,	R	6,413	1,608	76	8,097
Union,	U	42,529	7,592	449	50,570
Warren,	W	9,098	2,018	98	11,214
Non-Resident,	4	7,432	2,584	306	10,322
Special,		2	.....	.....	2
		*584,667	125,364	10,377	†720,408

*Passenger registrations, full rate,	531,689
Passenger registrations, half rate,	44,444
Duplicate tags,	7,520
Exchanges,	415
No fee registrations,	77
Undertaker registrations,	522
	584,667

† This includes duplicate tags, no fee registrations, and exchanges.

2019-10-25 17:34 GMT / http://hdl.handle.net/2027/uc1.b2968823
In the United States, Google-digitized / http://www.hathitrust.org/access_use#pd-us-google

Non-homogeneity is an opportunity for LLMs

The number of registrations issued by Counties is as follows:

	Letter.	Passenger Tags.	Commercial Registrations.	Omnibus Reg.	Total.
Atlantic,	A	20,433	5,783	738	26,954
Bergen,	B	55,576	8,668	580	64,824
Burlington,	N	16,643	3,578	123	20,344
Camden,	C	33,620	6,886	562	41,068
Cape May,	F	7,159	1,673	92	8,924
Cumberland,	Y	13,291	4,222	134	17,647
	E				
Essex,	1-E	115,295	23,042	2,933	141,270
	2-E				
Gloucester,	Z	13,146	3,248	139	16,533
Hudson,	H	58,559	13,674	1,528	73,761
Hunterdon,	J	7,236	1,767	79	9,082
Mercer,	L	27,803	6,710	438	34,951
Middlesex,	K	28,937	5,116	293	34,346
Monmouth,	M	33,665	8,163	558	42,386
Morris,	V	21,052	3,088	272	24,412
Ocean,	O	8.741	2,183	180	11,104
	N				
Passaic,	P	39,803	9,538	634	49,975
Salem,	S	7,185	2,022	30	9,237
Somerset,	I	11,049	2,201	135	13,385
Sussex,	R	6,413	1,608	76	8,097
Union,	U	42,529	7,592	449	50,570
Warren,	W	9,098	2,018	98	11,214
Non-Resident,	4	7,432	2,584	306	10,322
Special,		2	.....	.....	2
		*584,667	125,364	10,377	†720,408

*Passenger registrations, full rate,	531,689
Passenger registrations, half rate,	44,444
Duplicate tags,	7,520
Exchanges,	415
No fee registrations,	77
Undertaker registrations,	522
	584,667

† This includes duplicate tags, no fee registrations, and exchanges.

Table VI Motor Vehicle License Registration by Counties

1953 and 1954

County	1953 Autos	1953 Trucks	1953 Total	1954 Autos	1954 Trucks	1954 Total
Ada	29,877	6,490	36,367	32,000	7,179	39,179
Adams	1,017	509	1,526	993	553	1,546
Bannock	15,459	3,011	18,470	15,742	3,441	19,183
Bear Lake	2,563	886	3,449	2,598	978	3,576
Benewah	2,290	1,108	3,398	2,258	1,271	3,529
Bingham	7,451	2,436	9,887	7,601	3,301	10,902
Blaine	1,980	795	2,775	2,002	831	2,833
Boise	484	274	758	551	311	862
Bonner	5,868	2,306	8,174	5,666	2,432	8,098
Bonneville	12,570	3,340	15,910	13,110	3,804	16,914
Boundary	1,883	1,044	2,927	2,107	1,166	3,273
Butte	1,037	462	1,499	1,075	525	1,600
Camas	385	256	641	395	274	669
Canyon	19,345	5,367	24,712	20,041	6,673	26,714
Caribou	2,125	853	2,978	2,189	1,149	3,338
Cassia	4,732	1,759	6,491	4,755	2,261	7,016
Clark	437	202	639	496	225	721
Clearwater	2,894	1,173	4,067	3,059	1,355	4,414
Custer	1,000	622	1,622	1,010	618	1,628
Elmore	3,347	804	4,151	3,991	972	4,963
Franklin	3,707	1,089	4,796	3,638	1,407	5,045
Fremont	2,948	1,118	4,066	2,971	1,582	4,553
Gem	3,092	1,071	4,163	3,235	1,175	4,410
Gooding	3,490	1,425	4,915	3,433	1,562	4,995
Idaho	3,756	2,166	5,922	3,900	2,411	6,311
Jefferson	3,192	1,185	4,377	3,181	1,458	4,639
Jerome	3,902	1,726	5,628	3,966	1,919	5,885
Kootenai	11,347	3,027	14,374	11,619	3,414	15,033
Latah	6,922	2,486	9,408	6,875	2,891	9,766
Lemhi	1,982	995	2,977	2,060	1,022	3,082
Lewis	1,487	659	2,146	1,604	970	2,574
Lincoln	1,159	524	1,683	1,201	597	1,798
Madison	2,897	1,076	3,973	2,844	1,390	4,234
Minidoka	3,477	1,375	4,852	3,701	1,860	5,561
Nez Perce	9,784	2,917	12,701	10,038	3,408	13,446
Oneida	1,803	810	2,613	1,743	937	2,680
Owyhee	1,917	1,040	2,957	1,684	1,128	2,812
Payette	4,261	1,532	5,793	4,350	1,645	5,995
Power	1,301	597	1,898	1,376	1,023	2,399
Shoshone	8,345	1,875	10,220	7,964	1,961	9,925
Teton	971	428	1,399	988	517	1,505
Twin Falls	15,529	5,200	20,729	15,771	6,230	22,001
Valley	1,283	509	1,792	1,327	571	1,898
Washington	2,912	1,205	4,117	3,035	1,409	4,444
TOTALS	218,208	69,732	287,940	224,143	81,806	305,949

2019-10-25 17:34 GMT / http://hdl.handle.net/2027/uc1.b2968823 / http://www.hathitrust.org/access_use#pd-us-google In the United States, Google-digitized

Non-homogeneity is an opportunity for LLMs

The number of registrations issued by Counties is as follows:

	Letter.	Passenger Tags.	Commercial Registrations.	Omnibus Reg.	Total.
Atlantic,	A	20,433	5,783	738	26,954
Bergen,	B	55,576	8,668	580	64,824
Burlington,	N	16,643	3,578	123	20,344
Camden,	C	33,620	6,886	562	41,068
Cape May,	F	7,159	1,673	92	8,924
Cumberland,	Y	13,291	4,222	134	17,647
	E				
Essex,	1-E	115,295	23,042	2,933	141,270
	2-E				
Gloucester,	Z	13,146	3,248	139	16,533
Hudson,	H	58,559	13,674	1,528	73,761
Hunterdon,	J	7,236	1,767	79	9,082
Mercer,	L	27,803	6,710	438	34,951
Middlesex,	K	28,937	5,116	293	34,346
Monmouth,	M	33,665	8,163	558	42,386
Morris,	V	21,052	3,088	272	24,412
Ocean,	O	8.741	2,183	180	11,104
	N				
Passaic,	P	39,803	9.538	634	49,975
Salem,	S	7,185	2,022	30	9,237
Somerset,	I	11,049	2,201	135	13,385
Sussex,	R	6,413	1,608	76	8,097
Union,	U	42,529	7,592	449	50,570
Warren,	W	9,098	2,018	98	11,214
Non-Resident,	4	7,432	2,584	306	10,322
Special,		2	.....	.....	2
		*584,667	125,364	10,377	†720,408

*Passenger registrations, full rate,	531,689
Passenger registrations, half rate,	44,444
Duplicate tags,	7,520
Exchanges,	415
No fee registrations,	77
Undertaker registrations,	522
	584,667

† This includes duplicate tags, no fee registrations, and exchanges.

Table VI Motor Vehicle License Registration by Counties

1953 and 1954

County	1953 Autos	1953 Trucks	1953 Total	1954 Autos	1954 Trucks	1954 Total
Ada	29,877	6,490	36,367	32,000	7,179	39,179
Adams	1,017	509	1,526	993	553	1,546
Bannock	15,459	3,011	18,470	15,742	3,441	19,183
Bear Lake	2,563	886	3,449	2,598	978	3,576
Benewah	2,290	1,108	3,398	2,258	1,271	3,529
Bingham	7,451	2,436	9,887	7,601	3,301	10,902
Blaine	1,980	795	2,775	2,002	831	2,833
Boise	484	274	758	551	311	862
Bonner	5,868	2,306	8,174	5,666	2,432	8,098
Bonneville	12,570	3,340	15,910	13,110	3,804	16,914
Boundary	1,883	1,044	2,927	2,107	1,166	3,273
Butte	1,037	462	1,499	1,075	525	1,600
Camas	385	256	641	395	274	669
Canyon	19,345	5,367	24,712	20,041	6,673	26,714
Caribou	2,125	853	2,978	2,189	1,149	3,338
Cassia	4,732	1,759	6,491	4,755	2,261	7,016
Clark	437	202	639	496	225	721
Clearwater	2,894	1,173	4,067	3,059	1,355	4,414
Custer	1,000	622	1,622	1,010	618	1,628
Elmore	3,347	804	4,151	3,991	972	4,963
Franklin	3,707	1,089	4,796	3,638	1,407	5,045
Fremont	2,948	1,118	4,066	2,971	1,582	4,553
Gem	3,092	1,071	4,163	3,235	1,175	4,410
Gooding	3,490	1,425	4,915	3,433	1,562	4,995
Idaho	3,756	2,166	5,922	3,900	2,411	6,311
Jefferson	3,192	1,185	4,377	3,181	1,458	4,639
Jerome	3,902	1,726	5,628	3,966	1,919	5,885
Kootenai	11,347	3,027	14,374	11,619	3,414	15,033
Latah	6,922	2,486	9,408	6,875	2,891	9,766
Lemhi	1,982	995	2,977	2,060	1,022	3,082
Lewis	1,487	659	2,146	1,604	970	2,574
Lincoln	1,159	524	1,683	1,201	597	1,798
Madison	2,897	1,076	3,973	2,844	1,390	4,234
Minidoka	3,477	1,375	4,852	3,701	1,860	5,561
Nez Perce	9,784	2,917	12,701	10,038	3,408	13,446
Oneida	1,803	810	2,613	1,743	937	2,680
Owyhee	1,917	1,040	2,957	1,684	1,128	2,812
Payette	4,261	1,532	5,793	4,350	1,645	5,995
Power	1,301	597	1,898	1,376	1,023	2,399
Shoshone	8,345	1,875	10,220	7,964	1,961	9,925
Teton	971	428	1,399	988	517	1,505
Twin Falls	15,529	5,200	20,729	15,771	6,230	22,001
Valley	1,283	509	1,792	1,327	571	1,898
Washington	2,912	1,205	4,117	3,035	1,409	4,444
TOTALS	218,208	69,732	287,940	224,143	81,806	305,949

TABLE VI - M.V.R. BY COUNTY 75

SCHEDULE SHOWING BY COUNTIES TOTAL NUMBER OF AUTOMOBILES, TRUCKS, TRAILERS, MOTOR CYCLES AND DEALERS REGISTERED AND TRANSFERS MADE JANUARY 1 THRU DECEMBER 31, 1924. IN COMPARISON WITH REGISTRATIONS FOR YEAR 1923.

Counties	Automobiles 1924	Automobiles 1923	Trucks 1924	Trucks 1923	Trailers 1924	Trailers 1923	Motor Cycles 1924	Motor Cycles 1923	Dealers 1924	Dealers 1923	Transfers 1924	Transfers 1923
Abbeville	2,148	1,641	130	84	....	....	2	9	6	3	33	17
Aiken	4,151	3,182	415	321	11	14	16	12	10	10	75	68
Allendale	1,193	764	136	65	30	12	1	1	2	3	20	8
Anderson	8,512	7,046	730	629	3	1	12	13	37	33	250	202
Bamberg	1,465	1,143	159	112	31	17	1	....	10	7	21	15
Barnwell	1,729	1,212	164	116	6	4	5	6	9	9	30	12
Beaufort	928	755	214	159	....	1	1	2	6	6	24	20
Berkeley	821	548	108	72	23	18	....	....	1	1	38	16
Calhoun	1,070	806	113	71	31	9	1	5	3	3	21	13
Charleston	7,028	5,873	1,201	997	37	26	61	62	31	26	753	453
Cherokee	2,860	2,237	237	201	....	....	9	12	12	14	93	73
Chester	2,709	2,068	240	192	1	1	6	6	15	15	58	53
Chesterfield	3,278	2,573	361	235	7	5	5	4	14	15	83	70
Clarendon	1,671	1,267	162	115	21	9	3	4	6	5	38	27
Colleton	1,533	1,196	179	151	45	25	2	3	5	4	22	19
Darlington	3,662	2,990	302	263	15	15	6	11	19	24	96	90
Dillon	2,482	1,933	179	152	10	5	2	6	10	10	106	84
Dorchester	1,308	1,046	166	99	46	20	1	1	5	2	52	29
Edgefield	1,483	1,039	207	155	....	....	2	1	7	8	42	15
Fairfield	1,477	1,075	116	82	6	5	2	1	2	4	34	34
Florence	4,999	4,023	461	354	31	11	9	7	17	13	118	94
Georgetown	1,209	949	181	121	17	3	4	1	10	7	48	30
Greenville	12,889	10,167	1,400	1,044	4	3	45	66	69	39	445	305
Greenwood	3,612	3,012	416	366	2	4	14	11	12	13	49	60
Hampton	1,498	1,184	146	95	26	15	2	1	3	3	34	7
Horry	1,868	1,544	282	207	5	11	....	1	7	10	53	44
Jasper	477	389	45	40	10	8	....	....	2	1	6	3
Kershaw	2,447	1,878	287	230	2	3	4	4	8	14	62	71
Lancaster	1,960	1,603	169	108	3	6	7	10	12	7	37	38
Laurens	3,770	2,926	301	233	2	....	9	7	22	19	93	49
Lee	1,793	1,360	137	117	13	4	....	1	6	10	41	44
Lexington	4,006	3,024	481	349	33	32	9	12	11	7	199	135
McCormick	815	652	106	79	1	1	2	2	3	4	12	12
Marion	2,074	1,570	175	132	3	4	5	3	15	12	61	40
Marlboro	3,188	2,588	254	181	13	3	10	7	12	9	114	90
Newberry	3,196	2,606	378	265	12	19	10	9	9	17	114	101
Oconee	2,645	2,212	259	200	2	....	5	5	7	12	60	44
Orangeburg	5,078	3,953	550	416	150	98	9	9	23	21	112	75
Pickens	3,054	2,361	281	167	1	....	2	8	6	7	68	51
Richland	9,906	8,213	1,381	1,191	28	23	95	123	39	35	781	537
Saluda	1,375	1,008	169	115	6	3	2	2	4	1	24	15
Spartanburg	11,513	8,859	1,251	907	2	....	41	68	34	30	437	319
Sumter	3,371	2,708	372	280	30	6	21	11	17	15	105	71
Union	2,121	1,649	252	201	2	2	14	15	4	5	49	34
Williamsburg	1,622	1,256	154	105	19	9	1	2	7	6	18	28
York	5,033	4,148	387	322	2	....	17	15	23	25	173	187
Out of State	504	301	58	23	1	....	2	2	1	1	25	24
Total	147,530	116,537	15,852	12,119	743	455	477	561	593	545	5,227	3,826

SUMMARY

	1924	1923
Automobiles	147,530	116,537
Trucks	15,852	12,119
Trailers	743	455
Motor Cycles	477	561
Dealers	593	545
Demonstration Plates (Extra)	1,148	884
Transfers	5,227	3,826
Total Fees	$1,151,983.02	$902,608.69
Postage, etc.	12,169.83	9,888.04
Total Receipts	$1,164,152.85	$912,496.73

Digitized by Google
Original from UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
2019-10-25 17:34 GMT / http://hdl.handle.net/2027/uc1.b2968823 / http://www.hathitrust.org/access_use#pd-us-google In the United States, Google-digitized

Non-homogeneity is an opportunity for LLMs

Standard digitization tools are more likely to stumble with non standard layouts

The number of registrations issued by Counties is as follows:

	Letter.	Passenger Tags.	Commercial Registrations.	Omnibus Reg.	Total.
Atlantic,	A	20,433	5,783	738	26,954
Bergen,	B	55,576	8,668	580	64,824
Burlington,	N	16,643	3,578	123	20,344
Camden,	C	33,620	6,886	562	41,068
Cape May,	F	7,159	1,673	92	8,924
Cumberland,	Y	13,291	4,222	134	17,647
	E				
Essex,	1-E	115,295	23,042	2,933	141,270
	2-E				
Gloucester,	Z	13,146	3,248	139	16,533
Hudson,	H	58,559	13,674	1,528	73,761
Hunterdon,	J	7,236	1,767	79	9,082
Mercer,	L	27,803	6,710	438	34,951
Middlesex,	K	28,937	5,116	293	34,346
Monmouth,	M	33,665	8,163	558	42,386
Morris,	V	21,052	3,088	272	24,412
Ocean,	O	8.741	2,183	180	11,104
	N				
Passaic,	P	39,803	9,538	634	49,975
Salem,	S	7,185	2,022	30	9,237
Somerset,	I	11,049	2,201	135	13,385
Sussex,	R	6,413	1,608	76	8,097
Union,	U	42,529	7,592	449	50,570
Warren,	W	9,098	2,018	98	11,214
Non-Resident,	4	7,432	2,584	306	10,322
Special,		2	.....	.....	2
		*584,667	125,364	10,377	†720,408

*Passenger registrations, full rate,	531,689
Passenger registrations, half rate,	44,444
Duplicate tags,	7,520
Exchanges,	415
No fee registrations,	77
Undertaker registrations,	522
	584,667

† This includes duplicate tags, no fee registrations, and exchanges.

Table VI Motor Vehicle License Registration by Counties

1953 and 1954

County	1953 Autos	1953 Trucks	1953 Total	1954 Autos	1954 Trucks	1954 Total
Ada	29,877	6,490	36,367	32,000	7,179	39,179
Adams	1,017	509	1,526	993	553	1,546
Bannock	15,459	3,011	18,470	15,742	3,441	19,183
Bear Lake	2,563	886	3,449	2,598	978	3,576
Benewah	2,290	1,108	3,398	2,258	1,271	3,529
Bingham	7,451	2,436	9,887	7,601	3,301	10,902
Blaine	1,980	795	2,775	2,002	831	2,833
Boise	484	274	758	551	311	862
Bonner	5,868	2,306	8,174	5,666	2,432	8,098
Bonneville	12,570	3,340	15,910	13,110	3,804	16,914
Boundary	1,883	1,044	2,927	2,107	1,166	3,273
Butte	1,037	462	1,499	1,075	525	1,600
Camas	385	256	641	395	274	669
Canyon	19,345	5,367	24,712	20,041	6,673	26,714
Caribou	2,125	853	2,978	2,189	1,149	3,338
Cassia	4,732	1,759	6,491	4,755	2,261	7,016
Clark	437	202	639	496	225	721
Clearwater	2,894	1,173	4,067	3,059	1,355	4,414
Custer	1,000	622	1,622	1,010	618	1,628
Elmore	3,347	804	4,151	3,991	972	4,963
Franklin	3,707	1,089	4,796	3,638	1,407	5,045
Fremont	2,948	1,118	4,066	2,971	1,582	4,553
Gem	3,092	1,071	4,163	3,235	1,175	4,410
Gooding	3,490	1,425	4,915	3,433	1,562	4,995
Idaho	3,756	2,166	5,922	3,900	2,411	6,311
Jefferson	3,192	1,185	4,377	3,181	1,458	4,639
Jerome	3,902	1,726	5,628	3,966	1,919	5,885
Kootenai	11,347	3,027	14,374	11,619	3,414	15,033
Latah	6,922	2,486	9,408	6,875	2,891	9,766
Lemhi	1,982	995	2,977	2,060	1,022	3,082
Lewis	1,487	659	2,146	1,604	970	2,574
Lincoln	1,159	524	1,683	1,201	597	1,798
Madison	2,897	1,076	3,973	2,844	1,390	4,234
Minidoka	3,477	1,375	4,852	3,701	1,860	5,561
Nez Perce	9,784	2,917	12,701	10,038	3,408	13,446
Oneida	1,803	810	2,613	1,743	937	2,680
Owyhee	1,917	1,040	2,957	1,684	1,128	2,812
Payette	4,261	1,532	5,793	4,350	1,645	5,995
Power	1,301	597	1,898	1,376	1,023	2,399
Shoshone	8,345	1,875	10,220	7,964	1,961	9,925
Teton	971	428	1,399	988	517	1,505
Twin Falls	15,529	5,200	20,729	15,771	6,230	22,001
Valley	1,283	509	1,792	1,327	571	1,898
Washington	2,912	1,205	4,117	3,035	1,409	4,444
TOTALS	218,208	69,732	287,940	224,143	81,806	305,949

TABLE VI - M.V.R. BY COUNTY 75

SCHEDULE SHOWING BY COUNTIES TOTAL NUMBER OF AUTOMOBILES, TRUCKS, TRAILERS, MOTOR CYCLES AND DEALERS REGISTERED AND TRANSFERS MADE JANUARY 1 THRU DECEMBER 31, 1924. IN COMPARISON WITH REGISTRATIONS FOR YEAR 1923.

Counties	Automobiles 1924	Automobiles 1923	Trucks 1924	Trucks 1923	Trailers 1924	Trailers 1923	Motor Cycles 1924	Motor Cycles 1923	Dealers 1924	Dealers 1923	Transfers 1924	Transfers 1923
Abbeville	2,148	1,641	130	84	....	....	2	9	6	3	33	17
Aiken	4,151	3,182	415	321	11	14	16	12	10	10	75	68
Allendale	1,193	764	136	65	30	12	1	1	2	3	20	8
Anderson	8,512	7,046	730	629	3	1	12	13	37	33	250	202
Bamberg	1,465	1,143	159	112	31	17	1	....	10	7	21	15
Barnwell	1,729	1,212	164	116	6	4	5	6	9	9	30	12
Beaufort	928	755	214	159	....	1	1	2	6	6	24	20
Berkeley	821	548	108	72	23	18	....	....	1	1	38	16
Calhoun	1,070	806	113	71	31	9	1	5	3	3	21	13
Charleston	7,028	5,873	1,201	997	37	26	61	62	31	26	753	453
Cherokee	2,860	2,237	237	201	....	....	9	12	12	14	93	73
Chester	2,709	2,068	240	192	1	1	6	6	15	15	58	53
Chesterfield	3,278	2,573	361	235	7	5	5	4	14	15	83	70
Clarendon	1,671	1,267	162	115	21	9	3	4	6	5	38	27
Colleton	1,533	1,196	179	151	45	25	2	3	5	4	22	19
Darlington	3,662	2,990	302	263	15	15	6	11	19	24	96	90
Dillon	2,482	1,933	179	152	10	5	2	6	10	10	106	84
Dorchester	1,308	1,046	166	99	46	20	1	1	5	2	52	29
Edgefield	1,483	1,039	207	155	....	....	2	1	7	8	42	15
Fairfield	1,477	1,075	116	82	6	5	2	1	2	4	34	34
Florence	4,999	4,023	461	354	31	11	9	7	17	13	118	94
Georgetown	1,209	949	181	121	17	3	4	1	10	7	48	30
Greenville	12,889	10,167	1,400	1,044	4	3	45	66	69	39	445	305
Greenwood	3,612	3,012	416	366	2	4	14	11	12	13	49	60
Hampton	1,498	1,184	146	95	26	15	2	1	3	3	34	7
Horry	1,868	1,544	282	207	5	11	....	1	7	10	53	44
Jasper	477	389	45	40	10	8	....	....	2	1	6	3
Kershaw	2,447	1,878	287	230	2	3	4	4	8	14	62	71
Lancaster	1,960	1,603	169	108	3	6	7	10	12	7	37	38
Laurens	3,770	2,926	301	233	2	....	9	7	22	19	93	49
Lee	1,793	1,360	137	117	13	4	....	1	6	10	41	44
Lexington	4,006	3,024	481	349	33	32	9	12	11	7	199	135
McCormick	815	652	106	79	1	1	2	2	3	4	12	12
Marion	2,074	1,570	175	132	3	4	5	3	15	12	61	40
Marlboro	3,188	2,588	254	181	13	3	10	7	12	9	114	90
Newberry	3,196	2,606	378	265	12	19	10	9	9	17	114	101
Oconee	2,645	2,212	259	200	2	....	5	5	7	12	60	44
Orangeburg	5,078	3,953	550	416	150	98	9	9	23	21	112	75
Pickens	3,054	2,361	281	167	1	....	2	8	6	7	68	51
Richland	9,906	8,213	1,381	1,191	28	23	95	123	39	35	781	537
Saluda	1,375	1,008	169	115	6	3	2	2	4	1	24	15
Spartanburg	11,513	8,859	1,251	907	2	....	41	68	34	30	437	319
Sumter	3,371	2,708	372	280	30	6	21	11	17	15	105	71
Union	2,121	1,649	252	201	2	2	14	15	4	5	49	34
Williamsburg	1,622	1,256	154	105	19	9	1	2	7	6	18	28
York	5,033	4,148	387	322	2	....	17	15	23	25	173	187
Out of State	504	301	58	23	1	....	2	2	1	1	25	24
Total	147,530	116,537	15,852	12,119	743	455	477	561	593	545	5,227	3,826

SUMMARY

	1924	1923
Automobiles	147,530	116,537
Trucks	15,852	12,119
Trailers	743	455
Motor Cycles	477	561
Dealers	593	545
Demonstration Plates (Extra)	1,148	884
Transfers	5,227	3,826
Total Fees	$1,151,983.02	$902,608.69
Postage, etc.	12,169.83	9,888.04
Total Receipts	$1,164,152.85	$912,496.73

Digitized by Google
Original from UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
2019-10-25 17:34 GMT / http://hdl.handle.net/2027/uc1.b2968823 / http://www.hathitrust.org/access_use#pd-us-google
In the United States, Google-digitized

Non-homogeneity is an opportunity for LLMs

Standard digitization tools are more likely to stumble with non standard layouts

Even if table is successfully processed, homogenization is challenging

The number of registrations issued by Counties is as follows:

	Letter.	*Passenger Tags.*	*Commercial Registrations.*	*Omnibus Reg.*	*Total.*
Atlantic,	A	20,433	5,783	738	26,954
Bergen,	B	55,576	8,668	580	64,824
Burlington,	N	16,643	3,578	123	20,344
Camden,	C	33,620	6,886	562	41,068
Cape May,	F	7,159	1,673	92	8,924
Cumberland,	Y	13,291	4,222	134	17,647
	E				
Essex,	1-E	115,295	23,042	2,933	141,270
	2-E				
Gloucester,	Z	13,146	3,248	139	16,533
Hudson,	H	58,559	13,674	1,528	73,761
Hunterdon,	J	7,236	1,767	79	9,082
Mercer,	L	27,803	6,710	438	34,951
Middlesex,	K	28,937	5,116	293	34,346
Monmouth,	M	33,665	8,163	558	42,386
Morris,	V	21,052	3,088	272	24,412
Ocean,	O	8.741	2,183	180	11,104
	N				
Passaic,	P	39,803	9,538	634	49,975
Salem,	S	7,185	2,022	30	9,237
Somerset,	I	11,049	2,201	135	13,385
Sussex,	R	6,413	1,608	76	8,097
Union,	U	42,529	7,592	449	50,570
Warren,	W	9,098	2,018	98	11,214
Non-Resident,	4	7,432	2,584	306	10,322
Special,		2	.....	.....	2
		*584,667	125,364	10,377	†720,408

*Passenger registrations, full rate,	531,689
Passenger registrations, half rate,	44,444
Duplicate tags,	7,520
Exchanges,	415
No fee registrations,	77
Undertaker registrations,	522
	584,667

† This includes duplicate tags, no fee registrations, and exchanges.

