Vehicle_data_analysis

Yash Kharade

2022-08-10

Executive Summary

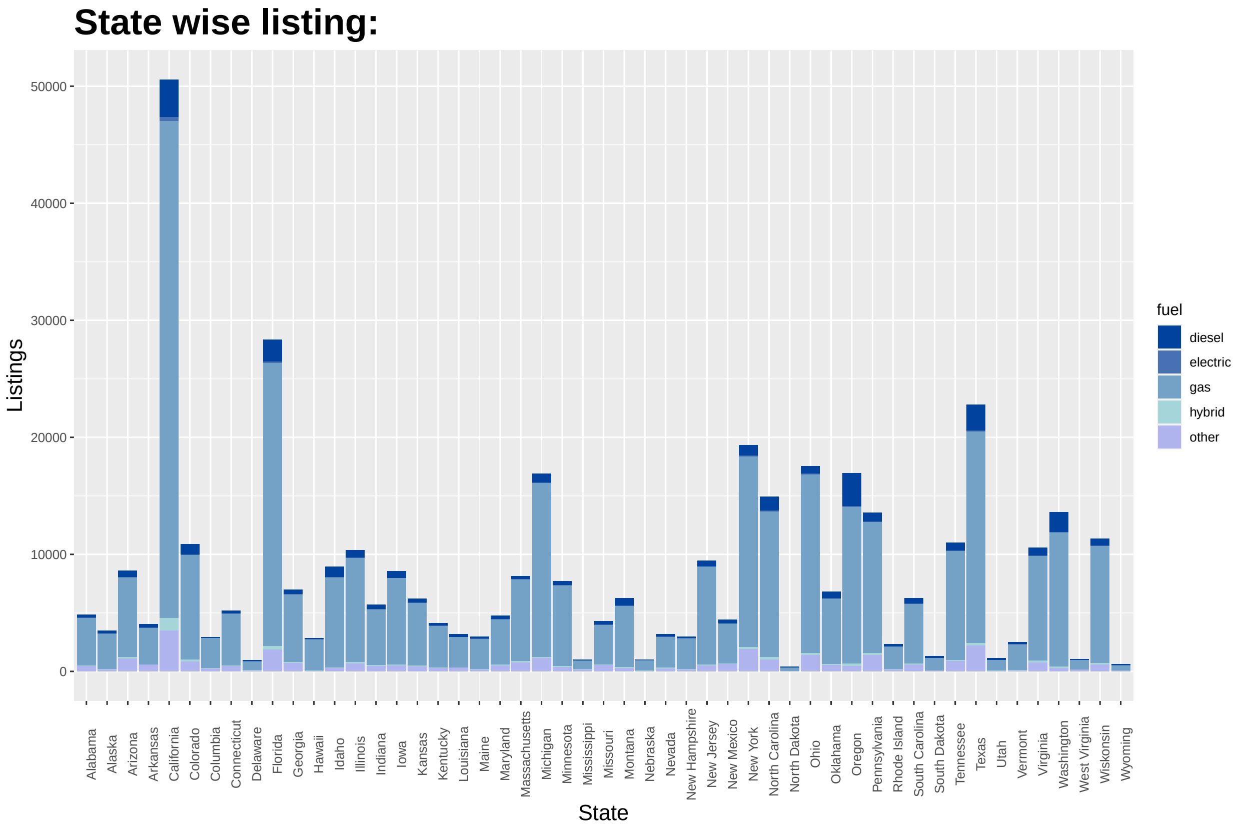
- The data was collected from craigslist from 4th April 2021 to 5th May 2021.
- Overall, there are a total of 426880 observations and 26 variables in the data.
- Another table that is used is statepop which comes with the 'usmap' package. This table shows the population of all 50 states in the USA.
- Every column before using was checked for duplicate values. There were around 92858 blanks in most of the columns.
- The large percentage of listed cars use gas as their fuel.
- California, Florida, Texas, and New York have the highest number of listings.
- Montana, Idaho, Delaware, and Oregon have the highest per capita listings.
- "Credit," "vehicle," "car," and " financing" are a few of the most common words used in the description column in the top 4 states.
- Average prices of vehicles have increased from 10000 in the year 2000 to 30000 in recent times.
- "Ford" followed by "Chevrolet" have the highest listings in the top 4 states.
- Sedan, SUV, pickup, and truck are the most listed type of vehicles across all the states.
- Listed cars which are manufactured between 1975 to 2008 are listed at a fair price.

