UNIVERSITY of HOUSTON

EDS6397 - INFORMATION VISUALIZATION

Spring 2025

A Visual Exploration of US Used-Car Sales Trends

Professor: Dr. Nwosu, Lucy PhD
Teaching Assistant: Onyinyechi C Ihesiulo

GROUP – 4

Lekha Chittajallu

Rishi Yedlapalli

Sai Praneeth Achanta

Sharath Kumar Reddy Kapu

Sree Sai Preetham Nandamuri

Vamshidhar Reddy Ankenapalle

Venkata Achyuth Kumar Sanagapalli

Venkata Sumanth Reddy Vangala

Victor Paul Buddha

Vishnu Sai Inakollu

ABSTRACT

This project explores a comprehensive dataset of U.S. car auction listings to uncover actionable insights using Tableau visual analytics. By analysing over 2,400 car entries across various brands, models, and states, we aimed to identify patterns in pricing, mileage, regional variations, and the impact of title status. Through meticulous data cleaning and exploratory analysis guided by Tufte's principles of data integrity and Colin Ware's perceptual guidelines, we built an interactive dashboard offering dynamic insights. Our findings highlight significant depreciation trends based on mileage, brand-based price stability, and geographic pricing disparities. The final visualization emphasizes user engagement, ethical storytelling, and real-world applicability for buyers, sellers, and data-driven businesses. This work underscores the power of visual analytics in turning raw auction data into strategic knowledge.

TABLE OF CONTENTS

1. INTRODUCTION

Objective:

The main objective of this project is to analyse trends and patterns in the U.S. used car market to uncover insights on pricing, mileage, model popularity, regional sales, and auction behaviour. Additionally, the analysis focuses on evaluating how title status influences vehicle demand, ultimately helping a potential buyer identify cars that offer the best features at the most affordable prices.

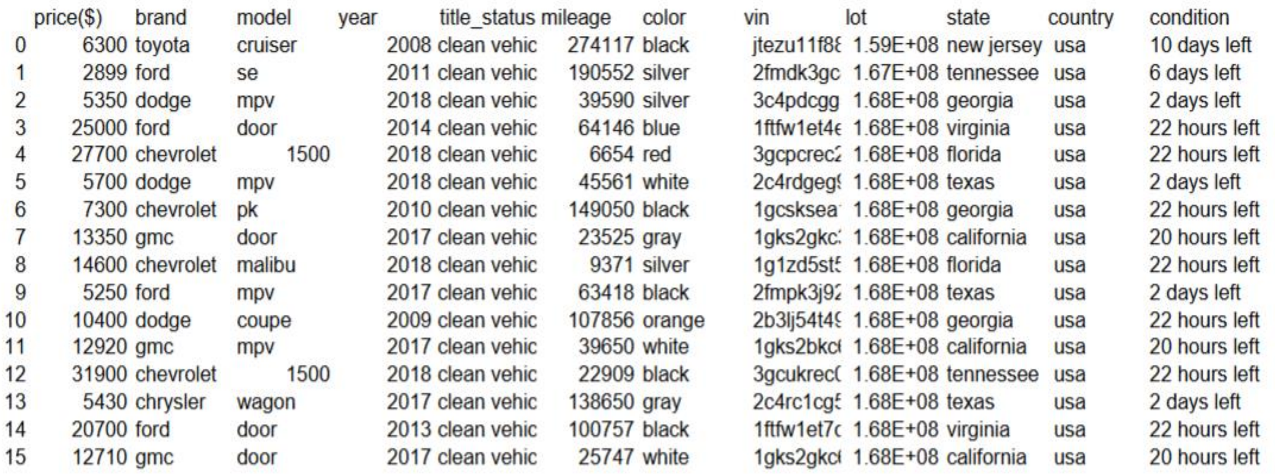
Dataset Overview:

Source: AuctionExport.com

Size: 2499 Records and 12 Attributes

Scope: USA cars auction listings with fields like brand, model, year, mileage, price, title status, location, colour, and time left.

Snapshot



	price($)	brand	model	year	title_status	mileage	color	vin	lot	state	country	condition
0	6300	toyota	cruiser	2008	clean vehic	274117	black	jtezu11f8	1.59E+08	new jersey	usa	10 days left
1	2899	ford	se	2011	clean vehic	190552	silver	2fmdk3gc	1.67E+08	tennessee	usa	6 days left
2	5350	dodge	mpv	2018	clean vehic	39590	silver	3c4pdcgg	1.68E+08	georgia	usa	2 days left
3	25000	ford	door	2014	clean vehic	64146	blue	1ftfw1et4	1.68E+08	virginia	usa	22 hours left
4	27700	chevrolet	1500	2018	clean vehic	6654	red	3gcpcrec	1.68E+08	florida	usa	22 hours left
5	5700	dodge	mpv	2018	clean vehic	45561	white	2c4rdgeg	1.68E+08	texas	usa	2 days left
6	7300	chevrolet	pk	2010	clean vehic	149050	black	1gcsksea	1.68E+08	georgia	usa	22 hours left
7	13350	gmc	door	2017	clean vehic	23525	gray	1gks2gkc	1.68E+08	california	usa	20 hours left
8	14600	chevrolet	malibu	2018	clean vehic	9371	silver	1g1zd5st	1.68E+08	florida	usa	22 hours left
9	5250	ford	mpv	2017	clean vehic	63418	black	2fmpk3j9	1.68E+08	texas	usa	2 days left
10	10400	dodge	coupe	2009	clean vehic	107856	orange	2b3lj54t4	1.68E+08	georgia	usa	22 hours left
11	12920	gmc	mpv	2017	clean vehic	39650	white	1gks2bkc	1.68E+08	california	usa	20 hours left
12	31900	chevrolet	1500	2018	clean vehic	22909	black	3gcukrec	1.68E+08	tennessee	usa	22 hours left
13	5430	chrysler	wagon	2017	clean vehic	138650	gray	2c4rc1cg	1.68E+08	texas	usa	2 days left
14	20700	ford	door	2013	clean vehic	100757	black	1ftfw1et7	1.68E+08	virginia	usa	22 hours left
15	12710	gmc	door	2017	clean vehic	25747	white	1gks2gkc	1.68E+08	california	usa	20 hours left

Figure 1. Snapshot of the raw data

Timeframe: All the data was in between the year 1973(include) to 2020(include). So, it covers almost 5 decades.

Variables: price($) | brand | model | year | title_status | mileage | color | vin | lot | state | country | condition, are all the available variables in the dataset.

Data Points:

price($) – 6300

brand – toyota

model – cruiser

year – 2008

title_status – clean vehicle

mileage – 274117

color - black

vin - jtezu11f88k007763

lot - 159348797

state – new jersey

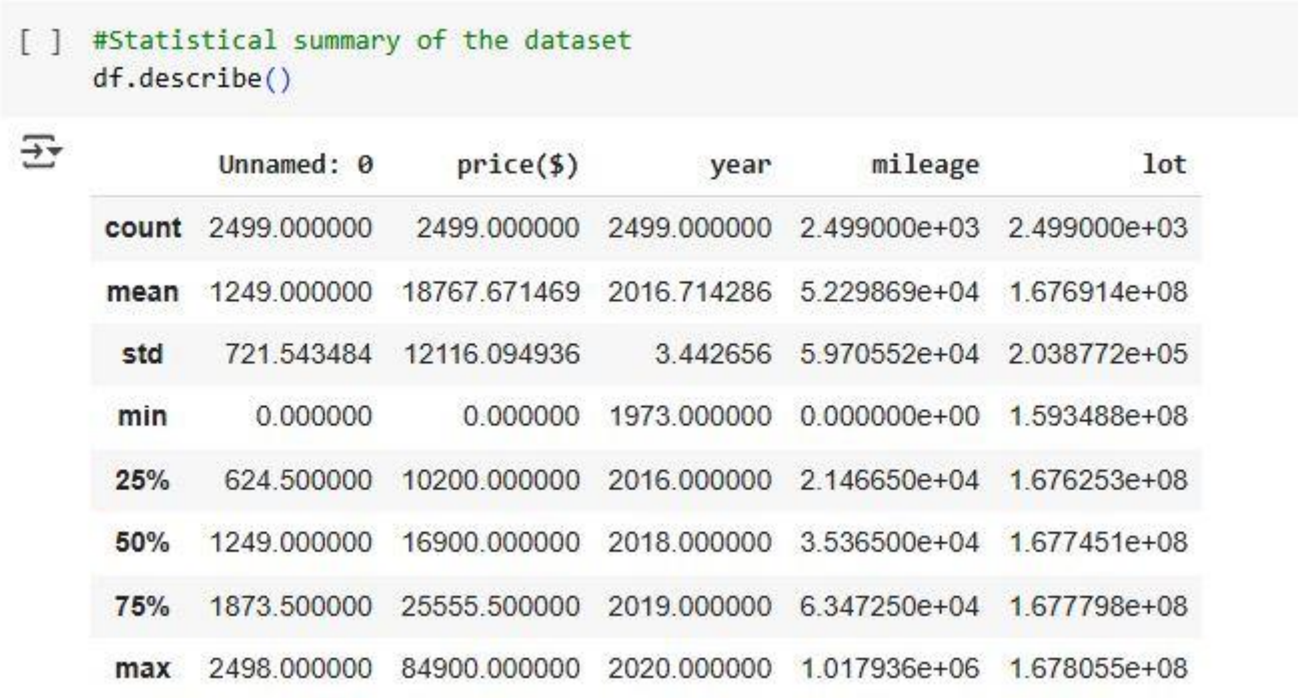
country - USA

condition – 10 days left

Data Cleaning:

As data science enthusiasts, we understand the critical importance of data quality. We believe that data visualizations are like the face of a person, while data quality represents the health of the heart. Just as a healthy heart reflects positively on a person's appearance, high-quality data leads to more accurate, engaging, and effective visualizations.

The dataset wasn't overly messy, so we started with fundamental pre-processing steps beginning with checking for null values and handling them appropriately to ensure a clean foundation for our analysis.

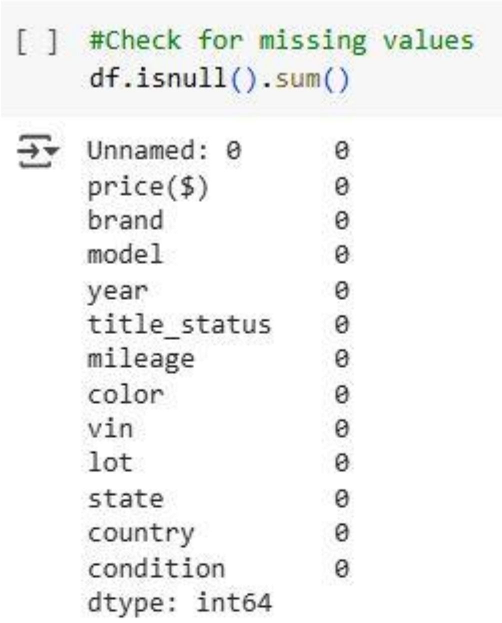


```
[ ] #Statistical summary of the dataset
    df.describe()
```

	Unnamed: 0	price($)	year	mileage	lot
count	2499.000000	2499.000000	2499.000000	2.499000e+03	2.499000e+03
mean	1249.000000	18767.671469	2016.714286	5.229869e+04	1.676914e+08
std	721.543484	12116.094936	3.442656	5.970552e+04	2.038772e+05
min	0.000000	0.000000	1973.000000	0.000000e+00	1.593488e+08
25%	624.500000	10200.000000	2016.000000	2.146650e+04	1.676253e+08
50%	1249.000000	16900.000000	2018.000000	3.536500e+04	1.677451e+08
75%	1873.500000	25555.500000	2019.000000	6.347250e+04	1.677798e+08
max	2498.000000	84900.000000	2020.000000	1.017936e+06	1.678055e+08

Figure 2. Statistical Summary of the Dataset

Null Values: Luckily, we have no NULL values to clean the data.

