Public Vehicle Data in the United States: A Practical Guide for Academic Researchers

Andreas Kraft*

September 2, 2025

Abstract

This guide provides an overview of publicly available vehicle transaction datasets in the United States, with the aim of helping researchers identify, access, and effectively use these resources for academic work. Although no centralized database exists—unlike in advertising or consumer packaged goods—state-level DMV records and federal sources offer rich information on vehicle sales, registrations, and attributes. The guide compiles details on frequently used DMV data, VIN decoding tools, leasing and loan information from asset-backed securities, as well as supplementary datasets on recalls, accidents, fueling infrastructure, and rideshare activity. By outlining the key features, limitations, and access procedures for each source, the guide seeks to lower barriers to entry and enable rigorous empirical research in economics, marketing, and related fields that rely on vehicle-level transaction data.

*Andreas Kraft (andreas.kraft@chicagobooth.edu) is an Assistant Professor of Marketing and Asness Faculty Fellow at the Booth School of Business, University of Chicago.

1 Introduction

Durable goods account for a substantial share of household expenditure, and for most U.S. consumers the purchase of a vehicle is the largest and most consequential durable transaction they will undertake. Vehicles are not only costly, long-lived, and frequently financed; they are also central to daily mobility, embedded in public infrastructure, and subject to regulation at multiple levels of government. These features make the vehicle market a particularly rich setting for studying a wide range of economic and marketing phenomena.

Unsurprisingly, vehicle markets have played a foundational role in empirical research. Classic contributions include Akerlof (1970), which introduced the concept of asymmetric information through the used-car market; Berry et al. (1995), which developed the random-coefficients logit model using new-car demand; and Rust (1987), which estimated a dynamic discrete-choice model of engine replacement using bus data.

Yet working with vehicle market data presents distinct challenges. Unlike consumer packaged goods or television advertising—domains where standardized, nationally representative datasets are readily accessible—vehicle transactions and ownership records in the United States are highly decentralized. Each state maintains its own registration and title databases through its Department of Motor Vehicles (DMV), while complementary information on safety recalls, technical attributes, financing terms, and usage is dispersed across federal agencies and local authorities. Although several commercial vendors compile and sell vehicle data (Murry and Schneider, 2016), a wide array of public datasets also exists, often free of charge or available for a modest fee. These sources remain underutilized despite their potential to support rigorous, policy-relevant research.

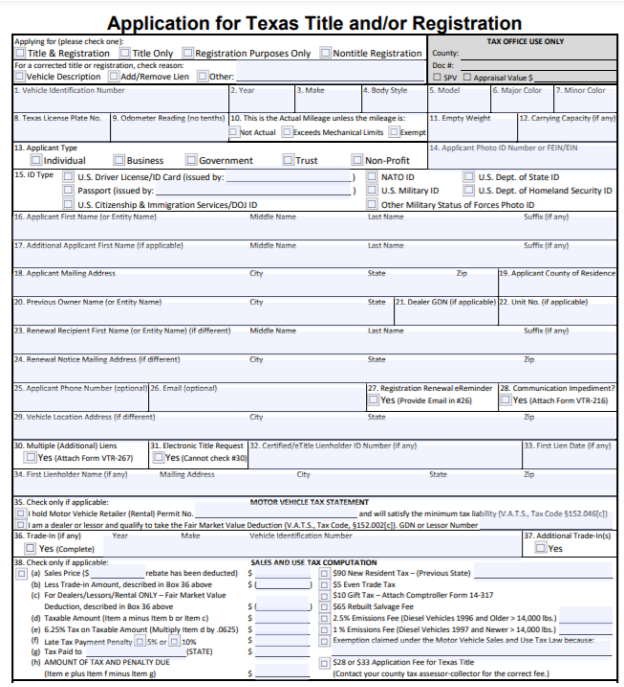
The aim of this guide is to lower barriers to the use of public vehicle data by cataloging available sources, identifying common challenges, and offering practical guidance. I begin with state DMV datasets, focusing in particular on Texas, which I have worked with myself. I then review additional resources that may be valuable depending on the research question, including VIN decoding services, fuel-economy ratings, safety recalls and investigations, leasing and loan data from asset-backed securities, fueling infrastructure, accident records, and data on taxi and rideshare activity.

This guide is intended for researchers who are either entering this domain for the first time or exploring vehicle data as a new empirical context. In addition to describing available data and

their structure, I highlight important idiosyncrasies and limitations that are not always evident in published work. By consolidating this information, my goal is to make vehicle market data more accessible and to encourage further creative and rigorous research in this area.

