Data Analysis and Visualization

Spring 2021 - Project

Pranav Patil

pp798@scarletmail.rutgers.edu

Problem Statement: -

1.Use Tableau to explore a dataset of your choice based on your HW3 assignment. Optional, you may also use python or any other language to prepare your dataset.

2.Develop three visualizations that tell a story. These should be carefully designed and annotated. Use Tableau to build them but you may have to annotate them yourself.

3.Create one or more interactive visualizations. Focus on quality rather than quantity.

4.Optional: May consider creating a dashboard or a High Dimensional Visualization.

5.Deliver a word/pdf document with your findings and include screenshots of your visualizations. Write two pages about the insights you obtained from the data and the findings your visualizations convey.

6.Upload a Tableau Packaged Workbook (TWBX) File with your pdf document.

7. You will present your project on May 12th to the entire class based on alphabetical order.

U.S. Used Cars Data

Dataset

The dataset that I have chosen for this project is the Used Cars Dataset (Vehicles listings from Craigslist.org). This dataset contains all the information that Craigslist provides on car sales within the United States. The data is limited to advertisements that were posted in 2020.

Dataset URL- https://www.kaggle.com/austinreese/craigslist-carstrucks-data

Exploring the Dataset

Initially, I connected the csv file to a new tableau workbook and further explored the dataset. I also kept the csv file open in Excel to get a better understanding about the data.

Variables and variable types: -

1. Unnamed – Integer – Serial number
2. id - Unique (Integer) - Listing ID
3. url – Unique (String) - Listing URL
4. region - String - Given Craigslist region
5. region_url - Unique – Region URL
6. price - Integer - Listing price
7. year - Integer - Year of manufacturing
8. manufacturer - String - Manufacturing company
9. model - String - Model of vehicle
10. condition - String - Condition of vehicle
11. cylinders - String - Number of cylinders
12. fuel - String - Type of fuel required
13. odometer - Integer - Miles traveled
14. title_status - String - Vehicle title status/existence
15. transmission - String - Transmission of vehicle
16. VIN – Unique (String - Alphanumeric) - Vehicle Identification Number
17. drive - String - Drive of vehicle
18. size - String - Size of vehicle
19. type - String - Type of vehicle
20. paint_color - String - Color of vehicle
21. image_url - String - Image URL
22. description - String - Listing description
23. state - String - Listing state
24. lat - Geographic - Latitude of listing
25. long - Geographic - Longitude of listing
26. posting_date - Integer - Listing posting date

Variable documentation of the relevant variables: -

No.	Variable Name	Variable type	Variable Definition	Variable Unit / Description
1	year	numerical	Year when the vehicle was manufactured.	Year can be obtained from VIN and can be compared with what the seller has claimed.
2	odometer	numerical	Reading on the odometer as reported by the seller.	Unit – miles. Reported variable since not all values in the dataset are verified and the seller might report a false value.
3	lat	numerical	Latitude of the region from which the ad was posted.	Geographical variable (coordinate)
4	long	numerical	Longitude of the region from which the ad was posted.	Geographical variable (coordinate)
5	region	categorical	City from the which the ad was posted.	Name of city. (string)
6	manufacturer	categorical	Manufacturer of the vehicle	Name of the manufacturer (string)
7	model	categorical	Model number of the vehicle	Name of the model of the vehicle (string)
8	condition	categorical	The condition of the vehicle put up for sale. This is a reported variable. There is no way to measure it because the seller enters the condition of the vehicle manually.	All values range between new, like new, excellent, good, fair, and salvage. It is not a measured variable
9	cylinders	categorical	Number of cylinders the vehicle has.	It could have been a numerical variable, but every record has “cylinders” mentioned in front of the number. Ex. “4 cylinders”
10	fuel	categorical	The type of fuel the vehicle runs on.	All values range between gas, diesel, hybrid, electric, and other fuel.
11	title_status	categorical	Variable to establish if the car is legally “clean” or not. This is a reported vehicle, but can easily be determined by using the VIN.	All values range between clean, missing, lien, salvage, rebuilt, and parts only.
12	transmission	categorical	Transmission type of the vehicle	All possible values are automatic, manual, or other
13	drive	categorical	The type of drivetrain of the vehicle.	All values range between rwd, fwd, and 4wd.
14	size	categorical	The size of the vehicle in terms of space inside the vehicle and/or passenger capacity.	All values range between full-size, mid-size, compact, and sub-compact.

15	type	categorical	The type of the vehicle as per the physical dimensions or appearance and other factors.	All values range between other, sedan, SUV, pickup, coupe, van, truck, mini-van, wagon, convertible, hatchback, bus, offroad.
16	paint_color	categorical	The color of the exterior of the car.	All values range between blue, red, silver, black, white, grey, orange, green, yellow, custom, brown, and purple.
16	id	categorical	The state from which the ad was posted.	The name of the state (within the US).
18	price	numerical	The most important variable in which I am interested. It is the price at which the seller has listed his vehicle for sale.	The values of this variable are highly diverse, starting from 0 till around almost 200000.

Exploratory Data Analysis (EDA)

For Exploratory Data Analysis (EDA) and data preprocessing, I used Spyder (Python 3.8). Using python, I could easily perform data cleaning, and so I decided to use Spyder. In the next few pages, I have included snapshots of code from Spyder and some visualizations used during EDA.

I. **Original Dataset Size**

The original dataset that I downloaded had **458213 rows and 26 columns (variables)**.

