**Used Car Market Analysis In US**

A Project Report

Presented To

DATA-230

Spring,2022

By

Jitendhar Reddy Adulla

May 22,2021

**ABSTRACT**

**Used Car Market Analysis In US**

There has been an increasing demand for the cars from the last few decades. There are various factors that are responsible for the increase in the demand and thereby creating a hike in the prices. This project deals on the visualization of the specifications and history of used cars data which is collected from the kaggle website. Data Cleaning is performed on the dataset later and various visualizations are performed to determine what factors are affecting the sales of cars. Visualizing this data aids in a rounded analysis on the car condition and the usage statistics. Car resale features and condition data is preprocessed, visualized and presented to the customer, which aids in the determination of resale price, car usage analysis and other such applications.

**Acknowledgments**

I would like to express my sincere gratitude to Professor Andrew H. Bond, who was been a great mentor and guide throughout my project . Thank you for sharing your experiences and insights with us. They were really helpful for me throughout the project and also will be very helpful in my future works. And I would also like to thank our TA Ritanjali Jena for helping me throughout the course period.

**Table Of Contents**

1. **Introduction 5**
   1. **Purpose Of this Document 5**
   2. **Intended Audience 5**
   3. **Scope 5**
   4. **Definitions 6**
2. **Background And Objectives 8**
3. **Organization 9**
   1. **Project Group 10**
   2. **Customer 10**
4. **Architecture And High Level Design 10**
5. **Development Process 10**
6. **Deliverables 14**
7. **Github 14**
8. **Visualizations 14**
9. **Dashboards 24**
10. **Project Plan 27**
11. **References 27**

**1. Introduction**

**1.1 Purpose Of This Document**

The purpose of this document is to provide the description of the various visualizations performed to analyze the market conditions of the used cars in the United States. It provides the details of the trend analysis of the sales of our dataset. This document includes details about organization, roles, deliverables, project risks, time plans and financial plans.

**1.2 Intended Audience**

This document shall be used in all phases of the project as a guideline. Intended audiences of this project are all project stakeholders:

* Professor Andrew Bond
* Jitendhar Reddy Adulla

**1.3 Scope**

This document defines the project plan of the analysis used car sales dataset . The overview includes objectives of the project, architecture and design of project, development process that is going to be used during the project, deliverables, assessment of possible risks and project plan that includes time schedule and activity plan.

**1.4 Definitions**

The following are definitions and type of attributes in the dataset :

1. vin: Type String. Vehicle Identification Number is a unique encoded string for every vehicle.
2. back\_legroom: Type String. Legroom in the rear seat.
3. bed: Type String. Category of bed size(open cargo area) in pickup truck.
4. bed\_height: Type String. Height of bed in inches
5. bed\_length: Type String. Length of bed in inches
6. body\_type: Type String. Body Type of the vehicle. Like Convertible, Hatchback, Sedan, etc.
7. cabin: Type String. Category of cabin size(open cargo area) in pickup trucks. Eg: Crew Cab, Extended Cab, etc.
8. city: Type String. city where the car is listed. Eg: Houston, San Antonio, etc.
9. cityfueleconomy: Type Float. Fuel economy in city traffic in km per liter
10. combinefueleconomy: Type Float. Combined fuel economy is a weighted average of City and Highway fuel economy in km per liter
11. daysonmarket: Type Integer. Days since the vehicle was first listed on the website.
12. dealer\_zip: Type Integer. Zipcode of the dealer
13. description: Type String. Vehicle description on the vehicle's listing page
14. engine\_cylinders: Type String. The engine configuration. Eg: I4, V6, etc.
15. enginedisplacement: Type Float. enginedisplacement is the measure of the cylinder volume swept by all of the pistons of a piston engine, excluding the combustion chambers.
16. engine\_type: Type String. The engine configuration. Eg: I4, V6, etc.
17. exterior\_color: Type String. Exterior color of the vehicle, usually a fancy one same as the brochure.
18. fleet: Type Boolean. Whether the vehicle was previously part of a fleet.
19. frame\_damaged: Type Boolean. Whether the vehicle has a damaged frame.
20. franchise\_dealer: Type Boolean. Whether the dealer is a franchise dealer.
21. franchise\_make: Type String. The company that owns the franchise.
22. front\_legroom: Type String. The legroom in inches for the passenger seat
23. fueltankvolume: Type String. Fuel tank's filling capacity in gallons
24. fuel\_type: Type String. Dominant type of fuel ingested by the vehicle.
25. has\_accidents: Type Boolean. Whether the vin has any accidents registered.
26. height: Type String. Height of the vehicle in inches
27. highwayfueleconomy: Type Float. Fuel economy in highway traffic in km per litre
28. horsepower: Type Float. Horsepower is the power produced by an engine.
29. interior\_color: Type String. Interior color of the vehicle, usually a fancy one same as the brochure.
30. isCab: Type Boolean. Whether the vehicle was previously a taxi/cab.
31. is\_certified: Type Boolean. Whether the vehicle is certified. Certified cars are covered through warranty period
32. is\_cpo: Type Boolean. Pre-owned cars certified by the dealer. Certified vehicles come with a manufacturer warranty for free repairs for a certain time period.
33. is\_new: Type Boolean. If True means the vehicle was launched less than 2 years ago.
34. is\_oemcpo: Type Boolean. Pre-owned cars certified by the manufacturer.
35. latitude: Type Float. Latitude from the geolocation of the dealership.
36. length: Type String. Length of the vehicle in inches
37. listeddate: Type String. The date the vehicle was listed on the website.
38. listing\_color: Type String. Dominant color group from the exterior color.
39. listing\_id: Unique Type Integer. Listing id from the website
40. longitude: Type Float. Longitude from the geolocation of the dealership.
41. mainpictureurl: Type String.

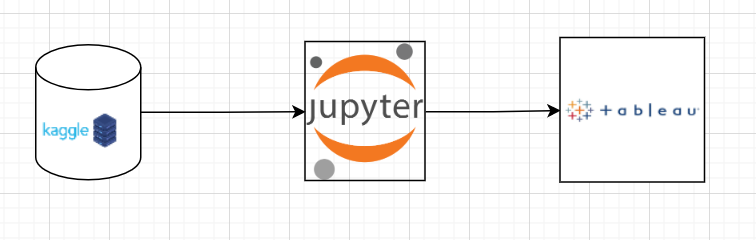
**2. Background and Objectives**

Car resale is a lucrative business, and many cars in the United States undergo the process of change in ownership. This can be due to various factors, such as the aging technology in the car, interest in obtaining newer car by the present owner, having an extensive ridership history and exhausting its prime period, purchasing cars to use them for or sell them after being used for commercial purposes such as taxis, transport vehicles, or for vintage collectibles as a memory of their historic value.

As such, their price determination becomes really hard when specific features aren’t highlighted to the customers. Car features go unnoticed whenever the end user purchases it, making it a lossy purchase sometimes.

To ensure the ease of understanding by all the users, the visualization of these features and specifications is required, making the pictographic representation diagrams clearer to identify and understand the distinctive features. This can facilitate applications like purchase feasibility study on the car using the data, and on other functions such as price negotiability.

**3. Architecture And High-Level Design**

****

Our dataset has been obtained from kaggle.com, the open-source site for multiple datasets on all kinds of information. The data set consists of information of used cars around the United States, between the years 1915 and 2021.

After collecting the data, I work on the data loading and the data cleaning processes. These processes are performed on the Jupyter notebook, using Pandas and numpy libraries for Python.

After data cleaning is performed, the data is loaded into the Tableau application, which is a visualization tool for implementing data visualization on various attributes on our cleaned dataset. Using these visualizations, for every correlated attribute, I build a dashboard that consists of pictographic representations.

**4. Organization**

**4.1 Project Group**

It is an individual project performed by Jitendhar Reddy Adulla.