2 Registrations and Transactions (DMV Datasets)

Vehicle sales are generally reported to the Department of Motor Vehicles (DMV) in each U.S. state. For example, when a consumer purchases or registers a vehicle in Texas, they must complete a title and registration form (Figure 1).



Application for Texas Title and/or Registration

Applying for (please check one): ☐ Title & Registration ☐ Title Only ☐ Registration Purposes Only ☐ Nontitle Registration

For a corrected title or registration, check reason: ☐ Vehicle Description ☐ Add/Remove Lien ☐ Other: _____

TAX OFFICE USE ONLY — County: _____ Doc #: _____ ☐ SPV ☐ Appraisal Value $ _____

1. Vehicle Identification Number | 2. Year | 3. Make | 4. Body Style | 5. Model | 6. Major Color | 7. Minor Color

8. Texas License Plate No. | 9. Odometer Reading (no tenths) | 10. This is the Actual Mileage unless the mileage is: ☐ Not Actual ☐ Exceeds Mechanical Limits ☐ Exempt | 11. Empty Weight | 12. Carrying Capacity (if any)

13. Applicant Type ☐ Individual ☐ Business ☐ Government ☐ Trust ☐ Non-Profit — 14. Applicant Photo ID Number or FEIN/TIN

15. ID Type ☐ U.S. Driver License/ID Card (issued by: _____) ☐ Passport (issued by: _____) ☐ U.S. Citizenship & Immigration Services/DOJ ID ☐ NATO ID ☐ U.S. Military ID ☐ Other Military Status of Forces Photo ID ☐ U.S. Dept. of State ID ☐ U.S. Dept. of Homeland Security ID

16. Applicant First Name (or Entity Name) | Middle Name | Last Name | Suffix (if any)

17. Additional Applicant First Name (if applicable) | Middle Name | Last Name | Suffix (if any)

18. Applicant Mailing Address | City | State | Zip | 19. Applicant County of Residence

20. Previous Owner Name (or Entity Name) | City | State | 21. Dealer GDN (if applicable) | 22. Unit No. (if applicable)

23. Renewal Recipient First Name (or Entity Name) (if different) | Middle Name | Last Name | Suffix (if any)

24. Renewal Notice Mailing Address (if different) | City | State | Zip

25. Applicant Phone Number (optional) | 26. Email (optional) | 27. Registration Renewal eReminder ☐ Yes (Provide Email in #26) | 28. Communication Impediment? ☐ Yes (Attach Form VTR-216)

29. Vehicle Location Address (if different) | City | State | Zip

30. Multiple (Additional) Liens ☐ Yes (Attach Form VTR-267) | 31. Electronic Title Request ☐ Yes (Cannot check #30) | 32. Certified/eTitle Lienholder ID Number (if any) | 33. First Lien Date (if any)

34. First Lienholder Name (if any) | Mailing Address | City | State | Zip

35. Check only if applicable: — MOTOR VEHICLE TAX STATEMENT
☐ I hold Motor Vehicle Retailer (Rental) Permit No. _____ and will satisfy the minimum tax liability (V.A.T.S., Tax Code §152.046[c])
☐ I am a dealer or lessor and qualify to take the Fair Market Value Deduction (V.A.T.S., Tax Code, §152.002[c]). GDN or Lessor Number _____

36. Trade-In (if any) ☐ Yes (Complete) | Year | Make | Vehicle Identification Number | 37. Additional Trade-In(s) ☐ Yes

38. Check only if applicable: — SALES AND USE TAX COMPUTATION
☐ (a) Sales Price ($ _____ rebate has been deducted) $ _____
(b) Less Trade-in Amount, described in Box 36 above $ (_____)
(c) For Dealers/Lessors/Rental ONLY – Fair Market Value Deduction, described in Box 36 above $ (_____)
(d) Taxable Amount (Item a minus Item b or Item c) $ _____
(e) 6.25% Tax on Taxable Amount (Multiply Item d by .0625) $ _____
(f) Late Tax Payment Penalty ☐ 5% or ☐ 10% $ _____
(g) Tax Paid to _____ (STATE) $ _____
(h) AMOUNT OF TAX AND PENALTY DUE (Item e plus Item f minus Item g) $ _____
☐ $90 New Resident Tax – (Previous State) _____
☐ $5 Even Trade Tax
☐ $10 Gift Tax – Attach Comptroller Form 14-317
☐ $65 Rebuilt Salvage Fee
☐ 2.5% Emissions Fee (Diesel Vehicles 1996 and Older > 14,000 lbs.) _____
☐ 1 % Emissions Fee (Diesel Vehicles 1997 and Newer > 14,000 lbs.) _____
☐ Exemption claimed under the Motor Vehicle Sales and Use Tax Law because: _____
☐ $28 or $33 Application Fee for Texas Title (Contact your county tax assessor-collector for the correct fee.)