Introduction

Data allows data analysts to visualize relationships between various variables present in the data. It helps data analysts to back their suggestions on a certain topic by data.
This report is an attempt to understand and visualize the vehicles dataset. Having a vehicle gives anyone the freedom to commute anywhere one needs to. Recently due to the chip shortage car companies are unable to roll out new vehicles as they used to before. Some reports say that the shortage couls stretch til 2024. Due to the large demand and low supply the prices of the cars have increased on an alarming rate.
Consumers in urgent need are moving towards the used car market. In an article named "Why are used Cars so expensive right now" published on the Honda website mshonda.com states that even the prices of the used cars have gone up. People are having a hard time deciding which car to buy.
This report aims to find the best value used cars from all the listed cars on carigslist from 4th April 2021 to 5th May 2021. The dataset has listings of cars manufactured in the early 1900's to recent times.

Findings:

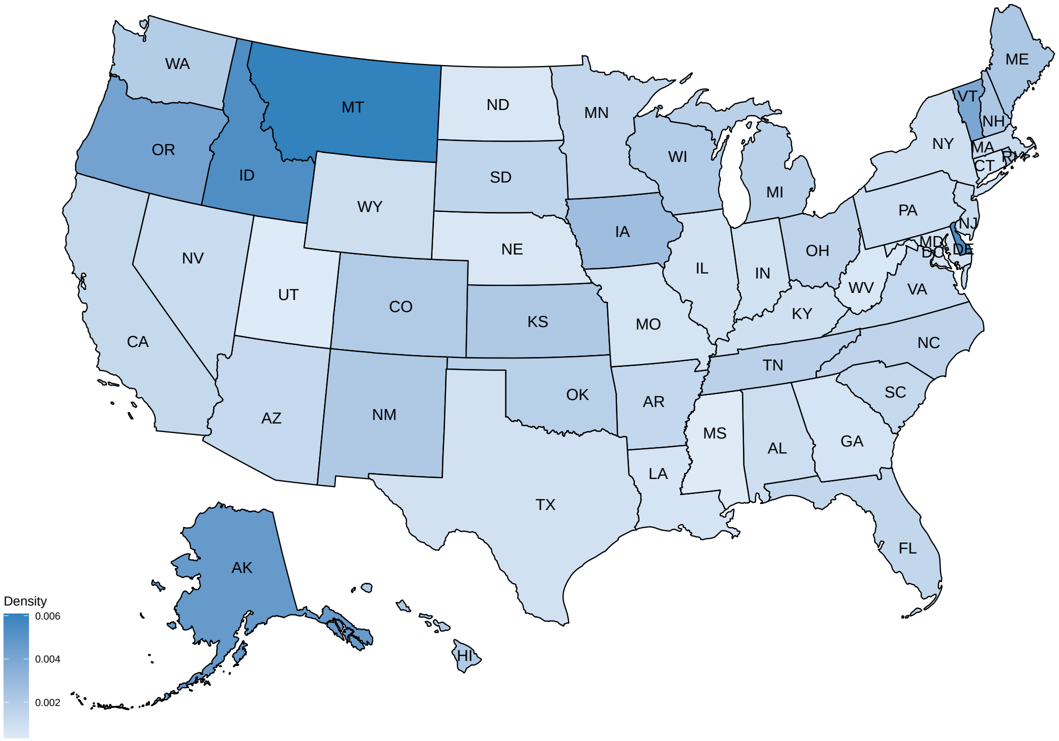
This graph shows state-wise listings of all 50 states. Each bar is divided into the type of fuel the listed vehicles use. A duplicate state column was created and named 'sfname' and replaced the abbreviations with full state names.