**4.2 Customer**

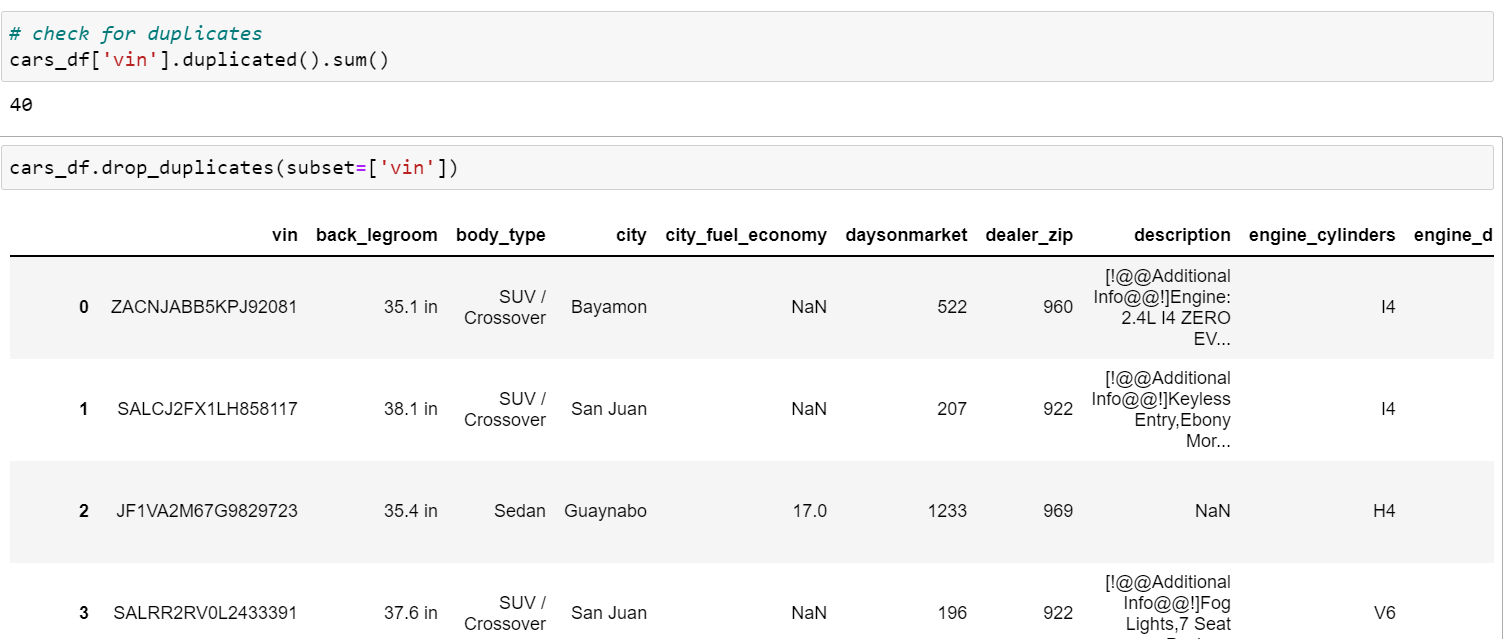
The targeted customers are listed below

1. Car Buyers
2. Car Sellers
3. Car Dealers

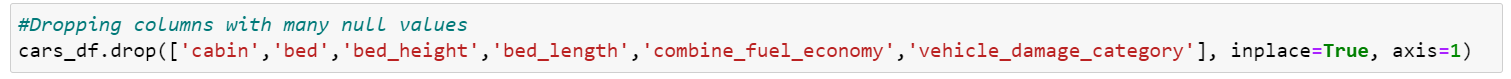
**5. Development Process**

I had chosen this dataset for the project, since it consisted of various attributes on the used cars data. The selection of about 3 million cars in the data gives us a well rounded information about the car usage, and the attributes like is\_cpo and is\_oemcpo determine the car’s sale quality as per the seller, which make them easier to classify for their sale.

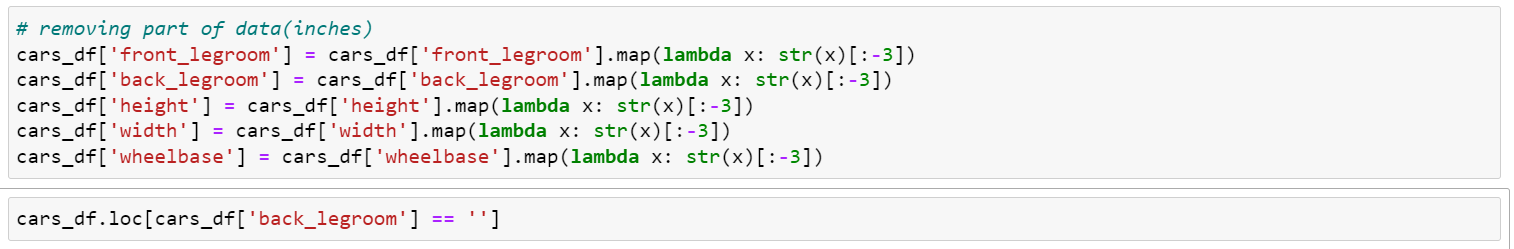
The data had been collected and cleaned, by normalizing some of the values, like making “vin” to be unique thereby evading any duplication of it. I found that there were around 40 duplicate values in the VIN attribute and I dropped all the duplicates using the drop\_duplicates function.



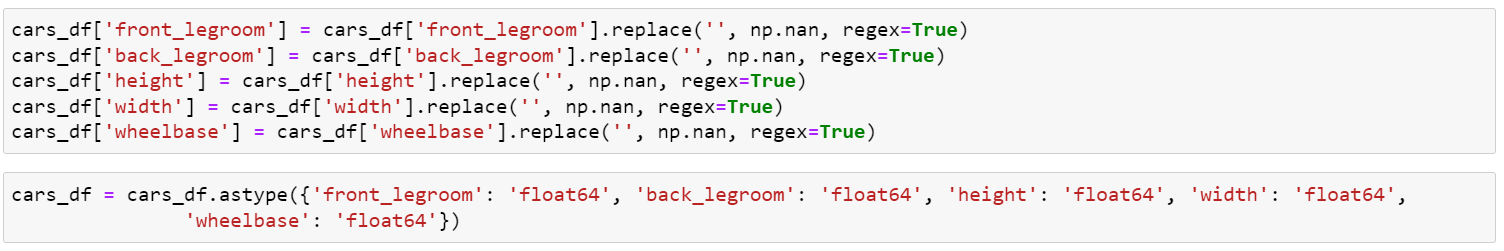
There were many null values in the columns like cabin,bed,bed\_height,etc. I dropped all these columns using the drop() function as they may impact our analysis because of the null values.



The values of the dimensions are reduced on their significant digits, to only include 3 digits. This helps in reducing the load of numerical calculation on the visualization.



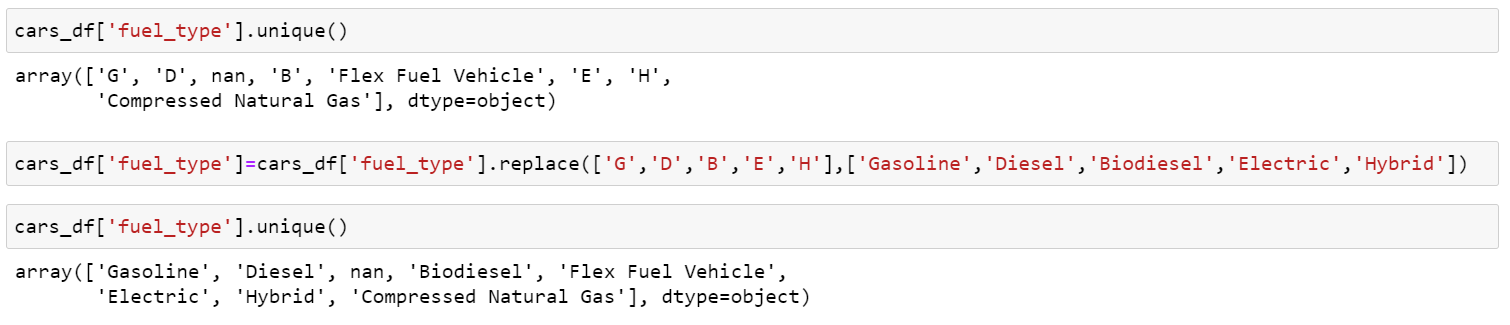
I then ended up replacing the missing dimensions with NaN instead of the blank string, which evaded the problem with empty values. I converted them to the float64 type from string, to perform better analysis over the numerical data.