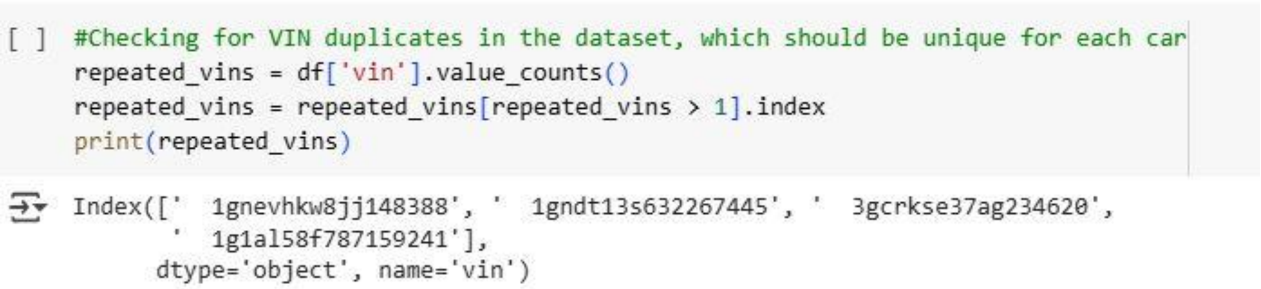


```
[ ] #Check for missing values
    df.isnull().sum()
```

```
Unnamed: 0      0
price($)        0
brand           0
model           0
year            0
title_status    0
mileage         0
color           0
vin             0
lot             0
state           0
country         0
condition       0
dtype: int64
```

Figure 3. Checking for NULL values

Duplicates: Since the Vehicle Identification Number (VIN) is a unique 17-character alphanumeric code assigned to each car, we checked for duplicate entries in the VIN field. Upon inspection, we identified 4 duplicate VINs. After reviewing the corresponding records, we removed these duplicates to maintain data integrity.

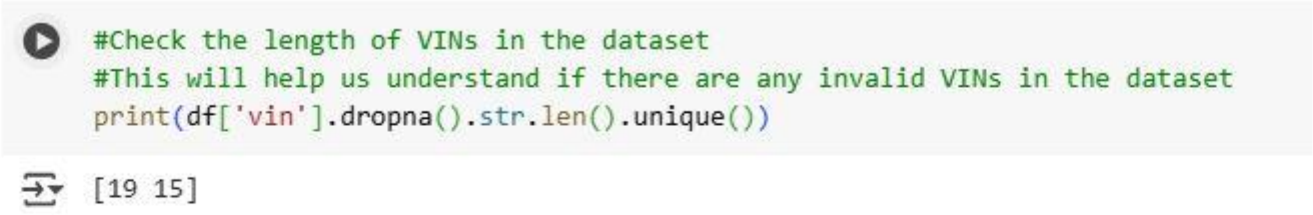


```
[ ]  #Checking for VIN duplicates in the dataset, which should be unique for each car
     repeated_vins = df['vin'].value_counts()
     repeated_vins = repeated_vins[repeated_vins > 1].index
     print(repeated_vins)
```

```
Index(['  1gnevhkw8jj148388', '  1gndt13s632267445', '  3gcrkse37ag234620',
       '  1g1al58f787159241'],
      dtype='object', name='vin')
```

Figure 4. Repeated VINs (Duplicates)

Invalid VINs: We also validated the length of each VIN, as it must be exactly 17 characters. Any entries with VINs shorter or longer than 17 characters were deemed invalid and subsequently removed from the dataset.



```
#Check the length of VINs in the dataset
#This will help us understand if there are any invalid VINs in the dataset
print(df['vin'].dropna().str.len().unique())
```

```
[19 15]
```

Figure 5. Invalid VINs

Categorical Attributes: When we observed all the unique models and brands in the datasets, we came to know that few of them were wrongly spelled. This might be a data entry error. As our dataset has less records, we decide to correct those using python.

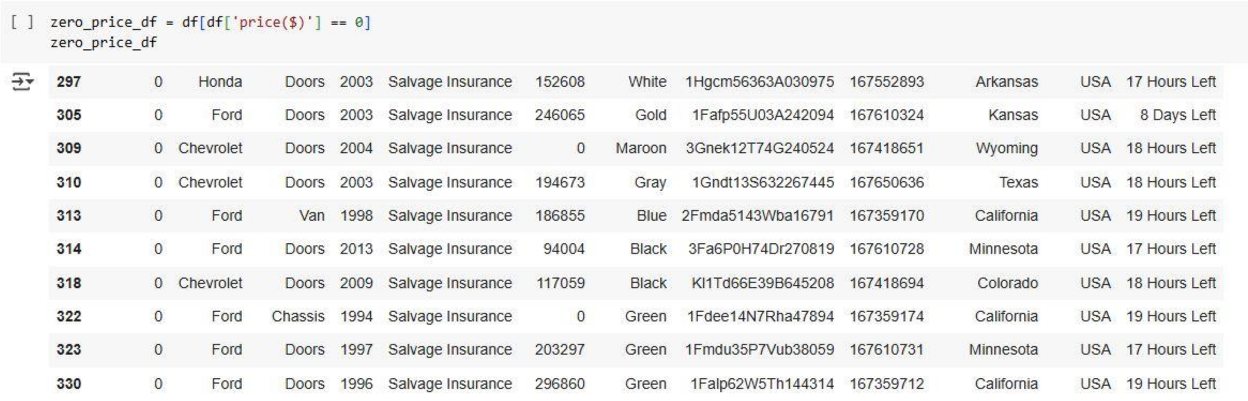
If we have more data records and feel like there are spelling mistakes in it, then we can use the machine learning techniques or NLP techniques to correct them

Uppercasing: Uppercased the brands and models that are less than or equal to 3 characters. This is a common practice to standardize the data and make it easier to analyse.

Also, uppercased the whole Country column.

Price ($) Column: We began our analysis by examining the 'Price ($)' column, focusing on entries where the price was listed as $0. Considering that the dataset pertains to auctioned vehicles, it's plausible for some listings to start at a base price of $0. After reviewing these entries, we determined that such values are acceptable within the context of auction data.

Mileage: However, upon further inspection of the same records, we identified instances where the mileage was recorded as 0. Given that all vehicles in the dataset are used, a mileage of zero is unrealistic. To maintain the dataset's integrity, we decided to remove records with mileage greater than or equal to 10, as they likely indicate data entry errors or missing information.



```
[ ] zero_price_df = df[df['price($)'] == 0]
    zero_price_df
```

297	0	Honda	Doors	2003	Salvage Insurance	152608	White	1Hgcm56363A030975	167552893	Arkansas	USA	17 Hours Left
305	0	Ford	Doors	2003	Salvage Insurance	246065	Gold	1Fafp55U03A242094	167610324	Kansas	USA	8 Days Left
309	0	Chevrolet	Doors	2004	Salvage Insurance	0	Maroon	3Gnek12T74G240524	167418651	Wyoming	USA	18 Hours Left
310	0	Chevrolet	Doors	2003	Salvage Insurance	194673	Gray	1Gndt13S632267445	167650636	Texas	USA	18 Hours Left
313	0	Ford	Van	1998	Salvage Insurance	186855	Blue	2Fmda5143Wba16791	167359170	California	USA	19 Hours Left
314	0	Ford	Doors	2013	Salvage Insurance	94004	Black	3Fa6P0H74Dr270819	167610728	Minnesota	USA	17 Hours Left
318	0	Chevrolet	Doors	2009	Salvage Insurance	117059	Black	Kl1Td66E39B645208	167418694	Colorado	USA	18 Hours Left
322	0	Ford	Chassis	1994	Salvage Insurance	0	Green	1Fdee14N7Rha47894	167359174	California	USA	19 Hours Left
323	0	Ford	Doors	1997	Salvage Insurance	203297	Green	1Fmdu35P7Vub38059	167610731	Minnesota	USA	17 Hours Left
330	0	Ford	Doors	1996	Salvage Insurance	296860	Green	1Falp62W5Th144314	167359712	California	USA	19 Hours Left

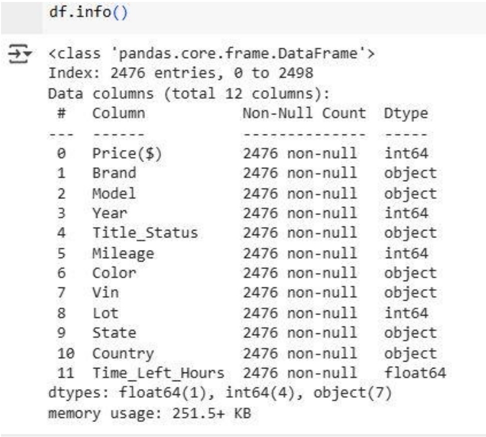
Figure 6. Snapshot of records having $0 price.

Condition: The 'Condition' column indicates the remaining time before a car is sold at auction, represented in a mix of minutes, hours, and days. To standardize these values, we used a Python function to convert all time formats into numeric values and then into hours.

Finally, we renamed the column to 'time_left_hours' to better reflect its purpose.

Outliers: Regarding outliers, we chose not to remove them. Given the nature of the attributes, these extreme values could indicate major players in the auction market. Additionally, since the dataset contains a limited number of records, eliminating the few outliers could result in the loss of valuable insights.