It is not surprising that most of the listed vehicles use gas as fuel. We can also see that California has a much higher number of listings than the rest of the states. California, Florida, Texas, and New York are the states with the highest number of listings. California has a little over 50000 total listings; the second highest is Florida, with a little under 30000 listings. We can see that California has a huge used car market.

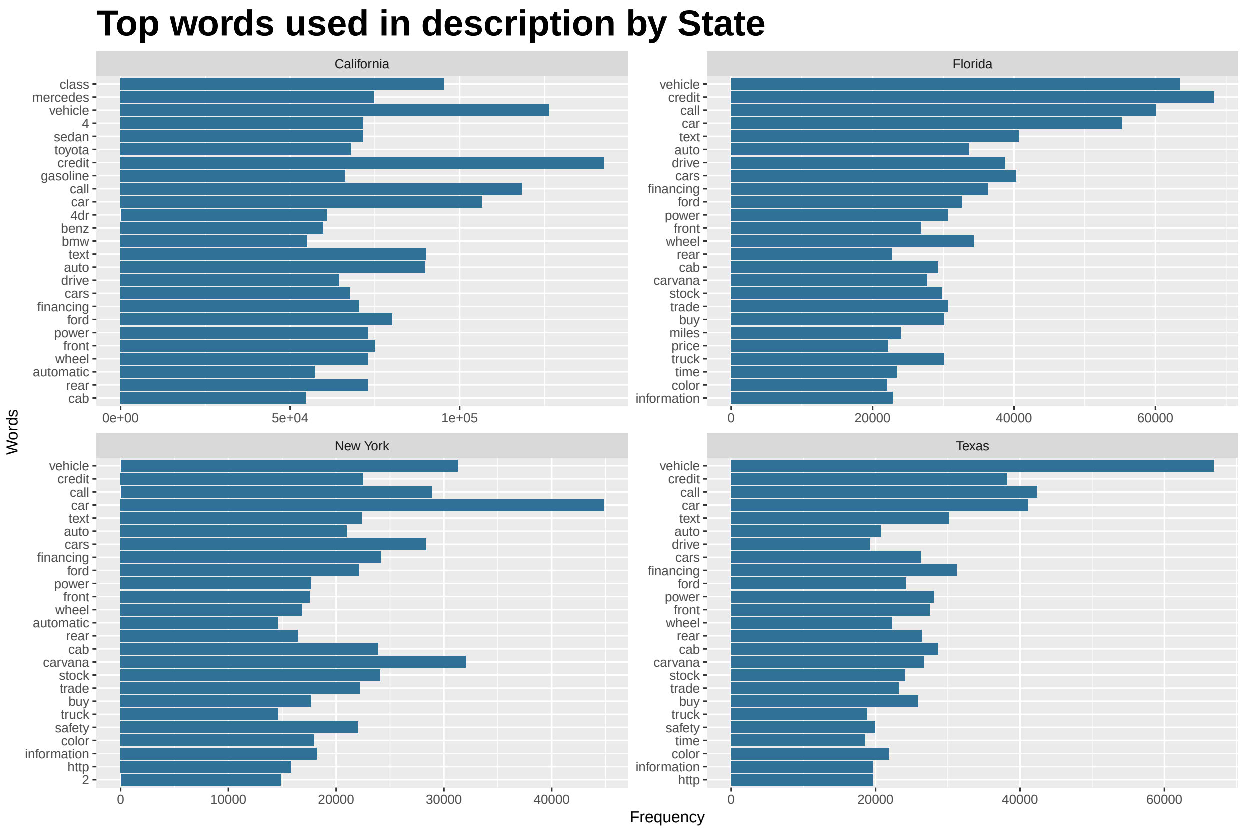
This graph shows us the state-wise per capita listing. We have used the 'population' column from the 'statepop' table to calculate the per capita(Count/Population).