Table VI Motor Vehicle License Registration by Counties

1953 and 1954

County	1953 Autos	1953 Trucks	1953 Total	1954 Autos	1954 Trucks	1954 Total
Ada	29,877	6,490	36,367	32,000	7,179	39,179
Adams	1,017	509	1,526	993	553	1,546
Bannock	15,459	3,011	18,470	15,742	3,441	19,183
Bear Lake	2,563	886	3,449	2,598	978	3,576
Benewah	2,290	1,108	3,398	2,258	1,271	3,529
Bingham	7,451	2,436	9,887	7,601	3,301	10,902
Blaine	1,980	795	2,775	2,002	831	2,833
Boise	484	274	758	551	311	862
Bonner	5,868	2,306	8,174	5,666	2,432	8,098
Bonneville	12,570	3,340	15,910	13,110	3,804	16,914
Boundary	1,883	1,044	2,927	2,107	1,166	3,273
Butte	1,037	462	1,499	1,075	525	1,600
Camas	385	256	641	395	274	669
Canyon	19,345	5,367	24,712	20,041	6,673	26,714
Caribou	2,125	853	2,978	2,189	1,149	3,338
Cassia	4,732	1,759	6,491	4,755	2,261	7,016
Clark	437	202	639	496	225	721
Clearwater	2,894	1,173	4,067	3,059	1,355	4,414
Custer	1,000	622	1,622	1,010	618	1,628
Elmore	3,347	804	4,151	3,991	972	4,963
Franklin	3,707	1,089	4,796	3,638	1,407	5,045
Fremont	2,948	1,118	4,066	2,971	1,582	4,553
Gem	3,092	1,071	4,163	3,235	1,175	4,410
Gooding	3,490	1,425	4,915	3,433	1,562	4,995
Idaho	3,756	2,166	5,922	3,900	2,411	6,311
Jefferson	3,192	1,185	4,377	3,181	1,458	4,639
Jerome	3,902	1,726	5,628	3,966	1,919	5,885
Kootenai	11,347	3,027	14,374	11,619	3,414	15,033
Latah	6,922	2,486	9,408	6,875	2,891	9,766
Lemhi	1,982	995	2,977	2,060	1,022	3,082
Lewis	1,487	659	2,146	1,604	970	2,574
Lincoln	1,159	524	1,683	1,201	597	1,798
Madison	2,897	1,076	3,973	2,844	1,390	4,234
Minidoka	3,477	1,375	4,852	3,701	1,860	5,561
Nez Perce	9,784	2,917	12,701	10,038	3,408	13,446
Oneida	1,803	810	2,613	1,743	937	2,680
Owyhee	1,917	1,040	2,957	1,684	1,128	2,812
Payette	4,261	1,532	5,793	4,350	1,645	5,995
Power	1,301	597	1,898	1,376	1,023	2,399
Shoshone	8,345	1,875	10,220	7,964	1,961	9,925
Teton	971	428	1,399	988	517	1,505
Twin Falls	15,529	5,200	20,729	15,771	6,230	22,001
Valley	1,283	509	1,792	1,327	571	1,898
Washington	2,912	1,205	4,117	3,035	1,409	4,444
TOTALS	218,208	69,732	287,940	224,143	81,806	305,949

TABLE VI - M.V.R. BY COUNTY 75

SCHEDULE SHOWING BY COUNTIES TOTAL NUMBER OF AUTOMOBILES, TRUCKS, TRAILERS, MOTOR CYCLES AND DEALERS REGISTERED AND TRANSFERS MADE JANUARY 1 THRU DECEMBER 31, 1924. IN COMPARISON WITH REGISTRATIONS FOR YEAR 1923.

Counties	Automobiles 1924	Automobiles 1923	Trucks 1924	Trucks 1923	Trailers 1924	Trailers 1923	Motor Cycles 1924	Motor Cycles 1923	Dealers 1924	Dealers 1923	Transfers 1924	Transfers 1923
Abbeville	2,148	1,641	130	84	....	....	2	9	6	3	33	17
Aiken	4,151	3,182	415	321	11	14	16	12	10	10	75	68
Allendale	1,193	764	136	65	30	12	1	1	2	3	20	8
Anderson	8,512	7,046	730	629	3	1	12	13	37	33	250	202
Bamberg	1,465	1,143	159	112	31	17	1	....	10	7	21	15
Barnwell	1,729	1,212	164	116	6	4	5	6	9	9	30	12
Beaufort	928	755	214	159	....	1	1	2	6	6	24	20
Berkeley	821	548	108	72	23	18	....	....	1	1	38	16
Calhoun	1,070	806	113	71	31	9	1	5	3	3	21	13
Charleston	7,028	5,873	1,201	997	37	26	61	62	31	26	753	453
Cherokee	2,860	2,237	237	201	....	....	9	12	12	14	93	73
Chester	2,709	2,068	240	192	1	1	6	6	15	15	58	53
Chesterfield	3,278	2,573	361	235	7	5	5	4	14	15	83	70
Clarendon	1,671	1,267	162	115	21	9	3	4	6	5	38	27
Colleton	1,533	1,196	179	151	45	25	2	3	5	4	22	19
Darlington	3,662	2,990	302	263	15	15	6	11	19	24	96	90
Dillon	2,482	1,933	179	152	10	5	2	6	10	10	106	84
Dorchester	1,308	1,046	166	99	46	20	1	1	5	2	52	29
Edgefield	1,483	1,039	207	155	....	....	2	1	7	8	42	15
Fairfield	1,477	1,075	116	82	6	5	2	1	2	4	34	34
Florence	4,999	4,023	461	354	31	11	9	7	17	13	118	94
Georgetown	1,209	949	181	121	17	3	4	1	10	7	48	30
Greenville	12,889	10,167	1,400	1,044	4	3	45	66	69	39	445	305
Greenwood	3,612	3,012	416	366	2	4	14	11	12	13	49	60
Hampton	1,498	1,184	146	95	26	15	2	1	3	3	34	7
Horry	1,868	1,544	282	207	5	11	....	1	7	10	53	44
Jasper	477	389	45	40	10	8	....	....	2	1	6	3
Kershaw	2,447	1,878	287	230	2	3	4	4	8	14	62	71
Lancaster	1,960	1,603	169	108	3	6	7	10	12	7	37	38
Laurens	3,770	2,926	301	233	2	....	9	7	22	19	93	49
Lee	1,793	1,360	137	117	13	4	....	1	6	10	41	44
Lexington	4,006	3,024	481	349	33	32	9	12	11	7	199	135
McCormick	815	652	106	79	1	1	2	2	3	4	12	12
Marion	2,074	1,570	175	132	3	4	5	3	15	12	61	40
Marlboro	3,188	2,588	254	181	13	3	10	7	12	9	114	90
Newberry	3,196	2,606	378	265	12	19	10	9	9	17	114	101
Oconee	2,645	2,212	259	200	2	....	5	5	7	12	60	44
Orangeburg	5,078	3,953	550	416	150	98	9	9	23	21	112	75
Pickens	3,054	2,361	281	167	1	....	2	8	6	7	68	51
Richland	9,906	8,213	1,381	1,191	28	23	95	123	39	35	781	537
Saluda	1,375	1,008	169	115	6	3	2	2	4	1	24	15
Spartanburg	11,513	8,859	1,251	907	2	....	41	68	34	30	437	319
Sumter	3,371	2,708	372	280	30	6	21	11	17	15	105	71
Union	2,121	1,649	252	201	2	2	14	15	4	5	49	34
Williamsburg	1,622	1,256	154	105	19	9	1	2	7	6	18	28
York	5,033	4,148	387	322	2	....	17	15	23	25	173	187
Out of State	504	301	58	23	1	....	2	2	1	1	25	24
Total	147,530	116,537	15,852	12,119	743	455	477	561	593	545	5,227	3,826

SUMMARY

	1924	1923
Automobiles	147,530	116,537
Trucks	15,852	12,119
Trailers	743	455
Motor Cycles	477	561
Dealers	593	545
Demonstration Plates (Extra)	1,148	884
Transfers	5,227	3,826
Total Fees	$1,151,983.02	$902,608.69
Postage, etc.	12,169.83	9,888.04
Total Receipts	$1,164,152.85	$912,496.73

Digitized by Google
Original from UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
2019-10-25 17:34 GMT / http://hdl.handle.net/2027/uc1.b2968823 / http://www.hathitrust.org/access_use#pd-us-google
In the United States, Google-digitized

Non-homogeneity is an opportunity for LLMs

Standard digitization tools are more likely to stumble with non standard layouts

Even if table is successfully processed, homogenization is challenging

Homogenization is the core LLM capability (paraphrase)

The number of registrations issued by Counties is as follows:

	Letter.	Passenger Tags.	Commercial Registrations.	Omnibus Reg.	Total.
Atlantic,	A	20,433	5,783	738	26,954
Bergen,	B	55,576	8,668	580	64,824
Burlington,	N	16,643	3,578	123	20,344
Camden,	C	33,620	6,886	562	41,068
Cape May,	F	7,159	1,673	92	8,924
Cumberland,	Y	13,291	4,222	134	17,647
	E				
Essex,	1-E	115,295	23,042	2,933	141,270
	2-E				
Gloucester,	Z	13,146	3,248	139	16,533
Hudson,	H	58,559	13,674	1,528	73,761
Hunterdon,	J	7,236	1,767	79	9,082
Mercer,	L	27,803	6,710	438	34,951
Middlesex,	K	28,937	5,116	293	34,346
Monmouth,	M	33,665	8,163	558	42,386
Morris,	V	21,052	3,088	272	24,412
Ocean,	O	8.741	2,183	180	11,104
	N				
Passaic,	P	39,803	9.538	634	49,975
Salem,	S	7,185	2,022	30	9,237
Somerset,	I	11,049	2,201	135	13,385
Sussex,	R	6,413	1,608	76	8,097
Union,	U	42,529	7,592	449	50,570
Warren,	W	9,098	2,018	98	11,214
Non-Resident,	4	7,432	2,584	306	10,322
Special,		2	.....	.....	2
		*584,667	125,364	10,377	†720,408

*Passenger registrations, full rate,	531,689
Passenger registrations, half rate,	44,444
Duplicate tags,	7,520
Exchanges,	415
No fee registrations,	77
Undertaker registrations,	522
	584,667

† This includes duplicate tags, no fee registrations, and exchanges.

Table VI Motor Vehicle License Registration by Counties

1953 and 1954

County	1953 Autos	1953 Trucks	1953 Total	1954 Autos	1954 Trucks	1954 Total
Ada	29,877	6,490	36,367	32,000	7,179	39,179
Adams	1,017	509	1,526	993	553	1,546
Bannock	15,459	3,011	18,470	15,742	3,441	19,183
Bear Lake	2,563	886	3,449	2,598	978	3,576
Benewah	2,290	1,108	3,398	2,258	1,271	3,529
Bingham	7,451	2,436	9,887	7,601	3,301	10,902
Blaine	1,980	795	2,775	2,002	831	2,833
Boise	484	274	758	551	311	862
Bonner	5,868	2,306	8,174	5,666	2,432	8,098
Bonneville	12,570	3,340	15,910	13,110	3,804	16,914
Boundary	1,883	1,044	2,927	2,107	1,166	3,273
Butte	1,037	462	1,499	1,075	525	1,600
Camas	385	256	641	395	274	669
Canyon	19,345	5,367	24,712	20,041	6,673	26,714
Caribou	2,125	853	2,978	2,189	1,149	3,338
Cassia	4,732	1,759	6,491	4,755	2,261	7,016
Clark	437	202	639	496	225	721
Clearwater	2,894	1,173	4,067	3,059	1,355	4,414
Custer	1,000	622	1,622	1,010	618	1,628
Elmore	3,347	804	4,151	3,991	972	4,963
Franklin	3,707	1,089	4,796	3,638	1,407	5,045
Fremont	2,948	1,118	4,066	2,971	1,582	4,553
Gem	3,092	1,071	4,163	3,235	1,175	4,410
Gooding	3,490	1,425	4,915	3,433	1,562	4,995
Idaho	3,756	2,166	5,922	3,900	2,411	6,311
Jefferson	3,192	1,185	4,377	3,181	1,458	4,639
Jerome	3,902	1,726	5,628	3,966	1,919	5,885
Kootenai	11,347	3,027	14,374	11,619	3,414	15,033
Latah	6,922	2,486	9,408	6,875	2,891	9,766
Lemhi	1,982	995	2,977	2,060	1,022	3,082
Lewis	1,487	659	2,146	1,604	970	2,574
Lincoln	1,159	524	1,683	1,201	597	1,798
Madison	2,897	1,076	3,973	2,844	1,390	4,234
Minidoka	3,477	1,375	4,852	3,701	1,860	5,561
Nez Perce	9,784	2,917	12,701	10,038	3,408	13,446
Oneida	1,803	810	2,613	1,743	937	2,680
Owyhee	1,917	1,040	2,957	1,684	1,128	2,812
Payette	4,261	1,532	5,793	4,350	1,645	5,995
Power	1,301	597	1,898	1,376	1,023	2,399
Shoshone	8,345	1,875	10,220	7,964	1,961	9,925
Teton	971	428	1,399	988	517	1,505
Twin Falls	15,529	5,200	20,729	15,771	6,230	22,001
Valley	1,283	509	1,792	1,327	571	1,898
Washington	2,912	1,205	4,117	3,035	1,409	4,444
TOTALS	218,208	69,732	287,940	224,143	81,806	305,949

TABLE VI - M.V.R. BY COUNTY 75

SCHEDULE SHOWING BY COUNTIES TOTAL NUMBER OF AUTOMOBILES, TRUCKS, TRAILERS, MOTOR CYCLES AND DEALERS REGISTERED AND TRANSFERS MADE JANUARY 1 THRU DECEMBER 31, 1924. IN COMPARISON WITH REGISTRATIONS FOR YEAR 1923.

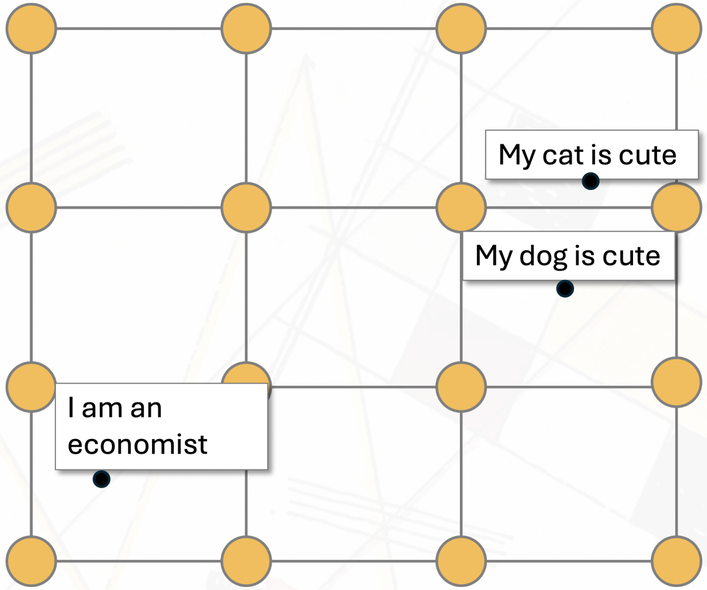
Counties	Automobiles 1924	Automobiles 1923	Trucks 1924	Trucks 1923	Trailers 1924	Trailers 1923	Motor Cycles 1924	Motor Cycles 1923	Dealers 1924	Dealers 1923	Transfers 1924	Transfers 1923
	2,148	1,641	130	84	....	....	2	9	6	3	33	17
	4,151	3,182	415	321	11	14	16	12	10	10	75	68
	1,193	764	136	65	30	12	1	1	2	3	20	8
	8,512	7,046	730	629	3	1	12	13	37	33	250	202
	1,465	1,143	159	112	31	17	1	....	10	7	21	15
	1,729	1,212	164	116	6	4	5	6	9	9	30	12
	928	755	214	159	....	1	1	2	6	6	24	20
	821	548	108	72	23	18	....	....	1	1	38	16
	1,070	806	113	71	31	9	1	5	3	3	21	13
Charleston	7,028	5,873	1,201	997	37	26	61	62	31	26	753	453
Cherokee	2,860	2,237	237	201	....	....	9	12	12	14	93	73
Chester	2,709	2,068	240	192	1	1	6	6	15	15	58	53
Chesterfield	3,278	2,573	361	235	7	5	5	4	14	15	83	70
Clarendon	1,671	1,267	162	115	21	9	3	4	6	5	38	27
Colleton	1,533	1,196	179	151	45	25	2	3	5	4	22	19
Darlington	3,662	2,990	302	263	15	15	6	11	19	24	96	90
Dillon	2,482	1,933	179	152	10	5	2	6	10	10	106	84
Dorchester	1,308	1,046	166	99	46	20	1	1	5	2	52	29
Edgefield	1,483	1,039	207	155	....	....	2	1	7	8	42	15
Fairfield	1,477	1,075	116	82	6	5	2	1	2	4	34	34
Florence	4,999	4,023	461	354	31	11	9	7	17	13	118	94
Georgetown	1,209	949	181	121	17	3	4	1	10	7	48	30
Greenville	12,889	10,167	1,400	1,044	4	3	45	66	69	39	445	305
Greenwood	3,612	3,012	416	366	2	4	14	11	12	13	49	60
Hampton	1,498	1,184	146	95	26	15	2	1	3	3	34	7
Horry	1,868	1,544	282	207	5	11	....	1	7	10	53	44
Jasper	477	389	45	40	10	8	....	....	2	1	6	3
Kershaw	2,447	1,878	287	230	2	3	4	4	8	14	62	71
Lancaster	1,960	1,603	169	108	3	6	7	10	12	7	37	38
Laurens	3,770	2,926	301	233	2	....	9	7	22	19	93	49
Lee	1,793	1,360	137	117	13	4	....	1	6	10	41	44
Lexington	4,006	3,024	481	349	33	32	9	12	11	7	199	135
McCormick	815	652	106	79	1	1	2	2	3	4	12	12
Marion	2,074	1,570	175	132	3	4	5	3	15	12	61	40
Marlboro	3,188	2,588	254	181	13	3	10	7	12	9	114	90
Newberry	3,196	2,606	378	265	12	19	10	9	9	17	114	101
Oconee	2,645	2,212	259	200	2	....	5	5	7	12	60	44
Orangeburg	5,078	3,953	550	416	150	98	9	9	23	21	112	75
Pickens	3,054	2,361	281	167	1	....	2	8	6	7	68	51
Richland	9,906	8,213	1,381	1,191	28	23	95	123	39	35	781	537
Saluda	1,375	1,008	169	115	6	3	2	2	4	1	24	15
Spartanburg	11,513	8,859	1,251	907	2	....	41	68	34	30	437	319
Sumter	3,371	2,708	372	280	30	6	21	11	17	15	105	71
Union	2,121	1,649	252	201	2	2	14	15	4	5	49	34
Williamsburg	1,622	1,256	154	105	19	9	1	2	7	6	18	28
York	5,033	4,148	387	322	2	....	17	15	23	25	173	187
Out of State	504	301	58	23	1	....	2	2	1	1	25	24
Total	147,530	116,537	15,852	12,119	743	455	477	561	593	545	5,227	3,826

SUMMARY

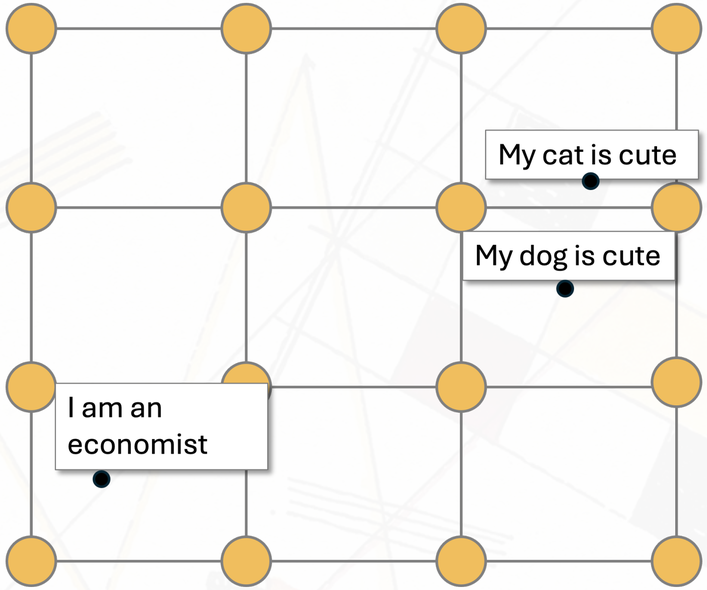
	1924	1923
Automobiles	147,530	116,537
Trucks	15,852	12,119
Trailers	743	455
Motor Cycles	477	561
Dealers	593	545
Demonstration Plates (Extra)	1,148	884
Transfers	5,227	3,826
Total Fees	$1,151,983.02	$902,608.69
Postage, etc.	12,169.83	9,888.04
Total Receipts	$1,164,152.85	$912,496.73

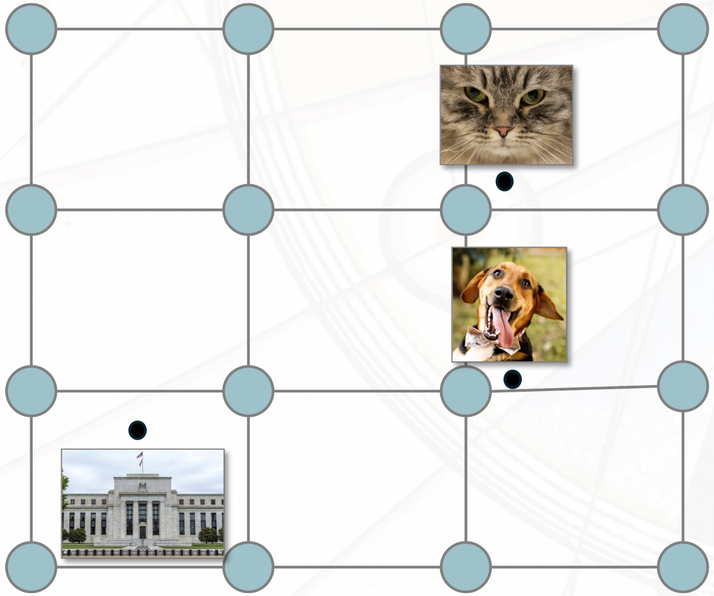
Digitized by Google
Original from UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
2019-10-25 17:34 GMT / http://hdl.handle.net/2027/uc1.b2968823 / http://www.hathitrust.org/access_use#pd-us-google
In the United States, Google-digitized

How does a (multimodal) LLM know how to process tables?