I removed the columns that did not have any significance towards analysis and visualization. The variables that I removed were as follows: -

1. ID - The unique ID of the listing. This variable is not required for analysis and visualization.
2. Unnamed – An unnamed column meant for indexing (serial number). This is an unnecessary variable.
3. url – The webpage of the advertisement listing.
4. region_url – The webpage of region in which the ad was listed. The region of the listing is already being used but the URL is unnecessary.
5. VIN – Vehicle Identification Number. It is only for identifying a vehicle uniquely and has no relevance for our analysis and visualization purposes.
6. Image_url – The image of the vehicle that has been put up for sale.
7. Description – The description that the seller/dealer of the car has given while posting the ad.
8. posting_date – The date on which the ad was posted for sale. I removed this variable because this dataset contains ads from 2020 only and it does not matter in the analysis.

```
df.columns
df2=df.copy()
df2=df2.drop(df2.columns[0],axis=1) #Dropping the unnamed column
df2=df2.drop(columns=['id','url','region_url','VIN','image_url','description','posting_date'])
```

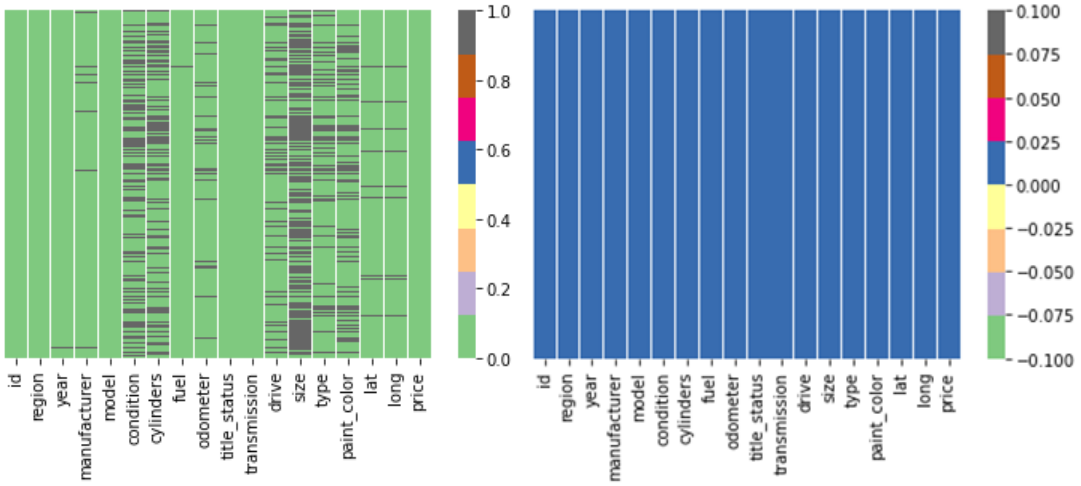
Thus, I removed 8 variables (out of which one was an unnamed variable) out of the total 26 variables. After removing the irrelevant variables, I **differentiated the variables as numerical and categorical.** Thus, the dataset now has **5 numerical variables and 13 categorical variables**.

II. **Finding and Dealing with Null Values**

The number of null values in each column were as below: -

```
id                0
region            0
year           1050
manufacturer  18220
model          4846
condition     192940
cylinders     171140
fuel            3237
odometer       55303
title_status    2577
transmission    2442
drive         134188
size          321348
type          112738
paint_color   140843
lat             7448
long            7448
price              0
dtype: int64
```

Thus, most of the columns had null values. To better understand the null-value data, I plotted a heatmap.



Heatmap: Null values in the original dataset

Heatmap: No null values present after applying an Iterative Imputer

Thus, from the first heatmap we can that there were an extremely high number of null values in the dataset, especially in the **size, condition, cylinders, type and paint_color columns.** Simply removing all the rows that had a null value would not work, since the dataset would be reduced to only a few thousand records if done so. This is because some rows had a null value in only one column, and every row that had even one null value in any of the column cannot be removed. So, to deal with the null values, I decided to use an **iterative imputer to estimate the null values**. So, the score on the entire dataset was estimated by filling missing values by mean and median. For this purpose, I used the **ExtraTreeRegressor since it had a low MSE (Mean Squared Error)** value as compared to other imputer methods. Thus, after filling the missing values, I plotted one more heatmap to make sure that there are **no null values in the dataset.**

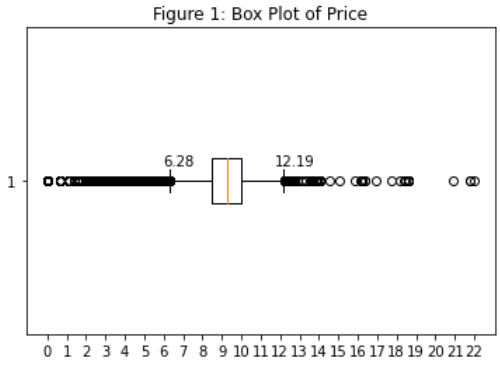
III. Finding and Dealing with Outliers

I defined the standard formula for finding outliers in every important variable **(price, odometer and year).**