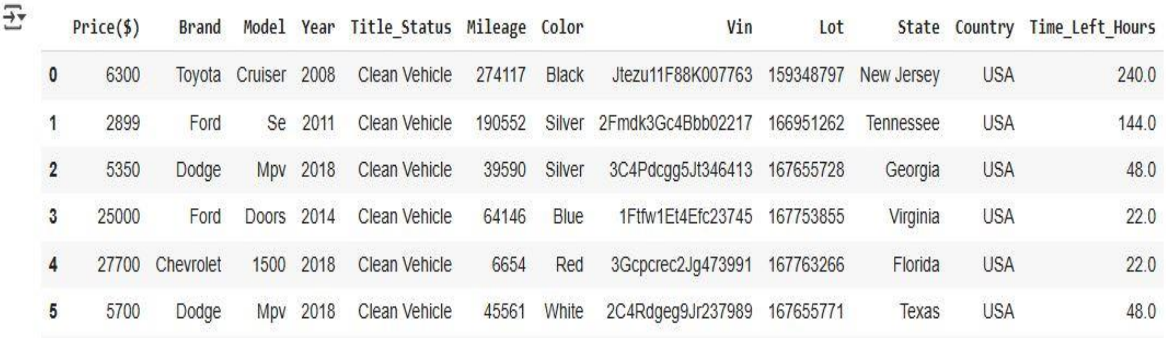
All the column names are title cased and the data frame is converted into a csv file.



```
df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 2476 entries, 0 to 2498
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Price($)        2476 non-null   int64
 1   Brand           2476 non-null   object
 2   Model           2476 non-null   object
 3   Year            2476 non-null   int64
 4   Title_Status    2476 non-null   object
 5   Mileage         2476 non-null   int64
 6   Color           2476 non-null   object
 7   Vin             2476 non-null   object
 8   Lot             2476 non-null   int64
 9   State           2476 non-null   object
 10  Country         2476 non-null   object
 11  Time_Left_Hours 2476 non-null   float64
dtypes: float64(1), int64(4), object(7)
memory usage: 251.5+ KB
```

Figure 7. Pre-processed data summary



	Price($)	Brand	Model	Year	Title_Status	Mileage	Color	Vin	Lot	State	Country	Time_Left_Hours
0	6300	Toyota	Cruiser	2008	Clean Vehicle	274117	Black	Jtezu11F88K007763	159348797	New Jersey	USA	240.0
1	2899	Ford	Se	2011	Clean Vehicle	190552	Silver	2Fmdk3Gc4Bbb02217	166951262	Tennessee	USA	144.0
2	5350	Dodge	Mpv	2018	Clean Vehicle	39590	Silver	3C4Pdcgg5Jt346413	167655728	Georgia	USA	48.0
3	25000	Ford	Doors	2014	Clean Vehicle	64146	Blue	1Ftfw1Et4Efc23745	167753855	Virginia	USA	22.0
4	27700	Chevrolet	1500	2018	Clean Vehicle	6654	Red	3Gcpcrec2Jg473991	167763266	Florida	USA	22.0
5	5700	Dodge	Mpv	2018	Clean Vehicle	45561	White	2C4Rdgeg9Jr237989	167655771	Texas	USA	48.0

Figure 8. Snapshot of the Pre-processed data

2. DATA VISUALIZATION

2.1 **Sankey Chart:** Title Status vs Price Category

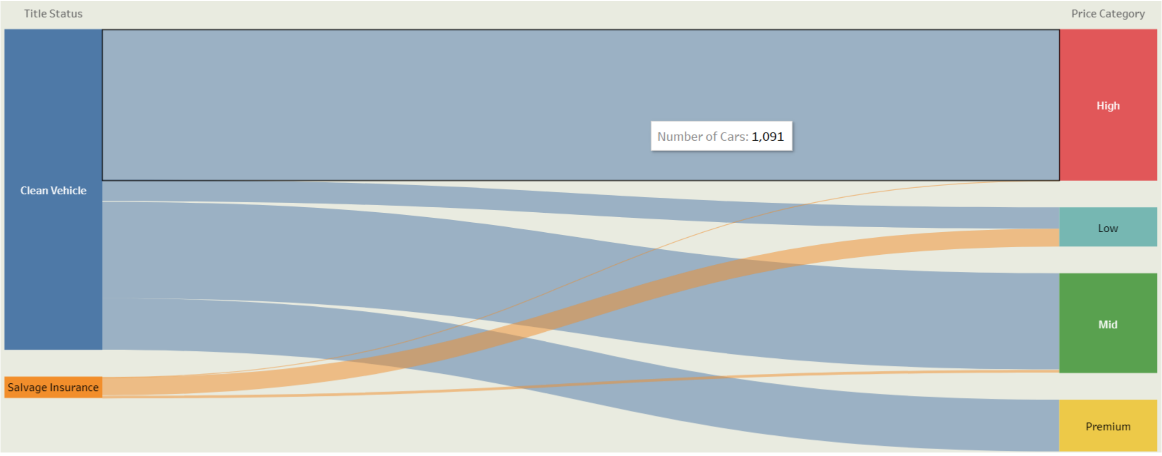


Figure 9. Sankey Chart: Title Status vs Price Category

Title: Flow of Vehicles from Title Status to Price Category

Marks:

Nodes

Source – Title Status (Clean Vehicle and Salvage Insurance Vehicles)

Target – Price Category (Low, Mid, High, and Premium)

Channels:

Links - Left-to-right horizontal alignment connects Title Status to Price Category, Links display the number of vehicles.

Colour:

- Each title status has a unique colour (e.g., blue for "Clean Vehicle", orange for "Salvage Insurance")
- Each price category has distinguishable colours (e.g., red for "High", green for "Mid", Cyan for “Low”, and Yellow for “Premium”)

Size (Thickness of Flows): Encodes number of vehicles from one category to another

Shape: Curved paths for visual continuity and flow representation

Insights:

1. Most vehicles fall into the High price category, with the majority coming from Clean Vehicle titles.
2. No Salvage Insurance cars are priced as Premium, highlighting a clear market distinction based on title status.
3. Clean Vehicles are distributed across all price categories, showing they offer a wider pricing range compared to salvage cars.

2.2 Box Plot: Price Distribution by Title Status

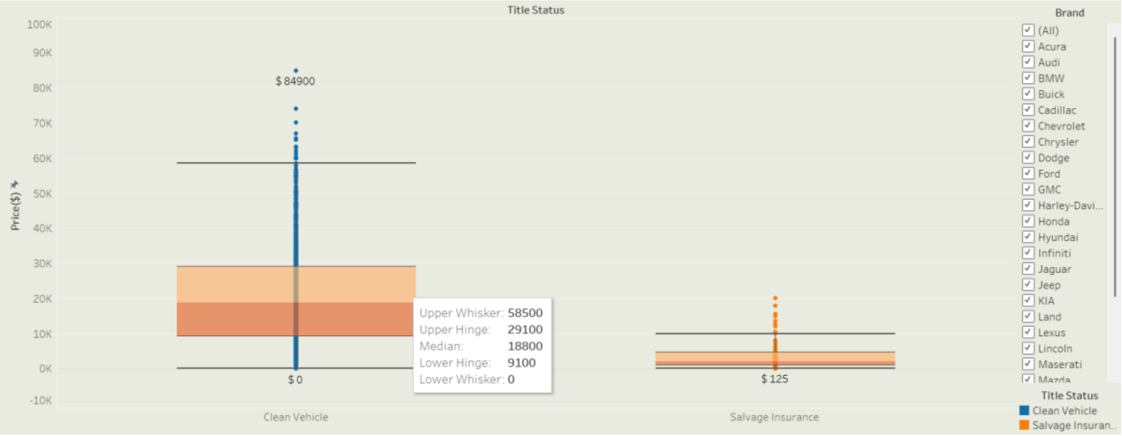


Figure 10. Box Plot: Price Distribution by Title Status

Title: Distribution of Vehicle Prices by Title Status

Marks: Box and whisker plots, Individual dots (outliers)

Channels:

- **Position (Y-axis):** Encodes Price ($) on a continuous quantitative scale
- **Grouping (X-axis):** By Title Status — “Clean Vehicle” vs. “Salvage Insurance”
- **Length of Box & Whiskers:** Represents interquartile range (IQR) and full data spread

Colour:

- Blue for Clean Vehicles
- Orange for Salvage Insurance

Size: Not Applicable

Shape: Box Plot, helps us to know the 5 number summary

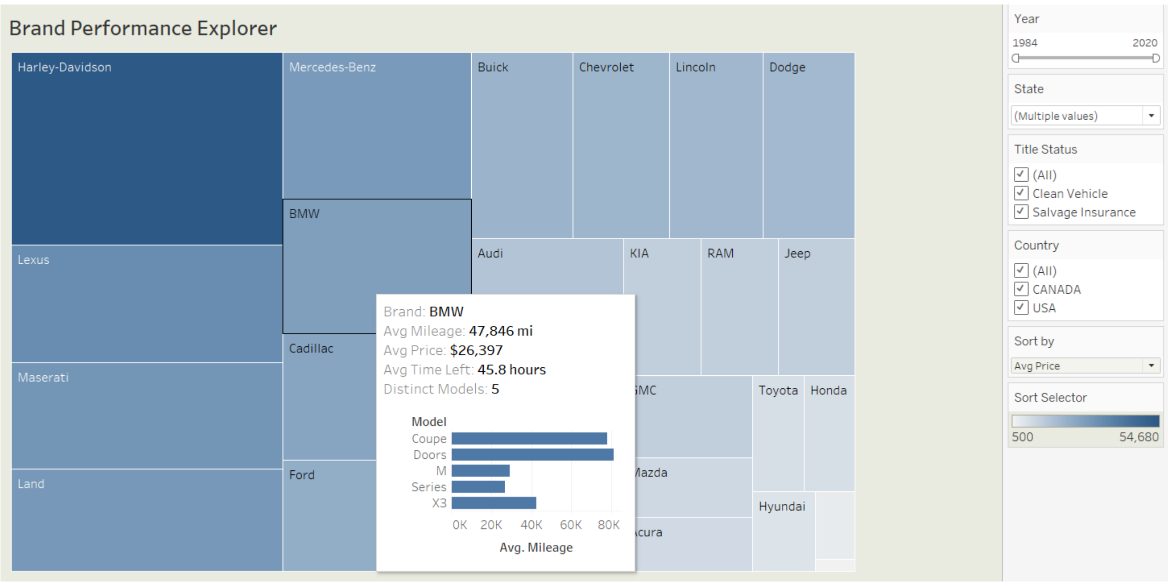
Interactive Features:

- Tooltips on hover reveal statistical values (e.g., median, hinges, whiskers)
- Filter pane (right) lets users slice data by brand for dynamic comparison

Insights:

1. Clean Vehicles have a broader and higher price distribution, with a median of $18,800 and upper whiskers reaching $58,500.
2. Salvage Insurance cars are priced significantly lower, with a tight range and much lower upper bounds.
3. Clean Vehicles have many high-value outliers, including cars priced up to $84,900, indicating a mix of both standard and luxury vehicles.