Figure 1: Title Form (Texas)

The form requires the buyer to provide information about themselves as well as the vehicle. In Texas, this includes the vehicle identification number (VIN), the make and model, and the model year.

2.1 Texas

I (Kraft and Rao, 2024), along with other researchers (e.g., Yavorsky et al., 2021; Liu, 2022; Sagl, 2023), have used the Texas DMV data, which are unusually rich in observable attributes and

cover both registrations and transactions. The data capture the universe of vehicle activity in the state, including millions of observations each year. For example, in 2018 alone there were roughly 19 million registration renewals, 5 million sales transactions, more than one million tax-exempt transfers, and about 200,000 new resident registrations (Table 1). For each transaction, the dataset includes consumer information (such as name and address), vehicle details (VIN, make, model, model year, and sales price), the trade-in allowance, and dealership identifiers.

SALES TAX CATEGORY	N
Registration Renewal	19,261,165
SALES/USE	4,759,319
EXEMPT	1,169,011
NEW RESIDENT	218,737
GIFT	188,641
EVEN TRADE	16,926

Table 1: Sales Tax Category Distribution of Texas DMV Dataset (Sample Year 2018)

A distinctive feature of the Texas data is the presence of a dealership identifier. This allows researchers to distinguish between consumer-to-consumer transactions and those conducted through dealerships. In 2018, about 30% of vehicle transactions occurred in the consumer-to-consumer market. To illustrate the variation in transaction prices across market segments, Figure 2 highlights the distinct pricing patterns observed in consumer-to-consumer versus dealership sales prices.