The city fuel economy attribute has some null values in some data points. The average city fuel economy from the existent values is used to replace the null values in the dataset, thereby normalizing the attribute.

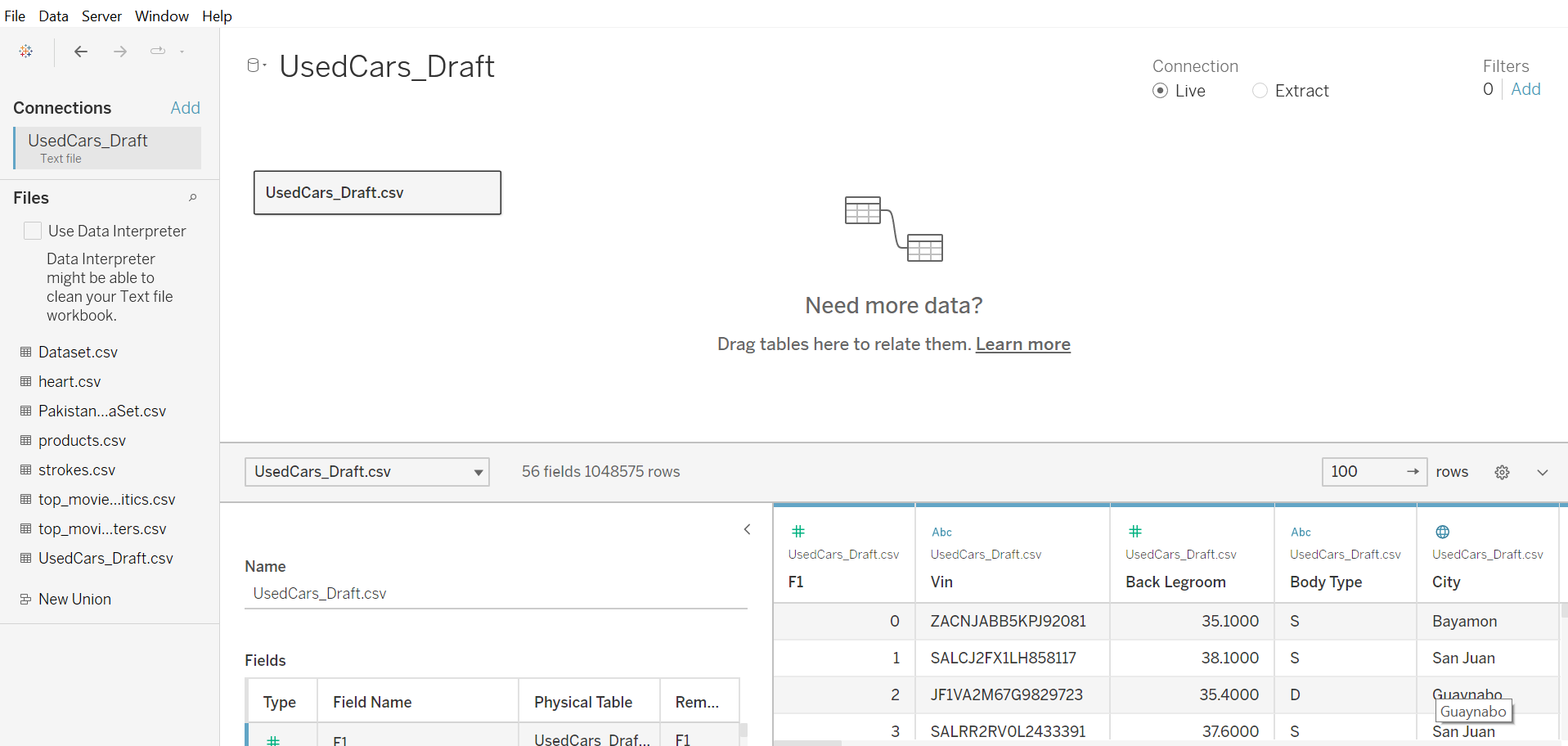


The fuel types were initially of mixed values, including separate character and string values. This inconsistency was handled by replacing the character values with the respective string values.



After all the cleaning is done , it's time to perform the analysis on the cleaned dataset.

Visualizations were performed using the Tableau application, which is a publicly available source for visualization. The following shows how the file looks when imported into the tableau .



The main risk concerned about this project is dataset size . As the dataset was around 10 gigabytes , it was taking longtime to run the queries in the jupyter notebook. Then I decided to perform part of the cleaning and save it into one file and later performed the remaining cleaning on the saved file which was cleaned earlier.

**6. Deliverables**

The major deliverables of this project are visualizations and dashboards showing the analysis of the car sales. There are other deliverables like abstract, code for data loading and cleaning, project presentation and report.

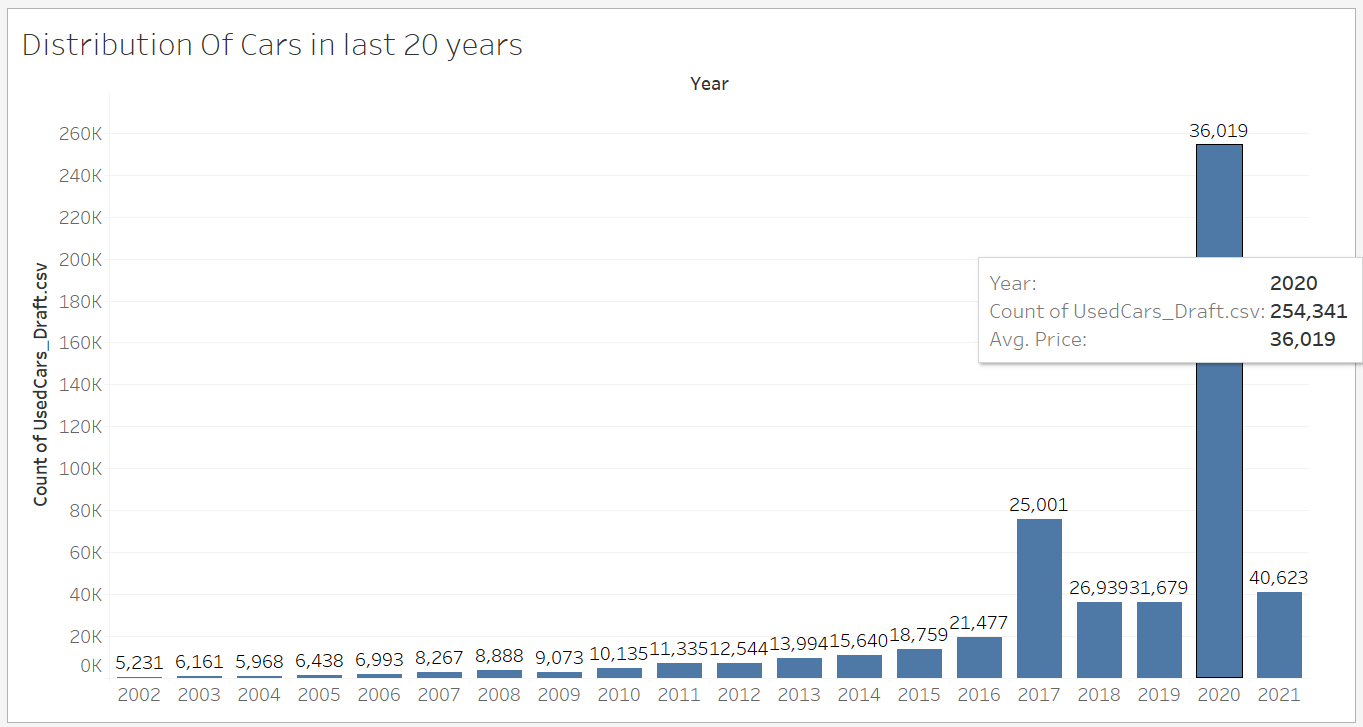
**7. Github**

All source code and finished documentation will be uploaded to Github repository.