2.3 Tree map: Brand Performance Explorer



Brand Performance Explorer

Brand: BMW
Avg Mileage: 47,846 mi
Avg Price: $26,397
Avg Time Left: 45.8 hours
Distinct Models: 5

Year
1984 2020

State
(Multiple values)

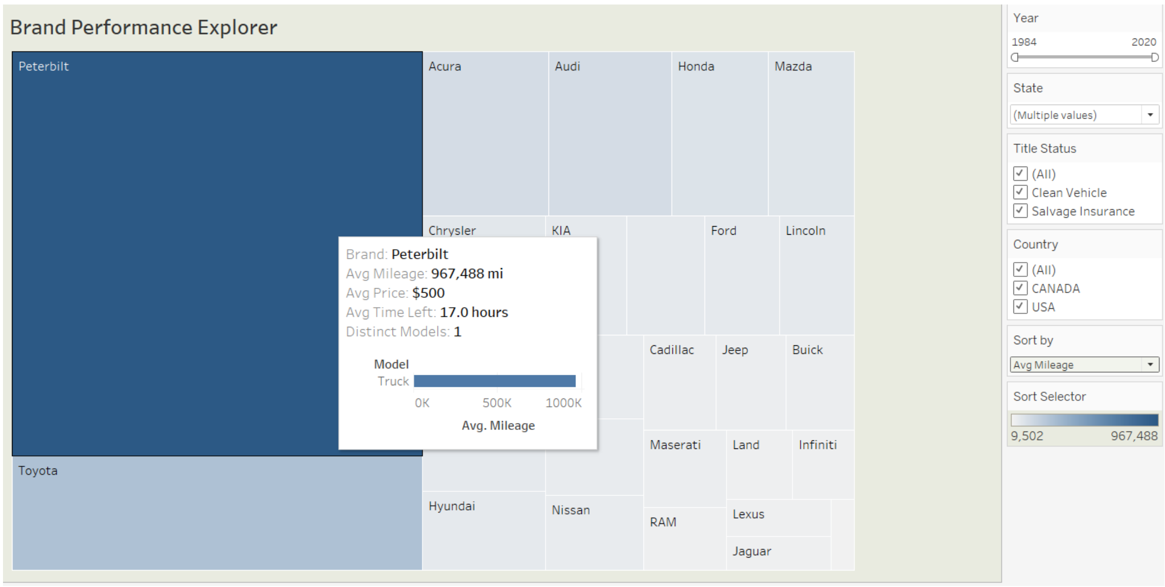
Title Status
☑ (All)
☑ Clean Vehicle
☑ Salvage Insurance

Country
☑ (All)
☑ CANADA
☑ USA

Sort by
Avg Price

Sort Selector
500 54,680

Figure 11. Tree map: Brand Performance Explorer based on Average Price



Brand Performance Explorer

Brand: Peterbilt
Avg Mileage: 967,488 mi
Avg Price: $500
Avg Time Left: 17.0 hours
Distinct Models: 1

Year
1984 2020

State
(Multiple values)

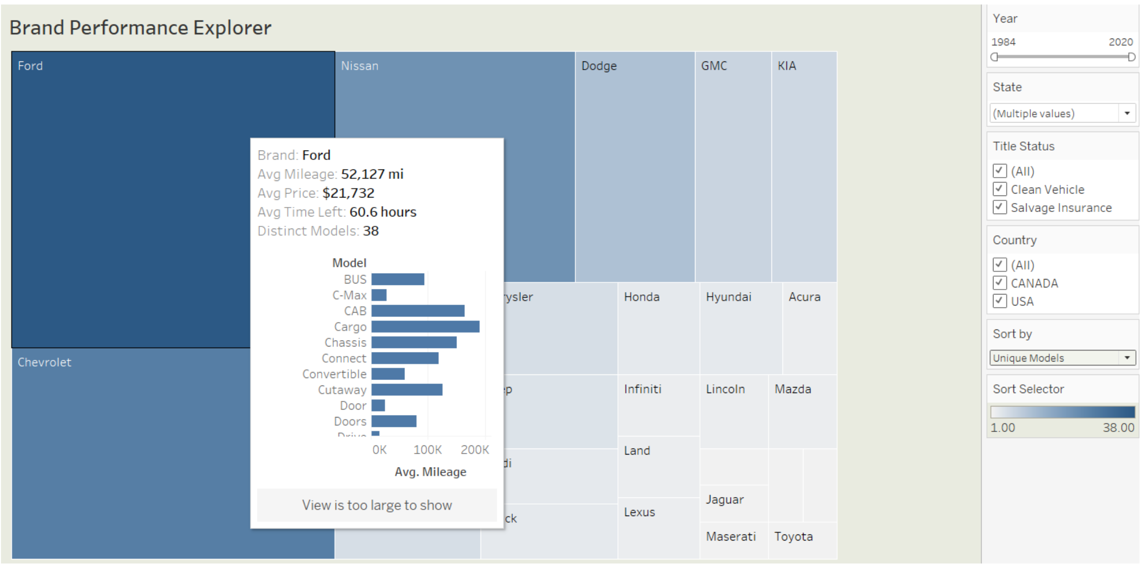
Title Status
☑ (All)
☑ Clean Vehicle
☑ Salvage Insurance

Country
☑ (All)
☑ CANADA
☑ USA

Sort by
Avg Mileage

Sort Selector
9,502 967,488

Figure 12. Tree map: Brand Performance Explorer based on Average Mileage



Brand Performance Explorer

Brand: Ford
Avg Mileage: 52,127 mi
Avg Price: $21,732
Avg Time Left: 60.6 hours
Distinct Models: 38

View is too large to show

Year
1984 2020

State
(Multiple values)

Title Status
☑ (All)
☑ Clean Vehicle
☑ Salvage Insurance

Country
☑ (All)
☑ CANADA
☑ USA

Sort by
Unique Models

Sort Selector
1.00 38.00

Figure 13. Tree map: Brand Performance Explorer based on Unique Models

Title: Brand Performance Explorer – Average Price, Mileage & Model Insights

Marks: Rectangles (Treemap blocks) for each brand

Channels:

Colour: Encodes the selected numeric value, e.g., Avg Price (dark = higher)

Size: Represents the average price or metric selected in the "Sort Selector"

Shape: Rectangular tree blocks

Position: Naturally grouped to fill space for visual comparison

Tooltip content (on hover):

- Brand name
- Average mileage, price, time left
- Count of distinct models
- Mini bar chart for average mileage by model

Parameter: The Sort Selector is a parameter control that lets users dynamically sort the treemap based on three key performance metrics:

- Average Price
- Average Mileage
- Number of Unique Models

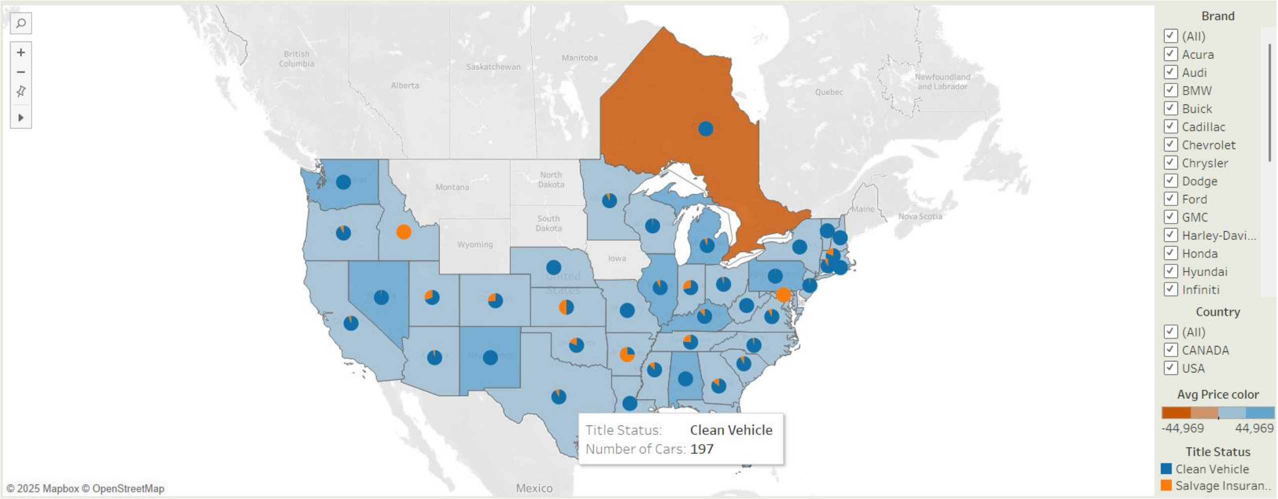
Interactive Features:

- Filters for Year, State, Title Status, Country
- Sort controls to rearrange based on price, mileage, etc.
- Sheet-in-Sheet Tooltip Feature:
  # A miniature bar chart (secondary sheet) embedded inside the tooltip
  # Displays model-level details like average mileage when hovering over a brand
  # Enhances exploratory analysis without adding extra dashboard space

Insights:

1. Harley-Davidson stands out as the highest average-priced brand, suggesting a niche but premium market presence.
2. Peterbilt dominates with an astonishing average mileage of 967,488 mi, indicating that its listings are likely heavy-duty commercial trucks, used extensively over long durations.
3. Ford leads with the highest variety, offering 38 distinct models, showcasing its diverse inventory and strong presence across vehicle categories like cargo, cab, convertible, and more.
4. With an average price of $26,397, average mileage of 47,846 mi, and 5 distinct models, BMW positions itself as a mid-to-high-end brand with moderate usage and model diversity — appealing to a broad segment of used car buyers.

2.4 Symbol and Glyph Map: Vehicle Titles & Price Trends Across Regions



Brand
☑ (All)
☑ Acura
☑ Audi
☑ BMW
☑ Buick
☑ Cadillac
☑ Chevrolet
☑ Chrysler
☑ Dodge
☑ Ford
☑ GMC
☑ Harley-Davi...
☑ Honda
☑ Hyundai
☑ Infiniti

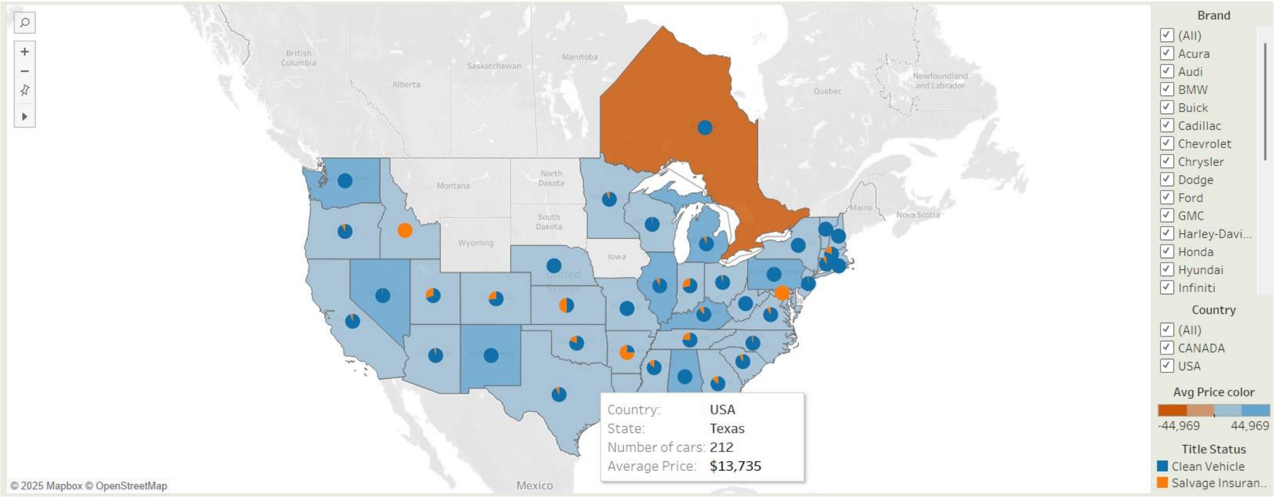
Country
☑ (All)
☑ CANADA
☑ USA

Avg Price color
-44,969 44,969

Title Status
■ Clean Vehicle
■ Salvage Insuran..