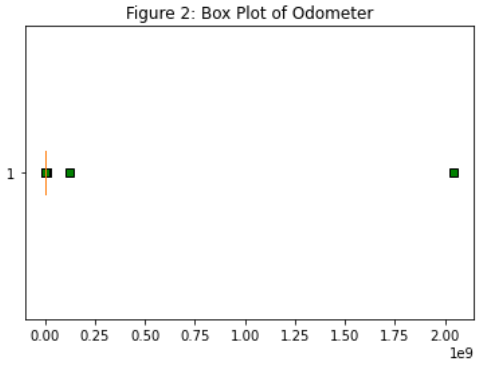
```
def outliers(arr,col):
    x=sorted(arr[col].values.ravel())
    L_25=25/100*(len(x)+1) #L_p where p=25%
    i_p=int(str(L_25).split(".")[0])
    f_p=int(str(L_25).split(".")[1])
    q1=x[i_p]+f_p*(x[i_p+1]-x[i_p])

    L_75=75/100*(len(x)+1) #L_p where p=75%
    i_p=int(str(L_75).split(".")[0])
    f_p=int(str(L_75).split(".")[1])
    q3=x[i_p]+f_p*(x[i_p+1]-x[i_p])

    #q1,q3=(arr[col].quantile([0.25,0.75]))

    IQR=q3-q1
    x1=q1-1.5*IQR
    x2=q3+1.5*IQR
    return (x1,x2)
```

After defining the formula, I plotted box plots for price, odometer, and year; and plotted a histogram for the year variable to find out the frequency and density of records in a year-wise manner. After plotting the outliers, I removed them.

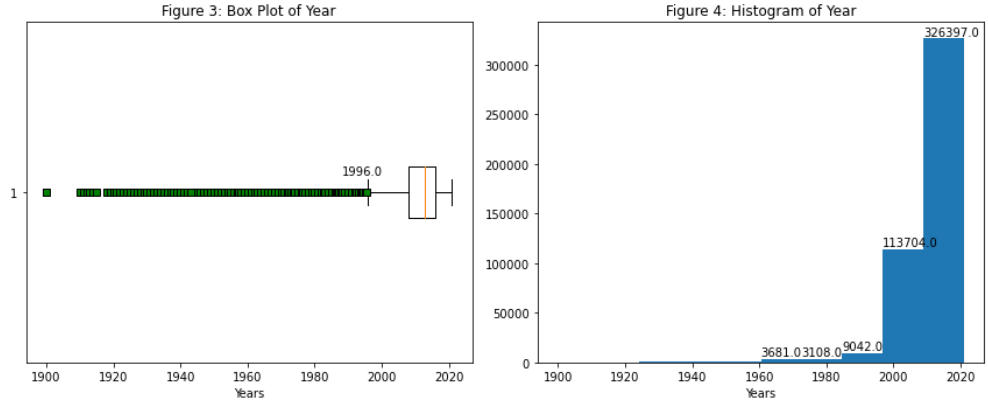
1. Price



2. Odometer



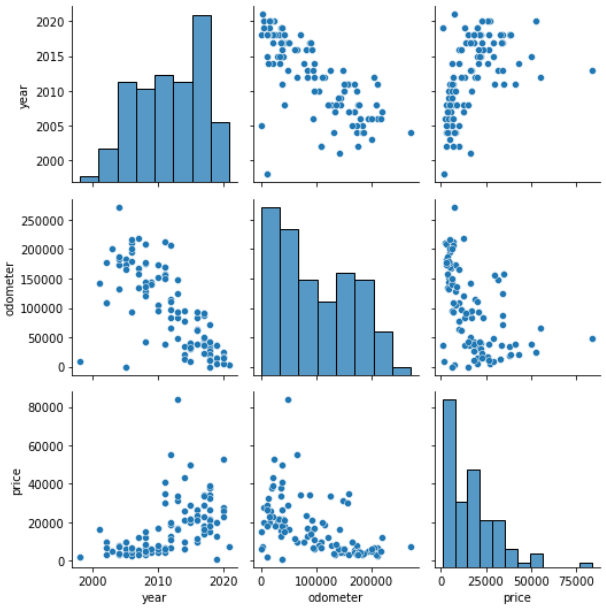
3. Year- Box plot and histogram



IV. Finding and plotting relationships between variables

To see distribution of single variables and the relationships between two variables, I used a pair plot. I took a small sample out of the dataset for this purpose. To plot the pair plot, I used the Seaborn library, which is a Python data visualization library based on matplotlib. The pair plot can be interpreted as follows: -

- The **histogram** along the diagonal shows the probability distribution of a single variable.
- The **scatter plots** in the upper triangle and lower triangle show the relationship between the 3 variables.



Finally, after completing EDA, I created a new csv file that had the same dataset, but without any unnecessary columns (variables), null values, and outliers. I then used this cleaned dataset for

creating visualizations. This dataset required a fair amount work before it was ready to use for visualization.