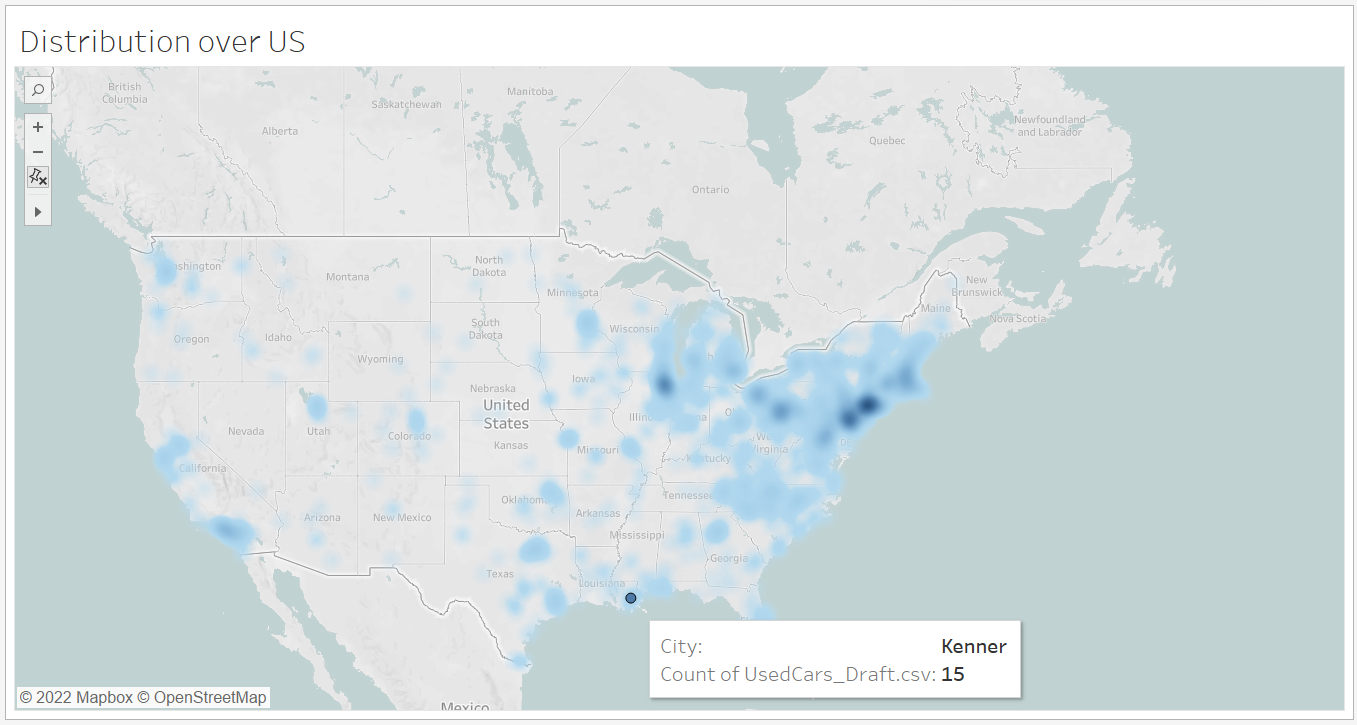
Link - <https://github.com/Jitendhar8384/Data-Visualization>

**8. Visualizations**

Various visualizations are performed on the dataset to show the analysis of the cars data between 1915 and 2021.



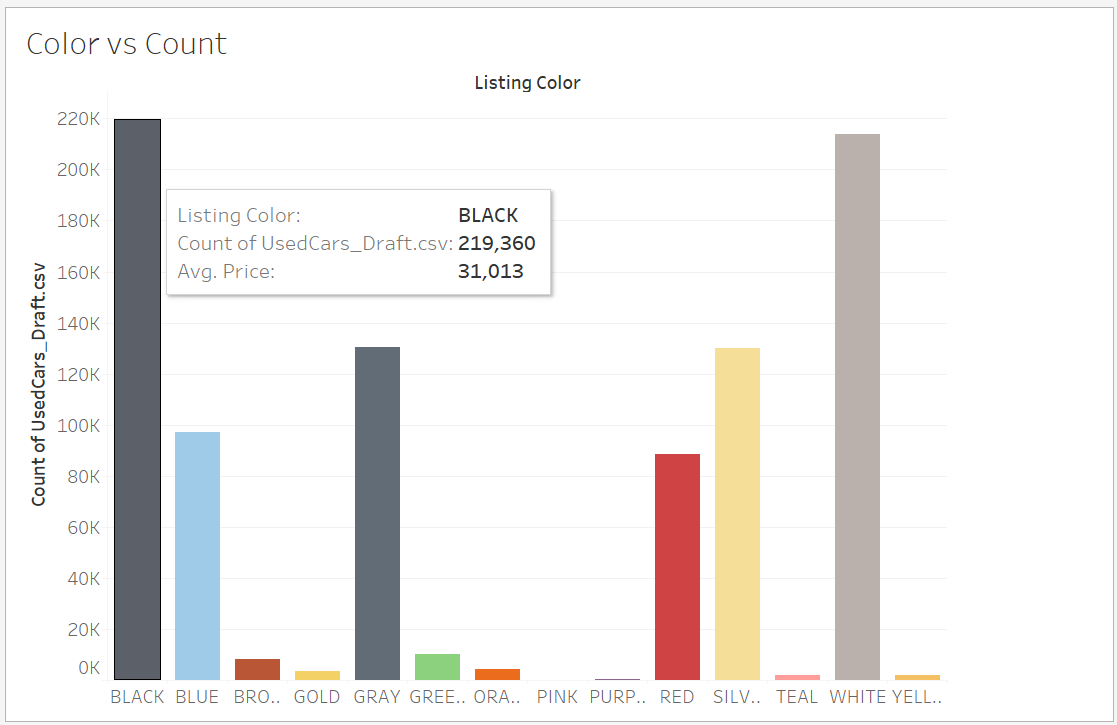
This visualization describes the distribution of the number of used cars sold each year. We can see that the number of cars sold in 2020 were very high due to the pandemic and the financial crisis thereafter. We can also see the average price of cars sold in the respective year with labels on the top of the bars.



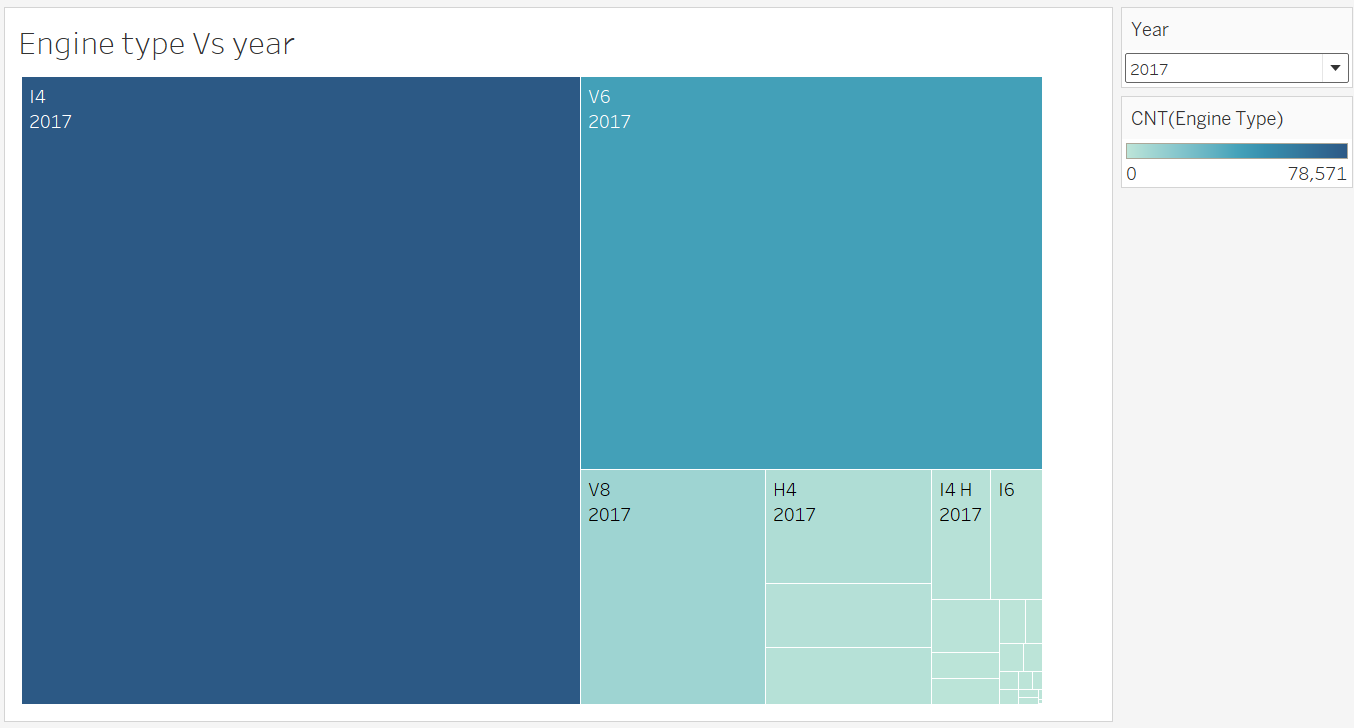
This map above talks about the places where most used cars were sold. There’s a high correlation with the number of people and the number of cars sold, as the map concentrates over the east coast and major cities have a strong highlighted presence. When we hover through the map , we can see the city along with the count of cars in that city sold so far.