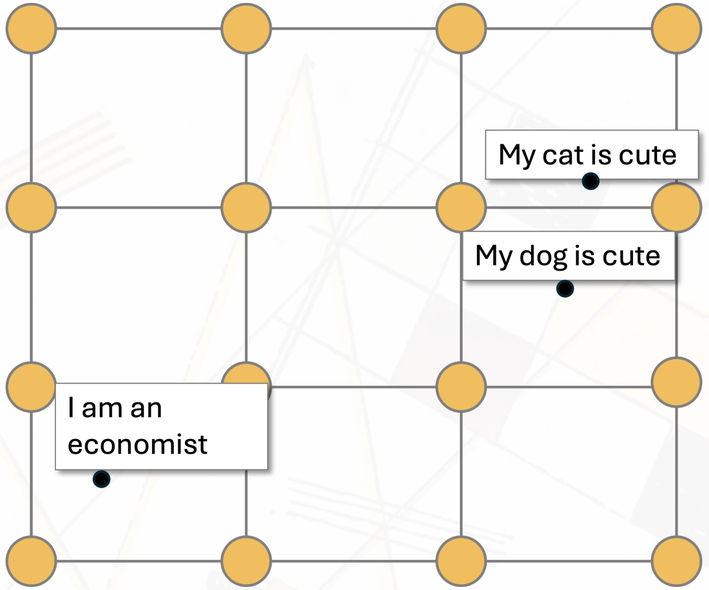
How does a (multimodal) LLM know how to process tables?

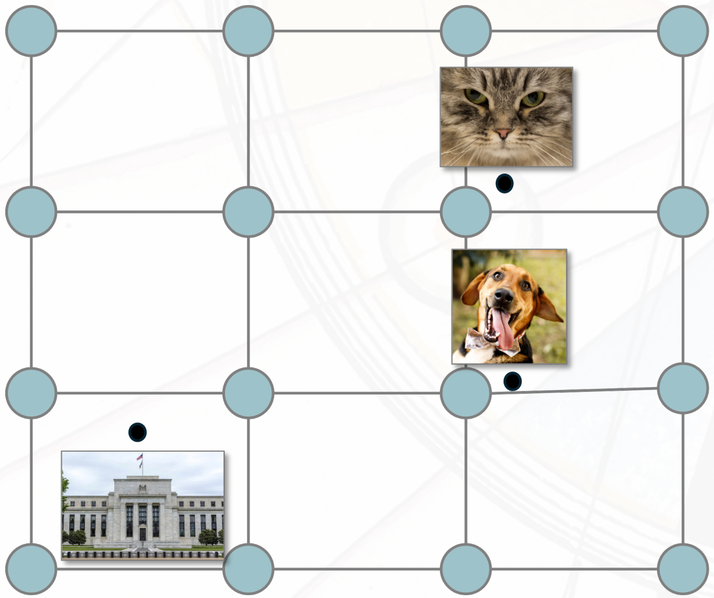




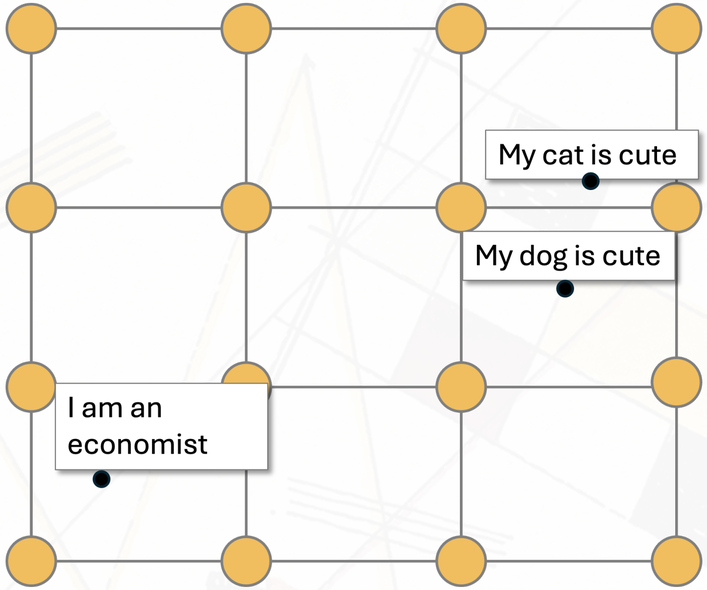
How does a (multimodal) LLM know how to process tables?

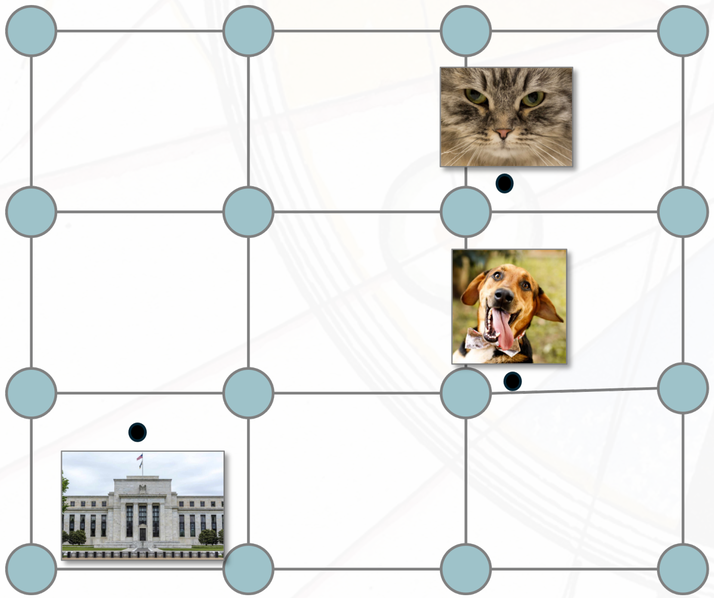
How does a (multimodal) LLM know how to process tables?

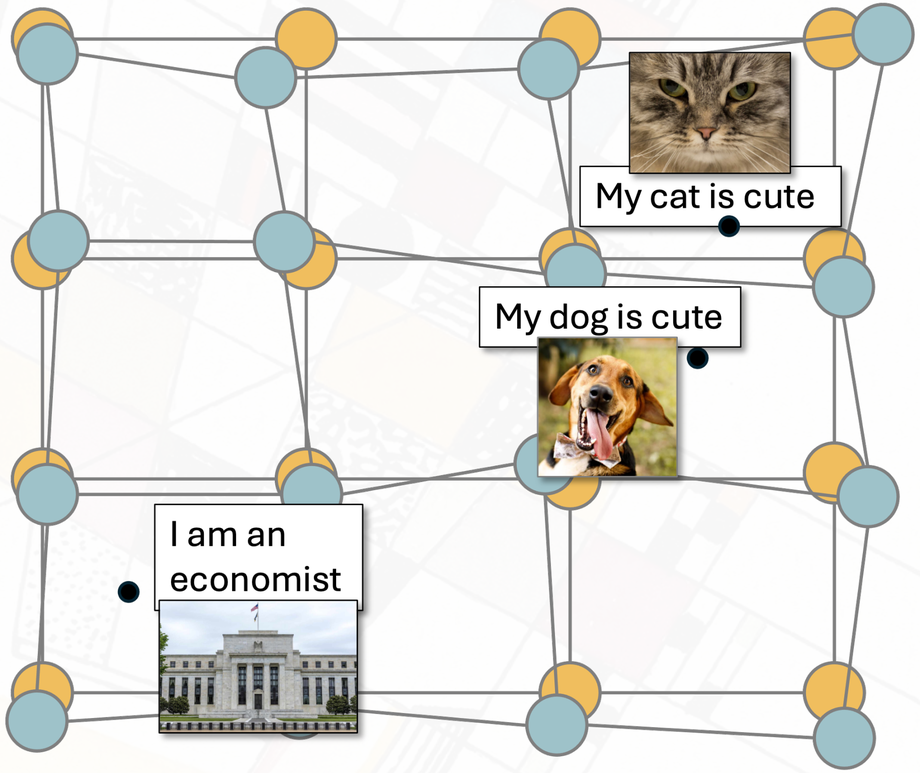




How does a (multimodal) LLM know how to process tables?







Mixed text + image data is plentiful, including table OCR

Mixed text + image data is plentiful, including table OCR

My cat is cute



My dog is cute



Mixed text + image data is plentiful, including table OCR






```
State,Population
California,39538223
Texas,29145505
Florida,21538187
New York,20201249
Pennsylvania,13002700
Wyoming,576851
Vermont,643077
Alaska,733391
North Dakota,779094
South Dakota,886667
```

Table 1.
States with the Largest and Smallest Resident Population: 2020 Census

State	Population
Largest Population	
California	39,538,223
Texas	29,145,505
Florida	21,538,187
New York	20,201,249
Pennsylvania	13,002,700
Smallest Population	
Wyoming	576,851
Vermont	643,077
Alaska	733,391
North Dakota	779,094
South Dakota	886,667

Source: U.S. Census Bureau, 2020 Census

We use Claude 3.5 Sonnet

Expected bill for the full project (~5,000 tables) is <$1,000

We use Claude 3.5 Sonnet

Claude 3 Opus works the same but 5x more expensive

Expected bill for the full project (~5,000 tables) is <$1,000

We use Claude 3.5 Sonnet

Claude 3 Opus works the same but 5x more expensive

Expected bill for the full project (~5,000 tables) is <$1,000

We use Claude 3.5 Sonnet

GPT-4o and Gemini 1.5 worked poorly

Claude 3 Opus works the same but 5x more expensive

Expected bill for the full project (~5,000 tables) is <$1,000

We use Claude 3.5 Sonnet

GPT-4o and Gemini 1.5 worked poorly

We didn’t try anything that came after

Functional Steps

Scans

Dataset

Functional Steps

Scans

Image layout parsing

Dataset

Functional Steps

Scans

Image layout parsing

Table layout parsing

Dataset

Functional Steps

Scans

Image layout parsing

Table layout parsing

OCR

Dataset

Functional Steps

Scans

Image layout parsing

Table layout parsing

OCR

Alignment

Dataset

Deep Learning Pipeline

Functional Steps

Scans

Image layout parsing

Table layout parsing

OCR

Alignment

Dataset

Deep Learning Pipeline	Functional Steps
	Scans
Layout parser	Image layout parsing
	Table layout parsing
	OCR
	Alignment
	Dataset

Deep Learning Pipeline	Functional Steps
	Scans
Layout parser	Image layout parsing
Layout parser	Table layout parsing
	OCR
	Alignment
	Dataset

Deep Learning Pipeline	Functional Steps
	Scans
Layout parser	Image layout parsing
Layout parser	Table layout parsing
OCR model	OCR
	Alignment
	Dataset

Deep Learning Pipeline	Functional Steps
	Scans
Layout parser	Image layout parsing
Layout parser	Table layout parsing
OCR model	OCR
Custom logic/NLP	Alignment
	Dataset

Deep Learning Pipeline	Functional Steps	LLM Pipeline
	Scans	
Layout parser	Image layout parsing	
Layout parser	Table layout parsing	
OCR model	OCR	
Custom logic/NLP	Alignment	
	Dataset	

Deep Learning Pipeline	Functional Steps	LLM Pipeline
	Scans	
Layout parser	Image layout parsing	Layout parser (a little)
Layout parser	Table layout parsing	
OCR model	OCR	
Custom logic/NLP	Alignment	
	Dataset	

Deep Learning Pipeline	Functional Steps	LLM Pipeline
	Scans	
Layout parser	Image layout parsing	Layout parser (a little)
Layout parser	Table layout parsing	LLM
OCR model	OCR	
Custom logic/NLP	Alignment	
	Dataset	

Deep Learning Pipeline	Functional Steps	LLM Pipeline
	Scans	
Layout parser	Image layout parsing	Layout parser (a little)
Layout parser	Table layout parsing	LLM
OCR model	OCR	LLM
Custom logic/NLP	Alignment	
	Dataset	

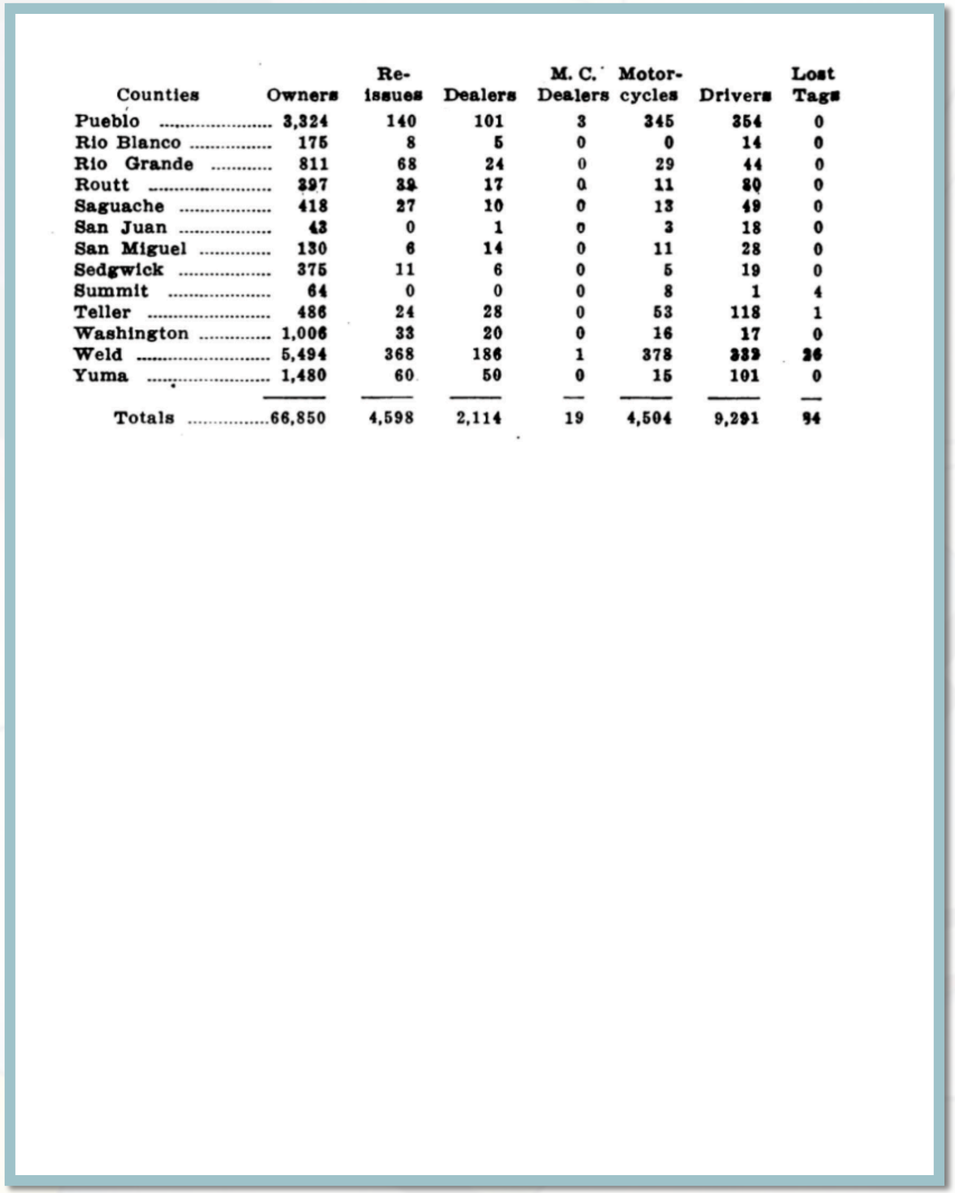
Deep Learning Pipeline	Functional Steps	LLM Pipeline
	Scans	
Layout parser	Image layout parsing	Layout parser (a little)
Layout parser	Table layout parsing	LLM
OCR model	OCR	LLM
Custom logic/NLP	Alignment	LLM
	Dataset	

Deep Learning Pipeline	Functional Steps	LLM Pipeline
	Scans	
Layout parser	Image layout parsing	Layout parser (a little)
Layout parser	Table layout parsing	LLM
OCR model	OCR	LLM
Custom logic/NLP	Alignment	LLM
	Dataset	

Requires tinkering in non-standard cases, steeper learning curve

Deep Learning Pipeline	Functional Steps	LLM Pipeline
	Scans	
Layout parser	Image layout parsing	Layout parser (a little)
Layout parser	Table layout parsing	LLM
OCR model	OCR	LLM
Custom logic/NLP	Alignment	LLM
	Dataset	
Requires tinkering in non-standard cases, steeper learning curve		Less tinkering and it’s done in English, flatter learning curve

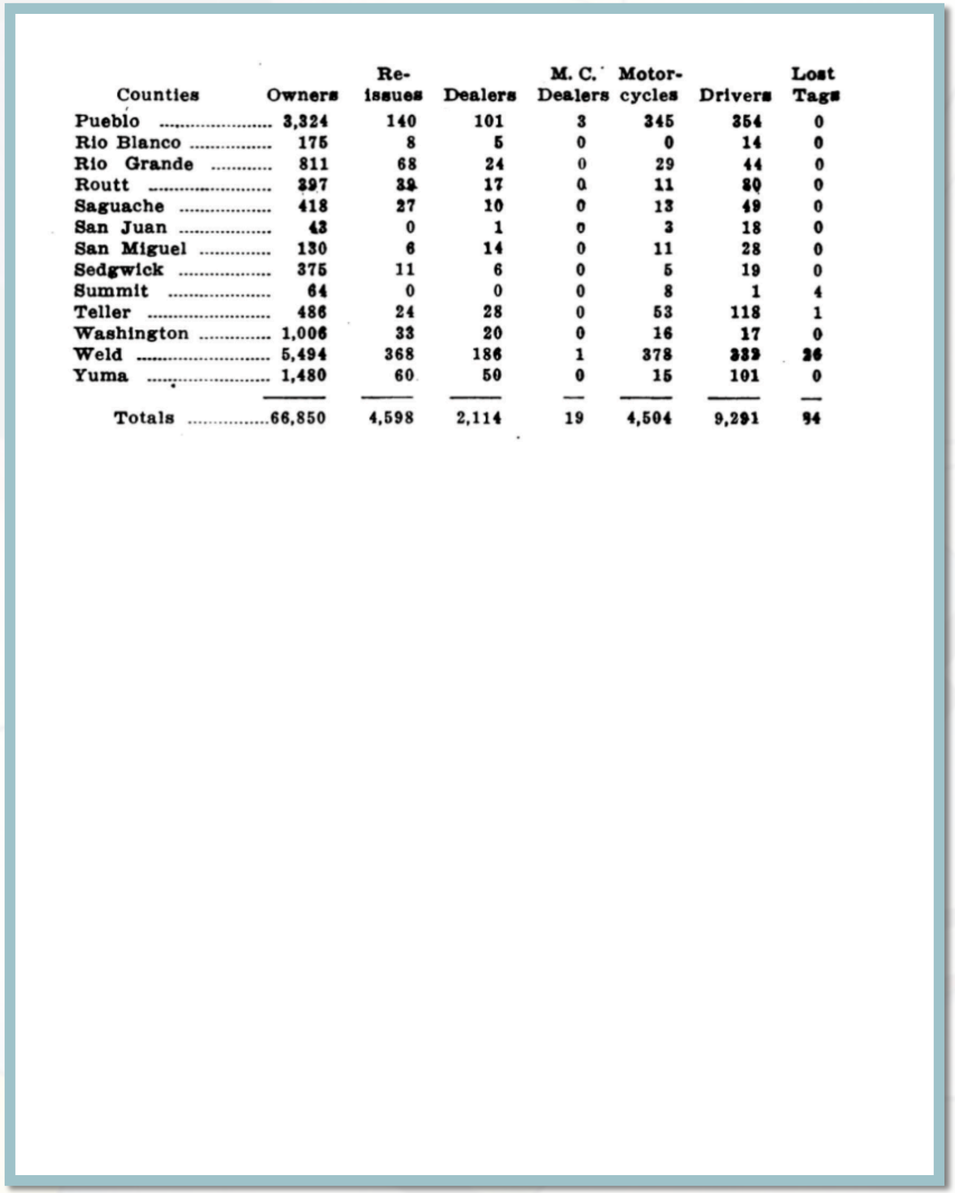
Step 1: Image Layout Parsing



Counties	Owners	Re-issues	Dealers	M. C. Dealers	Motor-cycles	Drivers	Lost Tags
Pueblo	3,324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	38	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1,006	33	20	0	16	17	0
Weld	5,494	368	186	1	378	333	26
Yuma	1,480	60	50	0	15	101	0
Totals	66,850	4,598	2,114	19	4,504	9,291	94

Raw image (Colorado, 1917)

Step 1: Image Layout Parsing



Counties	Owners	Re-issues	Dealers	M. C. Dealers	Motor-cycles	Drivers	Lost Tags
Pueblo	3,324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	39	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1,006	33	20	0	16	17	0
Weld	5,494	368	186	1	378	333	26
Yuma	1,480	60	50	0	15	101	0
Totals	66,850	4,598	2,114	19	4,504	9,291	94

Raw image (Colorado, 1917)

Step 1: Image Layout Parsing

Table 1

Counties	Owners	Re-issues	Dealers	M. C. Dealers	Motor-cycles	Drivers	Lost Tags
Pueblo	3,324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	39	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1,006	33	20	0	16	17	0
Weld	5,494	368	186	1	378	333	26
Yuma	1,480	60	50	0	15	101	0
Totals	66,850	4,598	2,114	19	4,504	9,291	94

Cropped image

Step 1: Image Layout Parsing (Special Case)

Step 1: Image Layout Parsing (Special Case)

County Registrations of States for 1919

Kansas

(Compiled by Copper Farm Motor from Kansas State License Bureau records)

Table 2

County	Cars and Trucks
Allen	2,738
Anderson	1,577
Atchison	2,414
Barber	1,466
Barton	2,782
Bourbon	2,360
Brown	3,315
Butler	7,376
Chase	1,076
Chautauqua	1,678
Cherokee	3,049
Cheyenne	1,088
Clark	787
Clay	2,462
Cloud	2,444
Coffey	1,946
Comanche	867
Cowley	4,781
Crawford	5,220
Decatur	1,254
Dickinson	4,526
Doniphan	1,761
Douglas	2,973
Edwards	1,278
Elk	1,080
Ellis	1,196
Ellsworth	1,571
Finney	960
Ford	1,823
Franklin	2,643
Geary	1,748
Gove	549
Graham	914
Grant	169
Gray	648
Greeley	187

Table 1

County	Cars and Trucks
Greenwood	1,822
Hamilton	306
Harper	2,270
Harvey	2,774
Haskell	208
Hodgeman	579
Jackson	2,141
Jefferson	2,039
Jewell	2,604
Johnson	2,399
Kearney	339
Kingman	2,083
Kiowa	1,003
Labette	3,344
Lane	435
Leavenworth	2,416
Lincoln	1,662
Linn	1,577
Logan	463
Lyon	3,106
McPherson	4,008
Marion	3,710
Marshall	4,008
Meade	804
Miami	2,646
Mitchell	2,187
Montgomery	5,896
Morris	1,793
Morton	389
Nemaha	2,984
Neosho	2,985
Ness	1,104
Norton	1,568
Osage	2,488
Osborne	2,057
Ottawa	1,649

Table 3

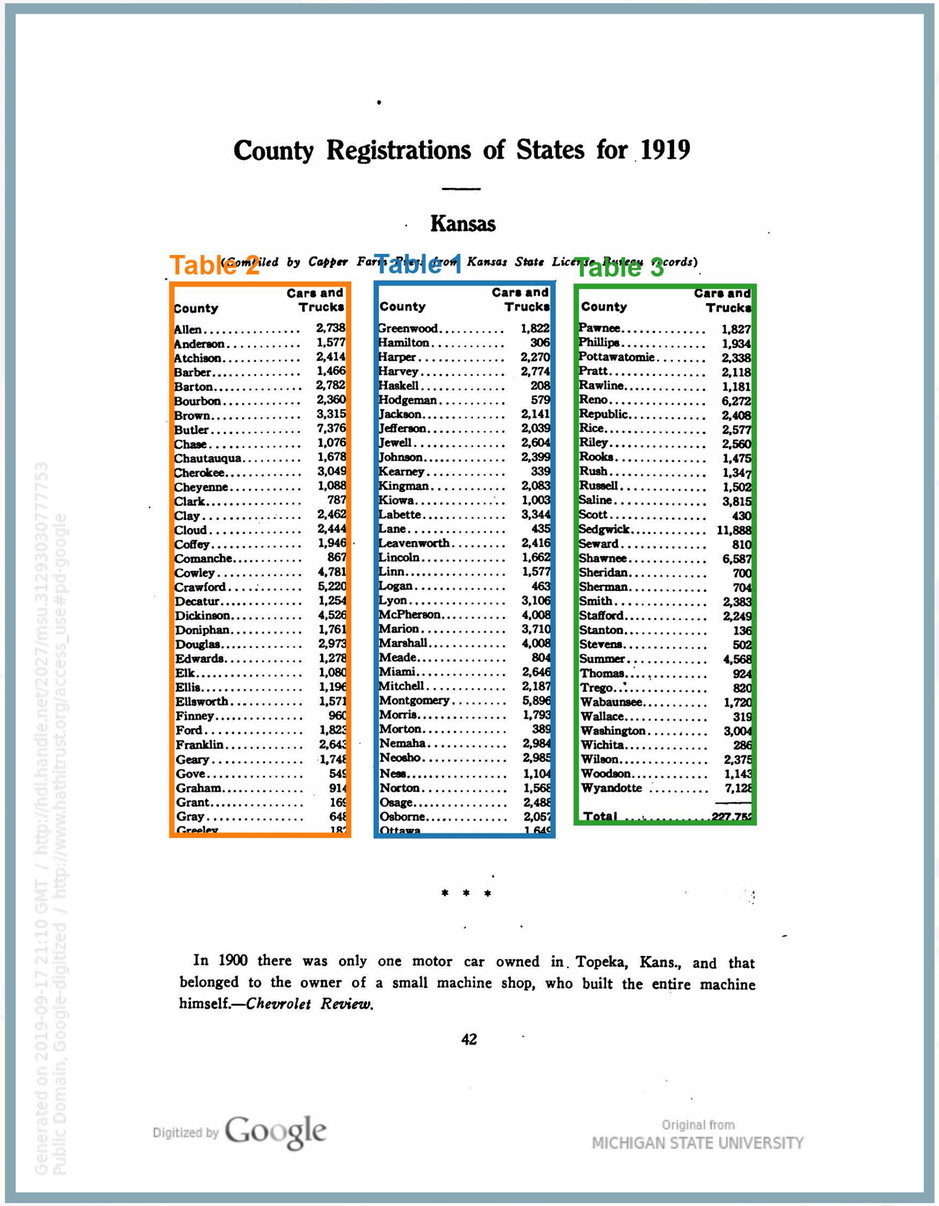
County	Cars and Trucks
Pawnee	1,827
Phillips	1,934
Pottawatomie	2,338
Pratt	2,118
Rawline	1,181
Reno	6,272
Republic	2,408
Rice	2,577
Riley	2,560
Rooks	1,475
Rush	1,347
Russell	1,502
Saline	3,815
Scott	430
Sedgwick	11,888
Seward	810
Shawnee	6,587
Sheridan	700
Sherman	704
Smith	2,383
Stafford	2,249
Stanton	136
Stevens	502
Summer	4,568
Thomas	924
Trego	820
Wabaunsee	1,720
Wallace	319
Washington	3,004
Wichita	286
Wilson	2,375
Woodson	1,143
Wyandotte	7,128
Total	227,755