An overview of the original dataset that I downloaded and the dataset after cleaning: -

```
df=pd.read_csv('C:/Users/prana/Desktop/Spring_Courses/DAV/Homeworks/HW3/vehicles.csv')
df=pd.DataFrame(df)
df.shape

df2=pd.read_csv('C:/Users/prana/Desktop/Spring_Courses/DAV/Homeworks/HW3/vehiclesFinal.csv')
df2=pd.DataFrame(df2)
df2.shape
```

```
In [7]: df.shape
Out[7]: (458213, 26)

In [8]: df=pd.read_csv('C:/Users/prana/Desktop/Spring_Courses/DAV/Homeworks/HW3/
vehicles.csv')

In [9]: df=pd.DataFrame(df)

In [10]: df.shape
Out[10]: (458213, 26)

In [11]: df2=pd.read_csv('C:/Users/prana/Desktop/Spring_Courses/DAV/Homeworks/HW3/
vehiclesFinal.csv')

In [12]: df2=pd.DataFrame(df2)

In [13]: df2.shape
Out[13]: (395980, 18)

In [14]:
```

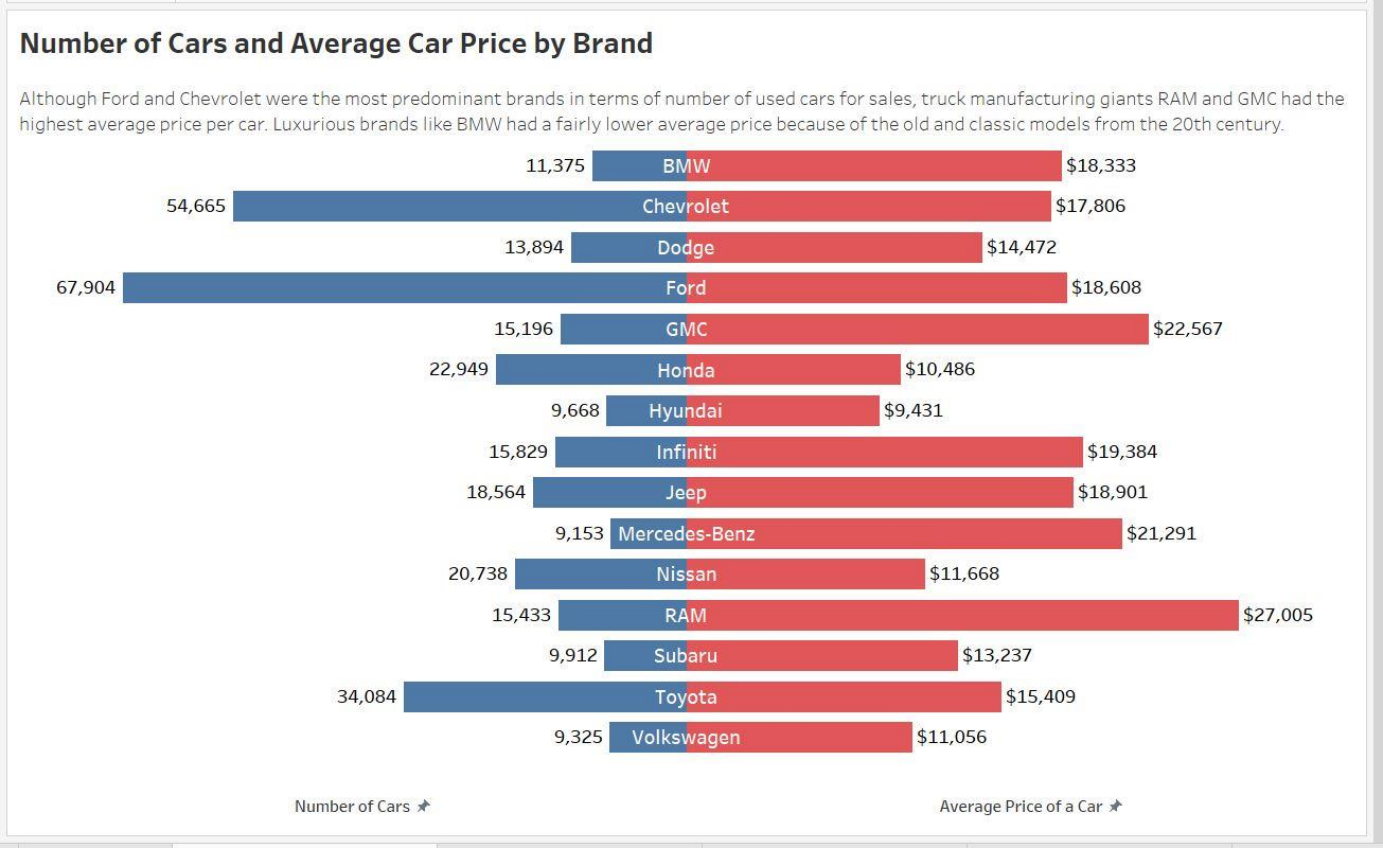
Before starting to develop visualizations, I wanted to make sure if I can tell a story using this dataset. After completing EDA, I had understood the dataset well and I believed that I could now tell a story using effective visualizations made from this dataset. After initially creating exploratory visualizations in HW3, I planned to create explanatory visualizations that told a story. This dataset supported my visualization idea: -

- It can tell interesting facts with various types of visualizations.
- The dataset is relevant and fresh since has data of US used car sales from 2020.
- It is simple and has a large number of records, which will help me visualize the data in a better way. I think that having a dataset with 400,000 records is advantageous over a small dataset for a number of reasons, one of them being that the analysis has better accuracy and it means more if applied in practical life.

In the next section, I have included screenshots of my visualizations. I have also written the about the insights I obtained from the data and the findings the visualizations convey.