Per capita listings:



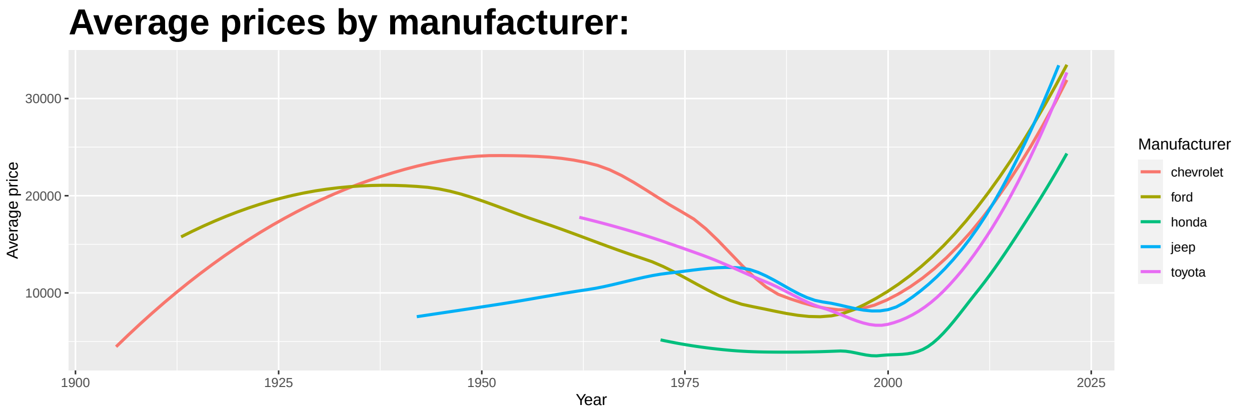
California may have the largest market but does not have a high per capita listing. Even Texas, Florida, and New York have low density. Montana, Idaho, Delaware, and Oregon have the highest per capita listings.

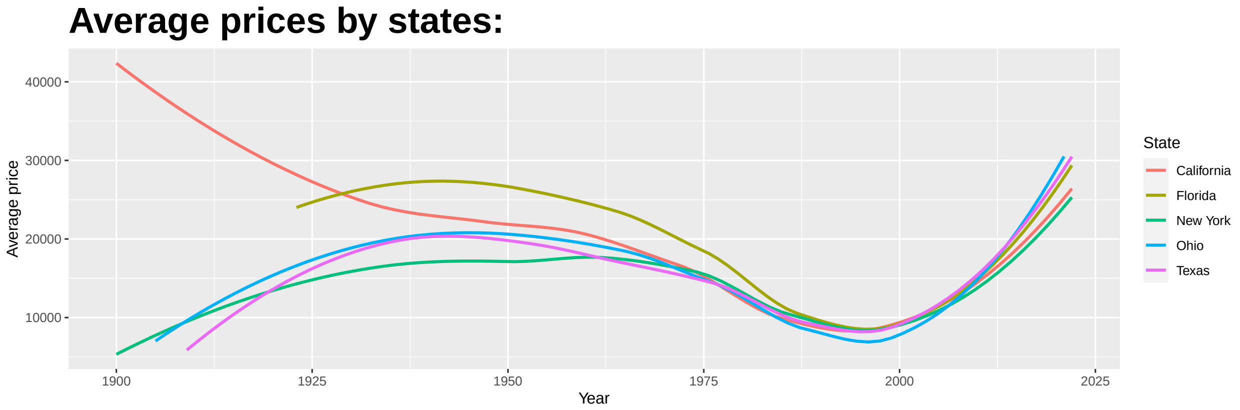
This plot has used the ‘description’ column to find the most common words used by the sellers in their descriptions. Each description was first tokenized and then counted. This plot only uses the top 4 states(California, Florida, New York, and Texas) for this plot.