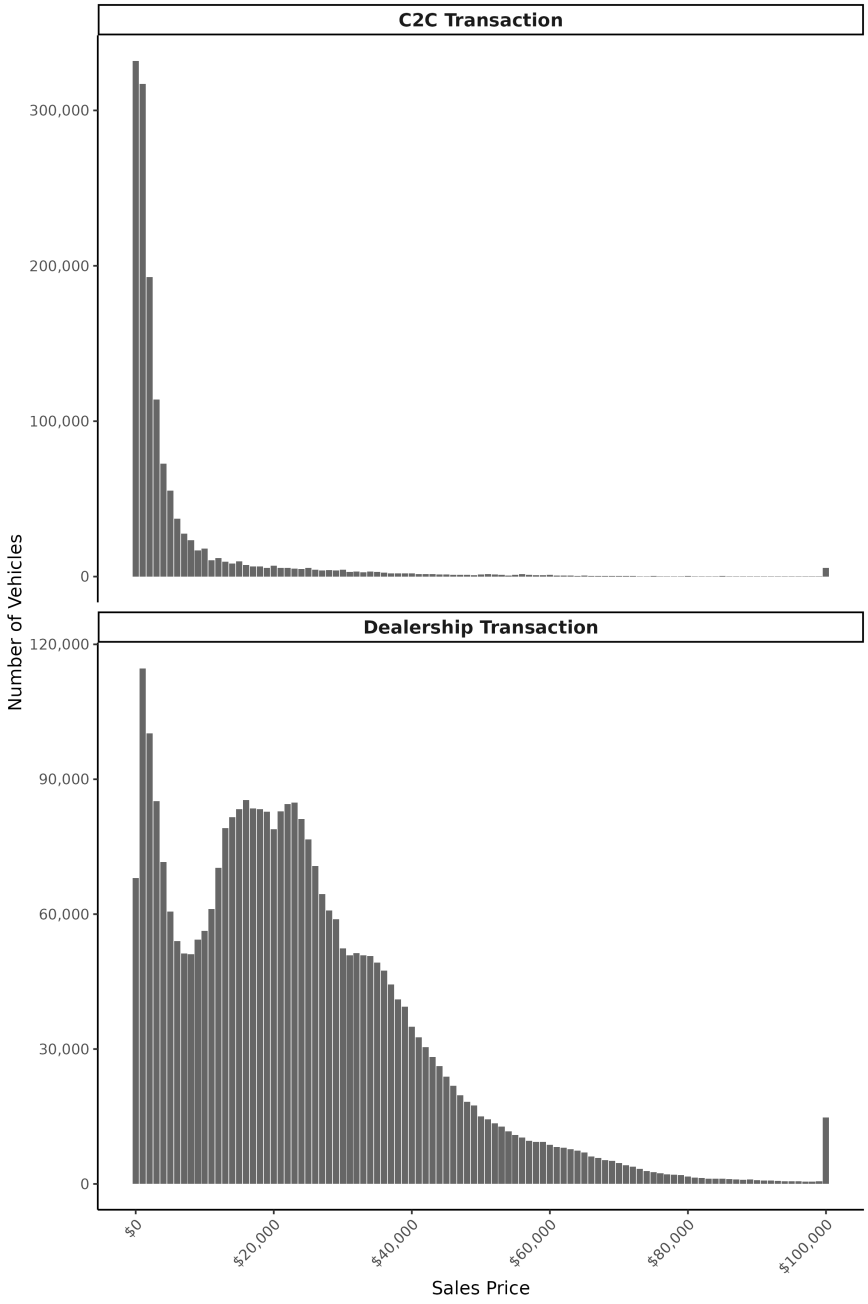


Figure 2: Distribution of prices in consumer-to-consumer and dealership transactions (2018).

Another important consideration is the availability of odometer readings, which are subject to regulatory exemptions. Until 2021, vehicles more than ten years old were not required to report odometer values. Beginning in 2021, the exemption threshold was extended to vehicles more than twenty years older than the current year. Because this change was applied prospectively, vehicles older than model year 2011 are exempt from reporting, while younger vehicles are not. This struc-

ture can introduce sample-selection concerns: using odometer readings as a conditioning variable effectively restricts the analysis to relatively new vehicles. However, the pattern of missingness varies both across vehicles sold at different points within a calendar year and between transactions occurring before and after the 2021 rule change. Figure 3 shows the percentage of missing odometer readings by vehicle year among cars sold in 2018, illustrating how missingness increases drastically after 10 years.

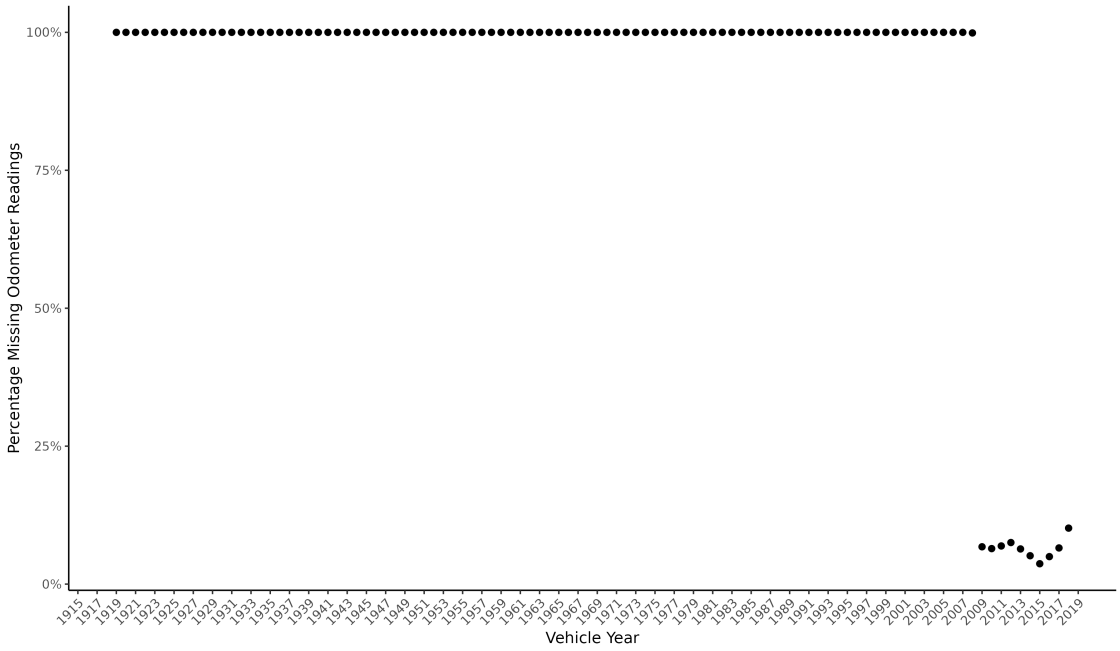


Figure 3: Percentage of vehicle transactions with missing odometer readings (2018).