Visualizations

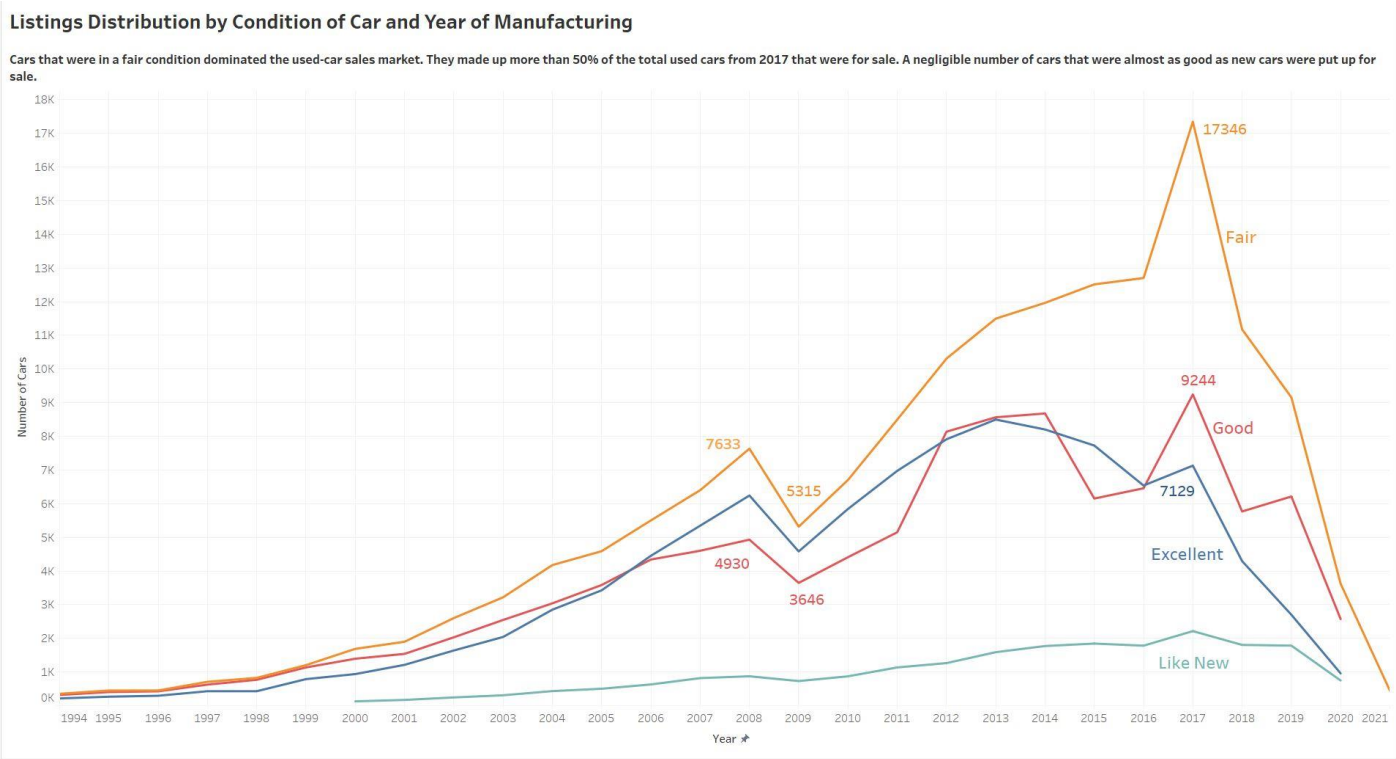
1. Dual-Axis Diverging Bar Chart



Findings and Insights: -

1. Ford, Chevrolet, and Toyota were the top three brands of used cars for sale in 2020 across the US. These brands had a large percentage of new as well as used cars in the US automobile industry for almost two decades.
2. The average price of used trucks was the highest amongst all brands- RAM and GMC, manufacturers of trucks, had the highest average price per vehicle. These brands made utility vehicles and trucks, so the average price was naturally higher than other cars.
3. Although BMW is considered a luxurious brand, the average price of BMW cars was considerably low. The main reason for this is that most of the used BMWs for sale were classics- the models were manufactured a very long time ago, and most of them dated from the late 1990s to early 2000s.
4. There was a huge discrepancy in the brand wise distribution of cars- the top 15 brands made up a significant percentage of the total used cars for sale in the US. However, the total number of brands was very high- there were around 50 different brands of used cars for sale.

2. Line Chart



Findings and Insights: -

1. The cars in a fair condition made up a large proportion of the total cars, whereas cars that were almost like new cars were extremely rare.

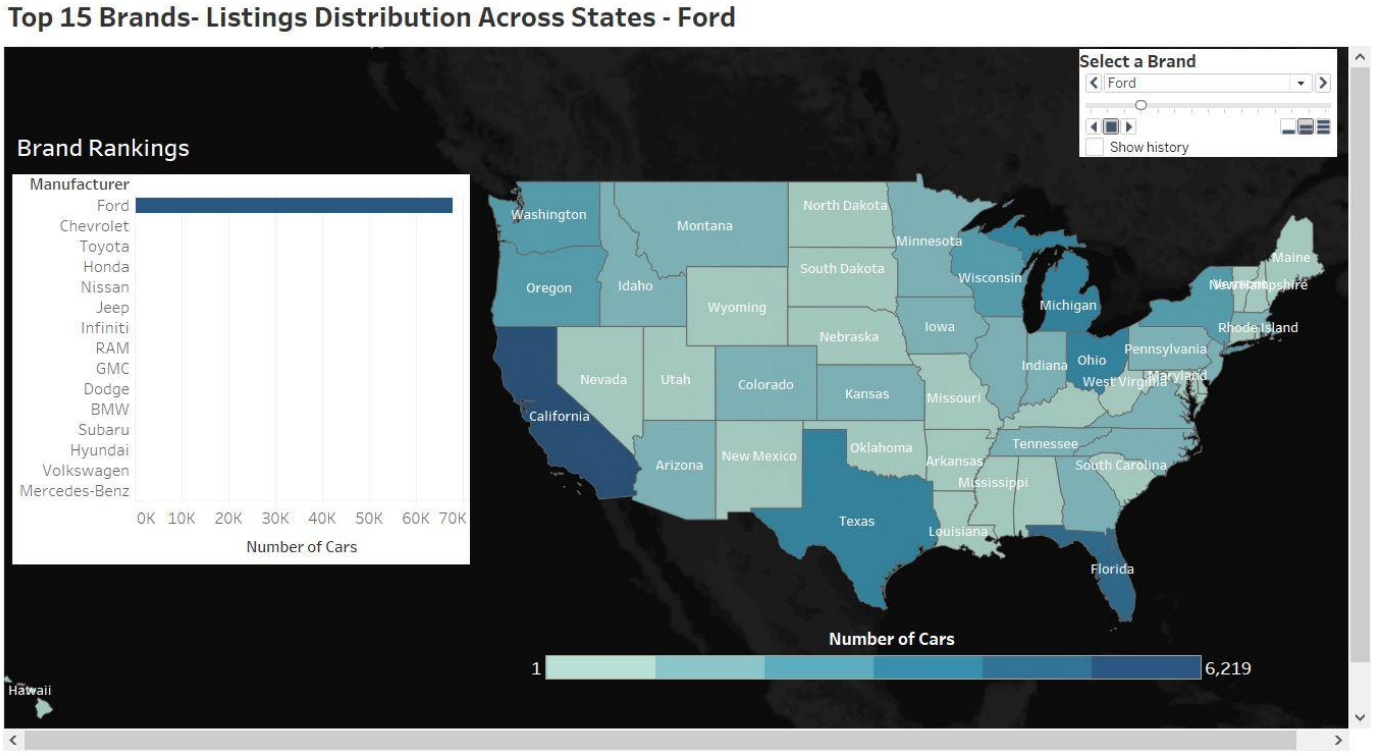
2. Out of the total number of used cars for sale, cars in a fair condition have always dominated the market, no matter what the year of manufacturing is.

3. The number of cars in excellent condition was steadily growing and almost keeping up with cars in a fair condition until 2011, when the number of cars in a fair condition sky-rocketed whereas the pace of excellent-conditioned cars only managed increase slightly.

4. A sudden decline in the number of listings can be seen from 2008 to 2009. This was because of The Great Recession that started in 2007 and ended in 2009. After 2009, a steady rise in number of listings was seen up until 2017.

5. An interesting point to consider was that the “condition” variable was a reported variable. The condition of the car is reported by the seller, and it might not always be true. But, if someone really wants to sell their car, they must put the true condition of the car, or the chances of their car being sold are less.

3. An Interactive Dashboard- Animated Map and Bar Chart



Findings and Insights: -

1. This dashboard consists of an interactive and animation-enabled map, and a dynamic bar chart.

2. On the left-hand side, there are rankings of the top-15 brands. They are ranked by the number of cars. On the right-hand side, the map shows the number of listings of every brand across all states in the US.

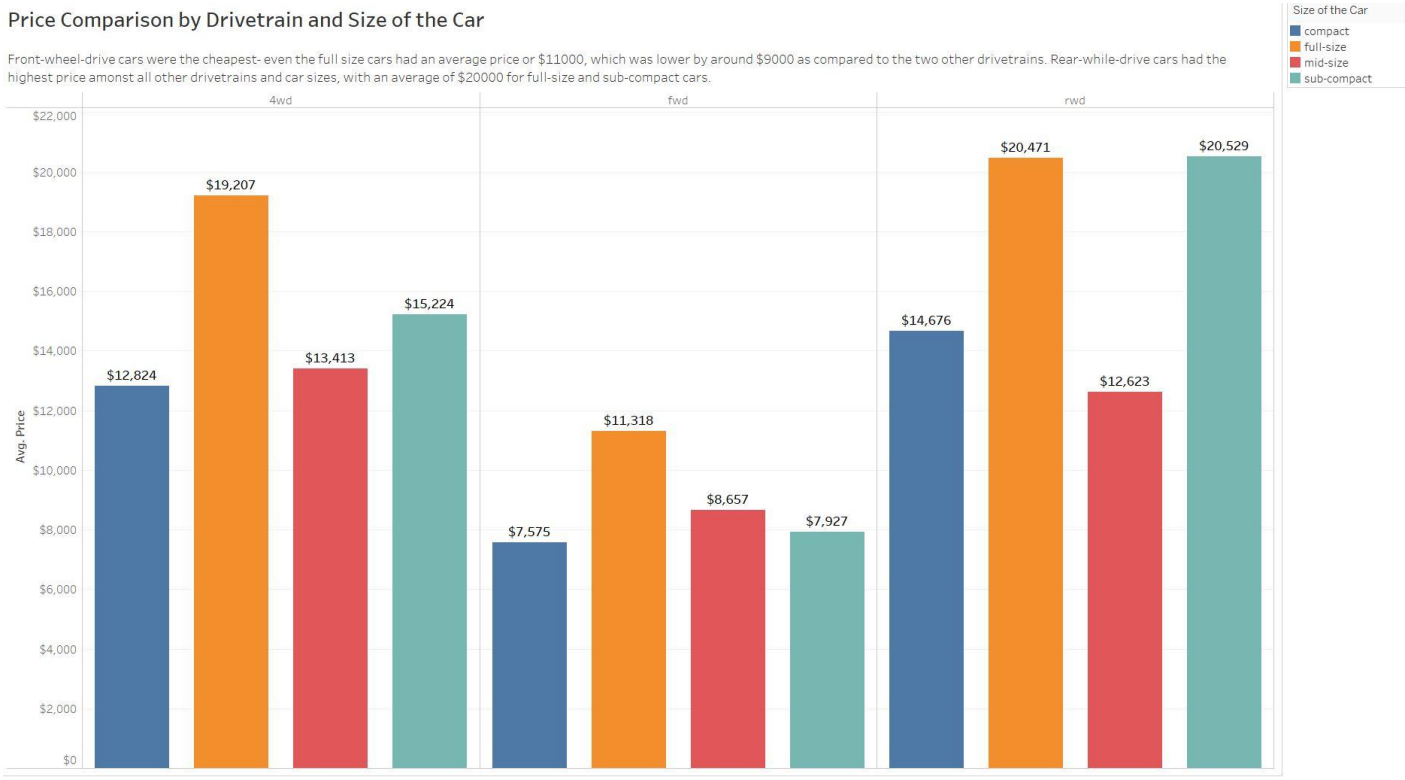
3. Ford is the leading brand of used cars for sale with the highest number of sales listings in California followed by Florida.

4. Mercedes-Benz, being a luxurious brand, does not have many listings, but it is interesting to find that some of the states have a high number of listings for the brand, whereas the brand is almost non-existent in other states.

5. California has the highest number of listed used cars for sale amongst all the states in the US, whereas Wyoming ranks as the lowest in this department.

6. North Dakota, South Dakota, Wyoming, and Nebraska are some of the states where the fewest listings for used-cars for sale are made across the US.

4. Side-by-Side Bar Charts



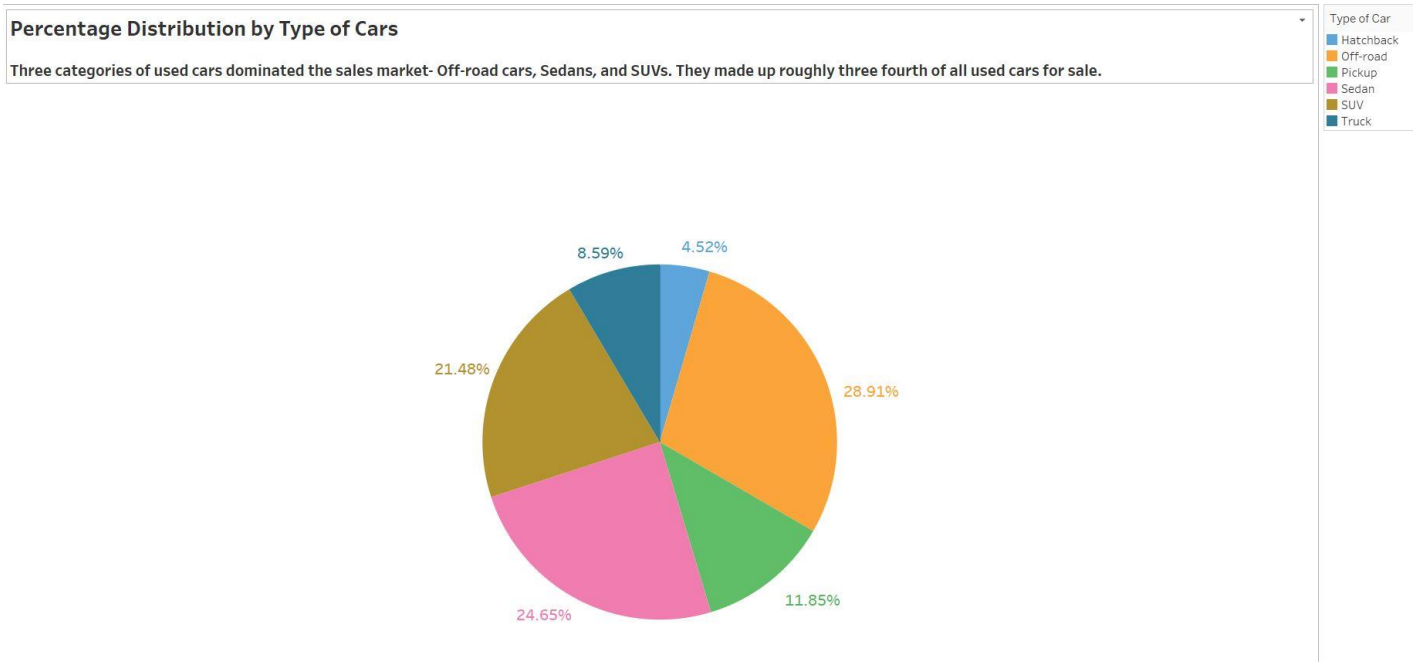
Findings and Insights: -

1. Front-wheel-drive used cars are the cheapest, which is explainable because they have a lower manufacturing cost. The opposite can be said for Rear-wheel-drive used cars- they have a higher manufacturing cost, and new Rwd cars cost a lot expensive as compared to the other two types. So, used Rwd drivetrain type cars are the costliest.

2. 4-wheel-drive cars is the most common type of drivetrain, and the price of these type of used cars lies roughly in between the other two-types. The most expensive 4wd used cars are the full-size cars.