Words such as: ‘vehicle’, ‘credit’, ‘financing’, ‘car’, ‘color’, ford etc are mentioned in all the 4 states. California has some unique frequent words such as: ‘Mercedes,’ ‘class, and ’BMW,’ suggesting that the percentage of ‘Mercedes’ and ‘BMW’ cars should be higher in California than in the rest of the states.

These plots calculate the average prices of vehicles from early 1900 to 2021 by different manufacturers(top 5) and state (top 5).

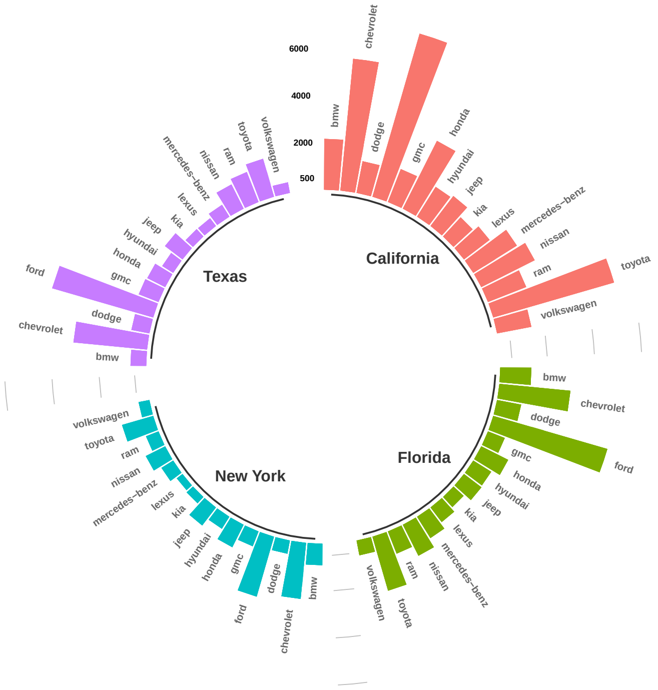




Here we can see that in the top graph, the lines are scattered as compared to the bottom graph. It is because every brand has a specific price range for its vehicles. Honda seems to be much cheaper than the rest of the brands. For Honda, the average price in the year 2000 is around 2500$, whereas the 2nd lowest(Toyota) is about 7500$ which is three times higher. In the state-wise plot, all the lines are close to each other, indicating that the prices do not change drastically in different states. It is no surprise that the cars manufactured after the year 2000 have a high price. The cars manufactured from 1925 to 1975 seem to have higher prices.