Access to the Texas DMV data is possible through two channels. First, it is sometimes feasible to share previously obtained datasets with other academic researchers.[1] Second, researchers can submit an open-records request directly to the Texas Department of Motor Vehicles. The DMV charges a fee that reflects the cost of labor involved in preparing the data.

2.1.1 New Mexico

I have also obtained data from the state of New Mexico, although I have not yet used it in research. This dataset contains detailed information on vehicle transactions, including the transaction date

[1] If you are an academic and interested in working with the 2014–2025 historical data, feel free to reach out to me. I can request the Texas DMV to grant you access. If approved by the DMV, you may use the data for your own research—no co-authorship expected or required.

and code, VIN, make and model, vehicle color, registration start and expiration dates, and geographic identifiers down to the ZIP code plus four digits. It also records vehicle characteristics (such as profile, ownership type, and odometer reading), acquisition type, dealer identifiers, and financial details including purchase price, taxable amount, and excise tax.

2.1.2 Other States

Comparable DMV data have also been successfully obtained and used in research on Virginia (Murry, 2017), Pennsylvania (Biglaiser and Li, 2018), and Ohio (Murry and Zhou, 2020). See those papers for descriptions of the data.

3 Other Publicly Available Data Sources

In addition to state DMV records, several other datasets may be of interest to researchers. These sources provide complementary information on vehicles that can be either merged by VIN or make/model/modelyear.

3.1 Vehicle Information

The DMV datasets described above can be combined with additional vehicle-level information. Some sources are outlined below.

3.1.1 VIN-Level Information

Every vehicle is assigned a unique Vehicle Identification Number (VIN). Since 1981, the U.S. National Highway Traffic Safety Administration (NHTSA) has required all vehicles sold in the United States to carry a standardized 17-character VIN. The first three digits identify the manufacturer; digits 4 through 8 capture vehicle attributes; digit 9 is a "check digit" that verifies authenticity; digit 10 indicates the model year; digit 11 identifies the plant code; and digits 12 through 17 are the sequential production number.