3. The cheapest cars are compact Fwd cars. These cars are generally of low engine capacity and do not have impressive features. Their horsepower is also on the lower end. Thus, it is expected that these types of used cars have a lower cost than the other types.

5. Pie Chart



Findings and Insights: -

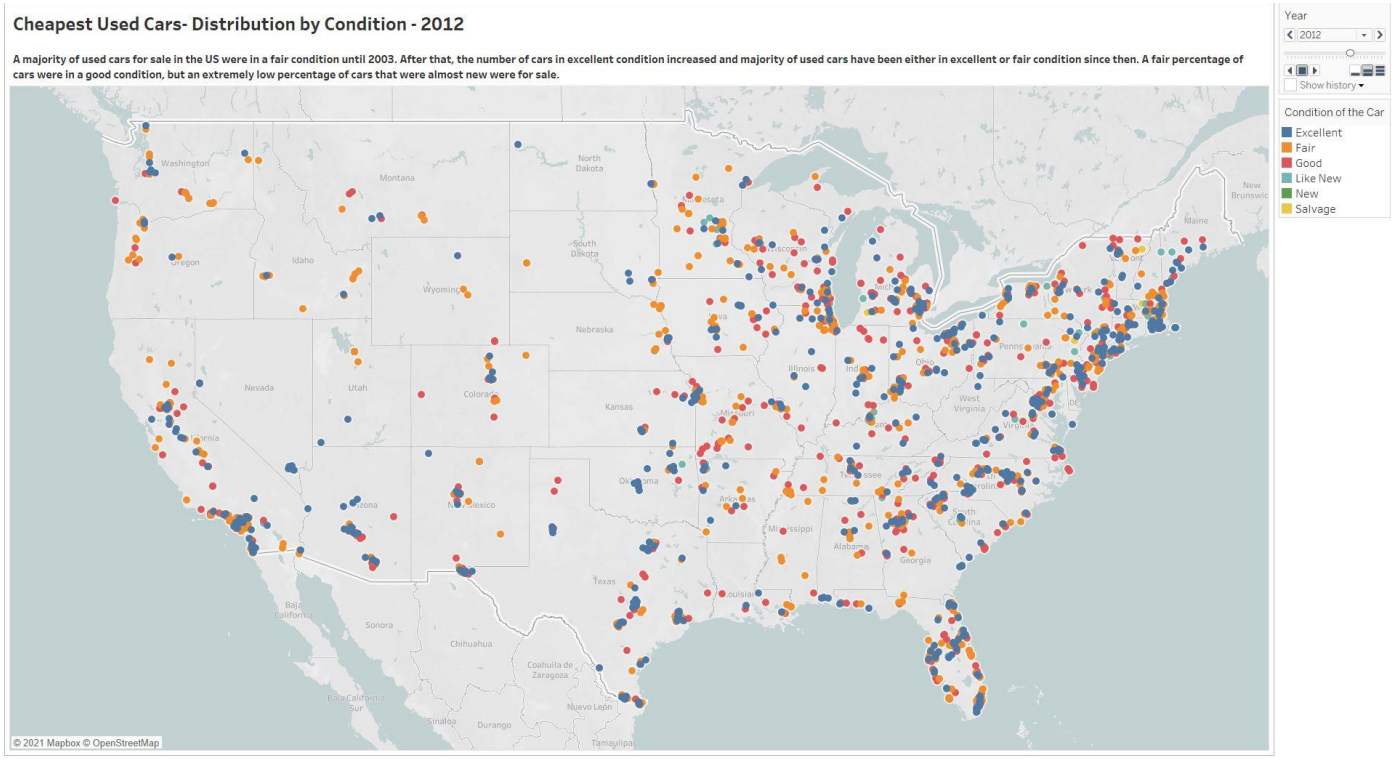
1. Three categories of used cars dominated the sales market- off-road cars, sedans, and SUVs. This can be explained by the fact that the automobile industry in the US majorly consists of consumer-cars; and sedans and SUVs are the most common types of consumer cars.

2. People owning off-road cars tend to sell their cars more often than people owning comparatively luxurious cars. Off-road cars also have a significant number of utility vehicles, which are very commonly listed for sale on Craigslist.

3. The top three categories of cars made up roughly three fourth of all used cars for sale. This is a huge number, especially because the total number of cars (listings) is around 400,000.

4. Hatchback cars made up a mere 4.5% percent of total used cars for sale, and trucks made up for around 9%. These were the lowest-listed types of cars on Craigslist.

6. Interactive Map



Findings and Insights: -

1. This visualization depicts the distribution of listings of cars in the price range of $500 to $5500. These are the cheapest available used cars for purchasing on Craigslist. The cars are filtered on this basis of the condition of the car as reported by the seller.

2. Newer used cars are costlier, older used cars are cheaper. A large number of used cars manufactured before 2014 were cheaper and the number of cheap used cars before 2014 for sale was extremely high as compared to cars manufactured after 2014.

3. The number of cheap used cars for sale started decreasing from 2014 till 2020.

4. A simple explanation for above insights is that as years passed, the average price of new as well as used cars has increased, and so such a pattern is visible.

5. An interesting finding is that, although when we earlier saw that the total number of used cars for sales is dominated by cars in fair condition, when it comes to cheap cars ranging between $500 and $5500, good-conditioned cars and excellent cars also dominate the race.