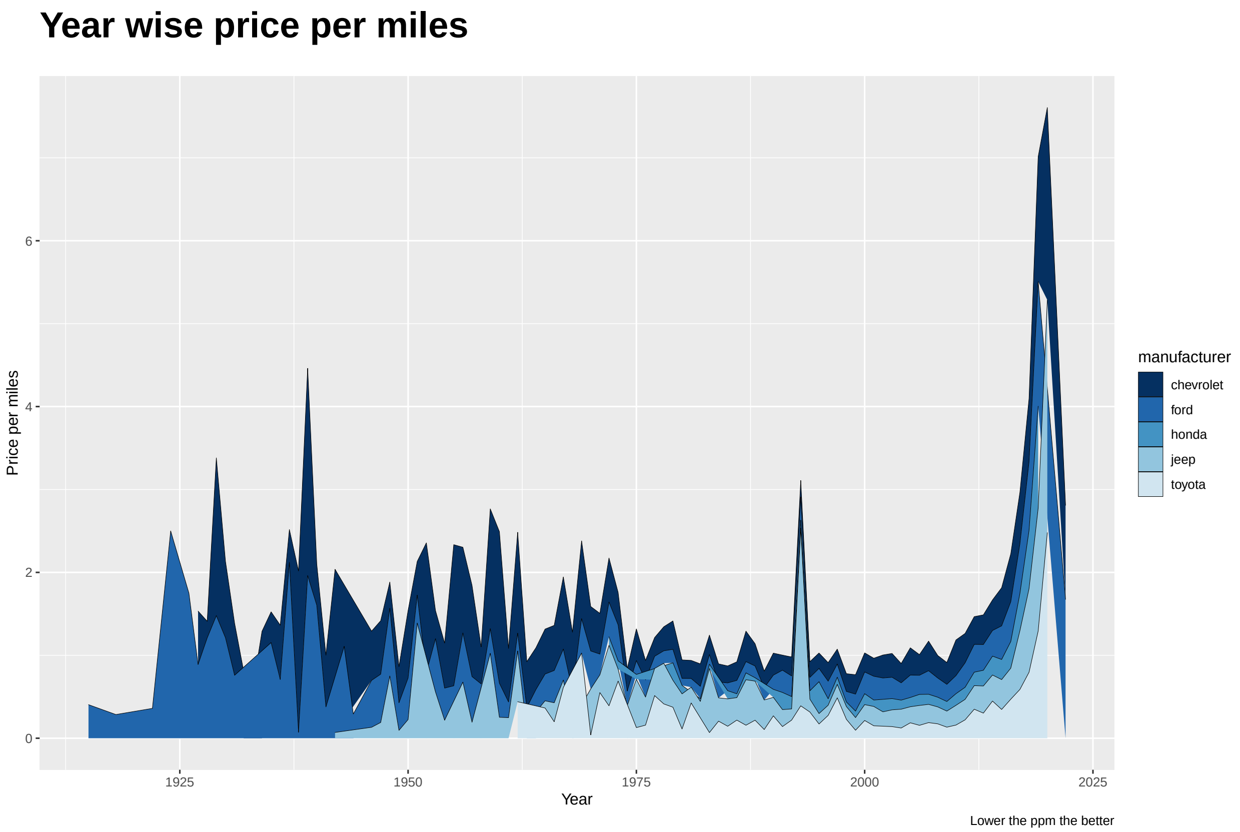
This is a circular bar plot showing the top 15 manufacturers of the top 4 states.

Statewise top 15 manufacturers.



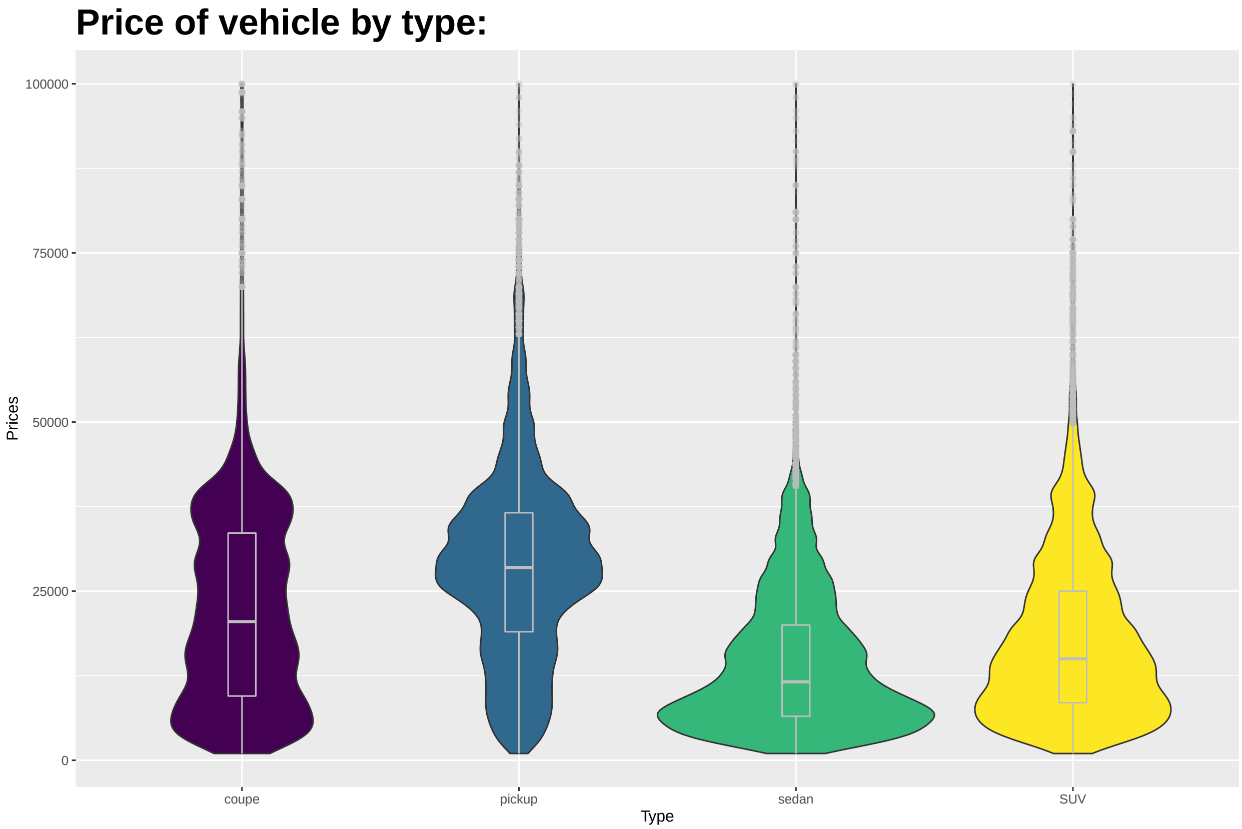
Ford and Chevrolet are dominant in California, Florida, and Texas. New York seems to have an even distribution amongst their top 15 manufacturers. As found in the description column analysis, ‘Mercedes’ and ‘BMW’ have a much higher presence in California than in other states.