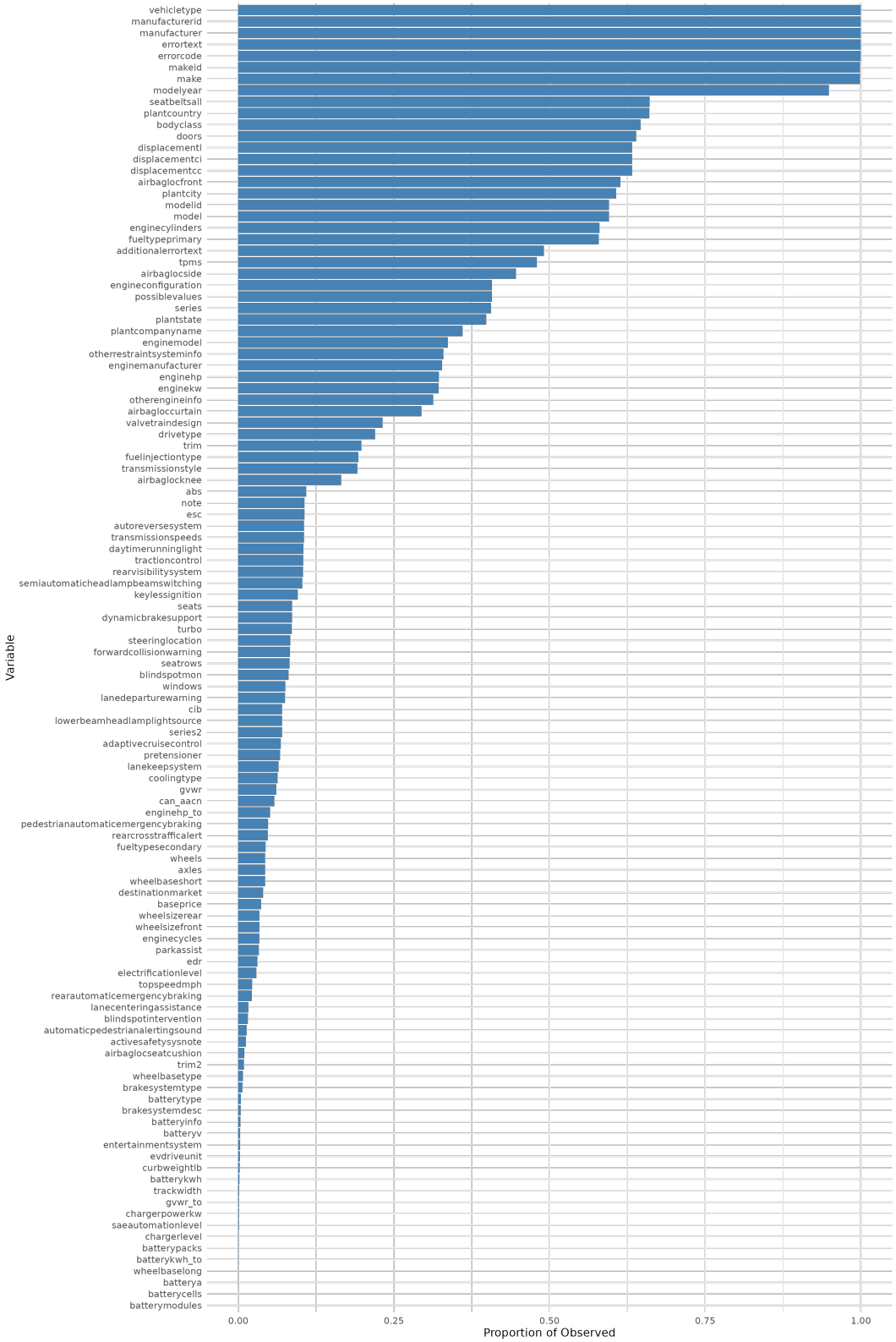


Figure 4: Proportion of variables observed for passenger vehicles in VIN data (Texas sample, 2014–2024).

Some DMV datasets include basic details such as vehicle class, model year, or make and model. However, the most complete source of VIN-based information is provided by NHTSA through its VIN decoding service.[2] This service is available as an API, a browser-based lookup tool, and a downloadable database.

When decoding VINs, it is useful to note that the sequential number (digits 12–17) and the check digit (digit 9) do not affect the decoded information. In practice, researchers can extract digits 1–11 and replace digit 9 with a wildcard, thereby reducing the number of VINs that need to be decoded.

The database contains up to 148 potential vehicle attributes, though most fields are missing for the majority of vehicles. The most reliably available attributes include manufacturer, make, model year, vehicle type, seatbelts, plant country and city, body class, doors, and engine displacement. To provide a sense of completeness, Figure 4 shows the proportion of fields available for passenger vehicle transactions in Texas between 2014 and 2024.

Notably, the NHTSA VIN data do not include all attributes potentially of interest to researchers. Missing variables include trim details, vehicle color, and fuel economy ratings.

3.1.2 Fuel Economy Information

Detailed fuel economy data can be obtained from the U.S. Department of Energy.[3] These datasets include standardized laboratory test results for city, highway, and combined fuel economy.

However, the data are not linked directly to VINs. Instead, each observation is identified by make, model, and model year. Merging these data with VIN-level transactions therefore requires matching on these variables. This process can introduce ambiguity—particularly for models with multiple trims or engine configurations—so researchers should exercise caution when assigning fuel economy values at the vehicle level.

3.1.3 Vehicle Recall, Investigation, and Complaints Data

In addition to its regulatory role, NHTSA collects and publishes a wide range of vehicle safety data.[4] This includes:

[2] See https://vpic.nhtsa.dot.gov/api/ for access to the API.
[3] See https://www.fueleconomy.gov/feg/download.shtml.
[4] See https://www.nhtsa.gov/nhtsa-datasets-and-apis.

- **Safety ratings** for new vehicles, based on standardized crash tests;
- **Consumer complaints** submitted to NHTSA about potential defects;
- **Defect investigations**, documenting issues under formal review; and
- **Recall data**, covering vehicles subject to safety-related recalls.

These datasets are valuable for studying product safety regulation, consumer responses to recalls, manufacturer compliance, and the relationship between safety concerns and market outcomes. They are regularly updated and available through both bulk downloads and APIs.