Title Status: Clean Vehicle
Number of Cars: 197

Figure 14. Symbol and Glyph Map: Vehicle Titles & Price Trends Across Regions



Brand
☑ (All)
☑ Acura
☑ Audi
☑ BMW
☑ Buick
☑ Cadillac
☑ Chevrolet
☑ Chrysler
☑ Dodge
☑ Ford
☑ GMC
☑ Harley-Davi...
☑ Honda
☑ Hyundai
☑ Infiniti

Country
☑ (All)
☑ CANADA
☑ USA

Avg Price color
-44,969 44,969

Title Status
■ Clean Vehicle
■ Salvage Insuran..

Country: USA
State: Texas
Number of cars: 212
Average Price: $13,735

Figure 15. Symbol and Glyph Map: Vehicle Titles & Price Trends Across Regions

Title: Geographic Distribution of Title Status and Average Price Across States

Marks:

- Filled maps (state colour)
- Dual-axis Pie Charts to represent title breakdown per state

Channels:

Colour: Represents average price, adjusted to show price difference by country

Shape: Pie

- Blue = Clean Vehicles
- Orange = Salvage Insurance

Size: Pie Size - Number of vehicles in that state

Tooltip: Displays title status and number of cars

Symbols (Pie Charts): Represent title status proportions using shape and colour

Glyphs (State Shading): Encode additional data, average price through colour gradients

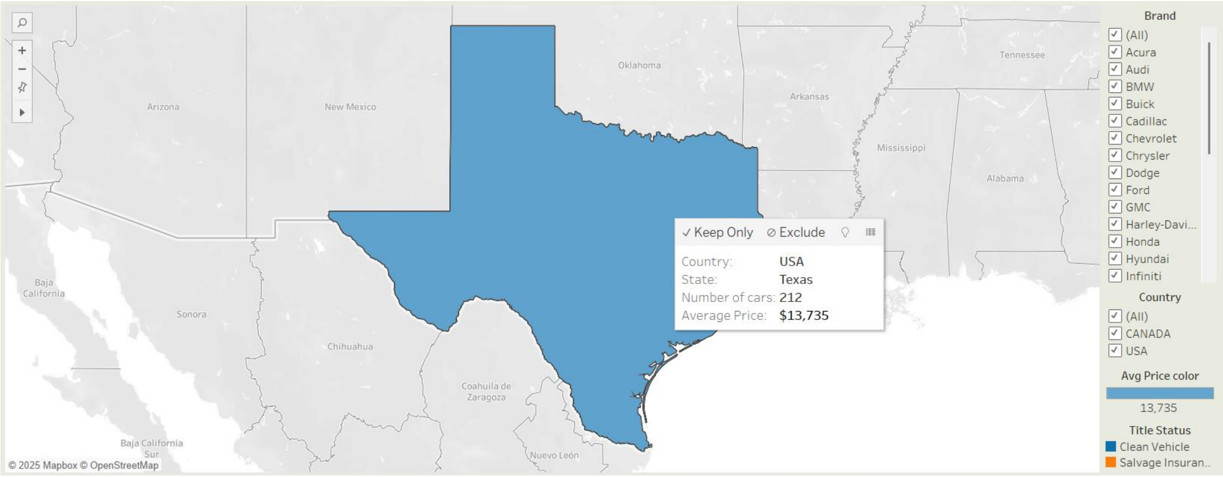
This dual encoding (symbols + glyphs) aligns with Colin Ware's perceptual principles, offering multi-variable insights in a single visual frame

Special Feature: We added a calculated attribute to differentiate between countries (USA vs. Canada). As a result, Canadian states like Ontario show negative average price values — not due to incorrect data, but by design, to visually separate and highlight regional pricing behaviours in the map.

Interactive Feature: "Filter Action + Map Zoom" in Tableau

How It Works:

- When a user clicks on a state, the entire map zooms in on that specific region
- The filter also optionally highlights or isolates the selected state's data
- This allows for focused geographic exploration without losing context



✓ Keep Only ⊘ Exclude

Country: USA
State: Texas
Number of cars: 212
Average Price: $13,735

Figure 16. Zoom Action

Insights:

1. Clean Vehicles dominate across almost all U.S. states, while Canadian provinces (e.g., Ontario) show a higher Salvage proportion and unique pricing behaviour due to country-based adjustment.
2. States with high vehicle volume (like Texas, California, and Florida) show strong dominance of Clean Vehicles, indicated by large blue segments in the pie charts suggesting these are key markets for used car auctions.
3. Salvage Insurance vehicles are more prominent in midwestern and southern states like Oklahoma, Louisiana, and Kentucky, hinting at possible regional differences in vehicle sourcing or insurance claim practices.
4. States in darker blue hues (e.g., California and New York) have higher average prices, while lighter or negatively coloured regions (like Quebec, Canada) indicate lower or adjusted prices, effectively highlighting cross-country market variations.

2.5 Bubble Chart: Car Colour Distribution

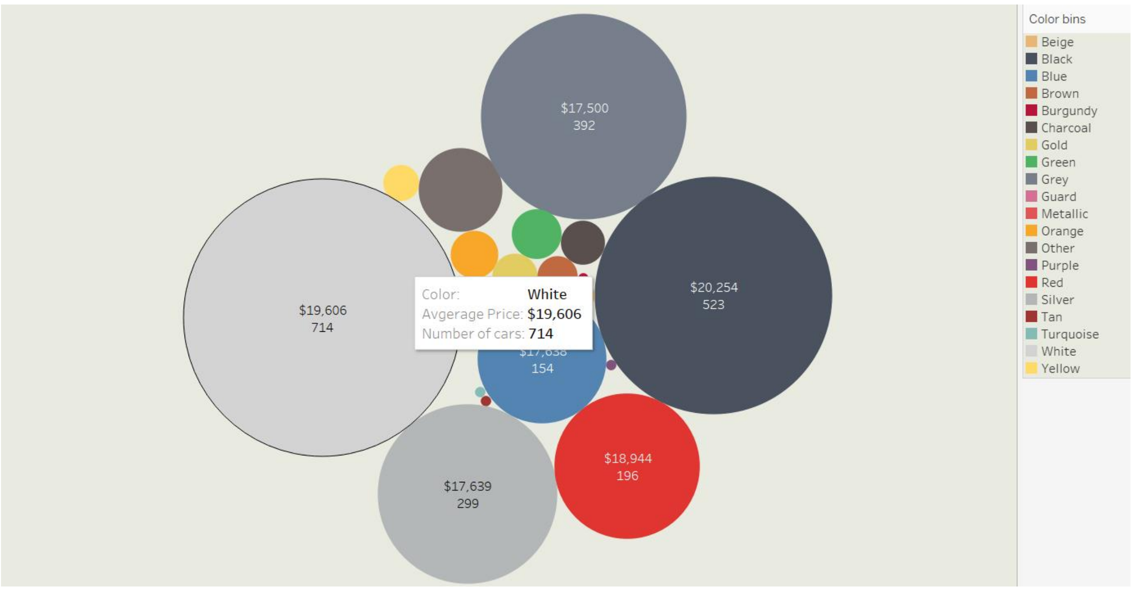


Figure 17. Bubble Chart: Car Colour Distribution

Title: Colour-wise Distribution of Vehicles with Average Price and Count

Marks: circles (Bubbles) representing car colours

Channels:

Position: Loosely packed for visibility (not spatially meaningful)

Size: Represents the number of cars for each colour

Colour: Matches the colour bin (e.g., Blue, Red, Grey)

Label/Text: Inside each bubble showing average price and vehicle count

Special feature (Colour Bins):

We created color bins by grouping similar color variations under a generic label for clarity. This ensures cleaner visual grouping, avoids clutter, and improves interpretability of color trends.

For example:

- "Light Blue", "Dark Blue", and "Sky Blue" are all grouped under "Blue"
- Similarly, "Dark Grey", "Silver Grey", and "Gunmetal" fall under "Grey"

Insights:

1. White is the most common vehicle color, with 714 cars and an average price of $19,606.
2. Black and Grey are also dominant in volume, with Grey having the highest average price at $20,254.
3. Red cars are moderately popular, while niche colors like Yellow, Green, and Burgundy have lower presence, suggesting limited market preference or availability.

2.6 **Line Trend:** Average Price and Average Mileage

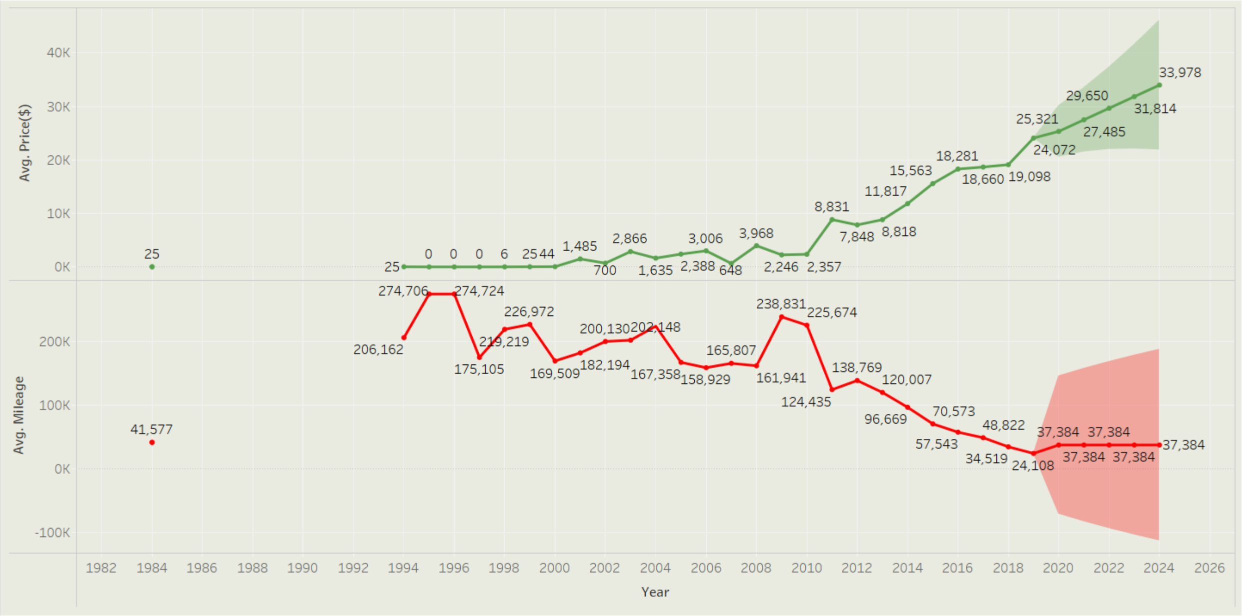


Figure 18. Line Trend: Average Price and Average Mileage

Title: Year-wise Trend of Average Price and Mileage with Forecast

Marks:

Lines:

- Green Line for Average Price ($)
- Red Line for Average Mileage

Circles: Represent actual yearly data points

Shaded bands: Represent the forecast with confidence interval (95%)

Channels:

Position (Y-axes): Two separate scales:

- Left Y-axis: Avg. Mileage
- Right Y-axis: Avg. Price

Colour:

- Green for picing trend
- Red for mileage trend
- Light green/red shaded area = forecast confidence interval

Tooltip: Hovering over a point reveals exact values for price and mileage per year

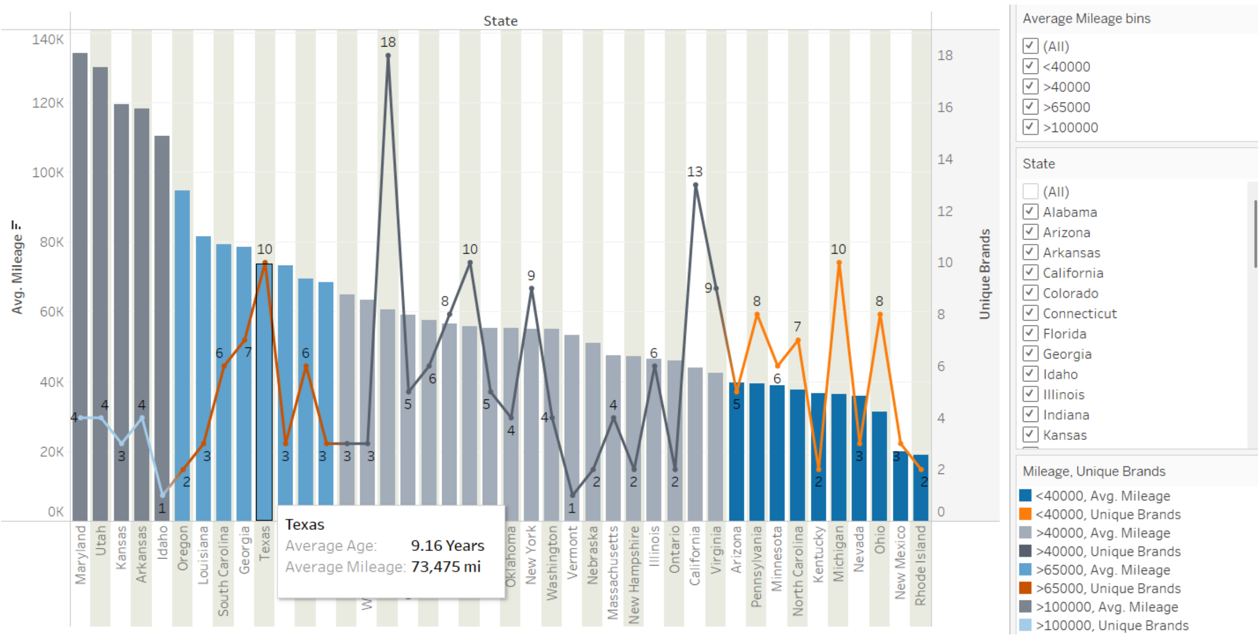
Interactive Feature: Forecast with 95% Confidence Interval

- We implemented a forecast model using Tableau's built-in exponential smoothing with a 95% confidence interval. This is shown as the shaded region beyond 2020, giving a visual range of future uncertainty.
- Dynamic: Users can hover to view forecasted average price and mileage
- Helps stakeholders understand expected trends and variability range

Insights:

1. Average price has steadily increased over the years, from around $25 in the 1980s to over $30,000 forecasted by 2025.
2. Average mileage has significantly dropped, falling from above 200,000 miles (1990s) to just around 37,000 miles in recent years, and is expected to stay flat.
3. The forecast bands indicate that while prices may rise further, mileage levels are stabilizing, possibly due to the increasing availability of newer or certified used vehicles in auctions.

2.7 Distribution of Average Mileage and unique brands among states



Texas
Average Age: 9.16 Years
Average Mileage: 73,475 mi

Average Mileage bins
☑ (All)
☑ <40000
☑ >40000
☑ >65000
☑ >100000

State
□ (All)
☑ Alabama
☑ Arizona
☑ Arkansas
☑ California
☑ Colorado
☑ Connecticut
☑ Florida
☑ Georgia
☑ Idaho
☑ Illinois
☑ Indiana
☑ Kansas

Mileage, Unique Brands
■ <40000, Avg. Mileage
■ <40000, Unique Brands
■ >40000, Avg. Mileage
■ >40000, Unique Brands
■ >65000, Avg. Mileage
■ >65000, Unique Brands
■ >100000, Avg. Mileage
■ >100000, Unique Brands

Figure 19. Distribution of Average Mileage and unique brands among states

Title: State-wise Average Mileage and Brand Variety Comparison

Marks:

Bars: Represent average mileage per state

Line: Represent the number of unique brands in each state

Tooltip: Displays average mileage, average age, and unique brand count for selected state

Channels:

Bar Height (Position): Encodes average mileage

Bar Colour: Encodes mileage bin, using categories like:

Darkish Grey - <40,000 mi

Light Blue - >40,000 mi

Light Grey - >65,000 mi

Dark Blue - >100,000 mi

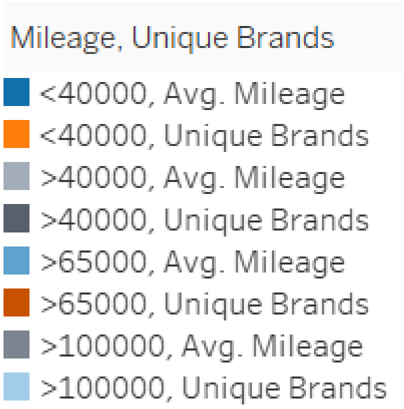


Figure 20. Average Mileage and Unique Brands

Colour: Each colour above is used to represent a Mileage bin, Unique Brands

Line Position (Y2-axis): Encodes count of unique brands

Bins: Average Mileage is divided into 4 bins based on the below conditions

Bin 1 - <40,000 mi
Bin 2 - >40,000 mi
Bin 3 - >65,000 mi
Bin 4 - >100,000 mi

Dual Axis: Two plots, Bar chart and Line chart are merged into a single axis, that is why we call it as dual axis plot.

Interactive Features:

- Mileage Bin Filter allows users to toggle specific ranges for focused comparison
- State Filter enables side-by-side inspection of selected regions
- Dual-axis interactivity makes it easy to view how brand diversity relates to average mileage within the same visual

Insights:

1. High-mileage states like Maryland, Utah, and Kansas exceed 120K mi, often with fewer brands, pointing to limited but heavily used inventory.
2. California leads in brand variety with 18 unique brands, despite a moderate average mileage reflecting a diverse and competitive auction market.
3. Texas stands out with 73,475 mi average mileage and 10 unique brands, making it one of the most balanced and active markets.
4. Low-mileage states like New Mexico, Ohio, and Rhode Island fall under the <40,000 mi bin and have fewer than 3 unique brands, indicating either limited auction activity or a high concentration of specific vehicle types.
5. States like Georgia and Michigan have similar mileage averages (~70,000 mi) but different brand diversity levels (Georgia with 3 brands vs. Michigan with 10), suggesting regional preferences or supply chain differences in used car availability.

2.8 Bar Chart - Average Price by Brand and Model

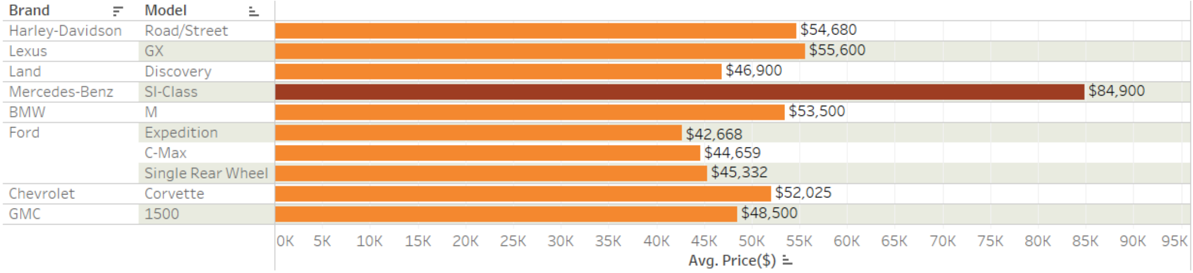


Figure 21. Bar Chart - Average Price by Brand and Model

Title: Top Models by Average Price – Brand and Model Breakdown

Marks:

Bars: Represent average price ($) per model

Colour Gradient: Based on the average price value (darker = higher)

Text Labels: Exact price values placed at the end of each bar

Channels:

Length of Bar (Position): Encodes Average Price ($

Colour Intensity: Represents relative ranking of price

Sorted Order: Highest average prices shown at the top

Tooltip (on hover): Displays detailed pricing info per model

Interactive Features:

Model and Brand Filters: Users can narrow down to specific models or manufacturers

Slider Control (Bottom Right):

Named "Top N Models by Average Price"

Dynamically adjusts how many top-priced models are shown (e.g., top 5, top 10)

Colour Legend: Visually reinforces price intensity from low ($0) to high ($84,900)

Insights:

1. Mercedes-Benz S-Class is the highest-priced model, averaging $84,900, clearly standing out with the darkest colour fill.
2. Lexus GX and Harley-Davidson Road/Street also rank in the top tier, averaging over $54,000, indicating their luxury or specialty status.
3. Ford's Expedition, C-Max, and Single Rear Wheel models fall in the mid-range (~$42K–$45K), showcasing affordable high-utility options compared to premium brands.
4. Chevrolet Corvette appears competitively priced at $52,025, combining sports appeal with relative affordability in the luxury space.

2.9 Horizontal Bar Chart showing Top N Vehicle Models

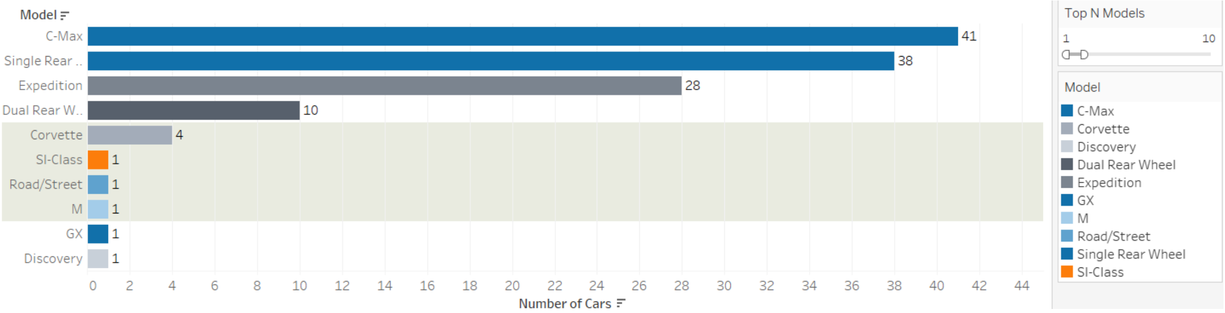


Figure 22. Horizontal Bar Chart showing Top N Vehicle Models

Title: Top N Vehicle Models by Number of Cars Listed

Marks:

Bars: Represent the number of cars listed per vehicle model
Bar Labels: Show exact counts
Colour: Encodes different vehicle models using distinct colours for easy differentiation

Channels:

Length of Bars (X-axis): Represents count of cars
Position (Y-axis): Lists vehicle models in descending order
Colour Legend: Helps quickly identify which bars belong to which model

Interactive Features:

- Top N Models Slider (Right Panel):
  Let's users dynamically control how many top models to display
  Ranges from Top 1 to Top 10, adjusting the bar chart in real-time
- Model Filter (Top Left): Allows selection/deselection of individual models
- Colour Coding: Colour legend updates as models change with Top N filter

Design Highlight:

This chart provides a ranked visual summary of vehicle models with the highest listing volume, helping users focus on market availability rather than price or mileage.