This graph shows us the actual value of the vehicle. The price per mile (ppm) is calculated by dividing the price by the odometer(number of miles). The lower the ppm, the better it is for the buyer.



It can be seen that cars manufactured between 1975 till around 2009 are listed at a fair price. The cars manufactured after 2010 seem to be costlier. Cars from 1925 to 1975 are cheaper than the 2010 to 2021 range but costlier than the 1975 to 2009 range.

This plot shows price of a vehicle by its vehicle type.



Sedan is the most economical type among the top 4 listed vehicles. Unsurprisingly pickup trucks are the costliest among all.

Conclusion and findings

It has been found that Montana has the highest per capita listings, although California has the largest market with above 50000 listings. Most of the listings are for gas-fueled cars. For a consumer deciding to buy a vehicle, a few things must be considered. If one wants an affordable car, they can opt for a Honda vehicle as the average price is way lower than the rest. It went from 2500$ in 2000 to around 23000$ in 2021 while the second lowest(Toyota) went from 7500$ in 2000 to 33000$ in 2021. Even though Honda vehicles seem cheaper than Toyota ones, Toyota has the best value for money than the rest of the manufacturers. Among the different types of vehicles, many sedan vehicles are around the 10000$ mark. The average of the rest of the vehicle types is above 15000$.
Suggestion: The consumer has two options if they want to buy a used car in the current market conditions. They can either buy a Toyota sedan or a Honda sedan (based on what they want, an affordable car vs. a quality car). They can also buy a Honda for now if it is an emergency and change it to a Toyota when the chip shortage problem is resolved, and the prices return to normal.

Appendix

Data dictionary(Vehicles dataset)

Type	Field Name	Considered using?
Integer	ID	No
Integer	ID	No
String	Url	No
String	Region	No
String	Region Url	No
Integer	Price	Yes
String	Manufacturer	Yes
String	Model	No
String	Condition	No
String	Cylinders	No
String	Fuel	Yes
Integer	Odometer	Yes
String	Title Status	No
String	Transmission	No
String	VIN	No
String	Drive	No
String	Size	No
String	Type	Yes
String	Paint	No
String	Description	Yes
String	State	Yes
String	Lat	No
String	Long	No
String	Posting Date	Yes

Data dictionary(statepop table)

Type	Field Name	Considered using?
Integer	fips	Yes
String	abbr	Yes
String	full	Yes
Integer	statepop_2015	Yes