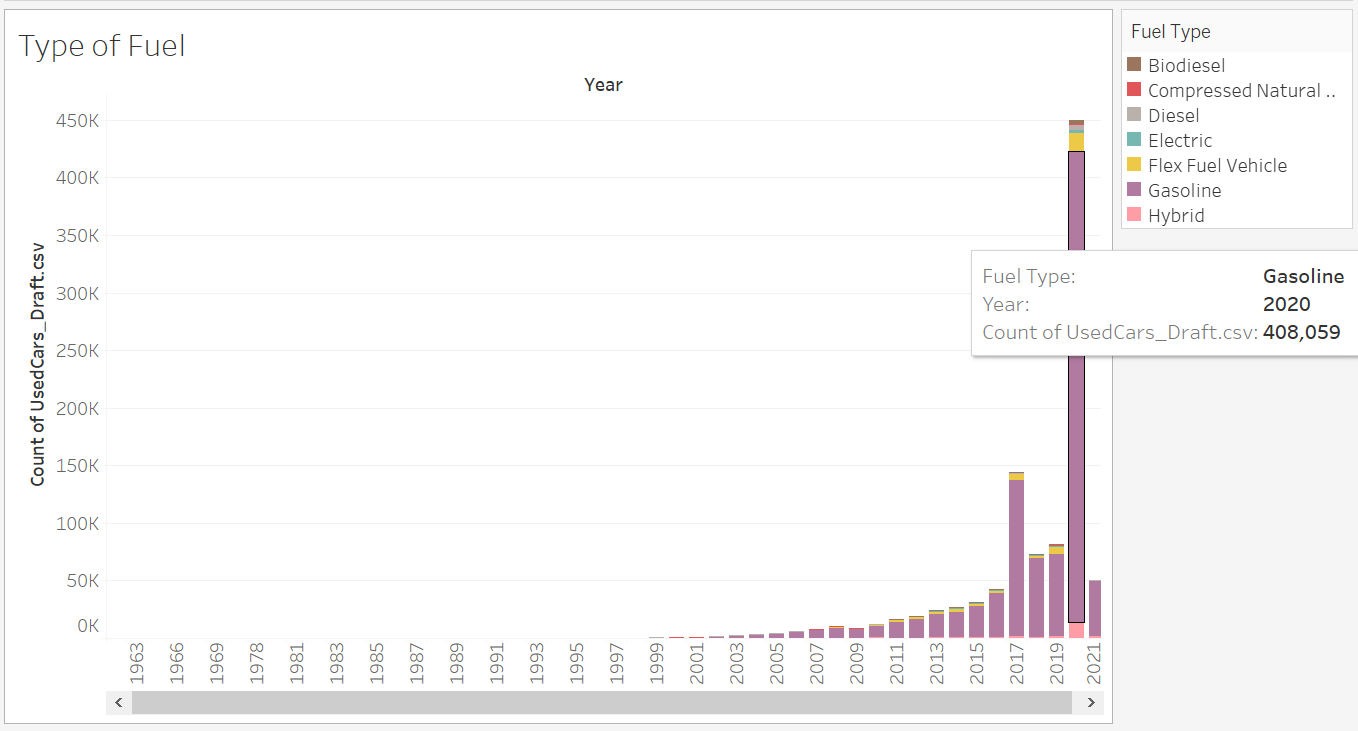
The Price versus Fuel Type estimate here gives us the size representation of the count of these used cars dictated by the seller for the year 2011. We can select the year by using the dropdown option on the right side to check the fuel type data of that particular year . As seen, electric has the lowest prices due to the fact that batteries deteriorate after the first owner, and make up to half of their initial prices. Biodiesel has the highest since the alternative fuels are seeing a rise nowadays. Hybrid stands midway between gasoline and electric cars, while flex fuel vehicles are also with a high price due to their ethanol blend compatibility. This stands out as the state-of-art latest technology that has a high demanding value.



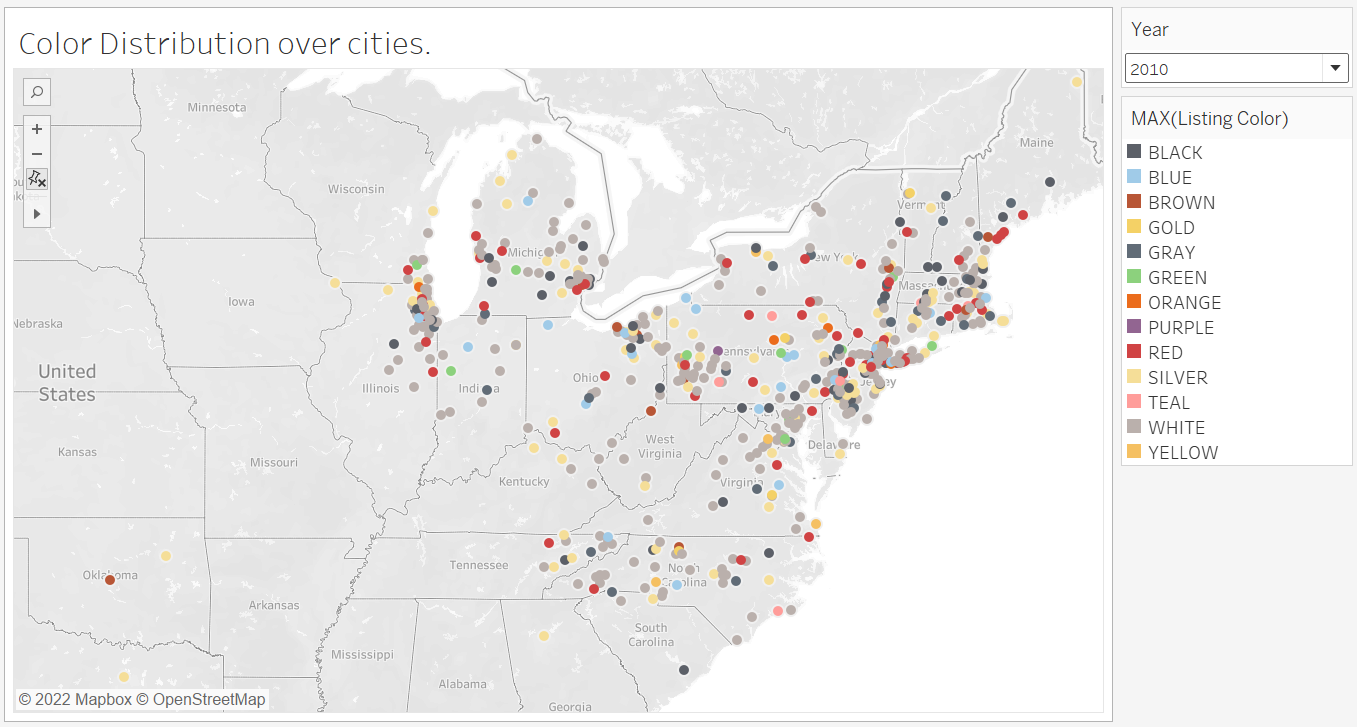
This bar graph shows the cars found of each distinct color in the used cars. As expected, most cars were of white or black in color, whereas the next highest were gray, silver, blue and red. When we hover through the colors , we can find the count and average price of the respective colors.