Insights:

1. Ford's C-Max and Single Rear Wheel models are the most frequently listed, with 41 and 38 cars, respectively — indicating high popularity or availability in the market.
2. Expedition and Dual Rear Wheel models also show strong presence, both Ford vehicles, reflecting Ford's dominance in auction volume.
3. Luxury or high-end models like Mercedes-Benz S-Class and Harley-Davidson Road/Street appear in the bottom of the Top 10 with only 1 car each, suggesting rarity and exclusivity.

2.10 Scatter Plot with Filters & Animated Timeline

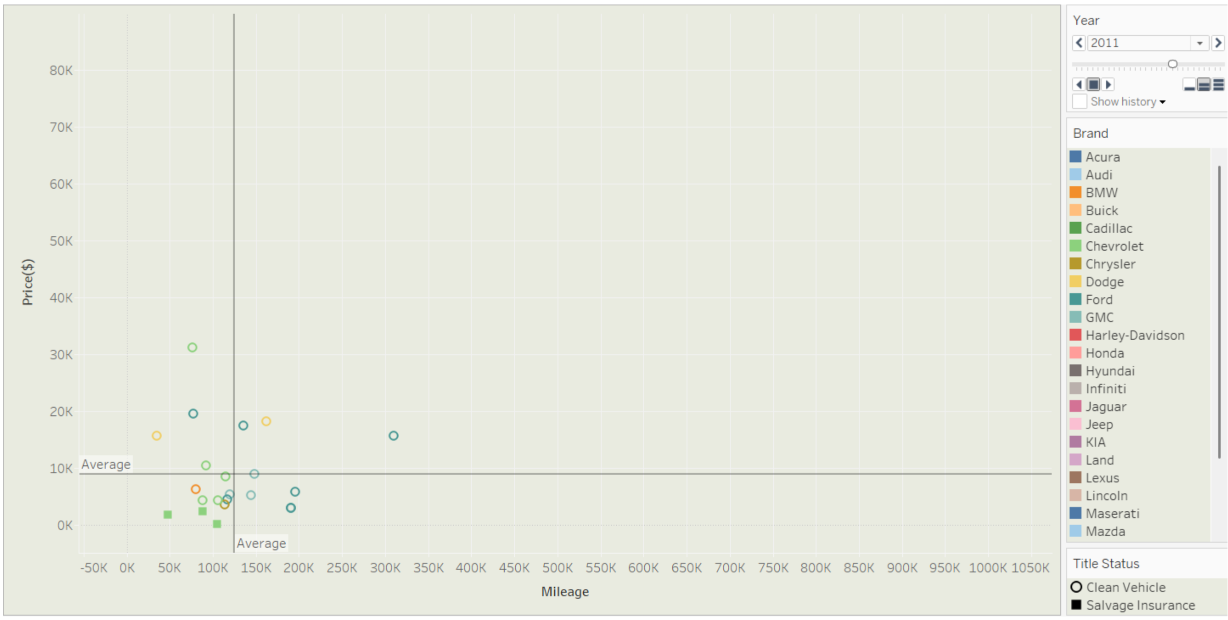


Figure 23. Scatter Plot with Filters & Animated Timeline

Title: Mileage vs. Price – Brand and Title Status Breakdown (Interactive by Year)

Marks:

Circles & Squares (Shapes): Represent individual vehicles

Colour: Encodes vehicle brand

Shape:

Circle - Clean Vehicle

Square - Salvage Insurance

Reference Lines: Show average price and average mileage

Tooltip (on hover): Displays brand, mileage, price, and title status

Channels:

X-axis: Mileage

Y-axis: Price

Colour Hue: Differentiates brands

Shape Type: Differentiates title status

Position (Scatter): Reflects price-mileage relationships across all entries

Interactive Features:

Year Slider:

- Allows users to explore year-specific trends
- As users move the slider, the scatter plot updates in real-time
- Useful for observing how price-mileage relationships change over time

Brand and Title Status Filters:

- Multi-select capability for exploring brand-specific or title-based performance
- Enables focused analysis for clean vs. salvage vehicles

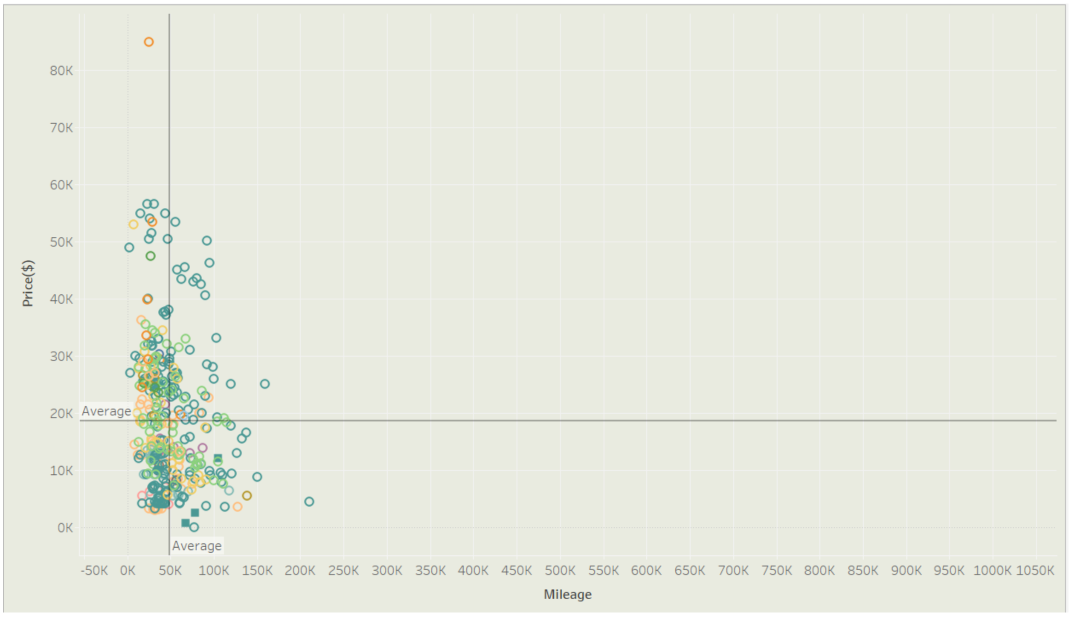


Figure 24. Scatterplot with Year Slider in 2017

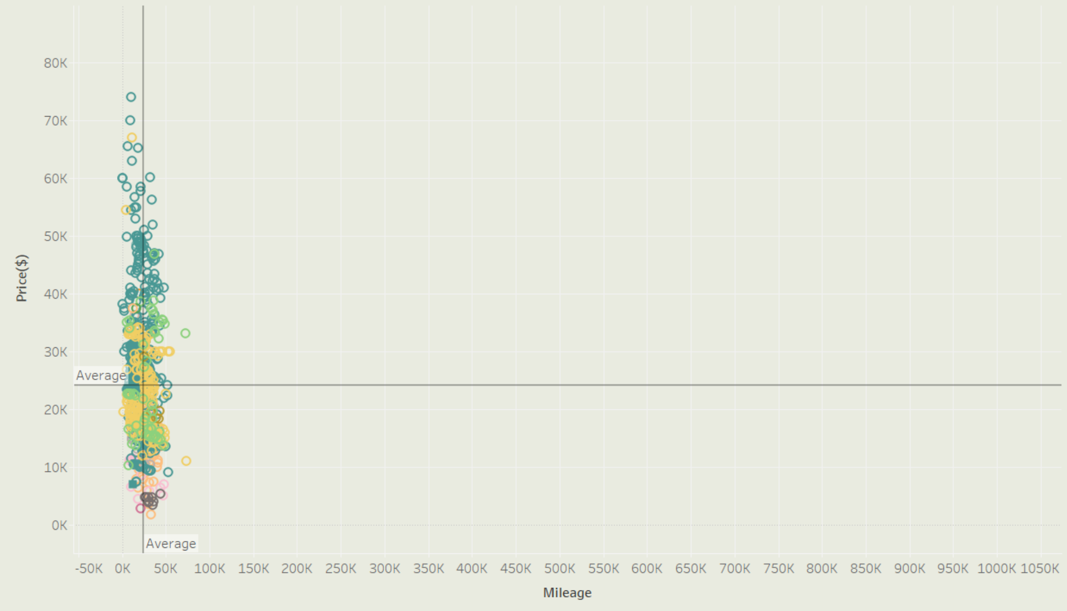


Figure 25. Scatterplot with Year Slider in 2019

Reference Lines:

- Clearly mark the mean mileage and price across all points
- Provides immediate visual cue for above/below average evaluation

Insights:

1. Most vehicles cluster around 100K–150K mileage and <$10,000 in price, aligning with expectations for 2011 models.
2. Clean Vehicles dominate the listings, shown by the abundance of circles over squares — suggesting greater resale demand or value.
3. Some outliers exceed 300K–500K mileage yet are still priced between $10K–$20K, possibly indicating heavy-duty or well-maintained models from brands like Ford or GMC.

3. HYPOTHESIS AND ANALYSIS DIRECTIONS

Hypothesis:

Seasonal Patterns Hypothesis: We expected to observe spikes in vehicle listings and prices during summer months and year-end, possibly driven by holiday demand.

Regional Variation Hypothesis: We hypothesized that average vehicle mileage and pricing vary significantly by state, driven by geography and market behaviour.

Time Series Forecasting Hypothesis: We projected that average price will continue to rise, while average mileage would decline, based on past sales trends.

Methodology:

To test these hypotheses, we used a combination of built-in Tableau functionalities and custom calculated fields:

Calculated Fields:
1. Age
Formula:

```
2025 - [Year]
```

Description: Calculates the age of each vehicle based on its manufacturing year.
Used In: Distribution of Avg Mileage and Unique Brands among States
• Displayed in the tooltip to show how old the cars are in each state, providing context for average mileage and brand diversity.

2. Average Mileage Bins
Formula:

```
IF AVG([Mileage]) > 100000 THEN ">100000"
ELSEIF AVG([Mileage]) > 65000 THEN ">65000"
ELSEIF AVG([Mileage]) > 40000 THEN ">40000"
ELSE "<40000"
END
```

Description: Categorizes the average mileage into bins to show variation across regions.
Used In: Distribution of Avg Mileage and Unique Brands among States
• Used as the color dimension to highlight mileage levels by state.

3. Average Price Color
Formula:

```
IF ATTR([Country]) = "CANADA" THEN -AVG([Price($)])
ELSE AVG([Price($)])
END
```

Description: Inverts color scale for Canadian data to separate it visually from U.S. listings while maintaining color consistency.

Used In: Map-Wise Distribution of Average Price

- Determines the color encoding for state-wise or country-wise average pricing.

4. Number of Cars

Formula:

```
COUNT([cleaned_usa_cars_dataset.csv])
```

Description: Counts the total number of vehicles in the dataset or per category.

Used In: Flow of Vehicle Status into Price Categories (Sankey Chart)

- Used to represent volume in each flow and shown in tooltips for category size reference.

5. Sort Selector

Formula:

```
CASE [Sort by]
WHEN "Avg Price" THEN AVG([Price($)])
WHEN "Avg Mileage" THEN AVG([Mileage])
WHEN "Unique Models" THEN COUNTD([Model])
END
```

Description: Allows dynamic sorting of the brand treemap based on user-selected metric from a parameter.

Used In: Brand Treemap

- Controls the treemap block size and sorting logic for interactive comparison.

6. Top N Models

Formula:

```
RANK_UNIQUE(AVG([Price($)]))
```

Description: Ranks vehicle models by their average price to filter or highlight the top N most expensive models.

Used In: Top Models Chart

- Supports filtering or labelling of high-priced models in visualizations.

7. Unique Brands

Formula:

```
COUNTD([Brand])
```

Description: Counts distinct brands available within each state.

Used In: Distribution of Average Mileage and Unique Brands among States

- Shown as a trend line to represent brand diversity across states, supporting regional comparison

8. Color Bins

Formula:

```
IF CONTAINS(LOWER([Color]), "red") THEN "Red"
ELSEIF CONTAINS(LOWER([Color]), "blue") THEN "Blue"
ELSEIF CONTAINS(LOWER([Color]), "green") THEN "Green"
ELSEIF CONTAINS(LOWER([Color]), "yellow") THEN "Yellow"
ELSEIF CONTAINS(LOWER([Color]), "orange") THEN "Orange"
ELSEIF CONTAINS(LOWER([Color]), "purple") THEN "Purple"
ELSEIF CONTAINS(LOWER([Color]), "pink") THEN "Pink"
ELSEIF CONTAINS(LOWER([Color]), "brown") THEN "Brown"
ELSEIF CONTAINS(LOWER([Color]), "white") THEN "White"
ELSEIF CONTAINS(LOWER([Color]), "grey") OR CONTAINS(LOWER([Color]), "gray")
THEN "Grey"
ELSEIF CONTAINS(LOWER([Color]), "black") THEN "Black"
ELSEIF CONTAINS(LOWER([Color]), "metallic") THEN "Metallic"
ELSEIF CONTAINS(LOWER([Color]), "silver") THEN "Silver"
ELSEIF CONTAINS(LOWER([Color]), "color") THEN "Other"
ELSE [Color]
END
```

Description: Categorizes vehicle colors into broader standardized color bins regardless of formatting variations (e.g., "Dark Red", "Gray Metallic", etc.), simplifying color analysis.
Used In: Car Color Distribution with Average Price and Count, Flow of Vehicle Status into Price Categories Based on Avg Mileage
• Used to group cars into visually consistent categories for color-based comparisons and flow analysis.

9. Price Category
Formula:

```
IF [Price($)] < 5000 THEN "Low"
ELSEIF [Price($)] < 15000 THEN "Mid"
ELSEIF [Price($)] < 30000 THEN "High"
ELSE "Premium"
END
```

Description: Classifies vehicles into pricing tiers: Low, Mid, High, and Premium, enabling stratified analysis based on affordability or market segment.
Used In: Flow of Vehicle Status into Price Categories Based on Average Mileage
• Used to segment vehicles by price and analyse how title status or mileage flows into pricing categories.

Forecasting in Tableau:

- We enabled forecasting with 95% confidence intervals for future trends of average price and mileage.
- Used exponential smoothing via Tableau’s built-in forecasting engine.

Map-Based Analysis:

- Built symbol and glyph maps showing regional differences in title status and pricing.
- Added a country differentiation field to distinguish U.S. and Canada data — and explain negative average prices for certain provinces like Ontario.

Interactive Parameterization:

- Created dynamic Sort Selector and Top N controls to let users slice and explore data across brand, mileage, and model-based hierarchies.

Insights and Patterns Uncovered:

- Time series plots revealed a steady growth in average vehicle prices and a decline in mileage, validated by forecast trends.

- Maps highlighted that southern states like Texas and Florida have higher brand variety and auction activity.
- Scatter plots with average lines helped detect outlier vehicles that are either overpriced for their mileage or unusually high-mileage but still costly, pointing to brand strength or use case (e.g., trucks, luxury models).
- The bubble chart on vehicle color showed that White, Black, and Grey dominate the market, with Grey cars demanding the highest price on average.

4. ETHICAL CONSIDERATIONS

Avoiding Misleading Visualizations
In building this dashboard and conducting our analysis, we ensured all visualizations were honest, transparent, and representative of the true nature of the data. Below are key ethical considerations we followed:

Proper Scaling and Axis Usage

- We avoided truncating axes (especially on scatter plots and line charts), which could distort trends or exaggerate differences.
- Dual-axis charts (e.g., mileage vs. brand count) were clearly labelled with separate Y-axes, and we maintained a consistent colour scheme to prevent misinterpretation.

Handling Outliers and Missing Data Transparently

- Outliers (e.g., mileage > 900K or price = $0) were not removed blindly. Instead, they were retained where relevant (e.g., Peterbilt trucks with extremely high mileage) to preserve insights about specific market segments.
- For missing or questionable values (e.g., blank price or VINs), we created data integrity checks and documented their treatment — either filtered out or flagged via calculated fields.

Fair Representation Through Binning and Grouping

- Custom mileage bins and color bins were created to group data for clarity but not to mask detail. For example:
  "Sky Blue," "Dark Blue," and "Light Blue" were grouped as "Blue" for readability — this was clearly noted in our methodology.
- We made sure bin thresholds were data-driven (e.g., quantiles) and not manipulated to exaggerate group differences.

Clear Use of Forecasting and Uncertainty

- In time series forecasting, we visualized 95% confidence intervals, making it clear that projections are estimates rather than certainties.
- Forecast bands were shaded and annotated to ensure users understood the range of expected variation.

Balanced Color Usage and Accessibility

- Colors were chosen to be intuitive and not emotionally manipulative (e.g., red for salvage not to imply danger, but clearly explained in the legend).
- Legends and color bins were included for every chart to aid interpretation and avoid hidden assumptions.

5. USER EXPERIENCE AND ENGAGEMENT

- **Dashboard Design**: The Tableau dashboard was designed with interactivity and clarity as primary goals. Filters and parameters allow users to dynamically explore the data, focusing on specific brands, models, or timeframes. Tooltips provide contextual details on demand, and clear visual encoding ensures that insights are intuitive and accessible.
- **Ethical Considerations**: Throughout the project, we adhered to ethical data visualization practices. Data accuracy and transparency were prioritized, and potential biases or misinterpretations were carefully addressed. The visualizations aim to empower users with reliable information, promoting informed decision-making in the used car market.
- **Storytelling**: The dashboard narrates a compelling story about the U.S. used car market. Each visualization builds upon the others, guiding the user through key findings on pricing trends, regional variations, and brand performance. The narrative is data-driven, providing actionable insights for buyers, sellers, and industry analysts.

6. CONTRIBUTIONS

Every member in the team contributed to **Data Pre-processing**.

Lekha Chittajallu	Distribution of Average Mileage and unique brands among states
Rishi Yedlapalli	Distribution of Vehicle Prices by Title Status
Sai Praneeth Achanta	Symbol map with Glyph
Sharath Kumar Reddy Kapu	Horizontal Bar Chart showing top N vehicle models
Sree Sai Preetham Nandamuri	Tree Map, Bar Chart with Average Price by Brand and Model
Vamshidhar Reddy Ankenapalle	Sankey Diagram & Presentation
Venkata Achyuth Kumar Sanagapalli	Forecasting Line Chart and Story
Venkata Sumanth Reddy Vangala	Heat Map, Bar Chart with Average Price by Brand and Model
Victor Paul Buddha	Colour-wise Distribution of Vehicles with Average Price and Count
Vishnu Sai Inakollu	Scatter Plot with Animation

7. CONCLUSION

- **Summary of Findings**: This project successfully transformed raw auction data into a dynamic visual analytics tool. Key insights include:
  - Mileage is a significant factor in price depreciation.
  - Certain brands (e.g., Mercedes-Benz) command premium prices.
  - Regional disparities exist in pricing and title status distribution.
  - The used car market exhibits seasonal and yearly trends that can inform buying and selling strategies.
- **Real-World Applications**: The visualizations have practical implications for various stakeholders:
  - **Buyers**: Can identify vehicles that offer the best value.
  - **Sellers**: Can optimize pricing strategies.
  - **Businesses**: Can make data-driven decisions on inventory management and market positioning.

8. FUTURE SCOPE

One of the key upgrades would be real-time monitoring of auction prices and alerts, enabling users to see live bidding activity, price changes, and time-sensitive offers. Alerts programmed by the user on price caps, low mileage, or imminent auction expirations would enable timely and strategic decision-making. Including a vehicle history score from services like Carfax or Auto Check would add trust and transparency by showing prior accident reports, service, and ownership history. This risk score would complement existing filters for filtering by surface-level details.
Furthermore, adding in an estimator of ownership cost would provide a fuller cost picture by adding in insurance, maintenance, and fuel costs by model and by location. This would allow comparison of automobiles based on long-term cost, not just starting price. Finally, a smart recommendation engine could browse user requirements and search histories and suggest listings suitable for their needs. By applying filtering logic or machine learning, the website could give a more tailored and simplified browsing experience.

9. REFERENCES

1. https://help.tableau.com/current/pro/desktop/en-us/forecasting.htm
2. https://www.auctionexport.com/?gad_source=1&gad_campaignid=21966310756&gbraid=0AAAAD3YprimAlv4Web9yvEWsu30Shpb-&gclid=Cj0KCQjwoNzABhDbARIsALfY8VOQ68ARPenlYMSaviIKGKUonvQDRESMT95uImU2TPcNtG0I91n3-fUaAqChEALw_wcB
3. https://www.tableau.com/blog/bring-your-data-life-viz-animations
4. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.dropna.html
5. https://scikit-learn.org/stable/modules/preprocessing.html

Link for Tableau Dashboard:

https://public.tableau.com/app/profile/venkata.achyuth.kumar.sanagapalli/viz/USUsedCarMarketVisualizer_Group4/PricingAnalysisDashboard?publish=yes