* * *

In 1900 there was only one motor car owned in Topeka, Kans., and that belonged to the owner of a small machine shop, who built the entire machine himself.—*Chevrolet Review.*

Digitized by Google
Original from MICHIGAN STATE UNIVERSITY
Generated on 2019-09-17 21:10 GMT / http://hdl.handle.net/2027/msu.31293030777753
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Step 1: Image Layout Parsing (Special Case)



County Registrations of States for 1919

Kansas

(Compiled by Copper Farm Motor from Kansas State License Bureau records)

Table 2

County	Cars and Trucks
Allen	2,738
Anderson	1,577
Atchison	2,414
Barber	1,466
Barton	2,782
Bourbon	2,360
Brown	3,315
Butler	7,376
Chase	1,076
Chautauqua	1,678
Cherokee	3,049
Cheyenne	1,088
Clark	787
Clay	2,462
Cloud	2,444
Coffey	1,946
Comanche	867
Cowley	4,781
Crawford	5,220
Decatur	1,254
Dickinson	4,526
Doniphan	1,761
Douglas	2,973
Edwards	1,278
Elk	1,080
Ellis	1,196
Ellsworth	1,571
Finney	960
Ford	1,823
Franklin	2,643
Geary	1,748
Gove	549
Graham	914
Grant	169
Gray	648
Greeley	187

Table 1

County	Cars and Trucks
Greenwood	1,822
Hamilton	306
Harper	2,270
Harvey	2,774
Haskell	208
Hodgeman	579
Jackson	2,141
Jefferson	2,039
Jewell	2,604
Johnson	2,399
Kearney	339
Kingman	2,083
Kiowa	1,003
Labette	3,344
Lane	435
Leavenworth	2,416
Lincoln	1,662
Linn	1,577
Logan	463
Lyon	3,106
McPherson	4,008
Marion	3,710
Marshall	4,008
Meade	804
Miami	2,646
Mitchell	2,187
Montgomery	5,896
Morris	1,793
Morton	389
Nemaha	2,984
Neosho	2,985
Ness	1,104
Norton	1,568
Osage	2,488
Osborne	2,057
Ottawa	1,649

Table 3

County	Cars and Trucks
Pawnee	1,827
Phillips	1,934
Pottawatomie	2,338
Pratt	2,118
Rawline	1,181
Reno	6,272
Republic	2,408
Rice	2,577
Riley	2,560
Rooks	1,475
Rush	1,347
Russell	1,502
Saline	3,815
Scott	430
Sedgwick	11,888
Seward	810
Shawnee	6,587
Sheridan	700
Sherman	704
Smith	2,383
Stafford	2,249
Stanton	136
Stevens	502
Summer	4,568
Thomas	924
Trego	820
Wabaunsee	1,720
Wallace	319
Washington	3,004
Wichita	286
Wilson	2,375
Woodson	1,143
Wyandotte	7,128
Total	227,755

* * *

In 1900 there was only one motor car owned in Topeka, Kans., and that belonged to the owner of a small machine shop, who built the entire machine himself.—*Chevrolet Review.*

Digitized by Google
Original from MICHIGAN STATE UNIVERSITY
Generated on 2019-09-17 21:10 GMT / http://hdl.handle.net/2027/msu.31293030777753
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Prompt: The image you see is a scanned record containing historical data on vehicle registrations. It contains {num_tables} tables, identified in different colors and labelled as {table_list}. You want to identify if any of the tables can be stacked on top of each other. Tables that can be stacked have the following traits:\n-They have the same number of columns\n-Their columns refer to the same thing.\nPlease output a list containing the numbers of the tables which can be stacked, in the order that they should be stacked. If none of the tables should be combined, output a blank list.

Step 1: Image Layout Parsing (Special Case)

County Registrations of States for 1919

Kansas

(Compiled by Copper Farm Motor from Kansas State Licenses Bureau records)

Table 2

County	Cars and Trucks
Allen	2,738
Anderson	1,577
Atchison	2,414
Barber	1,466
Barton	2,782
Bourbon	2,360
Brown	3,315
Butler	7,376
Chase	1,076
Chautauqua	1,678
Cherokee	3,049
Cheyenne	1,088
Clark	787
Clay	2,462
Cloud	2,444
Coffey	1,946
Comanche	867
Cowley	4,781
Crawford	5,220
Decatur	1,254
Dickinson	4,526
Doniphan	1,761
Douglas	2,973
Edwards	1,278
Elk	1,080
Ellis	1,196
Ellsworth	1,571
Finney	960
Ford	1,823
Franklin	2,643
Geary	1,748
Gove	549
Graham	914
Grant	169
Gray	648
Greeley	183

Table 1

County	Cars and Trucks
Greenwood	1,822
Hamilton	306
Harper	2,270
Harvey	2,774
Haskell	208
Hodgeman	579
Jackson	2,141
Jefferson	2,039
Jewell	2,604
Johnson	2,399
Kearney	339
Kingman	2,083
Kiowa	1,003
Labette	3,344
Lane	435
Leavenworth	2,416
Lincoln	1,662
Linn	1,577
Logan	463
Lyon	3,106
McPherson	4,008
Marion	3,710
Marshall	4,008
Meade	804
Miami	2,646
Mitchell	2,187
Montgomery	5,896
Morris	1,793
Morton	389
Nemaha	2,984
Neosho	2,985
Ness	1,104
Norton	1,568
Osage	2,488
Osborne	2,057
Ottawa	1,649

Table 3

County	Cars and Trucks
Pawnee	1,827
Phillips	1,934
Pottawatomie	2,338
Pratt	2,118
Rawline	1,181
Reno	6,272
Republic	2,408
Rice	2,577
Riley	2,560
Rooks	1,475
Rush	1,347
Russell	1,502
Saline	3,815
Scott	430
Sedgwick	11,888
Seward	810
Shawnee	6,587
Sheridan	700
Sherman	704
Smith	2,383
Stafford	2,249
Stanton	136
Stevens	502
Summer	4,568
Thomas	924
Trego	820
Wabaunsee	1,720
Wallace	319
Washington	3,004
Wichita	286
Wilson	2,375
Woodson	1,143
Wyandotte	7,128
Total	227,752

* * *

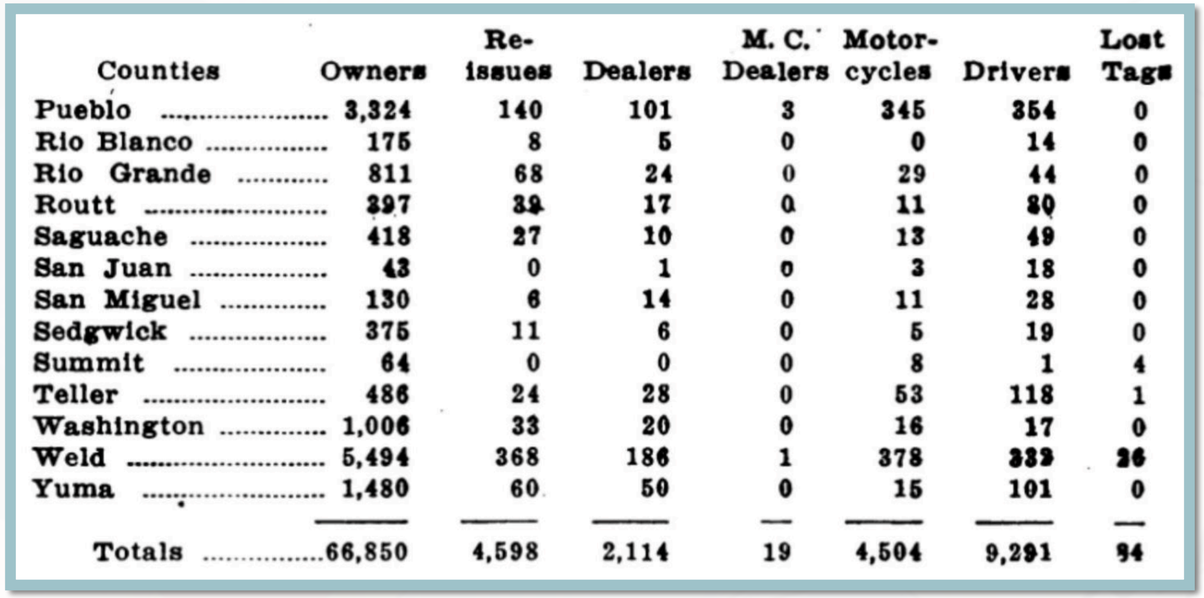
In 1900 there was only one motor car owned in Topeka, Kans., and that belonged to the owner of a small machine shop, who built the entire machine himself.—*Chevrolet Review.*

Digitized by Google
Original from MICHIGAN STATE UNIVERSITY
Generated on 2019-09-17 21:10 GMT / http://hdl.handle.net/2027/msu.31293030777753
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Keep all three tables

Prompt: The image you see is a scanned record containing historical data on vehicle registrations. It contains {num_tables} tables, identified in different colors and labelled as {table_list}. You want to identify if any of the tables can be stacked on top of each other. Tables that can be stacked have the following traits:\n-They have the same number of columns\n-Their columns refer to the same thing.\nPlease output a list containing the numbers of the tables which can be stacked, in the order that they should be stacked. If none of the tables should be combined, output a blank list.

Step 2: Table Layout Parsing + OCR with LLM



Counties	Owners	Re-issues	Dealers	M. C. Dealers	Motor-cycles	Drivers	Lost Tags
Pueblo	3,324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	32	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1,006	33	20	0	16	17	0
Weld	5,494	368	186	1	378	332	26
Yuma	1,480	60	50	0	15	101	0
Totals	66,850	4,598	2,114	19	4,504	9,291	94

Cropped image

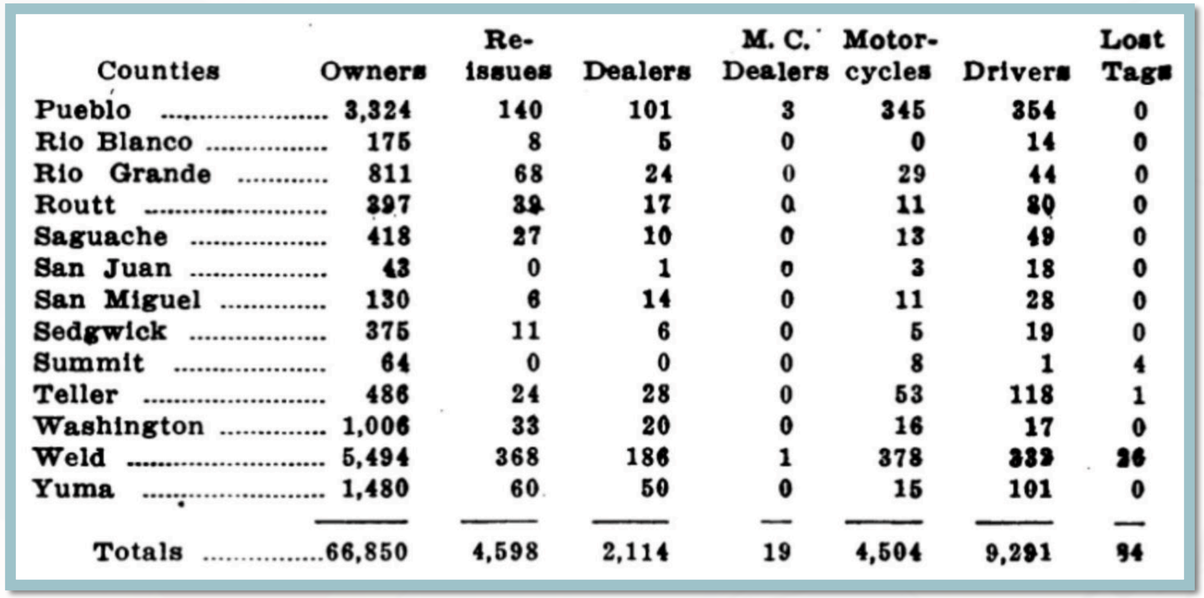
Step 2: Table Layout Parsing + OCR with LLM

Counties	Owners	Re-issues	Dealers	M. C. Dealers	Motor-cycles	Drivers	Lost Tags
Pueblo	3,324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	32	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1,006	33	20	0	16	17	0
Weld	5,494	368	186	1	378	332	26
Yuma	1,480	60	50	0	15	101	0
Totals	66,850	4,598	2,114	19	4,504	9,291	94

Cropped image

Step 2: Table Layout Parsing + OCR with LLM

Prompt: You are a researcher who carefully digitizes historical statistical tables. You look at scans from old books and put the tables in csv files. The tables can have multiple headers, make sure you merge them. Don't make up numbers because those are very important for your research. [TINKERING]. Output a csv table. Don't output any other text. Image to process is attached.



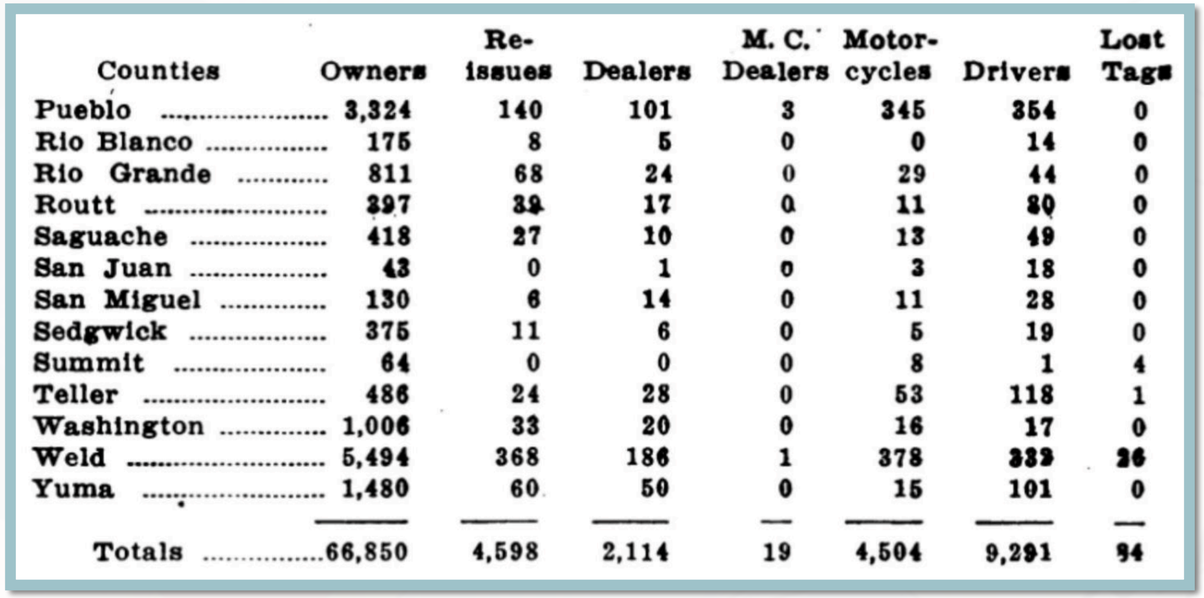
Counties	Owners	Re-issues	Dealers	M. C. Dealers	Motor-cycles	Drivers	Lost Tags
Pueblo	3,324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	32	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1,006	33	20	0	16	17	0
Weld	5,494	368	186	1	378	332	26
Yuma	1,480	60	50	0	15	101	0
Totals	66,850	4,598	2,114	19	4,504	9,291	94

Cropped image

Step 2: Table Layout Parsing + OCR with LLM

Prompt: You are a researcher who carefully digitizes historical statistical tables. You look at scans from old books and put the tables in csv files. The tables can have multiple headers, make sure you merge them. Don’t make up numbers because those are very important for your research. [TINKERING]. Output a csv table. Don’t output any other text. Image to process is attached.

Tinkering examples:
- The table might have multiple years. If it does, this means that likely there multiple headers for vehicle type and year. You want the columns in the output to look something like this: VehicleType_year
- The years can appear in any order.
- Empty rows in the image can be represented by multiple dots, encode them as a single blank space, don't add extra columns.
- If the column represents vehicle counts or other counts, it can't have decimal points. Only $ amount columns can have decimal points.



Counties	Owners	Re-issues	Dealers	M. C. Dealers	Motor-cycles	Drivers	Lost Tags
Pueblo	3,324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	32	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1,006	33	20	0	16	17	0
Weld	5,494	368	186	1	378	332	26
Yuma	1,480	60	50	0	15	101	0
Totals	66,850	4,598	2,114	19	4,504	9,291	94

Cropped image

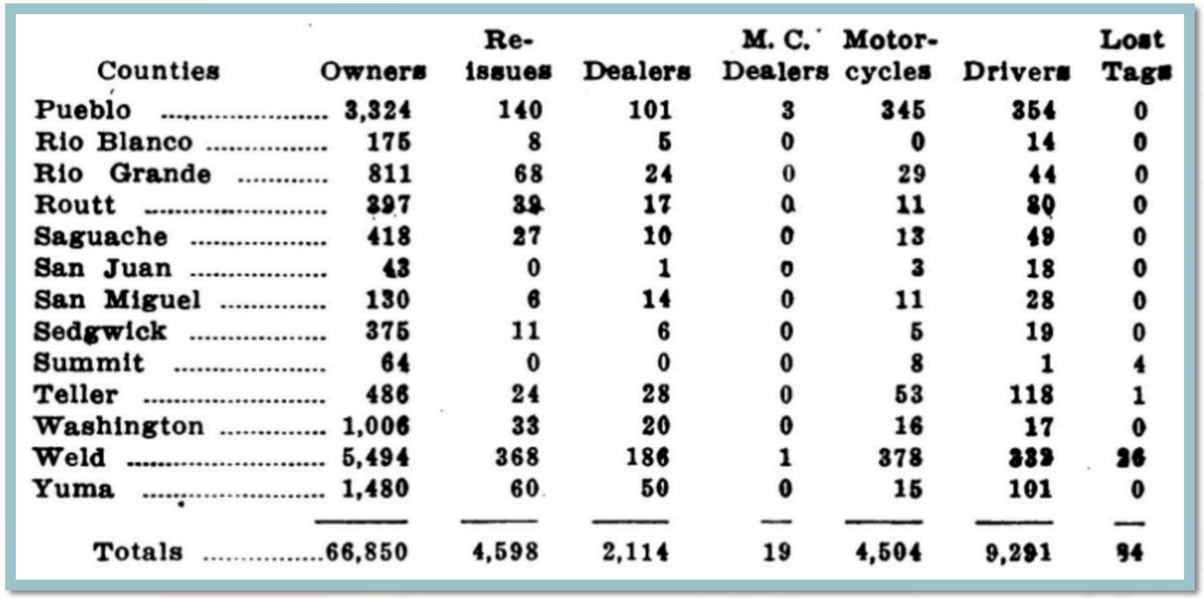
Step 2: Table Layout Parsing + OCR with LLM

Prompt: You are a researcher who carefully digitizes historical statistical tables. You look at scans from old books and put the tables in csv files. The tables can have multiple headers, make sure you merge them. Don't make up numbers because those are very important for your research. [TINKERING]. Output a csv table. Don't output any other text. Image to process is attached.

Tinkering examples:
- The table might have multiple years. If it does, this means that likely there multiple headers for vehicle type and year. You want the columns in the output to look something like this: VehicleType_year
- The years can appear in any order.
- Empty rows in the image can be represented by multiple dots, encode them as a single blank space, don't add extra columns.
- If the column represents vehicle counts or other counts, it can't have decimal points. Only $ amount columns can have decimal points.

Special case (multi-pdf table):
-The table you see is part of a bigger table. If the column names for the table are not provided, use the following column names in this order: {columns from first part of table}.

Step 2: Table Layout Parsing + OCR with LLM



Counties	Owners	Re-issues	Dealers	M. C. Dealers	Motor-cycles	Drivers	Lost Tags
Pueblo	3,324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	32	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1,006	33	20	0	16	17	0
Weld	5,494	368	186	1	378	332	26
Yuma	1,480	60	50	0	15	101	0
Totals	66,850	4,598	2,114	19	4,504	9,291	94

Cropped image

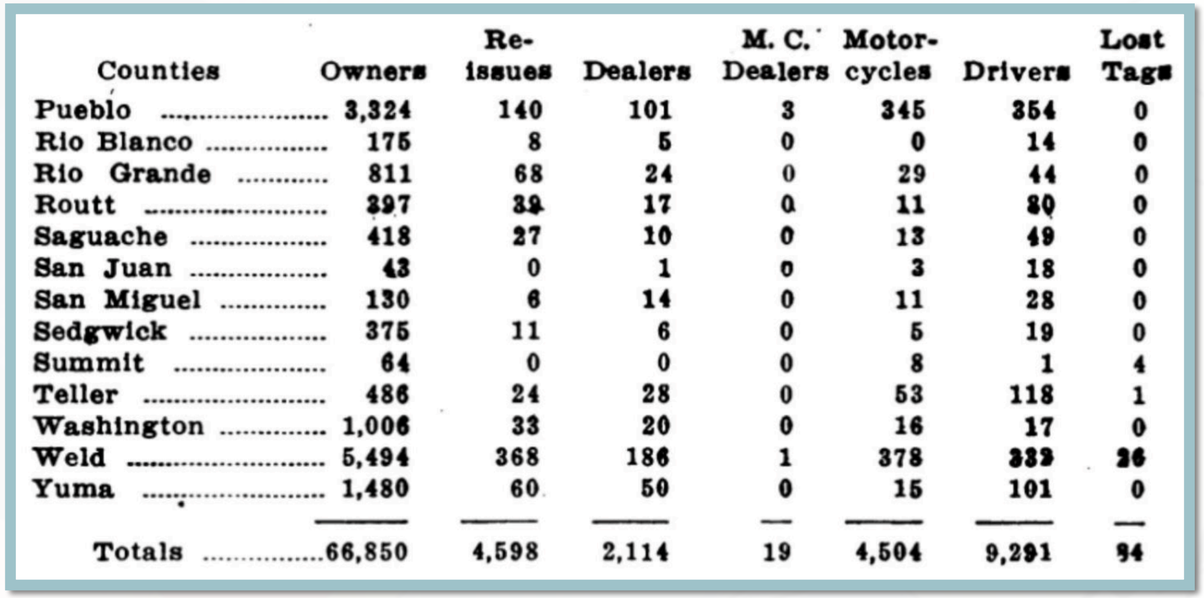
Prompt: You are a researcher who carefully digitizes historical statistical tables. You look at scans from old books and put the tables in csv files. The tables can have multiple headers, make sure you merge them. Don’t make up numbers because those are very important for your research. [TINKERING]. Output a csv table. Don’t output any other text. Image to process is attached.

LLM-processed table

Tinkering examples:
- The table might have multiple years. If it does, this means that likely there multiple headers for vehicle type and year. You want the columns in the output to look something like this: VehicleType_year
- The years can appear in any order.
- Empty rows in the image can be represented by multiple dots, encode them as a single blank space, don't add extra columns.
- If the column represents vehicle counts or other counts, it can't have decimal points. Only $ amount columns can have decimal points.

Special case (multi-pdf table):
-The table you see is part of a bigger table. If the column names for the table are not provided, use the following column names in this order: {columns from first part of table}.

Step 2: Table Layout Parsing + OCR with LLM



Counties	Owners	Re-issues	Dealers	M. C. Dealers	Motor-cycles	Drivers	Lost Tags
Pueblo	3,324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	38	17	0	11	30	0
Saguache	418	27	10	0	13	49	0
San Juan	48	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1,006	33	20	0	16	17	0
Weld	5,494	368	186	1	378	282	26
Yuma	1,480	60	50	0	15	101	0
Totals	66,850	4,598	2,114	19	4,504	9,291	94

Cropped image

Prompt: You are a researcher who carefully digitizes historical statistical tables. You look at scans from old books and put the tables in csv files. The tables can have multiple headers, make sure you merge them. Don't make up numbers because those are very important for your research. [TINKERING]. Output a csv table. Don't output any other text. Image to process is attached.

Tinkering examples:
- The table might have multiple years. If it does, this means that likely there multiple headers for vehicle type and year. You want the columns in the output to look something like this: VehicleType_year
- The years can appear in any order.
- Empty rows in the image can be represented by multiple dots, encode them as a single blank space, don't add extra columns.
- If the column represents vehicle counts or other counts, it can't have decimal points. Only $ amount columns can have decimal points.