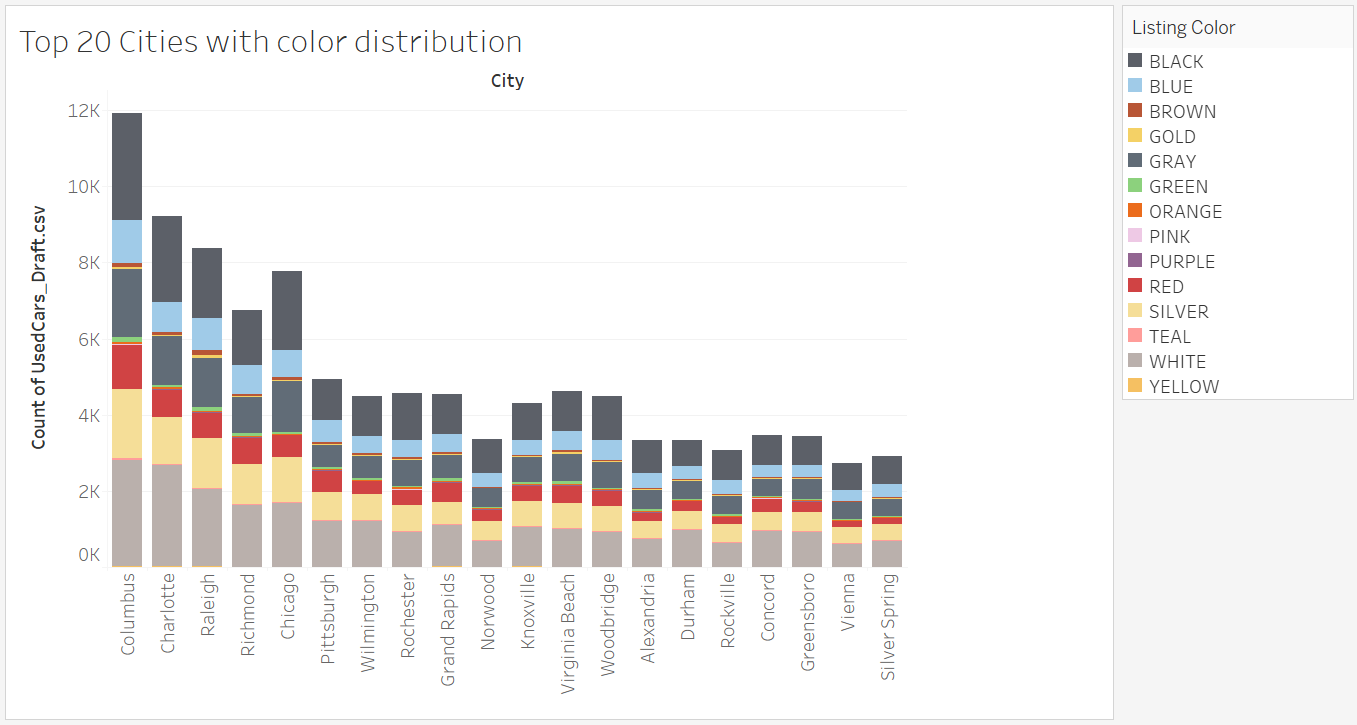
The proportional representation of engine types and years are given here, with almost half of the number of cars carrying the I4 engine made in 2017, while another 30% of the cars having V6 engines made in 2017. The next popular ones include V8 from 2017, H4 of 2017 and I4-H of 2017, which are placed here on the proportion of their market share in the used cars available. Similarly , we can find the data of other years with the help of the drop down menu on the right.



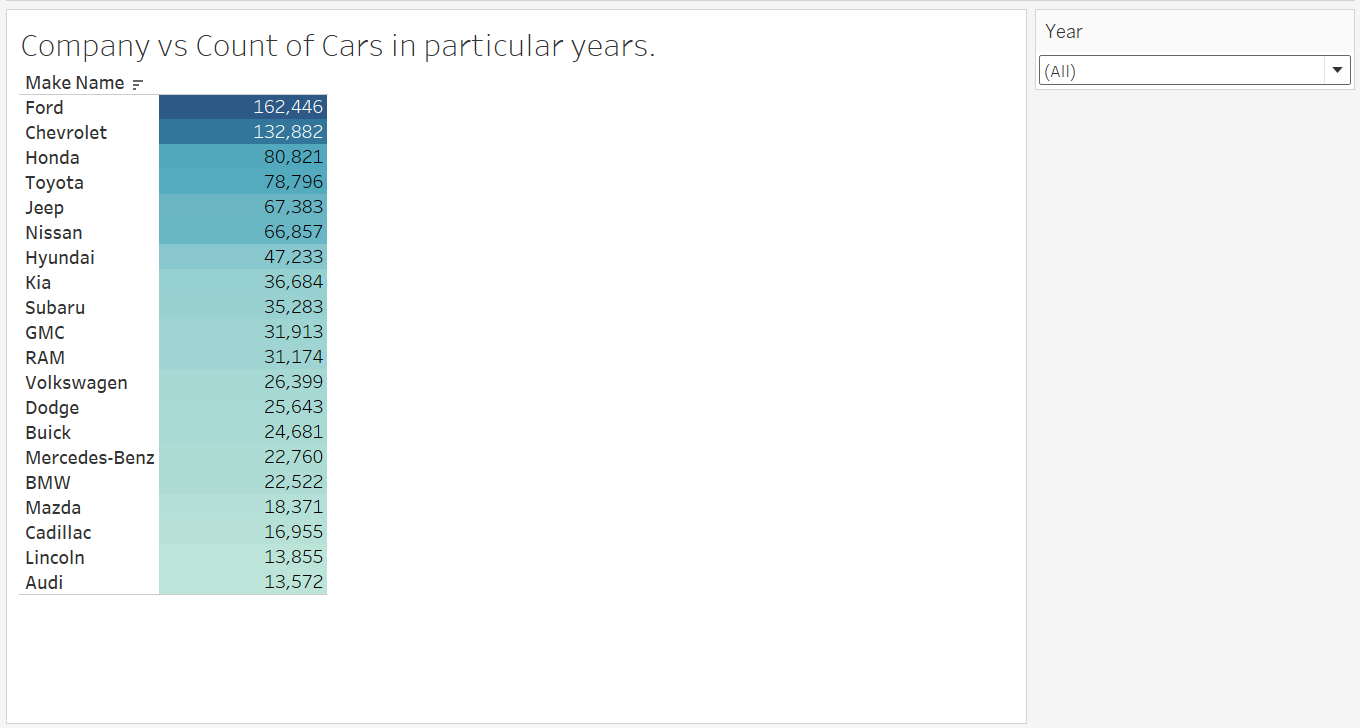
This histogram is of the type of fuels each engine uses. This chart above has features to show the fuel type, count of cars using that fuel and the year these cars were sold, when hovered over the histogram. This has a direct correlation with the total number of cars sold each year. The cars sold are mostly of the gasoline type of fuel, however there have been many Hybrid and Electric cars sold in 2020 due to their complete halt of usage in the lockdown. Diesel cars tend to be the least ones sold every year, and even flex fuel cars were sold a lot in 2020.