3.2 Lease & Loan Data

A substantial share of vehicle purchases in the U.S. are financed through loans or structured as leases. Yet, there is no comprehensive dataset that provides financing information at the VIN level across the entire market. One valuable alternative for researchers is the information disclosed through asset-backed securities (ABS). Since November 2016, the U.S. Securities and Exchange Commission (SEC) has required issuers of certain ABS to publicly report detailed information about the underlying assets. While much of this disclosure concerns real estate–backed instruments, an increasing number of vehicle loans and leases are securitized and included in these filings.

Although ABS data are not representative of the vehicle financing market—many securitized loans originate from captive finance arms or target specific consumer segments—they nonetheless offer several important advantages for research. First, they provide longitudinal tracking of loans and leases, enabling the study of repayment behavior, defaults, and contract terms over time. Second, the disclosures include information on borrowers, such as credit score, income, and geographic location—variables rarely available in other vehicle-level datasets. Third, each record contains rich detail on the financed vehicle, including its make, model, model year, whether it was new or used at purchase, and the terms of the financing agreement.

Taken together, these features make ABS data a valuable resource for studying consumer credit risk, the role of financing in vehicle purchases, and the interaction between vehicle characteristics and borrower behavior.

3.3 Vehicle Accident Data

Several U.S. states collect and publish detailed records of vehicle accidents, providing researchers with valuable information on traffic incidents, driver behavior, and safety outcomes. For instance, the state of Texas maintains a comprehensive crash database that covers all accidents reported by law enforcement agencies statewide.[5] These data are compiled by the Texas Department of Transportation (TxDOT) and made available for public use, subject to privacy restrictions.

The Texas dataset includes rich contextual information: the date, time, and location of each crash; the number and types of vehicles involved; contributing factors such as speeding, distraction, or intoxication; road and weather conditions; and the severity of injuries or fatalities. In some cases, the data also contain details about the vehicles (i.e., VIN).

Because collection practices and access policies differ by state, researchers should consult the relevant state agency for documentation, coverage, and request procedures.

3.4 Additional Relevant Data

There are a number of additional datasets that might be relevant for researchers interested in research regarding vehicle transactions.

3.4.1 Vehicle Dealership Data

In addition to transaction and registration records, researchers may also benefit from datasets capturing the universe of licensed dealerships. In Texas, the Department of Motor Vehicles maintains a publicly available database of active motor vehicle dealers.[6] This resource is updated regularly and provides detailed information on all currently licensed dealerships.

For historical coverage, records of closed dealerships are not posted online but can be obtained through the state's Open Records request process. When linked to vehicle transaction data, these dealership records enable the construction of dealership-level panels, making it possible to study entry and exit, competitive dynamics, and the evolution of dealership activity over time.

[5] See https://www.txdot.gov/data-maps/crash-reports-records/crash-data-analysis-statistics.html.
[6] See `https://texasdmv.my.salesforce-sites.com/dealers/motorvehicledealerliststaging`.

3.4.2 Fueling and Charging Locations

As of February 2025, the United States had roughly 84,000 alternative fueling stations and more than 150,000 conventional gasoline stations. For researchers interested in fuel infrastructure—whether to study energy transitions, transportation access, or patterns of vehicle adoption—several data sources provide valuable information, though coverage and availability differ between alternative and conventional fueling stations.

The U.S. Department of Energy maintains the most comprehensive, publicly available dataset on alternative fueling stations through its Alternative Fuels Data Center (AFDC).[7] The AFDC dataset covers a wide range of fuel types, including biodiesel, compressed natural gas (CNG), electricity (EV charging), ethanol (E85), hydrogen, liquefied natural gas (LNG), propane, and renewable diesel. It is updated regularly and includes detailed attributes for each station, such as geographic coordinates, fuel type(s) offered, access restrictions, operational status, hours of availability, accepted payment methods, and the presence of charging or fueling equipment. These features make the AFDC dataset particularly valuable for studying the diffusion of new fuel technologies, accessibility of charging infrastructure, and the effects of policy interventions on alternative energy adoption.

By contrast, no single comprehensive and publicly available dataset exists for conventional gasoline stations. This creates challenges for researchers studying the traditional fuel retail network. Some state-level administrative data may serve as proxies. For example, in Texas, gas stations must register and obtain a license to sell fuel. Researchers can use the state's licensing database[8] to compile a list of active fuel retailers. While this approach can help identify station locations, the licensing data generally lack details such as station size, fuel offerings, or hours of operation.

3.4.3 Taxis & Rideshare Data

Researchers interested in urban transportation, vehicle usage patterns, or the economics of mobility services may find valuable datasets related to taxi and rideshare activity. These data are typically collected at the city level, either by municipal agencies or through reporting requirements imposed on private operators. While coverage is not uniform across the U.S., several cities—most notably New York City and Chicago—make detailed records publicly available, offering rich opportunities

[7] See https://developer.nrel.gov/docs/transportation/alt-fuel-stations-v1/all/.
[8] See https://www.tdlr.texas.gov/LicenseSearch/licfile.asp.

for empirical work on transportation systems, pricing, access, and regulation.

The most extensive dataset comes from the New York City Taxi & Limousine Commission (TLC), which has released trip-level data for all yellow and green taxi rides since January 2009.[9] The dataset is updated monthly, with a two- to three-month lag between trip date and publication. Each record contains precise pick-up and drop-off times and locations, trip duration and distance, fare details (including surcharges and tips), rate type, payment method, and passenger count. This makes the data particularly useful for studying temporal and spatial ride patterns, demand elasticity, and responses to external shocks or policy changes.

Chicago provides a similar dataset, covering all taxi trips since 2013.[10] The structure closely resembles the NYC data, though researchers should be mindful of city-specific reporting standards and definitions.

In addition to traditional taxis, several cities also publish data on rideshare platforms. Since February 2019, the NYC TLC has released trip-level data for "High Volume For-Hire Vehicle" services, covering companies such as Uber and Lyft. These data mirror the structure of the taxi files, though certain variables are masked or aggregated for privacy. Chicago has made similar data available since 2018 through its Transportation Network Providers (TNP) reporting requirements.[11] These files include trip origins and destinations (by census tract), ride durations and distances, and fare components, though passenger-level details are not disclosed.

A more specialized dataset comes from RideAustin, a non-profit rideshare company that operated in Austin, Texas, between 2016 and 2017. When Uber and Lyft exited the Austin market in response to city regulations requiring fingerprint-based driver background checks, RideAustin rapidly expanded to fill the gap. The company publicly released detailed trip-level data from this period, including pick-up and drop-off coordinates, times, and fares.

4 Conclusion

This guide has reviewed a broad range of publicly available datasets relevant to the study of the U.S. vehicle market. While no single standardized source captures all vehicle transactions—as is

[9] See https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.
[10] See Taxi Trips 2013–2023 and Taxi Trips 2024–.
[11] See TNP Trips 2018–2022 and TNP Trips 2023–.

the case in some other consumer markets—researchers can draw on a patchwork of rich and detailed resources. These include state-level DMV records, federal agency databases, municipal transportation datasets, and regulatory disclosures. Collectively, these sources support empirical work across marketing, economics, transportation, and public policy by offering insight into consumer behavior, vehicle characteristics, financing, safety, infrastructure, and usage.

That said, the availability, structure, and quality of these data vary substantially. Researchers should therefore remain mindful of potential limitations in coverage, representativeness, and data accuracy when designing studies or interpreting results.

If you are a researcher using vehicle data and notice omissions, errors, or opportunities to improve this guide, I would welcome your feedback. Please feel free to reach out to me directly.

References

Akerlof, G. A. (1970): "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, 84, 488–500.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, 841–890.

Biglaiser, G. and F. Li (2018): "Middlemen: the good, the bad, and the ugly," *The RAND Journal of Economics*, 49, 3–22.

Kraft, A. and R. Rao (2024): "Behavioral skimming: Theory and evidence from resale markets," *Available at SSRN*.

Liu, J. (2022): "Who should be subsidized for electric vehicles? demand estimation and policy design under network effects," Tech. rep., Working paper.

Murry, C. (2017): "Advertising in vertical relationships: An equilibrium model of the automobile industry," *Available at SSRN 2549247*.

Murry, C. and H. S. Schneider (2016): "The economics of retail markets for new and used cars," in *Handbook on the Economics of Retailing and Distribution*, Edward Elgar Publishing, 343–367.

Murry, C. and Y. Zhou (2020): "Consumer search and automobile dealer colocation," *Management Science*, 66, 1909–1934.

Rust, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 999–1033.

Sagl, S. (2023): "Dispersion, discrimination, and the price of your pickup," Tech. rep., Working paper.

Yavorsky, D., E. Honka, and K. Chen (2021): "Consumer search in the US auto industry: The role of dealership visits," *Quantitative Marketing and Economics*, 19, 1–52.