Special case (multi-pdf table):
-The table you see is part of a bigger table. If the column names for the table are not provided, use the following column names in this order: {columns from first part of table}.

LLM-processed table

Counties	Owners	Reissues	Dealers	M. C. Dealers	Motorcycles	Drivers	Lost Tags
Pueblo	3324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	38	17	0	11	30	0
Saguache	418	27	10	0	13	49	0
San Juan	48	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1006	33	20	0	16	17	0
Weld	5494	368	186	1	378	282	26
Yuma	1480	60	50	0	15	101	0
Totals	66850	4598	2114	19	4504	9291	94

Step 3: Alignment

LLM-processed table

Counties	Owners	Reissues	Dealers	M. C. Dealers	Motorcycles	Drivers	Lost Tags
Pueblo	3324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	38	17	0	11	30	0
Saguache	418	27	10	0	13	49	0
San Juan	48	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1006	33	20	0	16	17	0
Weld	5494	368	186	1	378	282	26
Yuma	1480	60	50	0	15	101	0
Totals	66850	4598	2114	19	4504	9291	94

Step 3: Alignment

LLM-processed table

Counties	Owners	Reissues	Dealers	M. C. Dealers	Motorcycles	Drivers	Lost Tags
Pueblo	3324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	38	17	0	11	30	0
Saguache	418	27	10	0	13	49	0
San Juan	48	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1006	33	20	0	16	17	0
Weld	5494	368	186	1	378	282	26
Yuma	1480	60	50	0	15	101	0
Totals	66850	4598	2114	19	4504	9291	94

+

Counties	Owners	Re-Issue	Dealers	Motorcycles Dealers	Motorcycles	Drivers	Lost Tags
Pueblo	3324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	39	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1006	33	20	0	16	17	0
Weld	5494	368	186	1	378	339	26
Yuma	1480	60	50	0	15	101	0
Totals	66850	4598	2114	19	4504	9291	94

Human-processed gold standard table

Step 3: Alignment

Step 3: Alignment

LLM-processed table

Counties	Owners	Reissues	Dealers	M. C. Dealers	Motorcycles	Drivers	Lost Tags
Pueblo	3324	140	101	3	345	354	0
Rio Blanco							0
Rio Grande							0
Routt							0
Saguache							0
San Juan							0
San Miguel							0
Sedgwick							0
Summit							4
Teller							1
Washington							0
Weld							6
Yuma							0
Totals							4

Column alignment prompt:

You are a researcher aligning the column names of OCR'd tables with column names processed by humans. You will be given two lists of column names: Assess and True. Please output a python dictionary where every column in Assess is assigned a corresponding column in True. Don't output any other text.

Here are some things you should know:
- All columns in True must be in Assess
- There might be columns in Assess that are not in True. In that case, assign NoCorrespondingColumnX to that column, where X is the number of column with that label.
- Some columns represent number (no/num) of vehicles and some columns represent dollar amounts. Match them accordingly.

Assess: {cols_yhat}
True: {cols_y}

Counties	Owners	Re-Issue	Dealers	Motorcycles Dealers	Motorcycles	Drivers	Lost Tags
Pueblo	3324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	39	17	0	11	80	0
Saguache	418	27	10	0	13	49	0
San Juan	43	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1006	33	20	0	16	17	0
Weld	5494	368	186	1	378	339	26
Yuma	1480	60	50	0	15	101	0
Totals	66850	4598	2114	19	4504	9291	94

Human-processed gold standard table

Step 3: Alignment

LLM-processed table

Counties	Owners	Reissues	Dealers	M. C. Dealers	Motorcycles	Drivers	Lost Tags
Pueblo	3324	140	101	3	345	354	0
Rio Blanco							0
Rio Grande							0
Routt							0
Saguache							0
San Juan							0
San Miguel							0
Sedgwick							0
Summit							4
Teller							1
Washington							0
Weld							6
Yuma							0
Totals							4

Column alignment prompt:

You are a researcher aligning the column names of OCR'd tables with column names processed by humans. You will be given two lists of column names: Assess and True. Please output a python dictionary where every column in Assess is assigned a corresponding column in True. Don't output any other text.

Here are some things you should know:
- All columns in True must be in Assess
- There might be columns in Assess that are not in True. In that case, assign NoCorrespondingColumnX to that column, where X is the number of column with that label.
- Some columns represent number (no/num) of vehicles and some columns represent dollar amounts. Match them accordingly.

Assess: {cols_yhat}
True: {cols_y}

Counties	Owners	Re-Issue	Dealers	Motorcycles Dealers	Motorcycles	Drivers	Lost Tags
Pueblo					345	354	0
Rio Blanco					0	14	0
Rio Grande					29	44	0
Routt					11	80	0
Saguache					13	49	0
San Juan					3	18	0
San Miguel					11	28	0
Sedgwick					5	19	0
Summit					8	1	4
Teller					53	118	1
Washington					16	17	0
Weld					378	339	26
Yuma					15	101	0
Totals					4504	9291	94

County alignment prompt:

You are a researcher aligning the county names of OCR'd tables with county names processed by humans. You will be given two lists of counties names: Assess and True. Please output a python dictionary where every county in Assess is assigned a corresponding county in True. Don't output any other text.

Here are some things you should know:
- Every county in True can appear in Assess only once.
- Assign None to the county in Assess if it's not in True.

Assess: {str_counties_yhat}
True: {str_counties_y}

Step 3: Alignment

LLM-processed table

Counties	Owners	Reissues	Dealers	M. C. Dealers	Motorcycles	Drivers	Lost Tags
Pueblo	3324	140	101	3	345	354	0
Rio Blanco							0
Rio Grande							0
Routt							0
Saguache							0
San Juan							0
San Miguel							0
Sedgwick							0
Summit							4
Teller							1
Washington							0
Weld							6
Yuma							0
Totals							4

Column alignment prompt:
You are a researcher aligning the column names of OCR'd tables with column names processed by humans. You will be given two lists of column names: Assess and True. Please output a python dictionary where every column in Assess is assigned a corresponding column in True. Don't output any other text.

Here are some things you should know:
- All columns in True must be in Assess
- There might be columns in Assess that are not in True. In that case, assign NoCorrespondingColumnX to that column, where X is the number of column with that label.
- Some columns represent number (no/num) of vehicles and some columns represent dollar amounts. Match them accordingly.

Assess: {cols_yhat}
True: {cols_y}

Counties	Owners	Re-Issue	Dealers	Motorcycles Dealers	Motorcycles	Drivers	Lost Tags
Pueblo					345	354	0
Rio Blanco					0	14	0
Rio Grande					29	44	0
Routt					11	80	0
Saguache					13	49	0
San Juan					3	18	0
San Miguel					11	28	0
Sedgwick					5	19	0
Summit					8	1	4
Teller					53	118	1
Washington					16	17	0
Weld					378	339	26
Yuma					15	101	0
Totals					4504	9291	94

County alignment prompt:
You are a researcher aligning the county names of OCR'd tables with county names processed by humans. You will be given two lists of counties names: Assess and True. Please output a python dictionary where every county in Assess is assigned a corresponding county in True. Don't output any other text.

Here are some things you should know:
- Every county in True can appear in Assess only once.
- Assign None to the county in Assess if it's not in True.

Assess: {str_counties_yhat}
True: {str_counties_y}

Counties	Owners	Re-Issue	Dealers	Motorcycles Dealers	Motorcycles	Drivers	Lost Tags
Pueblo	3324	140	101	3	345	354	0
Rio Blanco	175	8	5	0	0	14	0
Rio Grande	811	68	24	0	29	44	0
Routt	397	38	17	0	11	30	0
Saguache	418	27	10	0	13	49	0
San Juan	48	0	1	0	3	18	0
San Miguel	130	6	14	0	11	28	0
Sedgwick	375	11	6	0	5	19	0
Summit	64	0	0	0	8	1	4
Teller	486	24	28	0	53	118	1
Washington	1006	33	20	0	16	17	0
Weld	5494	368	186	1	378	282	26
Yuma	1480	60	50	0	15	101	0
Totals	66850	4598	2114	19	4504	9291	94

Aligned table (same columns and counties as gold standard)

Step 3: Alignment (Special Case)

Step 3: Alignment (Special Case)

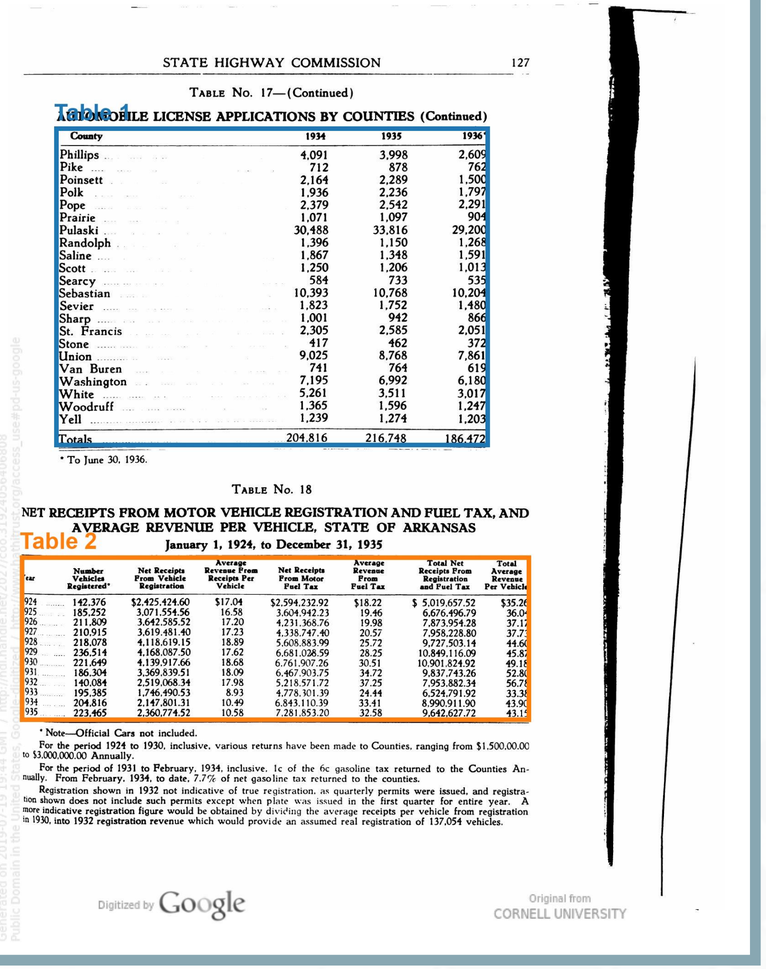


TABLE No. 17—(Continued)

Table 1

AUTOMOBILE LICENSE APPLICATIONS BY COUNTIES (Continued)

County	1934	1935	1936*
Phillips	4,091	3,998	2,609
Pike	712	878	762
Poinsett	2,164	2,289	1,500
Polk	1,936	2,236	1,797
Pope	2,379	2,542	2,291
Prairie	1,071	1,097	904
Pulaski	30,488	33,816	29,200
Randolph	1,396	1,150	1,268
Saline	1,867	1,348	1,591
Scott	1,250	1,206	1,013
Searcy	584	733	535
Sebastian	10,393	10,768	10,204
Sevier	1,823	1,752	1,480
Sharp	1,001	942	866
St. Francis	2,305	2,585	2,051
Stone	417	462	372
Union	9,025	8,768	7,861
Van Buren	741	764	615
Washington	7,195	6,992	6,180
White	5,261	3,511	3,017
Woodruff	1,365	1,596	1,247
Yell	1,239	1,274	1,203
Totals	204,816	216,748	186,472

* To June 30, 1936.

TABLE No. 18

NET RECEIPTS FROM MOTOR VEHICLE REGISTRATION AND FUEL TAX, AND AVERAGE REVENUE PER VEHICLE, STATE OF ARKANSAS

Table 2

January 1, 1924, to December 31, 1935

Year	Number Vehicles Registered*	Net Receipts From Vehicle Registration	Average Revenue From Receipts Per Vehicle	Net Receipts From Motor Fuel Tax	Average Revenue From Fuel Tax	Total Net Receipts From Registration and Fuel Tax	Total Average Revenue Per Vehicle
924	142,376	$2,425,424.60	$17.04	$2,594,232.92	$18.22	$ 5,019,657.52	$35.26
925	185,252	3,071,554.56	16.58	3,604,942.23	19.46	6,676,496.79	36.04
926	211,809	3,642,585.52	17.20	4,231,368.76	19.98	7,873,954.28	37.18
927	210,915	3,619,481.40	17.23	4,338,747.40	20.57	7,958,228.80	37.7
928	218,078	4,118,619.15	18.89	5,608,883.99	25.72	9,727,503.14	44.60
929	236,514	4,168,087.50	17.62	6,681,028.59	28.25	10,849,116.09	45.87
930	221,649	4,139,917.66	18.68	6,761,907.26	30.51	10,901,824.92	49.18
931	186,304	3,369,839.51	18.09	6,467,903.75	34.72	9,837,743.26	52.80
932	140,084	2,519,068.34	17.98	5,218,571.72	37.25	7,953,882.34	56.78
933	195,385	1,746,490.53	8.93	4,778,301.39	24.44	6,524,791.92	33.38
934	204,816	2,147,801.31	10.49	6,843,110.39	33.41	8,990,911.90	43.90
935	223,465	2,360,774.52	10.58	7,281,853.20	32.58	9,642,627.72	43.15

* Note—Official Cars not included.

For the period 1924 to 1930, inclusive, various returns have been made to Counties, ranging from $1,500.00.00 to $3,000,000.00 Annually.

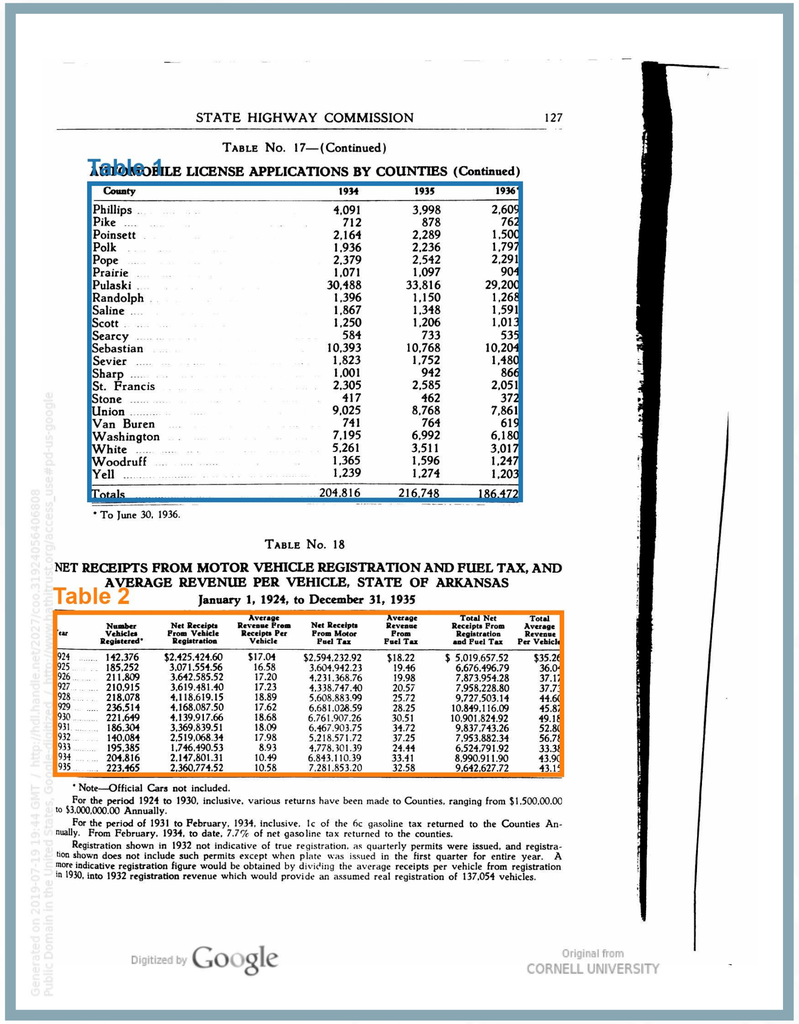
For the period of 1931 to February, 1934, inclusive, 1c of the 6c gasoline tax returned to the Counties Annually. From February, 1934, to date, 7.7% of net gasoline tax returned to the counties.

Registration shown in 1932 not indicative of true registration, as quarterly permits were issued, and registration shown does not include such permits except when plate was issued in the first quarter for entire year. A more indicative registration figure would be obtained by dividing the average receipts per vehicle from registration in 1930, into 1932 registration revenue which would provide an assumed real registration of 137,054 vehicles.

Digitized by Google
Original from CORNELL UNIVERSITY
Generated on 2019-07-19 19:44 GMT / http://hdl.handle.net/2027/coo.31924056406808
Public Domain in the United States, Google-digitized / http://www.hathitrust.org/access_use#pd-us-google

Image layout parsing identified 2 tables

Step 3: Alignment (Special Case)



STATE HIGHWAY COMMISSION 127

TABLE No. 17—(Continued)

Table 1

AUTOMOBILE LICENSE APPLICATIONS BY COUNTIES (Continued)

County	1934	1935	1936*
Phillips	4,091	3,998	2,609
Pike	712	878	762
Poinsett	2,164	2,289	1,500
Polk	1,936	2,236	1,797
Pope	2,379	2,542	2,291
Prairie	1,071	1,097	904
Pulaski	30,488	33,816	29,200
Randolph	1,396	1,150	1,268
Saline	1,867	1,348	1,591
Scott	1,250	1,206	1,013
Searcy	584	733	535
Sebastian	10,393	10,768	10,204
Sevier	1,823	1,752	1,480
Sharp	1,001	942	866
St. Francis	2,305	2,585	2,051
Stone	417	462	372
Union	9,025	8,768	7,861
Van Buren	741	764	619
Washington	7,195	6,992	6,180
White	5,261	3,511	3,017
Woodruff	1,365	1,596	1,247
Yell	1,239	1,274	1,203
Totals	204,816	216,748	186,472

* To June 30, 1936.

TABLE No. 18

NET RECEIPTS FROM MOTOR VEHICLE REGISTRATION AND FUEL TAX, AND AVERAGE REVENUE PER VEHICLE, STATE OF ARKANSAS

Table 2

January 1, 1924, to December 31, 1935

Year	Number Vehicles Registered*	Net Receipts From Vehicle Registration	Average Revenue From Receipts Per Vehicle	Net Receipts From Motor Fuel Tax	Average Revenue From Fuel Tax	Total Net Receipts From Registration and Fuel Tax	Total Average Revenue Per Vehicle
924	142,376	$2,425,424.60	$17.04	$2,594,232.92	$18.22	$ 5,019,657.52	$35.26
925	185,252	3,071,554.56	16.58	3,604,942.23	19.46	6,676,496.79	36.04
926	211,809	3,642,585.52	17.20	4,231,368.76	19.98	7,873,954.28	37.17
927	210,915	3,619,481.40	17.23	4,338,747.40	20.57	7,958,228.80	37.73
928	218,078	4,118,619.15	18.89	5,608,883.99	25.72	9,727,503.14	44.60
929	236,514	4,168,087.50	17.62	6,681,028.59	28.25	10,849,116.09	45.87
930	221,649	4,139,917.66	18.68	6,761,907.26	30.51	10,901,824.92	49.18
931	186,304	3,369,839.51	18.09	6,467,903.75	34.72	9,837,743.26	52.80
932	140,084	2,519,068.34	17.98	5,218,571.72	37.25	7,953,882.34	56.78
933	195,385	1,746,490.53	8.93	4,778,301.39	24.44	6,524,791.92	33.38
934	204,816	2,147,801.31	10.49	6,843,110.39	33.41	8,990,911.90	43.90
935	223,465	2,360,774.52	10.58	7,281,853.20	32.58	9,642,627.72	43.15

* Note—Official Cars not included.

For the period 1924 to 1930, inclusive, various returns have been made to Counties, ranging from $1,500.00.00 to $3,000,000.00 Annually.

For the period of 1931 to February, 1934, inclusive, 1c of the 6c gasoline tax returned to the Counties Annually. From February, 1934, to date, 7.7% of net gasoline tax returned to the counties.

Registration shown in 1932 not indicative of true registration, as quarterly permits were issued, and registration shown does not include such permits except when plate was issued in the first quarter for entire year. A more indicative registration figure would be obtained by dividing the average receipts per vehicle from registration in 1930, into 1932 registration revenue which would provide an assumed real registration of 137,054 vehicles.

Digitized by Google
Original from CORNELL UNIVERSITY
Generated on 2019-07-19 19:44 GMT / http://hdl.handle.net/2027/coo.31924056406808
Public Domain in the United States, Google-digitized / http://www.hathitrust.org/access_use#pd-us-google

Image layout parsing identified 2 tables

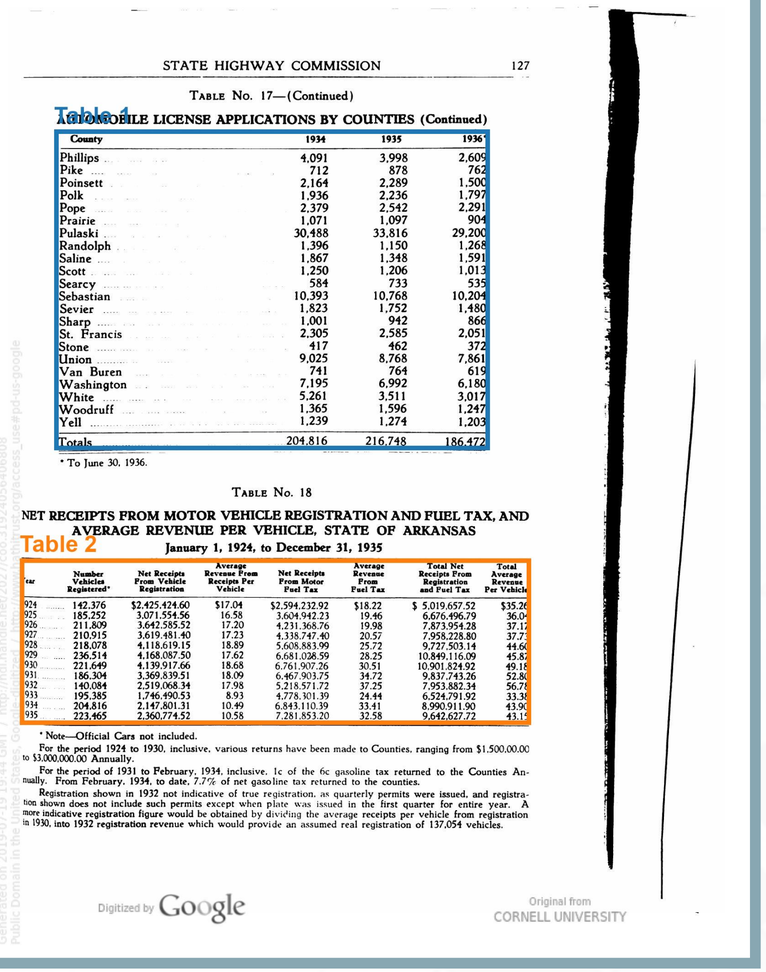
LLM-processed Table 1

County	1934	1935	1936
Phillips	4091	3998	2609
Pike	712	878	762
Poinsett	2164	2289	1500
Polk	1936	2236	1797
Pope	2379	2542	2291
Prairie	1071	1097	904
Pulaski	30488	33816	29200
Randolph	1396	1150	1268
Saline	1867	1348	1591
Scott	1250	1206	1013
Searcy	584	733	535
Sebastian	10393	10768	10204
Sevier	1823	1752	1480
Sharp	1001	942	866
St. Francis	2305	2585	2051
Stone	417	462	372
Union	9025	8768	7861
Van Buren	741	764	619
Washington	7195	6992	6180
White	5261	3511	3017
Woodruff	1365	1596	1247
Yell	1239	1274	1203
Totals	204816	216748	186472

LLM-processed Table 2

Year	Number_Vehicles_Registered	Net_Receipts_From_Vehicle_Registration	Average_Revenue_From_Receipts_Per_Vehicle	Net_Receipts_From_Motor_Fuel_Tax	Average_Revenue_From_Fuel_Tax	Total_Net_Receipts_From_Registration_and_Fuel_Tax	Total_Average_Revenue_Per_Vehicle
1924	142376	2425424.6	17.04	2594232.92	18.22	5019657.52	35.26
1925	185252	3071554.56	16.58	3604942.23	19.46	6676496.79	36.04
1926	211809	3642585.52	17.2	4231368.76	19.98	7873954.28	37.17
1927	210915	3619481.4	17.23	4338747.4	20.57	7958228.8	37.73
1928	218073	4118619.15	18.89	5608883.99	25.72	9727503.14	44.6
1929	236514	4168087.5	17.62	6681028.59	28.25	10349116.09	45.87
1930	221649	4139917.66	18.68	6761907.26	30.51	10901824.92	49.18
1931	186304	3369839.51	18.09	6467903.75	34.72	9837743.26	52.8
1932	140084	2519068.34	17.98	5218571.72	37.25	7953882.34	56.78
1933	195385	1746490.53	8.93	4778301.39	24.44	6524791.92	33.38
1934	204816	2147801.31	10.49	6843110.39	33.41	8990911.9	43.9
1935	223465	2360774.52	10.58	7281853.2	32.58	9642627.72	43.15

Step 3: Alignment (Special Case)



STATE HIGHWAY COMMISSION 127

TABLE No. 17—(Continued)

Table 1

AUTOMOBILE LICENSE APPLICATIONS BY COUNTIES (Continued)

County	1934	1935	1936*
Phillips	4,091	3,998	2,609
Pike	712	878	762
Poinsett	2,164	2,289	1,500
Polk	1,936	2,236	1,797
Pope	2,379	2,542	2,291
Prairie	1,071	1,097	904
Pulaski	30,488	33,816	29,200
Randolph	1,396	1,150	1,268
Saline	1,867	1,348	1,591
Scott	1,250	1,206	1,013
Searcy	584	733	535
Sebastian	10,393	10,768	10,204
Sevier	1,823	1,752	1,480
Sharp	1,001	942	866
St. Francis	2,305	2,585	2,051
Stone	417	462	372
Union	9,025	8,768	7,861
Van Buren	741	764	619
Washington	7,195	6,992	6,180
White	5,261	3,511	3,017
Woodruff	1,365	1,596	1,247
Yell	1,239	1,274	1,203
Totals	204,816	216,748	186,472

* To June 30, 1936.

TABLE No. 18

NET RECEIPTS FROM MOTOR VEHICLE REGISTRATION AND FUEL TAX, AND AVERAGE REVENUE PER VEHICLE, STATE OF ARKANSAS

Table 2

January 1, 1924, to December 31, 1935

Year	Number Vehicles Registered*	Net Receipts From Vehicle Registration	Average Revenue From Receipts Per Vehicle	Net Receipts From Motor Fuel Tax	Average Revenue From Fuel Tax	Total Net Receipts From Registration and Fuel Tax	Total Average Revenue Per Vehicle
1924	142,376	$2,425,424.60	$17.04	$2,594,232.92	$18.22	$ 5,019,657.52	$35.26
1925	185,252	3,071,554.56	16.58	3,604,942.23	19.46	6,676,496.79	36.04
1926	211,809	3,642,585.52	17.20	4,231,368.76	19.98	7,873,954.28	37.17
1927	210,915	3,619,481.40	17.23	4,338,747.40	20.57	7,958,228.80	37.73
1928	218,078	4,118,619.15	18.89	5,608,883.99	25.72	9,727,503.14	44.60
1929	236,514	4,168,087.50	17.62	6,681,028.59	28.25	10,849,116.09	45.87
1930	221,649	4,139,917.66	18.68	6,761,907.26	30.51	10,901,824.92	49.18
1931	186,304	3,369,839.51	18.09	6,467,903.75	34.72	9,837,743.26	52.80
1932	140,084	2,519,068.34	17.98	5,218,571.72	37.25	7,953,882.34	56.78
1933	195,385	1,746,490.53	8.93	4,778,301.39	24.44	6,524,791.92	33.38
1934	204,816	2,147,801.31	10.49	6,843,110.39	33.41	8,990,911.90	43.90
1935	223,465	2,360,774.52	10.58	7,281,853.20	32.58	9,642,627.72	43.15

* Note—Official Cars not included.

For the period 1924 to 1930, inclusive, various returns have been made to Counties, ranging from $1,500.00.00 to $3,000,000.00 Annually.

For the period of 1931 to February, 1934, inclusive, 1c of the 6c gasoline tax returned to the Counties Annually. From February, 1934, to date, 7.7% of net gasoline tax returned to the counties.

Registration shown in 1932 not indicative of true registration, as quarterly permits were issued, and registration shown does not include such permits except when plate was issued in the first quarter for entire year. A more indicative registration figure would be obtained by dividing the average receipts per vehicle from registration in 1930, into 1932 registration revenue which would provide an assumed real registration of 137,054 vehicles.

Digitized by Google
Original from CORNELL UNIVERSITY
Generated on 2019-07-19 19:44 GMT / http://hdl.handle.net/2027/coo.31924056406808
Public Domain in the United States, Google-digitized / http://www.hathitrust.org/access_use#pd-us-google

Image layout parsing identified 2 tables

LLM-processed Table 1

County	1934	1935	1936
Phillips	4091	3998	2609
Pike	712	878	762
Poinsett	2164	2289	1500
Polk	1936	2236	1797
Pope	2379	2542	2291
Prairie	1071	1097	904
Pulaski	30488	33816	29200
Randolph	1396	1150	1268
Saline	1867	1348	1591
Scott	1250	1206	1013
Searcy	584	733	535
Sebastian	10393	10768	10204
Sevier	1823	1752	1480
Sharp	1001	942	866
St. Francis	2305	2585	2051
Stone	417	462	372
Union	9025	8768	7861
Van Buren	741	764	619
Washington	7195	6992	6180
White	5261	3511	3017
Woodruff	1365	1596	1247
Yell	1239	1274	1203
Totals	204816	216748	186472

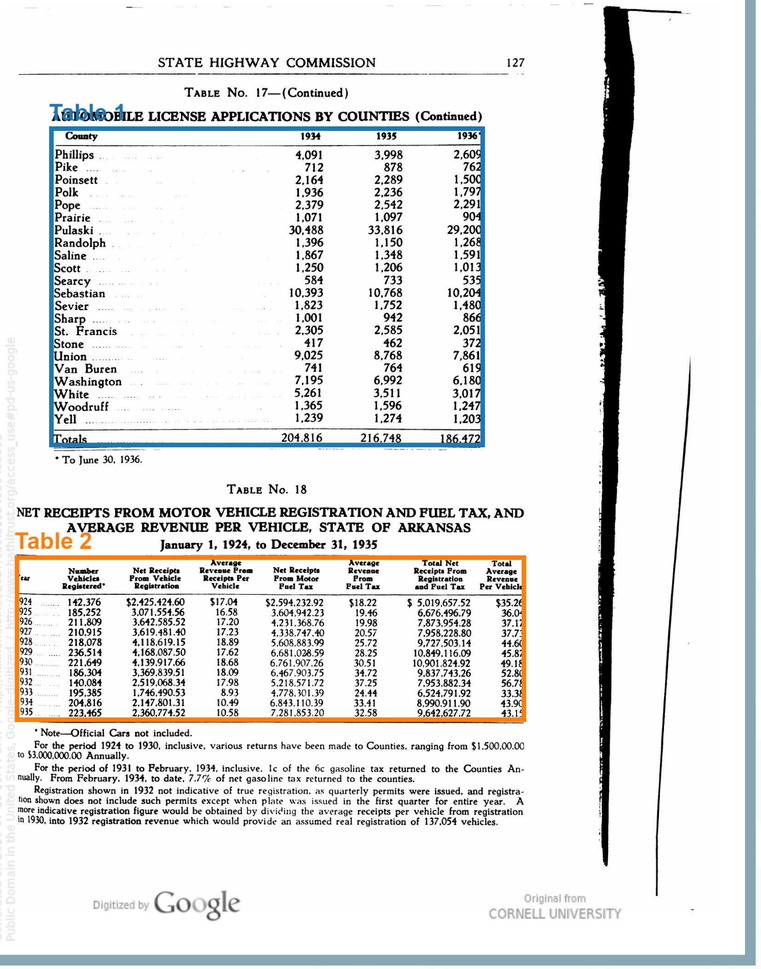
Gold Standard Table

Counties	Automobile License Applications; Year: 1934	Automobile License Applications; Year: 1935	Automobile License Applications; Year: 1936 (6 months)
Phillips	4091	3998	2609
Pike	712	878	762
Poinsett	2164	2289	1500
Polk	1936	2236	1797
Pope	2379	2542	2291
Prairie	1071	1097	904
Pulaski	30488	33816	29200
Randolph	1396	1150	1268
Saline	1867	1348	1591
Scott	1250	1206	1013
Searcy	584	733	535
Sebastian	10393	10768	10204
Sevier	1823	1752	1480
Sharp	1001	942	866
St. Francis	2305	2585	2051
Stone	417	462	372
Union	9025	8768	7861
Van Buren	741	764	619
Washington	7195	6992	6180
White	5261	3511	3017
Woodruff	1365	1596	1247
Yell	1239	1274	1203
Totals	204816	216748	186472

LLM-processed Table 2

Year	Number_Vehicles_Registered	Net_Receipts_From_Vehicle_Registration	Average_Revenue_From_Receipts_Per_Vehicle	Net_Receipts_From_Motor_Fuel_Tax	Average_Revenue_From_Fuel_Tax	Total_Net_Receipts_From_Registration_and_Fuel_Tax	Total_Average_Revenue_Per_Vehicle
1924	142376	2425424.6	17.04	2594232.92	18.22	5019657.52	35.26
1925	185252	3071554.56	16.58	3604942.23	19.46	6676496.79	36.04
1926	211809	3642585.52	17.2	4231368.76	19.98	7873954.28	37.17
1927	210915	3619481.4	17.23	4338747.4	20.57	7958228.8	37.73
1928	218073	4118619.15	18.89	5608883.99	25.72	9727503.14	44.6
1929	236514	4168087.5	17.62	6681028.59	28.25	10349116.09	45.87
1930	221649	4139917.66	18.68	6761907.26	30.51	10901824.92	49.18
1931	186304	3369839.51	18.09	6467903.75	34.72	9837743.26	52.8
1932	140084	2519068.34	17.98	5218571.72	37.25	7953882.34	56.78
1933	195385	1746490.53	8.93	4778301.39	24.44	6524791.92	33.38
1934	204816	2147801.31	10.49	6843110.39	33.41	8990911.9	43.9
1935	223465	2360774.52	10.58	7281853.2	32.58	9642627.72	43.15

Step 3: Alignment (Special Case)



STATE HIGHWAY COMMISSION 127

TABLE No. 17—(Continued)

Table 1

AUTOMOBILE LICENSE APPLICATIONS BY COUNTIES (Continued)

County	1934	1935	1936*
Phillips	4,091	3,998	2,609
Pike	712	878	762
Poinsett	2,164	2,289	1,500
Polk	1,936	2,236	1,797
Pope	2,379	2,542	2,291
Prairie	1,071	1,097	904
Pulaski	30,488	33,816	29,200
Randolph	1,396	1,150	1,268
Saline	1,867	1,348	1,591
Scott	1,250	1,206	1,013
Searcy	584	733	535
Sebastian	10,393	10,768	10,204
Sevier	1,823	1,752	1,480
Sharp	1,001	942	866
St. Francis	2,305	2,585	2,051
Stone	417	462	372
Union	9,025	8,768	7,861
Van Buren	741	764	619
Washington	7,195	6,992	6,180
White	5,261	3,511	3,017
Woodruff	1,365	1,596	1,247
Yell	1,239	1,274	1,203
Totals	204,816	216,748	186,472

* To June 30, 1936.

TABLE No. 18

NET RECEIPTS FROM MOTOR VEHICLE REGISTRATION AND FUEL TAX, AND AVERAGE REVENUE PER VEHICLE, STATE OF ARKANSAS

Table 2

January 1, 1924, to December 31, 1935

Year	Number Vehicles Registered*	Net Receipts From Vehicle Registration	Average Revenue From Receipts Per Vehicle	Net Receipts From Motor Fuel Tax	Average Revenue From Fuel Tax	Total Net Receipts From Registration and Fuel Tax	Total Average Revenue Per Vehicle
924	142,376	$2,425,424.60	$17.04	$2,594,232.92	$18.22	$ 5,019,657.52	$35.26
925	185,252	3,071,554.56	16.58	3,604,942.23	19.46	6,676,496.79	36.04
926	211,809	3,642,585.52	17.20	4,231,368.76	19.98	7,873,954.28	37.17
927	210,915	3,619,481.40	17.23	4,338,747.40	20.57	7,958,228.80	37.73
928	218,078	4,118,619.15	18.89	5,608,883.99	25.72	9,727,503.14	44.60
929	236,514	4,168,087.50	17.62	6,681,028.59	28.25	10,849,116.09	45.87
930	221,649	4,139,917.66	18.68	6,761,907.26	30.51	10,901,824.92	49.18
931	186,304	3,369,839.51	18.09	6,467,903.75	34.72	9,837,743.26	52.80
932	140,084	2,519,068.34	17.98	5,218,571.72	37.25	7,953,882.34	56.78
933	195,385	1,746,490.53	8.93	4,778,301.39	24.44	6,524,791.92	33.38
934	204,816	2,147,801.31	10.49	6,843,110.39	33.41	8,990,911.90	43.90
935	223,465	2,360,774.52	10.58	7,281,853.20	32.58	9,642,627.72	43.15

* Note—Official Cars not included.

For the period 1924 to 1930, inclusive, various returns have been made to Counties, ranging from $1,500.00.00 to $3,000,000.00 Annually.

For the period of 1931 to February, 1934, inclusive, 1c of the 6c gasoline tax returned to the Counties Annually. From February, 1934, to date, 7.7% of net gasoline tax returned to the counties.

Registration shown in 1932 not indicative of true registration, as quarterly permits were issued, and registration shown does not include such permits except when plate was issued in the first quarter for entire year. A more indicative registration figure would be obtained by dividing the average receipts per vehicle from registration in 1930, into 1932 registration revenue which would provide an assumed real registration of 137,054 vehicles.

Digitized by Google
Original from CORNELL UNIVERSITY
Generated on 2019-07-19 19:44 GMT / http://hdl.handle.net/2027/coo.31924056406808
Public Domain in the United States, Google-digitized / http://www.hathitrust.org/access_use#pd-us-google

Image layout parsing identified 2 tables

LLM-processed Table 1

County	1934	1935	1936
Phillips	4091	3998	2609
Pike	712	878	762
Poinsett	2164	2289	1500
Polk	1936	2236	1797
Pope	2379	2542	2291
Prairie	1071	1097	904
Pulaski	30488	33816	29200
Randolph	1396	1150	1268
Saline	1867	1348	1591
Scott	1250	1206	1013
Searcy	584	733	535
Sebastian	10393	10768	10204
Sevier	1823	1752	1480
Sharp	1001	942	866
St. Francis	2305	2585	2051
Stone	417	462	372
Union	9025	8768	7861
Van Buren	741	764	619
Washington	7195	6992	6180
White	5261	3511	3017
Woodruff	1365	1596	1247
Yell	1239	1274	1203
Totals	204816	216748	186472

Gold Standard Table

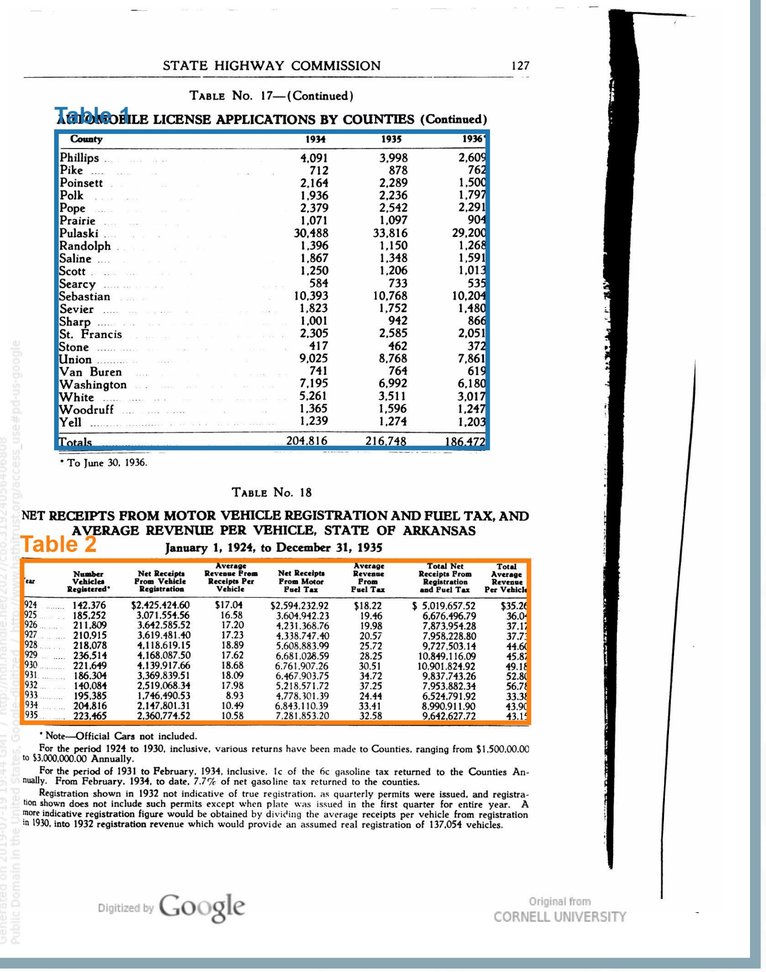
Counties	Automobile License Applications; Year: 1934	Automobile License Applications; Year: 1935	Automobile License Applications; Year: 1936 (6 months)
Phillips	4091	3998	2609
Pike	712	878	762
Poinsett	2164	2289	1500
Polk	1936	2236	1797
Pope	2379	2542	2291
Prairie	1071	1097	904
Pulaski	30488	33816	29200
Randolph	1396	1150	1268
Saline	1867	1348	1591
Scott	1250	1206	1013
Searcy	584	733	535
Sebastian	10393	10768	10204
Sevier	1823	1752	1480
Sharp	1001	942	866
St. Francis	2305	2585	2051
Stone	417	462	372
Union	9025	8768	7861
Van Buren	741	764	619
Washington	7195	6992	6180
White	5261	3511	3017
Woodruff	1365	1596	1247
Yell	1239	1274	1203
Totals	204816	216748	186472

Prompt: You are a researcher trying to pair related tables together. Your main table has the following column headers: {yhat_column_headers}. \n\nYou have two other tables with the following column headers:\n\n1. {y_column_headers_1}\n\n2. {y_column_headers_2}.\n\nPlease output the number of the table whose column headers are closest to the main table's column headers. Output the number only, with no text before. After the number, output an explanation.

LLM-processed Table 2

Year	Number_Vehicles_Registered	Net_Receipts_From_Vehicle_Registration	Average_Revenue_From_Receipts_Per_Vehicle	Net_Receipts_From_Motor_Fuel_Tax	Average_Revenue_From_Fuel_Tax	Total_Net_Receipts_From_Registration_and_Fuel_Tax	Total_Average_Revenue_Per_Vehicle
1924	142376	2425424.6	17.04	2594232.92	18.22	5019657.52	35.26
1925	185252	3071554.56	16.58	3604942.23	19.46	6676496.79	36.04
1926	211809	3642585.52	17.2	4231368.76	19.98	7873954.28	37.17
1927	210915	3619481.4	17.23	4338747.4	20.57	7958228.8	37.73
1928	218073	4118619.15	18.89	5608883.99	25.72	9727503.14	44.6
1929	236514	4168087.5	17.62	6681028.59	28.25	10349116.09	45.87
1930	221649	4139917.66	18.68	6761907.26	30.51	10901824.92	49.18
1931	186304	3369839.51	18.09	6467903.75	34.72	9837743.26	52.8
1932	140084	2519068.34	17.98	5218571.72	37.25	7953882.34	56.78
1933	195385	1746490.53	8.93	4778301.39	24.44	6524791.92	33.38
1934	204816	2147801.31	10.49	6843110.39	33.41	8990911.9	43.9
1935	223465	2360774.52	10.58	7281853.2	32.58	9642627.72	43.15

Step 3: Alignment (Special Case)



STATE HIGHWAY COMMISSION 127

TABLE No. 17—(Continued)

Table 1

AUTOMOBILE LICENSE APPLICATIONS BY COUNTIES (Continued)

County	1934	1935	1936*
Phillips	4,091	3,998	2,609
Pike	712	878	762
Poinsett	2,164	2,289	1,500
Polk	1,936	2,236	1,797
Pope	2,379	2,542	2,291
Prairie	1,071	1,097	904
Pulaski	30,488	33,816	29,200
Randolph	1,396	1,150	1,268
Saline	1,867	1,348	1,591
Scott	1,250	1,206	1,013
Searcy	584	733	535
Sebastian	10,393	10,768	10,204
Sevier	1,823	1,752	1,480
Sharp	1,001	942	866
St. Francis	2,305	2,585	2,051
Stone	417	462	372
Union	9,025	8,768	7,861
Van Buren	741	764	615
Washington	7,195	6,992	6,180
White	5,261	3,511	3,017
Woodruff	1,365	1,596	1,247
Yell	1,239	1,274	1,203
Totals	204,816	216,748	186,472

* To June 30, 1936.

TABLE No. 18

NET RECEIPTS FROM MOTOR VEHICLE REGISTRATION AND FUEL TAX, AND AVERAGE REVENUE PER VEHICLE, STATE OF ARKANSAS

Table 2

January 1, 1924, to December 31, 1935

Year	Number Vehicles Registered*	Net Receipts From Vehicle Registration	Average Revenue From Receipts Per Vehicle	Net Receipts From Motor Fuel Tax	Average Revenue From Fuel Tax	Total Net Receipts From Registration and Fuel Tax	Total Average Revenue Per Vehicle
924	142,376	$2,425,424.60	$17.04	$2,594,232.92	$18.22	$ 5,019,657.52	$35.26
925	185,252	3,071,554.56	16.58	3,604,942.23	19.46	6,676,496.79	36.04
926	211,809	3,642,585.52	17.20	4,231,368.76	19.98	7,873,954.28	37.17
927	210,915	3,619,481.40	17.23	4,338,747.40	20.57	7,958,228.80	37.73
928	218,078	4,118,619.15	18.89	5,608,883.99	25.72	9,727,503.14	44.60
929	236,514	4,168,087.50	17.62	6,681,028.59	28.25	10,849,116.09	45.87
930	221,649	4,139,917.66	18.68	6,761,907.26	30.51	10,901,824.92	49.18
931	186,304	3,369,839.51	18.09	6,467,903.75	34.72	9,837,743.26	52.80
932	140,084	2,519,068.34	17.98	5,218,571.72	37.25	7,953,882.34	56.78
933	195,385	1,746,490.53	8.93	4,778,301.39	24.44	6,524,791.92	33.38
934	204,816	2,147,801.31	10.49	6,843,110.39	33.41	8,990,911.90	43.90
935	223,465	2,360,774.52	10.58	7,281,853.20	32.58	9,642,627.72	43.15

* Note—Official Cars not included.

For the period 1924 to 1930, inclusive, various returns have been made to Counties, ranging from $1,500.00.00 to $3,000,000.00 Annually.

For the period of 1931 to February, 1934, inclusive, 1c of the 6c gasoline tax returned to the Counties Annually. From February, 1934, to date, 7.7% of net gasoline tax returned to the counties.

Registration shown in 1932 not indicative of true registration, as quarterly permits were issued, and registration shown does not include such permits except when plate was issued in the first quarter for entire year. A more indicative registration figure would be obtained by dividing the average receipts per vehicle from registration in 1930, into 1932 registration revenue which would provide an assumed real registration of 137,054 vehicles.

Digitized by Google
Original from CORNELL UNIVERSITY
Generated on 2019-07-19 19:44 GMT / http://hdl.handle.net/2027/coo.31924056406808
Public Domain in the United States, Google-digitized / http://www.hathitrust.org/access_use#pd-us-google

Image layout parsing identified 2 tables

LLM-processed Table 1

County	1934	1935	1936
Phillips	4091	3998	2609
Pike	712	878	762
Poinsett	2164	2289	1500
Polk	1936	2236	1797
Pope	2379	2542	2291
Prairie	1071	1097	904
Pulaski	30488	33816	29200
Randolph	1396	1150	1268
Sevier	1823	1752	1480
Sharp	1001	942	866
St. Francis	2305	2585	2051
Stone	417	462	372
Union	9025	8768	7861
Van Buren	741	764	619
Washington	7195	6992	6180
White	5261	3511	3017
Woodruff	1365	1596	1247
Yell	1239	1274	1203
Totals	204816	216748	186472

Keep Table 1 only

Gold Standard Table

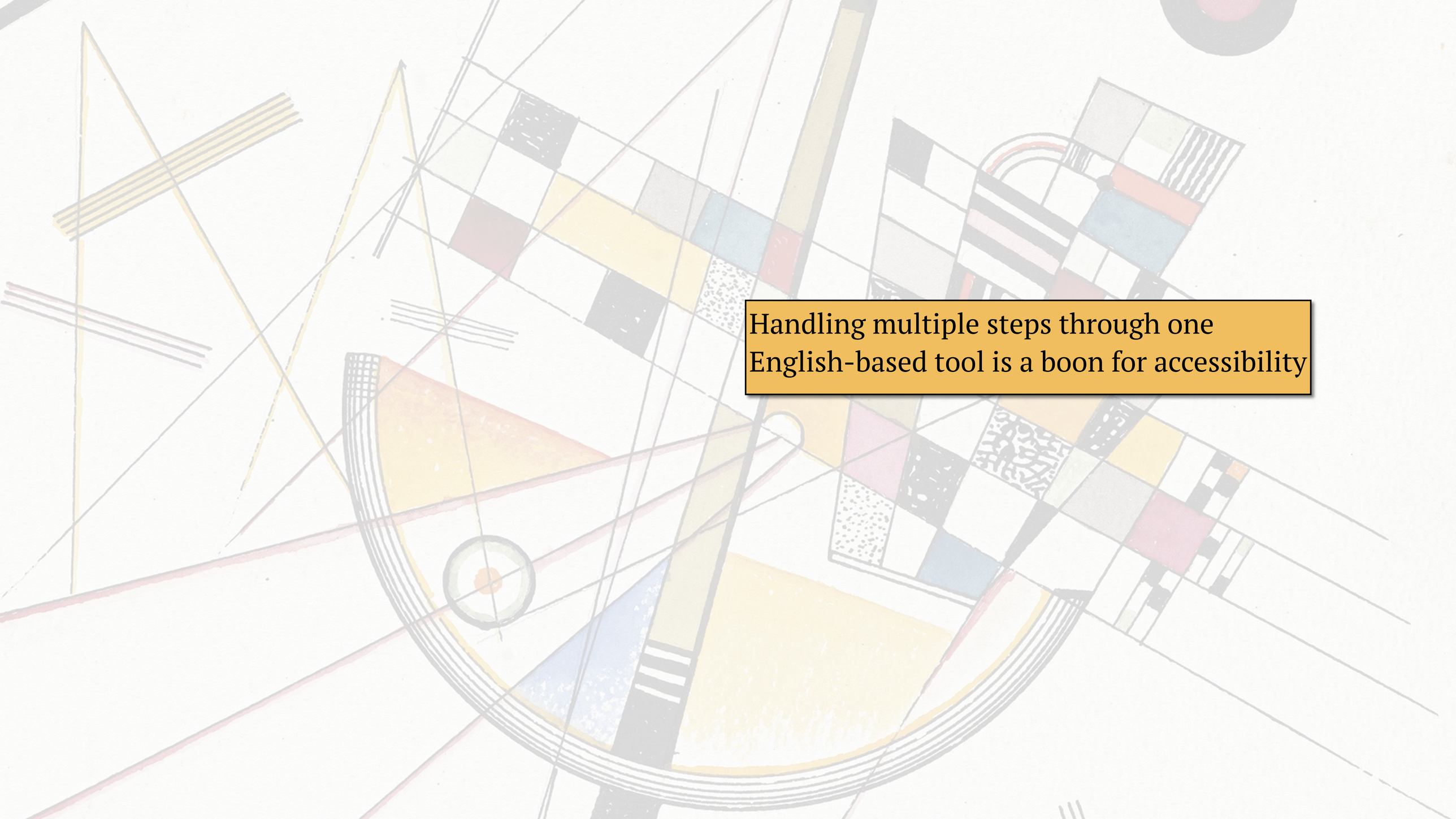
Counties	Automobile License Applications; Year: 1934	Automobile License Applications; Year: 1935	Automobile License Applications; Year: 1936 (6 months)
Phillips	4091	3998	2609
Pike	712	878	762
Poinsett	2164	2289	1500
Polk	1936	2236	1797
Pope	2379	2542	2291
Prairie	1071	1097	904
Pulaski	30488	33816	29200
Randolph	1396	1150	1268
Saline	1867	1348	1591
Scott	1250	1206	1013
Searcy	584	733	535
Sebastian	10393	10768	10204
Sevier	1823	1752	1480
Sharp	1001	942	866
St. Francis	2305	2585	2051
Stone	417	462	372
Union	9025	8768	7861
Van Buren	741	764	619
Washington	7195	6992	6180
White	5261	3511	3017
Woodruff	1365	1596	1247
Yell	1239	1274	1203
Totals	204816	216748	186472

Prompt: You are a researcher trying to pair related tables together. Your main table has the following column headers: {yhat_column_headers}. \n\nYou have two other tables with the following column headers:\n\n1. {y_column_headers_1}\n\n2. {y_column_headers_2}.\n\nPlease output the number of the table whose column headers are closest to the main table's column headers. Output the number only, with no text before. After the number, output an explanation.

LLM-processed Table 2

Year	Number_Vehicles_Registered	Net_Re m_Vehi ration	ge_Revenu Receipts	Net_Receipts m_Motor_F ax	venu uel_Tax	Total_Net_Receipts_From_Registration_and_Fuel_Tax	Total_Average_Revenue_Per_Vehicle
1924	142376	24254	04		18.22	5019657.52	35.26
1925	185252	3071554.5			19.46	6676496.79	36.04
1926	211809	3642585.52			19.98	7873954.28	37.17
1927	210915	3619481.4		4	20.57	7958228.8	37.73
1928	218073	4118619.15		3.99	25.72	9727503.14	44.6
1929	236514	4168087.5		028.59	28.25	10349116.09	45.87
1930	221649	4139917.66		.26	30.51	10901824.92	49.18
1931	186304	3369839.51			34.72	9837743.26	52.8
1932	140084	2519068.34			37.25	7953882.34	56.78
1933	195385	1746490		3	24.44	6524791.92	33.38
1934	204816	2147		0.49	33.41	8990911.9	43.9
1935	223465	23	10.58	72	32.58	9642627.72	43.15

One English-based tool does many steps



Handling multiple steps through one English-based tool is a boon for accessibility

LLM-based digitization pipeline provides **solid** quality for economic research

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median error:
Median % error:

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median	error:	1 unit
	% error:	~0%

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median	error:	1 unit
	% error:	~0%

Symmetric errors 🎉

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median error: 1 unit

Median % error: ~0%

Symmetric errors 🎉

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median	error:	1 unit
	% error:	~0%

Symmetric errors 🎉

Median	absolute error:
	absolute % error:

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median	error:	1 unit
	% error:	~0%

Symmetric errors 🎉

Median	absolute error:	90 units
	absolute % error:	2%

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median	error:	1 unit
	% error:	~0%

Symmetric errors 🎉

Median	absolute error:	90 units
	absolute % error:	2%

How will this affect my inference? 🧐

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median	error:	1 unit	**Symmetric errors** 🎉
	% error:	~0%	
Median	absolute error:	90 units	**How will this affect my inference?** 🧐
	absolute % error:	2%	

	(1)	(2)	(3)	(4)	(5)	(6)
	log(y)	log(y)	log(y)	log(\hat{y})	log(y)	log(\hat{y})
log(\hat{y})	0.993***	0.990***				
	(0.001)	(0.001)				
log(Popula~)			0.960***	0.955***	0.928***	0.923***
			(0.003)	(0.003)	(0.002)	(0.002)
Exc. Outli~s						
Year FE		\checkmark			\checkmark	\checkmark
N	50,318	50,318	50,165	50,165	50,165	50,165
R2	0.990	0.990	0.646	0.637	0.870	0.861

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median	error:	1 unit
	% error:	~0%

Symmetric errors 🎉

Median	absolute error:	90 units
	absolute % error:	2%

How will this affect my inference? 🧐

	(1) log(y)	(2) log(y)	(3) log(y)	(4) log(\hat{y})	(5) log(y)	(6) log(\hat{y})
log(\hat{y})	0.993***	0.990***				
	(0.001)	(0.001)				
log(Popula~)			0.960***	0.955***	0.928***	0.923***
			(0.003)	(0.003)	(0.002)	(0.002)
Exc. Outli~s						
Year FE		\checkmark			\checkmark	\checkmark
N	50,318	50,318	50,165	50,165	50,165	50,165
R2	0.990	0.990	0.646	0.637	0.870	0.861

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median | error: 1 unit | % error: ~0%

Symmetric errors 🎉

Median | absolute error: 90 units | absolute % error: 2%

How will this affect my inference? 🧐

	(1) log(y)	(2) log(y)	(3) log(y)	(4) log(\hat{y})	(5) log(y)	(6) log(\hat{y})
log(\hat{y})	0.993***	0.990***				
	(0.001)	(0.001)				
log(Popula~)			0.960***	0.955***	0.928***	0.923***
			(0.003)	(0.003)	(0.002)	(0.002)
Exc. Outli~s						
Year FE		\checkmark			\checkmark	\checkmark
N	50,318	50,318	50,165	50,165	50,165	50,165
R2	0.990	0.990	0.646	0.637	0.870	0.861

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median | error: 1 unit | % error: ~0%

Symmetric errors 🎉

Median | absolute error: 90 units | absolute % error: 2%

How will this affect my inference? 🧐

	(1) log(y)	(2) log(y)	(3) log(y)	(4) log(\hat{y})	(5) log(y)	(6) log(\hat{y})
log(\hat{y})	0.993***	0.990***				
	(0.001)	(0.001)				
log(Popula~)			0.960***	0.955***	0.928***	0.923***
			(0.003)	(0.003)	(0.002)	(0.002)
Exc. Outli~s						
Year FE		\checkmark			\checkmark	\checkmark
N	50,318	50,318	50,165	50,165	50,165	50,165
R2	0.990	0.990	0.646	0.637	0.870	0.861

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Correlation with gold standard is 99%

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median error: 1 unit
Median % error: ~0%

Symmetric errors 🎉

Median absolute error: 90 units
Median absolute % error: 2%

How will this affect my inference? 🧐

	(1) log(y)	(2) log(y)	(3) log(y)	(4) log(\hat{y})	(5) log(y)	(6) log(\hat{y})
log(\hat{y})	0.993***	0.990***				
	(0.001)	(0.001)				
log(Popula~)			0.960***	0.955***	0.928***	0.923***
			(0.003)	(0.003)	(0.002)	(0.002)
Exc. Outli~s						
Year FE		\checkmark			\checkmark	\checkmark
N	50,318	50,318	50,165	50,165	50,165	50,165
R2	0.990	0.990	0.646	0.637	0.870	0.861

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Correlation with gold standard is 99%

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median | error: | 1 unit
Median | absolute error: | 90 units

% error: ~0%

absolute % error: 2%

Symmetric errors 🎉

How will this affect my inference? 🧐

	(1) log(y)	(2) log(y)	(3) log(y)	(4) log(\hat{y})	(5) log(y)	(6) log(\hat{y})
log(\hat{y})	0.993***	0.990***				
	(0.001)	(0.001)				
log(Popula~)			0.960***	0.955***	0.928***	0.923***
			(0.003)	(0.003)	(0.002)	(0.002)
Exc. Outli~s						
Year FE		\checkmark			\checkmark	\checkmark
N	50,318	50,318	50,165	50,165	50,165	50,165
R2	0.990	0.990	0.646	0.637	0.870	0.861

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Correlation with gold standard is 99%

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median error: 1 unit
Median % error: ~0%

Symmetric errors 🎉

Median absolute error: 90 units
Median absolute % error: 2%

How will this affect my inference? 🧐

	(1) log(y)	(2) log(y)	(3) log(y)	(4) log(\hat{y})	(5) log(y)	(6) log(\hat{y})
log(\hat{y})	0.993***	0.990***				
	(0.001)	(0.001)				
log(Popula~)			0.960***	0.955***	0.928***	0.923***
			(0.003)	(0.003)	(0.002)	(0.002)
Exc. Outli~s						
Year FE		\checkmark			\checkmark	\checkmark
N	50,318	50,318	50,165	50,165	50,165	50,165
R2	0.990	0.990	0.646	0.637	0.870	0.861

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Correlation with gold standard is 99%

Effect on R^2 of population regression is ~1 p.p.

LLM-based digitization pipeline provides **solid** quality for economic research

10% of observations have errors, but they tend to be small

Median error: 1 unit
Median % error: ~0%

Symmetric errors 🎉

Median absolute error: 90 units
Median absolute % error: 2%

How will this affect my inference? 🧐

	(1) log(y)	(2) log(y)	(3) log(y)	(4) log(\hat{y})	(5) log(y)	(6) log(\hat{y})
log(\hat{y})	0.993***	0.990***				
	(0.001)	(0.001)				
log(Popula~)			0.960***	0.955***	0.928***	0.923***
			(0.003)	(0.003)	(0.002)	(0.002)
Exc. Outli~s						
Year FE		\checkmark			\checkmark	\checkmark
N	50,318	50,318	50,165	50,165	50,165	50,165
R2	0.990	0.990	0.646	0.637	0.870	0.861
Standard errors in parentheses						
* p<0.05, ** p<0.01, *** p<0.001						

Measurement error effects on inference are minimal 🥳

Correlation with gold standard is 99%

Effect on R^2 of population regression is ~1 p.p.

Evaluation happens at multiple levels

Multi-level evaluation enables fine-tuning

Evaluation happens at multiple levels

Multi-level evaluation enables fine-tuning

Table-level comparison with gold standard to help easily see where the pipeline fails

Counties	Owners	Re-Issue	Dealers	Motorcycles Dealers	Motorcycles	Drivers	Lost Tags
Pueblo							
Rio Blanco							
Rio Grande							
Routt		39.0 -> 38 (-2.56%)				80.0 -> 30 (-62.5%)	
Saguache							
San Juan	43.0 -> 48 (+11.63%)						
San Miguel							
Sedgwick							
Summit							
Teller							
Washington							
Weld						339.0 -> 282 (-16.81%)	

Evaluation happens at multiple levels

Multi-level evaluation enables fine-tuning

Table-level comparison with gold standard to help easily see where the pipeline fails

Counties	Owners	Re-Issue	Dealers	Motorcycles Dealers	Motorcycles	Drivers	Lost Tags
Pueblo							
Rio Blanco							
Rio Grande							
Routt		39.0 -> 38 (-2.56%)				80.0 -> 30 (-62.5%)	
Saguache							
San Juan	43.0 -> 48 (+11.63%)						
San Miguel							
Sedgwick							
Summit							
Teller							
Washington							
Weld						339.0 -> 282 (-16.81%)	

state	filename	sample	obs_table	error_type	mean	std	min	25%	50%	75%	max
CO	1917_Counties_Part2	Full	91	diff	-1.144444	7.9726514	-57	0	0	0	5
CO	1917_Counties_Part2	Full	91	diff_pct	-0.010644	0.0809176	-0.625	0	0	0	0.1162791
CO	1917_Counties_Part2	Full	91	diff_abs	1.2555556	7.9557217	0	0	0	0	57
CO	1917_Counties_Part2	Full	91	diff_pct_ab	0.0141676	0.0803672	0	0	0	0	0.625
CO	1917_Counties_Part2	Autos/Totals	13	diff	0.3846154	1.3867505	0	0	0	0	5
CO	1917_Counties_Part2	Autos/Totals	13	diff_pct	0.0089445	0.03225	0	0	0	0	0.1162791
CO	1917_Counties_Part2	Autos/Totals	13	diff_abs	0.3846154	1.3867505	0	0	0	0	5
CO	1917_Counties_Part2	Autos/Totals	13	diff_pct_ab	0.0089445	0.03225	0	0	0	0	0.1162791
CO	1917_Counties_Part2	Autos/Totals - Error Cells	13	diff	5		5	5	5	5	5
CO	1917_Counties_Part2	Autos/Totals - Error Cells	13	diff_pct	0.1162791		0.1162791	0.1162791	0.1162791	0.1162791	0.1162791
CO	1917_Counties_Part2	Autos/Totals - Error Cells	13	diff_abs	5		5	5	5	5	5
CO	1917_Counties_Part2	Autos/Totals - Error Cells	13	diff_pct_ab	0.1162791		0.1162791	0.1162791	0.1162791	0.1162791	0.1162791

Error stats within a table

Evaluation happens at multiple levels

Multi-level evaluation enables fine-tuning

Table-level comparison with gold standard to help easily see where the pipeline fails

Counties	Owners	Re-Issue	Dealers	Motorcycles Dealers	Motorcycles	Drivers	Lost Tags
Pueblo							
Rio Blanco							
Rio Grande							
Routt		39.0 -> 38 (-2.56%)				80.0 -> 30 (-62.5%)	
Saguache							
San Juan	43.0 -> 48 (+11.63%)						
San Miguel							
Sedgwick							
Summit							
Teller							
Washington							
Weld						339.0 -> 282 (-16.81%)	

state	filename	sample	obs_table	error_type	mean	std	min	25%	50%	75%	max
CO	1917_Counties_Part2	Full	91	diff	-1.144444	7.9726514	-57	0	0	0	5
CO	1917_Counties_Part2	Full	91	diff_pct	-0.010644	0.0809176	-0.625	0	0	0	0.1162791
CO	1917_Counties_Part2	Full	91	diff_abs	1.2555556	7.9557217	0	0	0	0	57
CO	1917_Counties_Part2	Full	91	diff_pct_ab	0.0141676	0.0803672	0	0	0	0	0.625
CO	1917_Counties_Part2	Autos/Totals	13	diff	0.3846154	1.3867505	0	0	0	0	5
CO	1917_Counties_Part2	Autos/Totals	13	diff_pct	0.0089445	0.03225	0	0	0	0	0.1162791
CO	1917_Counties_Part2	Autos/Totals	13	diff_abs	0.3846154	1.3867505	0	0	0	0	5
CO	1917_Counties_Part2	Autos/Totals	13	diff_pct_ab	0.0089445	0.03225	0	0	0	0	0.1162791
CO	1917_Counties_Part2	Autos/Totals - Error Cells	13	diff	5		5	5	5	5	5
CO	1917_Counties_Part2	Autos/Totals - Error Cells	13	diff_pct	0.1162791		0.1162791	0.1162791	0.1162791	0.1162791	0.1162791
CO	1917_Counties_Part2	Autos/Totals - Error Cells	13	diff_abs	5		5	5	5	5	5
CO	1917_Counties_Part2	Autos/Totals - Error Cells	13	diff_pct_ab	0.1162791		0.1162791	0.1162791	0.1162791	0.1162791	0.1162791

Error stats within a table

Errors by table

state	filename	num_obs	num_errors	pct_errors	mn_diff	mn_diff_pct	mn_diff_abs	mn_diff_pct_abs	...
AK	1956_Counties	14	0	0	0	0	0	0	...
AL	1927-1928_Counties_Part1	25	0	0	0	0	0	0	...
AL	1927-1928_Counties_Part2	26	0	0	0	0	0	0	...
AL	1927-1928_Counties_Part3	16	0	0	0	0	0	0	...
AL	1930-1931_Counties_Part1	50	1	0.02	-27.5	-0.019985465	27.5	0.019985465	...
AL	1930-1931_Counties_Part2	58	1	0.01724138	17.241379	0.006271873	17.24137931	0.006271873	...
AL	1930-1931_Counties_Part3	22	11	0.5	-1413.091	-0.49933894	1413.090909	0.49933894	...
AR	1913-1926_Part1	338	14	0.04142012	17.461538	0.023185802	21.41420118	0.025622301	...
AR	1913-1926_Part2	338	13	0.03846154	-17.39645	-0.003817044	20.21893491	0.00797901	...
AR	1913-1926_Part3	299	16	0.05351171	2.0167785	0.016514797	2.439597315	0.017545016	...

Aggregate impact of errors is limited

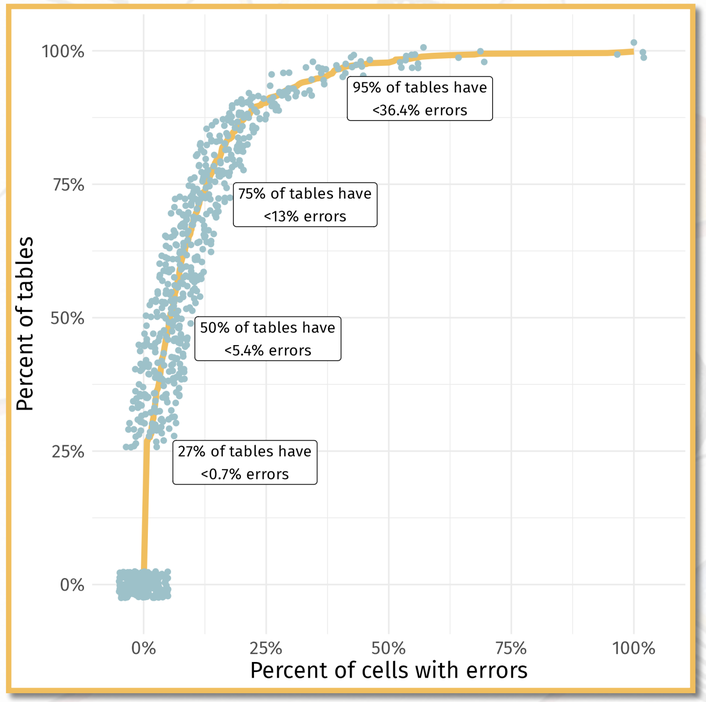
Aggregate impact of errors is limited

Unit of observation: Table
Plot type: Cumulative distribution, jitter for visual effect
Error definition: Gold standard value minus LLM-pipeline output value
Note: Mean computations include observations without errors

Aggregate impact of errors is limited

Unit of observation: Table
Plot type: Cumulative distribution, jitter for visual effect
Error definition: Gold standard value minus LLM-pipeline output value
Note: Mean computations include observations without errors

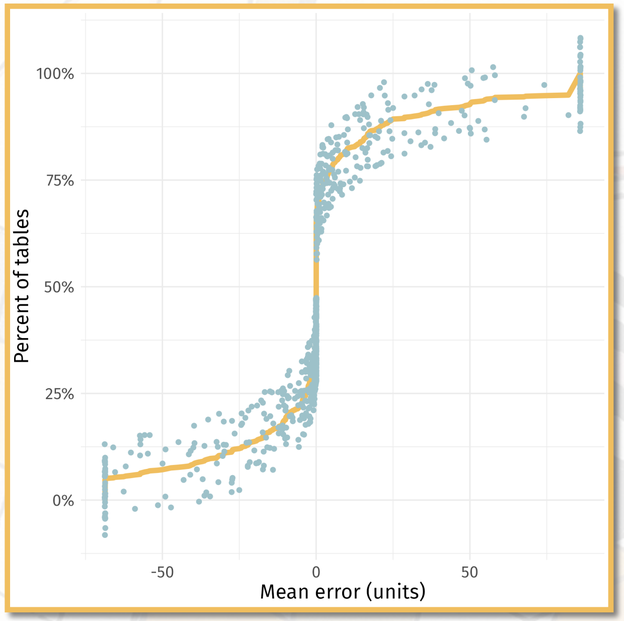
Percentage of cells with errors within a table



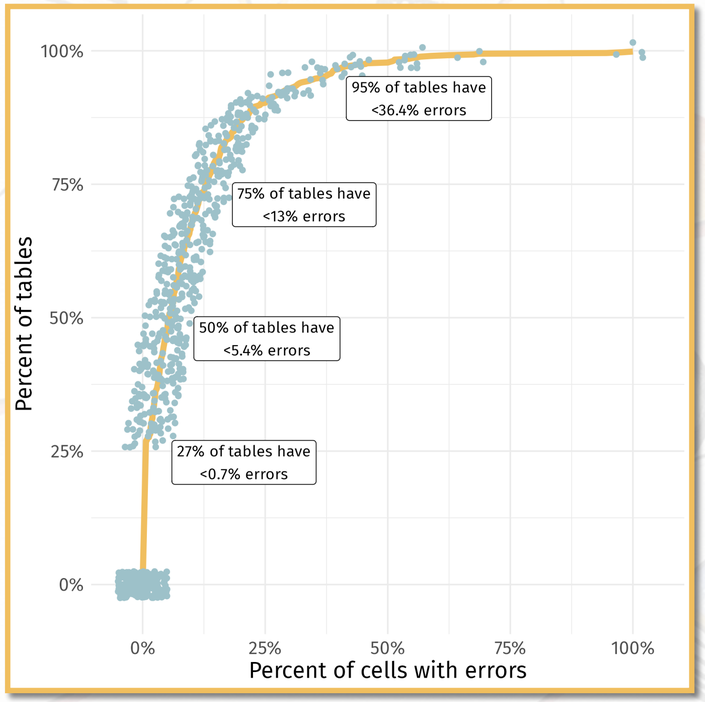
Aggregate impact of errors is limited

Unit of observation: Table
Plot type: Cumulative distribution, jitter for visual effect
Error definition: Gold standard value minus LLM-pipeline output value
Note: Mean computations include observations without errors

Mean error within a table



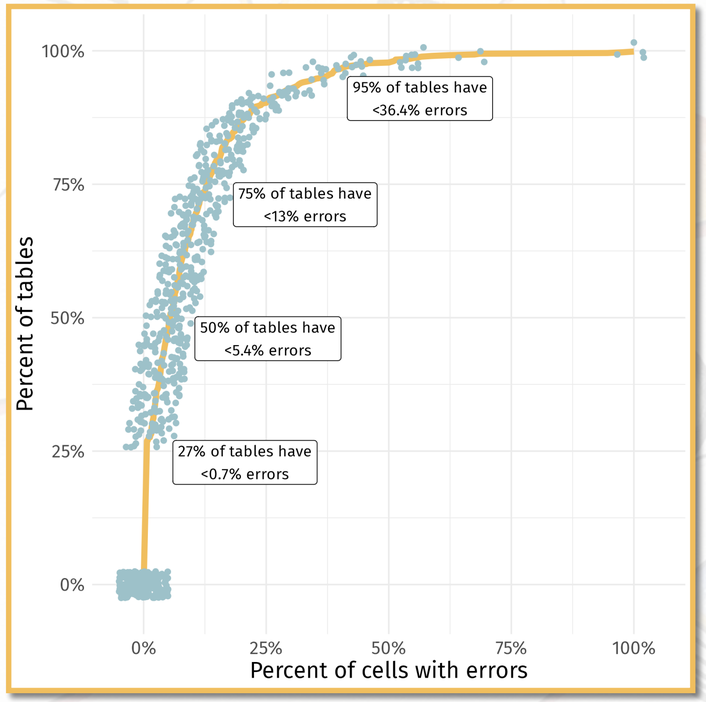
Percentage of cells with errors within a table



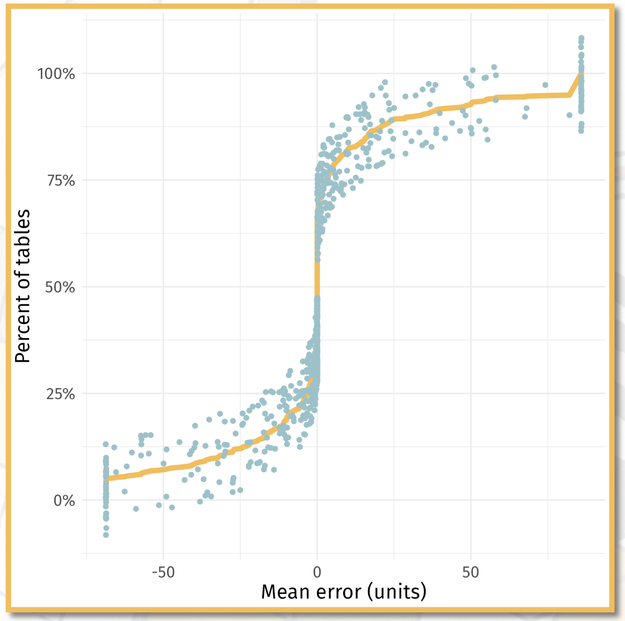
Aggregate impact of errors is limited

Unit of observation: Table
Plot type: Cumulative distribution, jitter for visual effect
Error definition: Gold standard value minus LLM-pipeline output value
Note: Mean computations include observations without errors

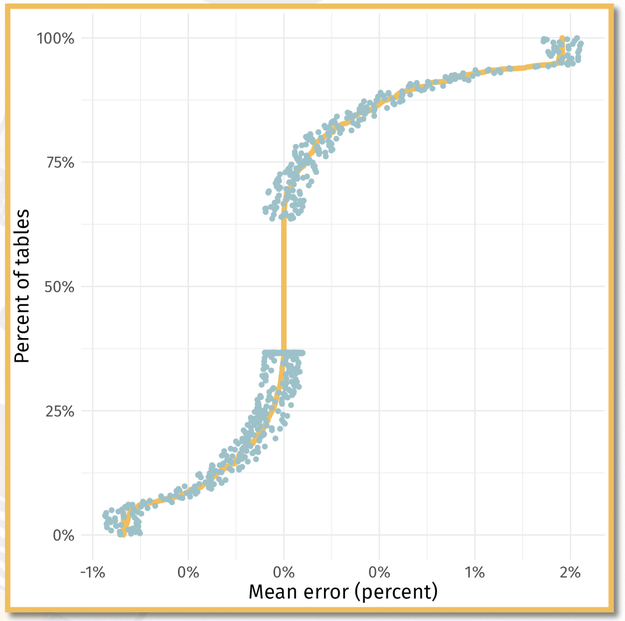
Percentage of cells with errors within a table



Mean error within a table



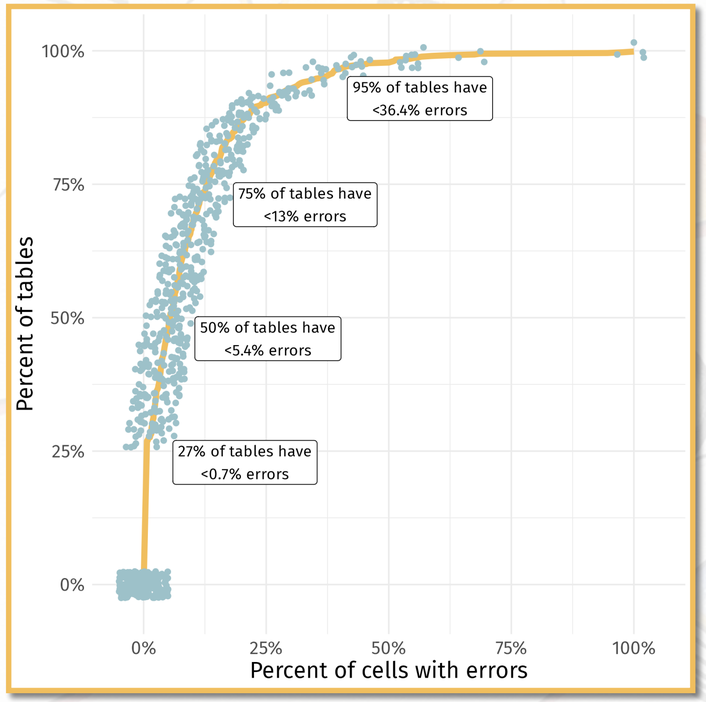
Mean % error within a table



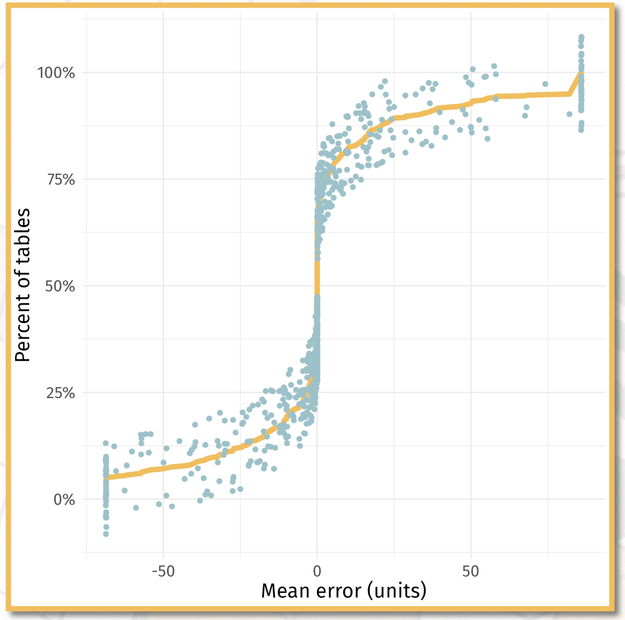
Aggregate impact of errors is limited

Unit of observation: Table
Plot type: Cumulative distribution, jitter for visual effect
Error definition: Gold standard value minus LLM-pipeline output value
Note: Mean computations include observations without errors

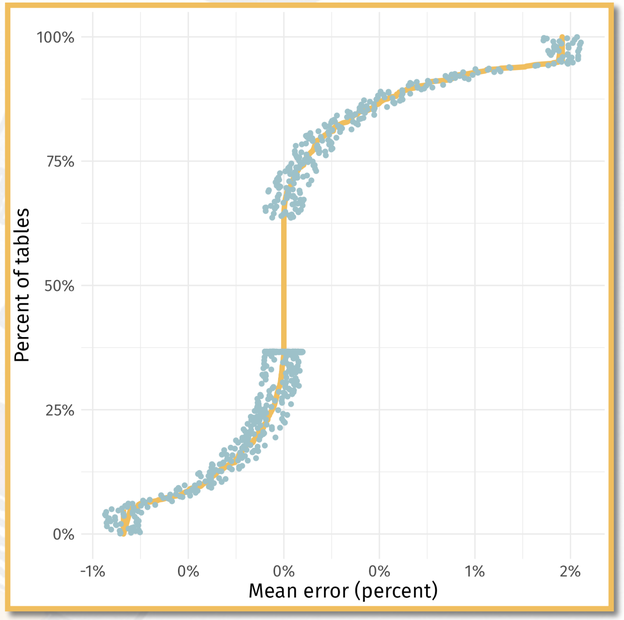
Percentage of cells with errors within a table



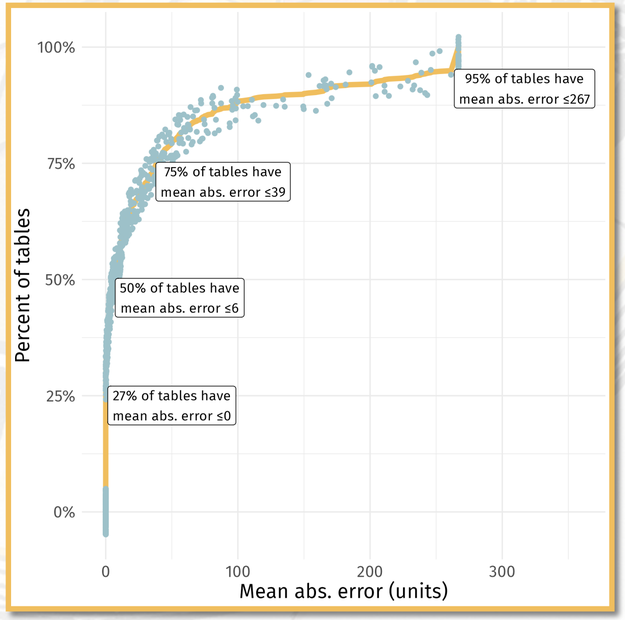
Mean error within a table



Mean % error within a table



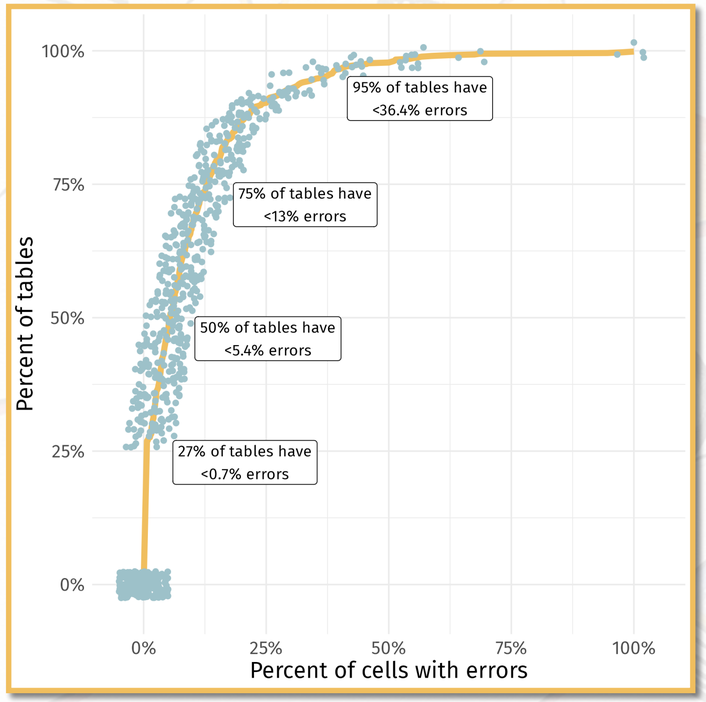
Mean abs. error within a table



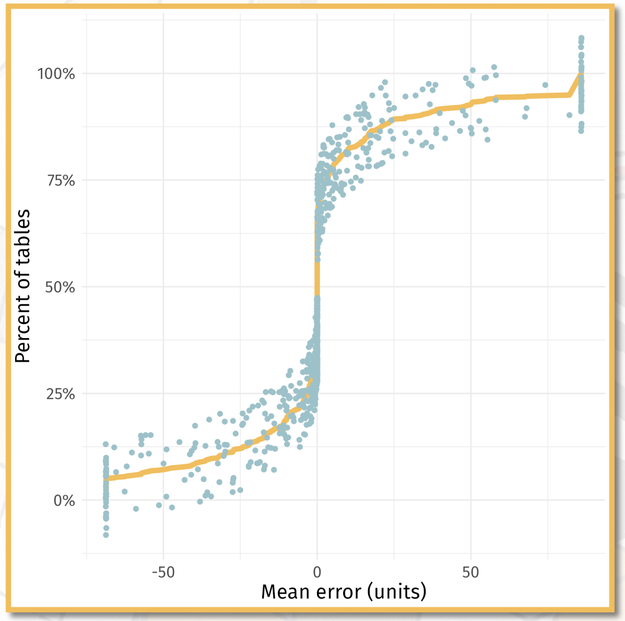
Aggregate impact of errors is limited

Unit of observation: Table
Plot type: Cumulative distribution, jitter for visual effect
Error definition: Gold standard value minus LLM-pipeline output value
Note: Mean computations include observations without errors

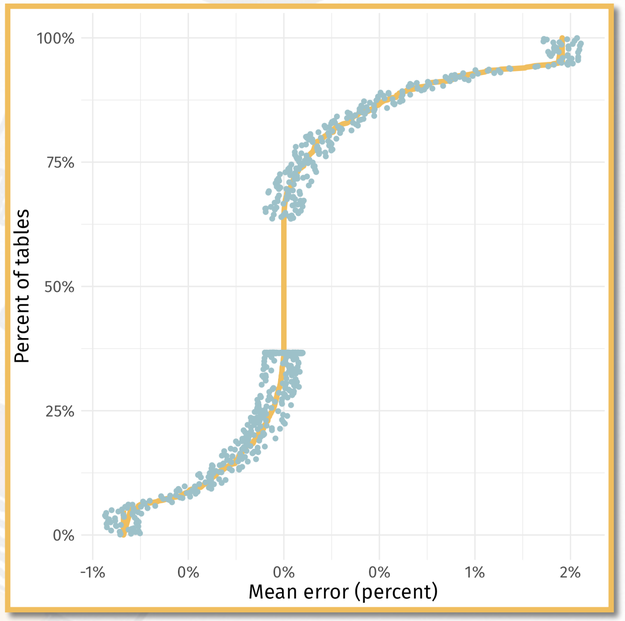
Percentage of cells with errors within a table



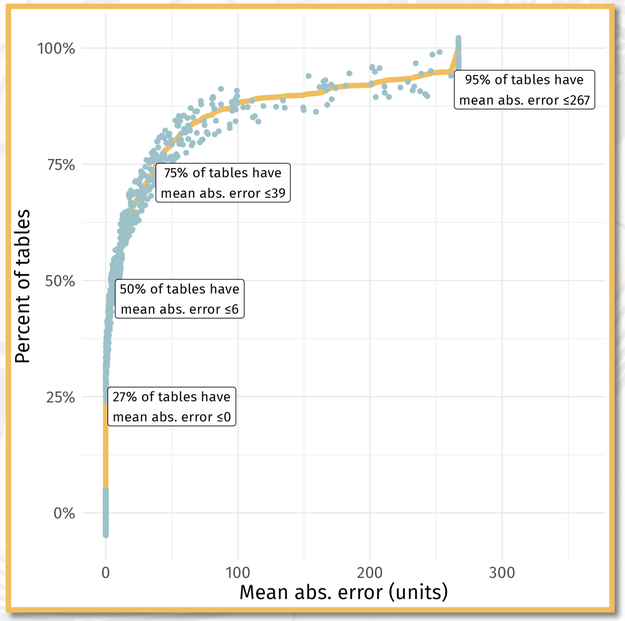
Mean error within a table



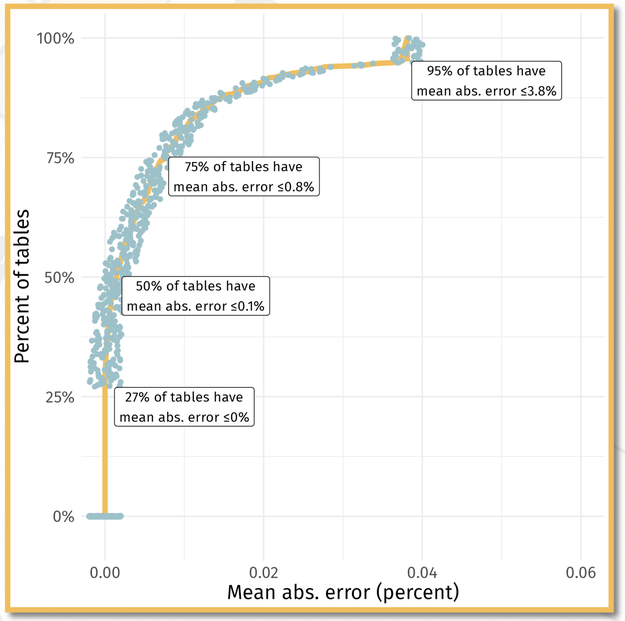
Mean % error within a table



Mean abs. error within a table



Mean abs. % error within a table



Tool doesn’t guarantee accuracy, multilevel evaluation is necessary



If you have money, it’s OK to outsource

💰😄



If you have money, it's OK to outsource



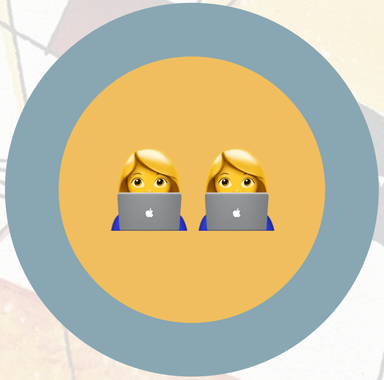
Layout Parser

A unified toolkit for Deep Learning Based Document Image Analysis

PyPI package v0.3.4 | license Apache-2.0 | downloads 328k/month

paper 2103.15348 | website layout-parser.github.io | doc layout-parser.readthedocs.io

If you're willing to tool up, that's always good





If you have money, it's OK to outsource

💰😁

Layout Parser

A unified toolkit for Deep Learning Based Document Image Analysis

PyPI package v0.3.4 | license Apache-2.0 | downloads 328k/month

paper 2103.15348 | website layout-parser.github.io | doc layout-parser.readthedocs.io

If you're willing to tool up, that's always good

👩‍💻👩‍💻



If you have money, it's OK to outsource

💰😄

Layout Parser

A unified toolkit for Deep Learning Based Document Image Analysis

PyPI package v0.3.4 | license Apache-2.0 | downloads 328k/month

paper 2103.15348 | website layout-parser.github.io | doc layout-parser.readthedocs.io

If you're willing to tool up, that's always good

👩‍💻👩‍💻

🗽

🤩🤩🤩

LLMs democratize historical dataset creation



If you have money, it's OK to outsource

💰😄

Layout Parser

A unified toolkit for Deep Learning Based Document Image Analysis

PyPI package v0.3.4 | license Apache-2.0 | downloads 328k/month

paper 2103.15348 | website layout-parser.github.io | doc layout-parser.readthedocs.io

If you're willing to tool up, that's always good

👩‍💻👩‍💻

🗽

🤩🤩🤩

LLMs democratize historical dataset creation

Good quality of outputs



If you have money, it's OK to outsource

💰😄

Layout Parser

A unified toolkit for Deep Learning Based Document Image Analysis

PyPI package v0.3.4 | license Apache-2.0 | downloads 328k/month

paper 2103.15348 | website layout-parser.github.io | doc layout-parser.readthedocs.io

If you're willing to tool up, that's always good

👩‍💻👩‍💻

🗽

🤩🤩🤩

LLMs democratize historical dataset creation

Good quality of outputs

Low costs



If you have money, it's OK to outsource

💰😁

Layout Parser

A unified toolkit for Deep Learning Based Document Image Analysis

PyPI package v0.3.4 | license Apache-2.0 | downloads 328k/month

paper 2103.15348 | website layout-parser.github.io | doc layout-parser.readthedocs.io

If you're willing to tool up, that's always good

👩‍💻👩‍💻

🗽

🤩🤩🤩

LLMs democratize historical dataset creation

Good quality of outputs

Low costs

Tool operated in English



If you have money, it's OK to outsource

💰😄

Layout Parser

A unified toolkit for Deep Learning Based Document Image Analysis

PyPI package v0.3.4 | license Apache-2.0 | downloads 328k/month

paper 2103.15348 | website layout-parser.github.io | doc layout-parser.readthedocs.io

If you're willing to tool up, that's always good

👩‍💻👩‍💻

🗽

🤩🤩🤩

LLMs democratize historical dataset creation

Good quality of outputs

Low costs

Tool operated in English

Single tool from (cropped) scans to cohesive dataset


Thank
you!

Who is this for?



If you have a huge RA budget, feel free to keep using them

Personally, I'm shocked at the cost/efficiency here and think it can be a major accessibility moment

Or you can mix and match of course, LLM is especially good for homogenization

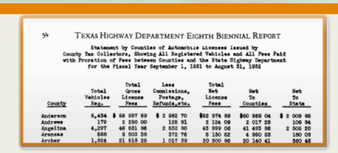
Layout Parser

A unified toolkit for Deep Learning Based Document Image Analysis

PyPI package v0.3.4 | license Apache-2.0 | downloads 328k/month

paper 2103.15348 | website layout-parser.github.io | doc layout-parser.readthedocs.io

If the deep learning pipeline gives you what you need, keep using it



TEXAS HIGHWAY DEPARTMENT EIGHTH BIENNIAL REPORT

Statement by Counties of Automobile Licenses issued by County Tax Collectors, Showing All Registered Vehicles and All Fees Paid with Proration of Fees between Counties and the State Highway Department for the Fiscal Year September 1, 1931 to August 31, 1932

County	Total Vehicles Reg.	Total Gross License Fees	Less Commissions, Postage, Refunds, etc.	Total Net License Fees	Net to Counties	Net to State
Anderson	5,484	$ 55 937 59	$ 2 962 70	$52 974 89	$50 065 04	$ 2 909 85
Andrews	179	2 250 00	125 91	2 124 09	2 017 28	106 81
Angelina	4,297	46 531 86	2 532 80	43 999 06	41 605 86	2 393 20
Aransas	588	5 503 38	372 76	5 130 62	4 850 58	280 04
Archer	1,924	21 518 25	1 017 39	20 500 86	20 140 42	360 44

VM Transcribe this table into markdown?

Here's the table transcribed into markdown format:

County	Total Vehicles Reg.	Total Gross License Fees	Less Commissions, Postage, Refunds
Anderson	5,484	$ 55,937.59	$ 2,962.70
Andrews	179	2,250.00	125.91
Angelina	4,297	46,531.86	2,532.80
Aransas	588	5,503.38	372.76
Archer	1,924	21,518.25	1,017.39

Copy Retry

Claude can make mistakes. Please double-check responses.

We hope that LLM pipelines brings the fun of cohesive dataset creation to new people due to low cost, flexibility and flatter learning curve

Outlier detection reliably identifies tables with major problems

What's a major problem?

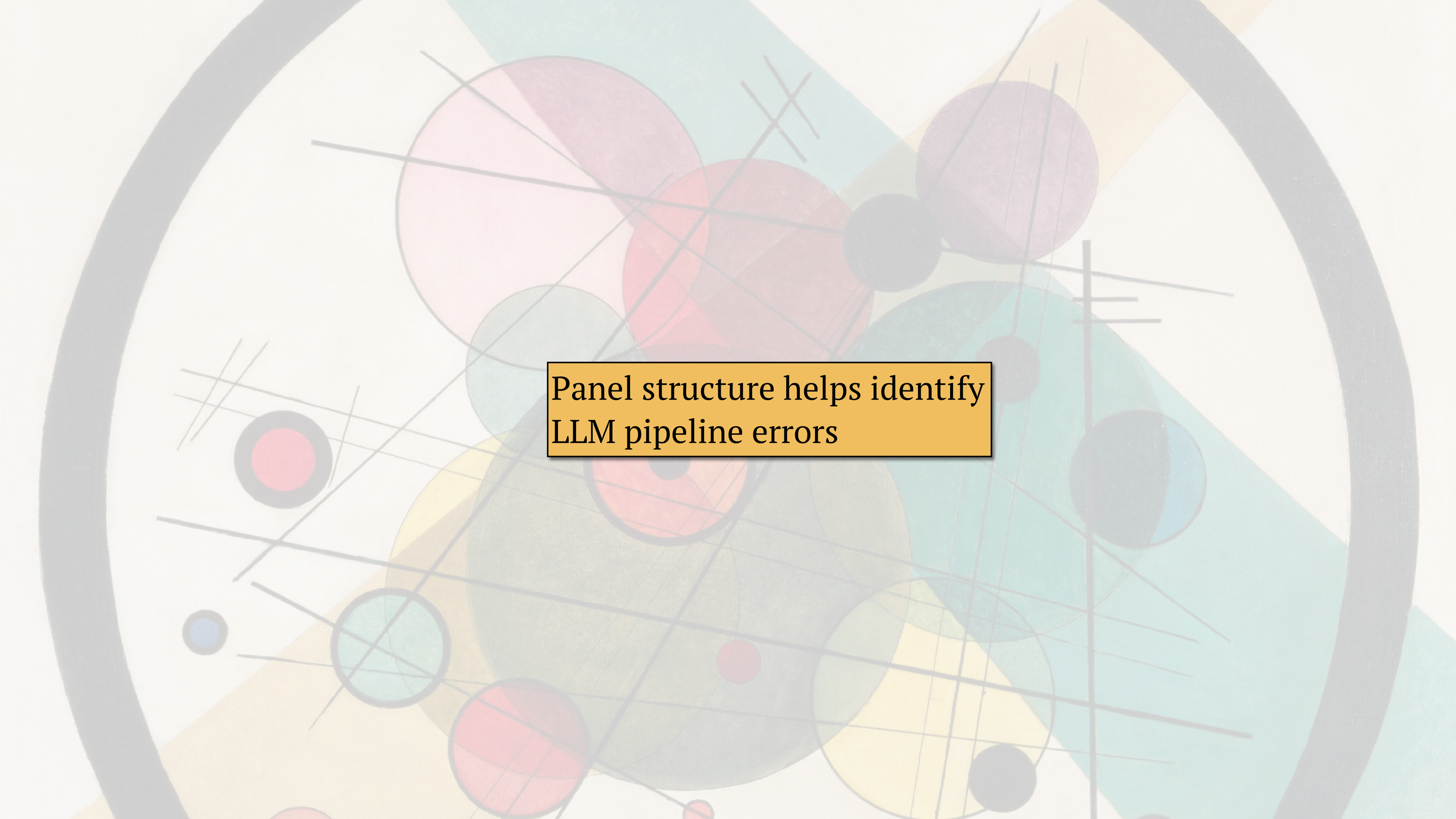
- A lot of errors **AND**
- % errors are large **AND NOT**
- Unit errors are small

17 major problems

Outlier types:

- Time series: if jump relative to t-1 and t+1 of > p_{ts} (default 150%)
- Field-specific: if number of cars < p_f (default 30%) of total vehicles number of cars > total vehicles
- Repeat files: values for the same county year differ by > p_r (default 50%)

Outliers detected	Major Problem	
	TRUE	FALSE
TRUE	**16**	16
FALSE	**1**	697


Panel structure helps identify
LLM pipeline errors