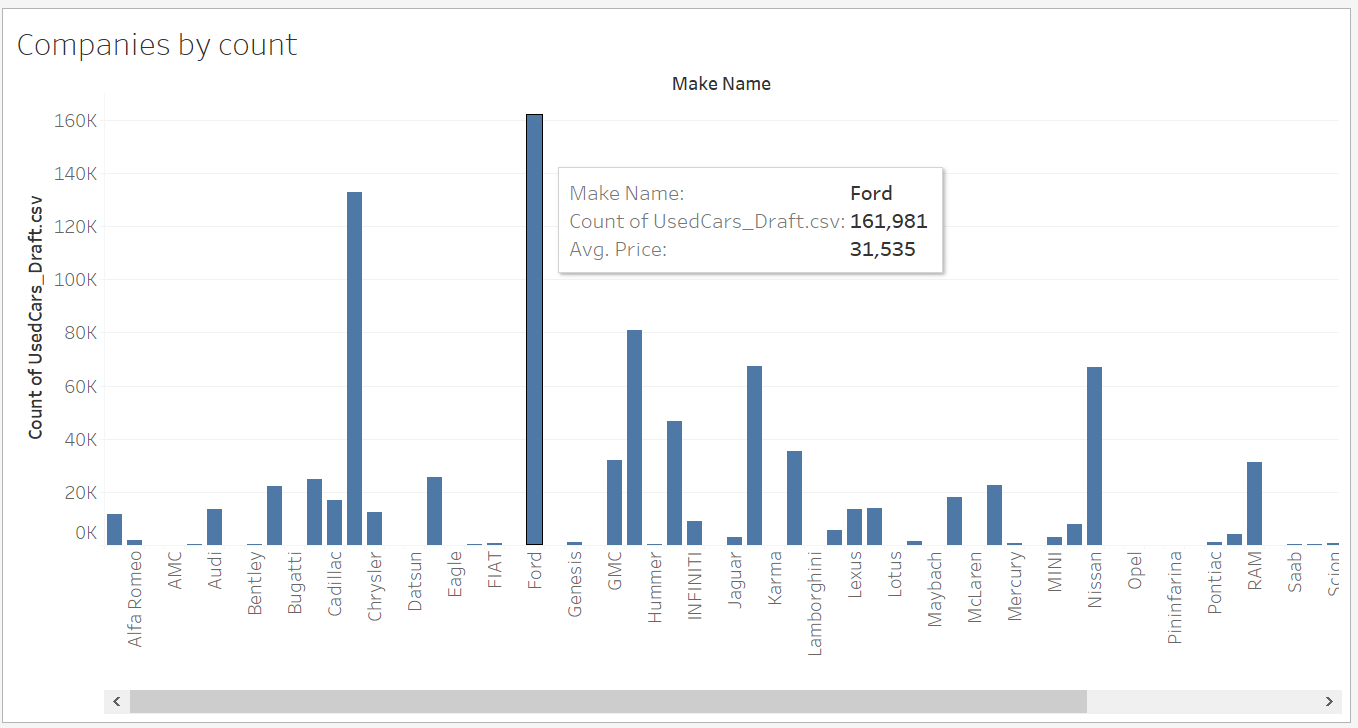
The color distribution of the majority of the cars in the cities has been mapped in the above map. Most densely populated cities have cars in black and white colors, while the rest of the colors are equally present in each one of them. Each city displays the majority color of the cars sold.



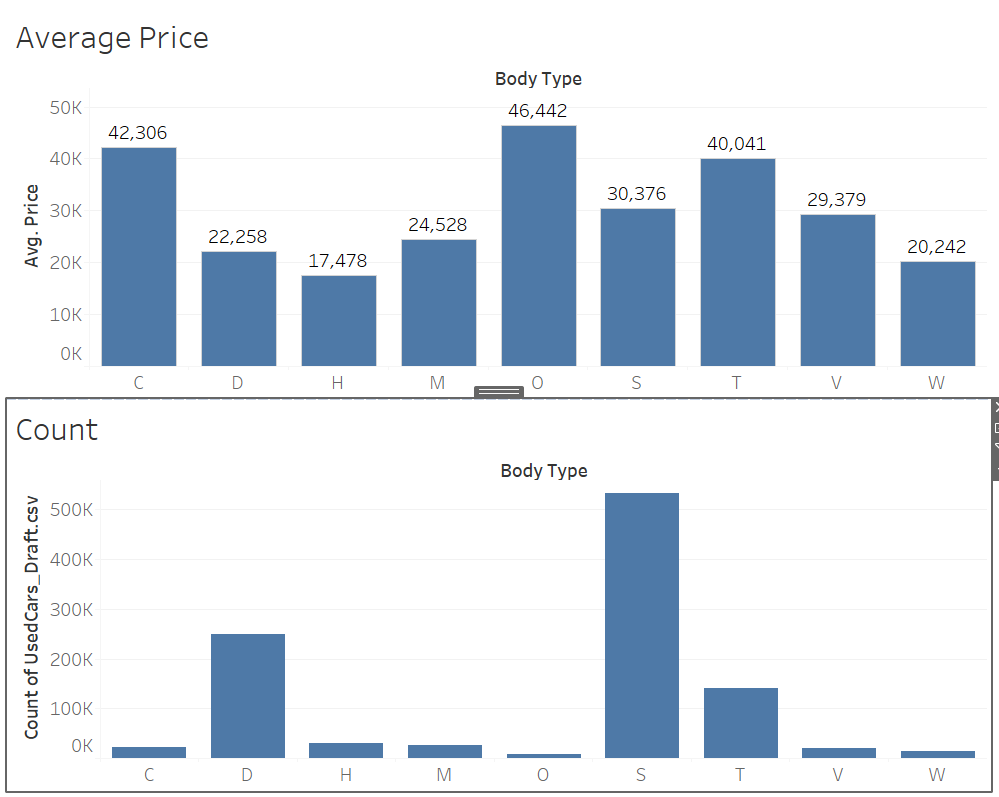
The above histogram shows the top 20 cities with the most equal color distribution. As such, most of the cities have black and white covering a quarter each of cars sold in those cities, while the other major colors like gray, brown,silver cover the rest.



This table shows the most cars sold of the top 20 companies in a particular year. We can choose the year on the right, and the total cars in the used vehicles are found on the left. We can see that Ford and Chevrolet lead the total number of cars made by a huge margin to their next competitors Honda, Toyota, Jeep and Nissan.



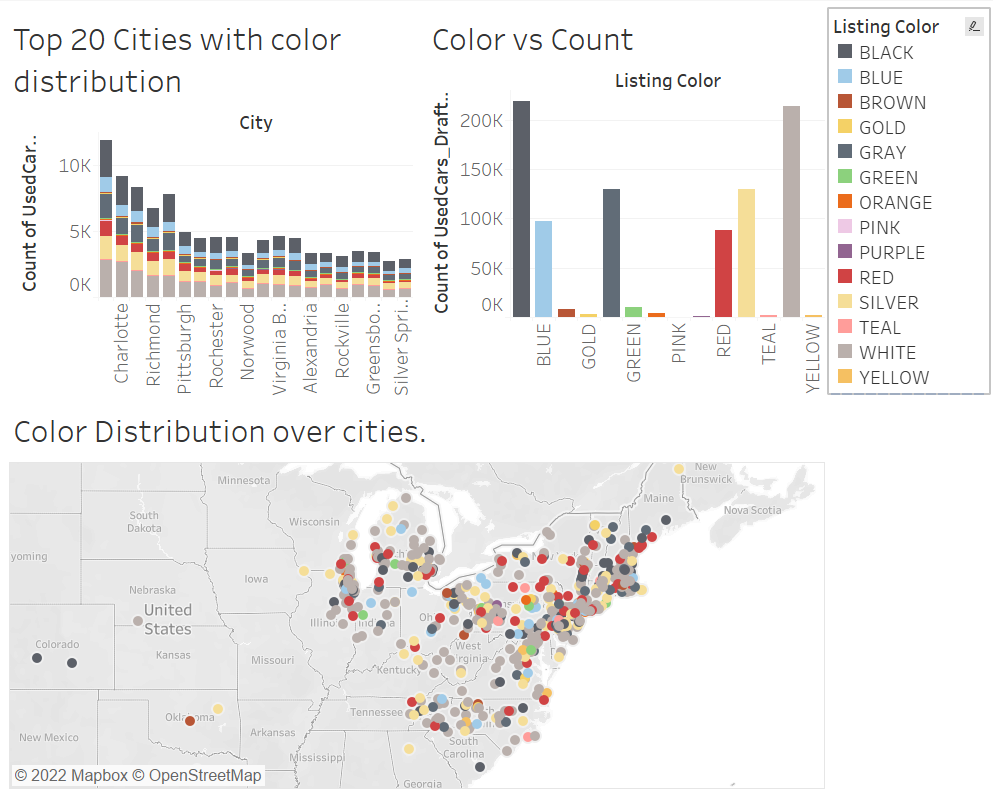
This histogram shows the same data as in the previous diagram, but for all the cars. Count of the cars sold by each company are on the y-axis, while the company names are on the x-axis. When we hover through the companies , we can see the exact count and average price of the cars sold by that company.



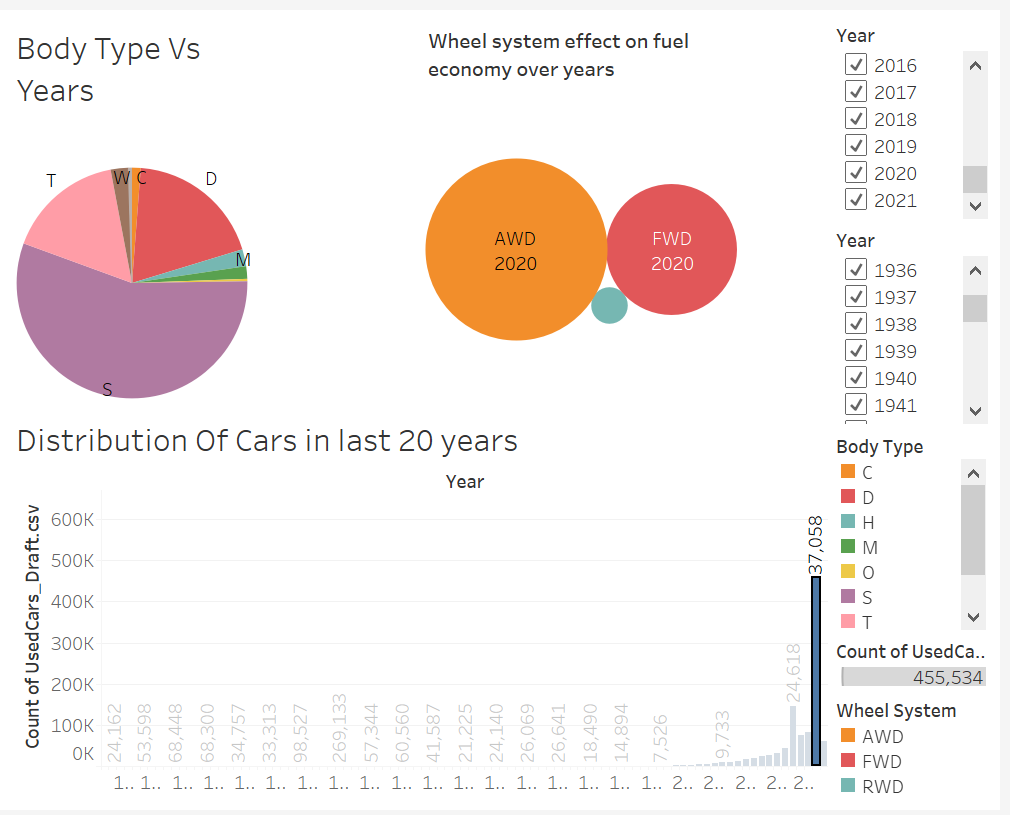
These histograms display the count and average price of the cars with respect to their body types. Here, C stands for coupe, D for Sedan, H for Hatchback, M for Minivan, O for Convertible, S for SUV/Crossover, T for Truck, V for Van and W for Wagon. We can see that convertible cars are having the highest average price and hatchbacks are having the lowest. SUVs are having the highest count while convertible cars are having least count.

**9. Dashboard**

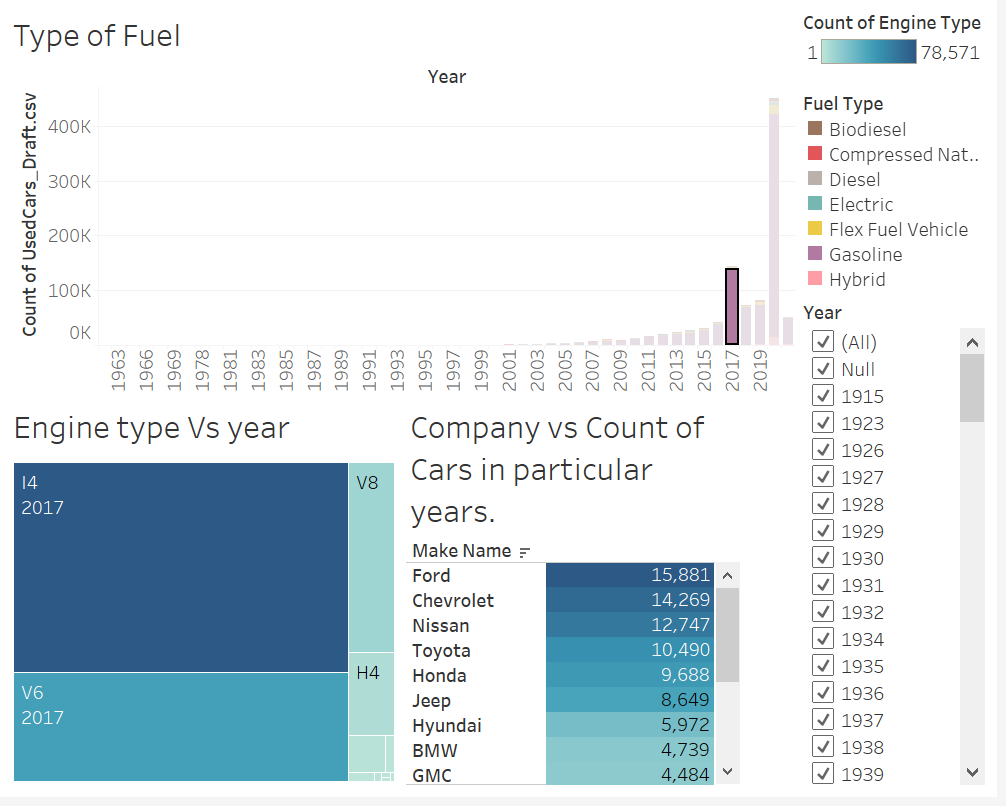
I had built various dashboards with respect to the different features of a car. These dashboards are further explained in the detail,



The dashboard above consists of the color distribution diagrams, with the top left being cities of color distribution, top right being colors versus the total count. The bottom one is the mapped distribution of what color of cars are present in each city. The top left and bottom are thus directly correlated to the top right chart. When each color is hovered in the first graph, the count is given and where the color is majority is also identified.



This given dashboard is given for the body type versus the year metrics. When the years are selected in the bottom chart, the count of cars are given as a pie chart of their body types, which is placed on top left. On top right, the cars using the wheelsystem with respect to their count and year are displayed. This is thus correlated with the total distribution of used cars in each year.



The dashboard given here correlates the type of fuel, type of engine and count of companies per selected year. The fuel type of a particular year can be selected on the top chart, which displays the engine type and top 20 companies with respect to count.

We can see that the cars using gasoline are mostly from the i4 engine type in the year 2017 and ford company shows the highest sales.

**10. Project Plan**

All the steps in the project were performed and completed on the scheduled date as shown below

|  |  |
| --- | --- |
| **Milestone** | **Submission Date** |
| Project Proposal | 02/20/22 |
| Data Collection | 03/01/22 |
| Data Cleaning | 03/10/22 |
| Data Exploration | 03/15/22 |
| Data Visualization | 04/01/22 |
| Dashboard Building | 04/11/22 |
| Presentation | 05/10/22 |
| Report Writing | 05/15/22 |
| Project Submission | 05/22/22 |

**11. References**

https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset?select=used\_cars\_data.csv

https://www.kaggle.com/code/monika171/u-s-used-car-price-prediction-project-in-progress

https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho