***Managing With Analytics-IPM 652***

**Project: Craigslist Used Cars**

Lily Sunil Karmakar

**Executive Summary**

Generating recommendations for consumer to consumer used car websites using a craigslist used cars dataset which was captured data in two-month intervals from October 2018 to March 2020 and includes every used vehicle entry within the United States on Craigslist in that time period.

**Dataset Cleansing**

We removed cars that have - more than 100,000 miles, are manufactured before the year 2000, have titles other than clean, rebuilt, lien, or salvage.

**Findings**

We can see most of the listings are in the Eastern half of the US, and we see a very dense aggregation of listings on both coasts as well as in Illinois, Indiana, and Ohio. The following states have more than 1 million listings: CA, CO, MO, MD, MN, OH, WA, OR, and MI.

**Ethical Considerations**

Ethical considerations to investigate before making our recommendations would be - discriminating against manufacturers and geographical locations. The top 5 manufacturers are either American or Japanese companies, as well as the best places to sell cars are located on the coasts. Making decisions focused to this demographic will leave out a majority.

**Recommendations**

It would be wise to work with dealerships or mechanics that specialize in inspecting domestic and Japanese cars. For operations, we have identified 7 areas to establish warehouses to efficiently and effectively transport cars. In addition, we have generated both online and offline advertising recommendations to get the most desirable cars to the right consumers.

**Introduction**

We are analyzing a craigslist used car dataset which contains all the relevant information that is shared when an individual post a car for sale on craigslist. The dataset contains the following information: manufacturer, type, transmission, color, odometer, size, price, state, and condition

**Recommendations for**

There are many potential groups of consumers that could gain valuable insight from the information that is derived from this dataset, but we feel that where it can be most impactful is for consumer to consumer used car websites like Carvana. Our goal of this analysis is to identify, areas and car types that we can recommend to websites like these for them to better target and advertise their product offerings.

**Popularity**

The first thing we are addressing is the popularity of these data points across the United States, we feel that if we can understand what types of cars are being put on the market and where, we will be able to provide recommendations on what cars you can expect.

**Price Point**

Understanding what the average sale price is based on condition, odometer, and car builds will put into perspective what the return on investment may be for these online retailers. Marketing for the ideal car for certain locations will be more worth it for a car that will sell quickly and for a higher price.

**Map**

As we know there are a lot of different preferences for people all over the country, understanding that a pickup truck might be more likely to be purchased in Texas then in Massachusetts would be valuable information for those who are selecting how much to advertise on car types across the country.

**DATA**

The data we have analyzed in this project is a compilation of used cars data from Craigslist listings. The data was scraped from craigslist by a Kaggle user[[1]](#footnote-1) in two-month intervals from October 2018 to March 2020 and includes every used vehicle entry within the united states on Craigslist in that time period.

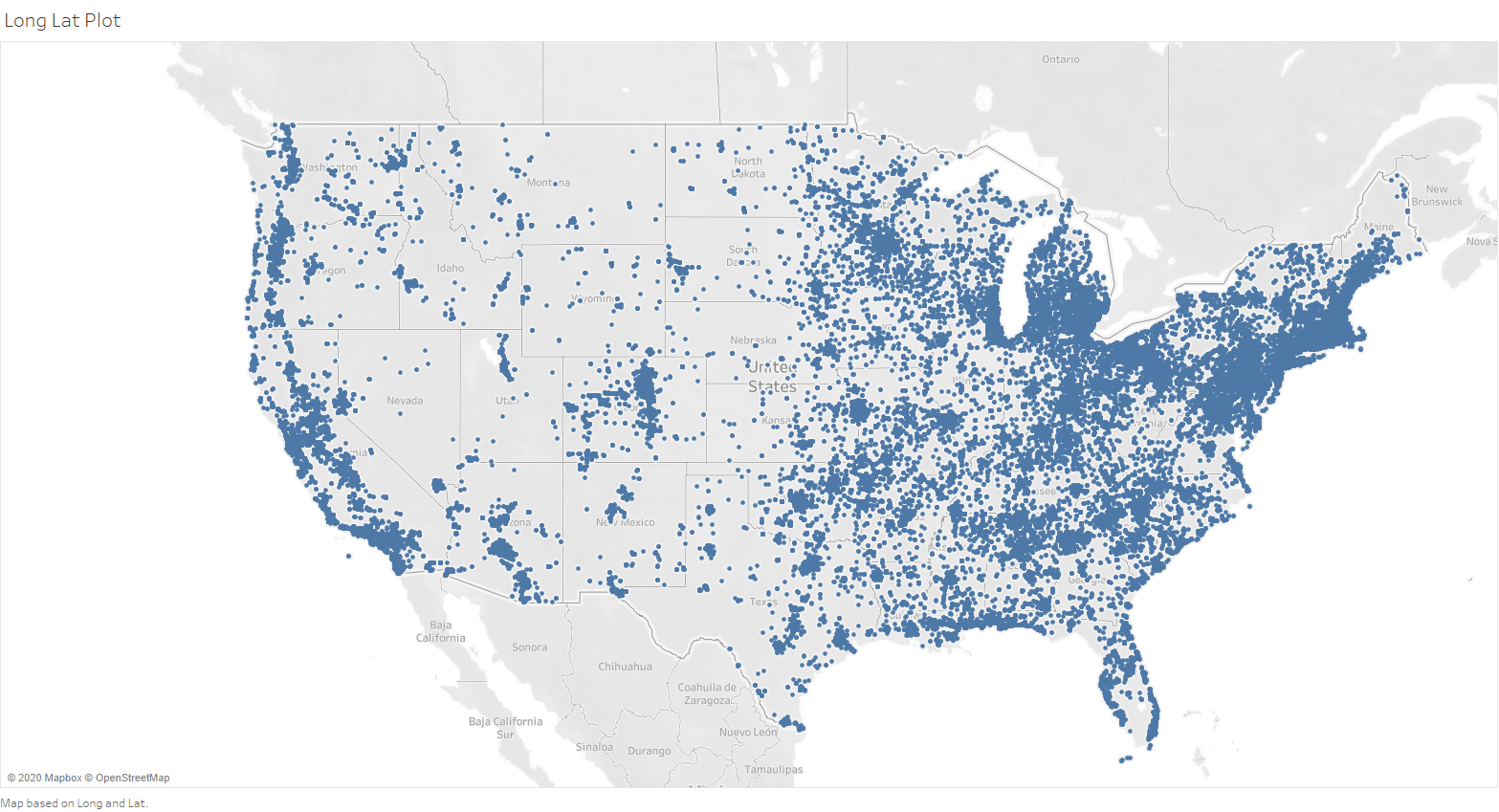
The data was scraped from only the used cars section of craigslist, as the website allows listing of both used and new cars. Any car listed, which was manufactured prior to 2019, and which had an odometer reading of 0 miles was also considered a used car for this project.

The dataset contains around 25 columns of descriptive information about the listings such as –Listing price, Year of production, Odometer reading of the vehicle, region of listing (with latitude and longitude), size of vehicle, condition of vehicle, model, manufacturer, and title status – to name a few. Not so useful columns such as Craigslist listing ID, vehicle VIN numbers and vehicle listing descriptions were eliminated to clean up the dataset and help in analyzing it better.

Some of the key dataset characteristics are further categorized into sub-groups to help aid in filtering out the dataset. Such as, the number of cylinders in a vehicle, the drive of the vehicle, type of vehicle (coupe, truck, sedan), and size (compact, mid-size, etc.). This depth in the dataset helped in finding trends and demands in different part of the country.

Since the recommendations of our analysis is geared towards online car reseller Carvana, which does not accept cars above 100,000 miles and cars manufactured before the year 2000 among other things, the dataset had to be cleansed to filter out those listings which did not fall into the Carvana listing criteria.

Pictured below is a heat map of all the listings of the dataset post-cleansing



As it can be observed, a considerable number of listings in the dataset are in the Northeast, Midwest and Southwest regions.

**Methodology**

As this data was directed pulled from Craigslist, the original data is extremely dirty, extremely comprehensive and extremely large at 1.1 GB with 10 million listings. We applied some data filtering and cleaning to get rid of the dimensions that we do not want as well as listings that Businesses like Carvana that would not be interested in. We did not want to clean the data too heavily, because cleaning data would essentially degrade the integrity of the data. Since we are not doing any regression analysis, we could make some compromise on the data cleanliness side.

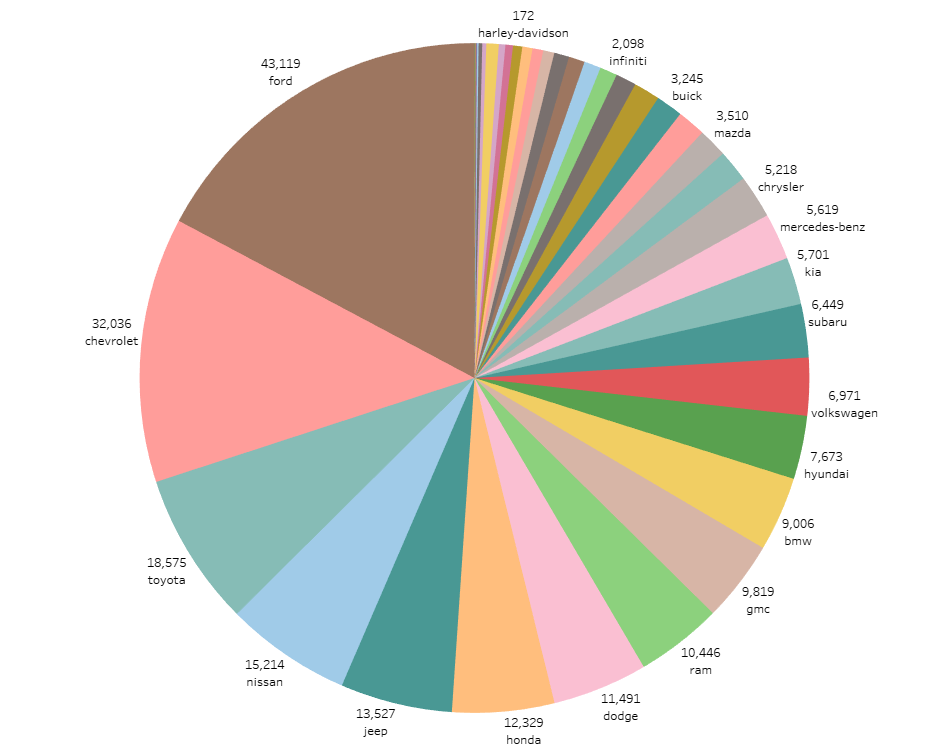
The first thing we did is to decrease the dimensions of the dataset. The original dataset includes dimensions like Region URL and listing URL, which did not aid our analysis process. Therefore, we had to delete those dimensions out of the dataset. This significantly reduced the size of the dataset to around 100 MB.

The next task was to delete any listing that is not complete. We filtered out all cases from Canada, because Carvana does not operate in Canada. We also conducted market research and eliminated the cars that is not suitable to sale on Carvana’s website, which eventually reduced the dataset to 40MB with 5 million listings.

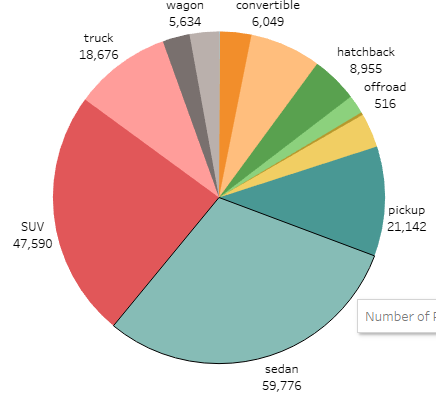
**FINDINGS**

**Popularity**

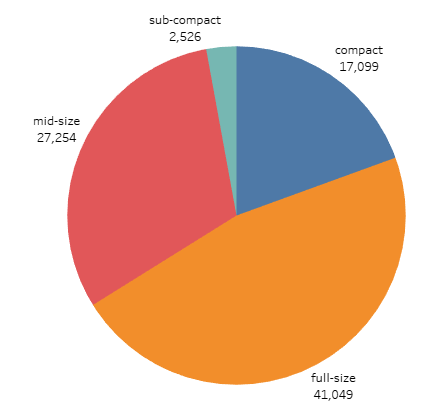
* **Manufacturer**- Ford, Chevy, and Toyota make up nearly 40% of car listings on Craigslist.



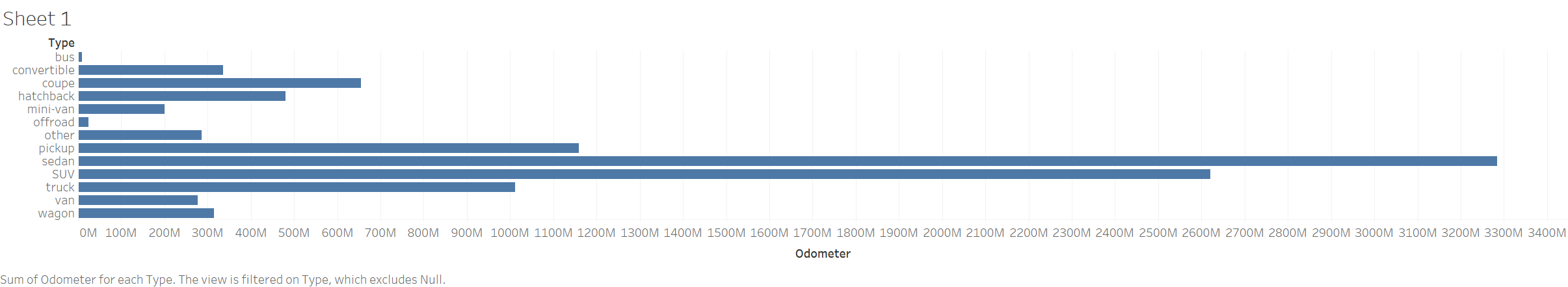
* **Type**



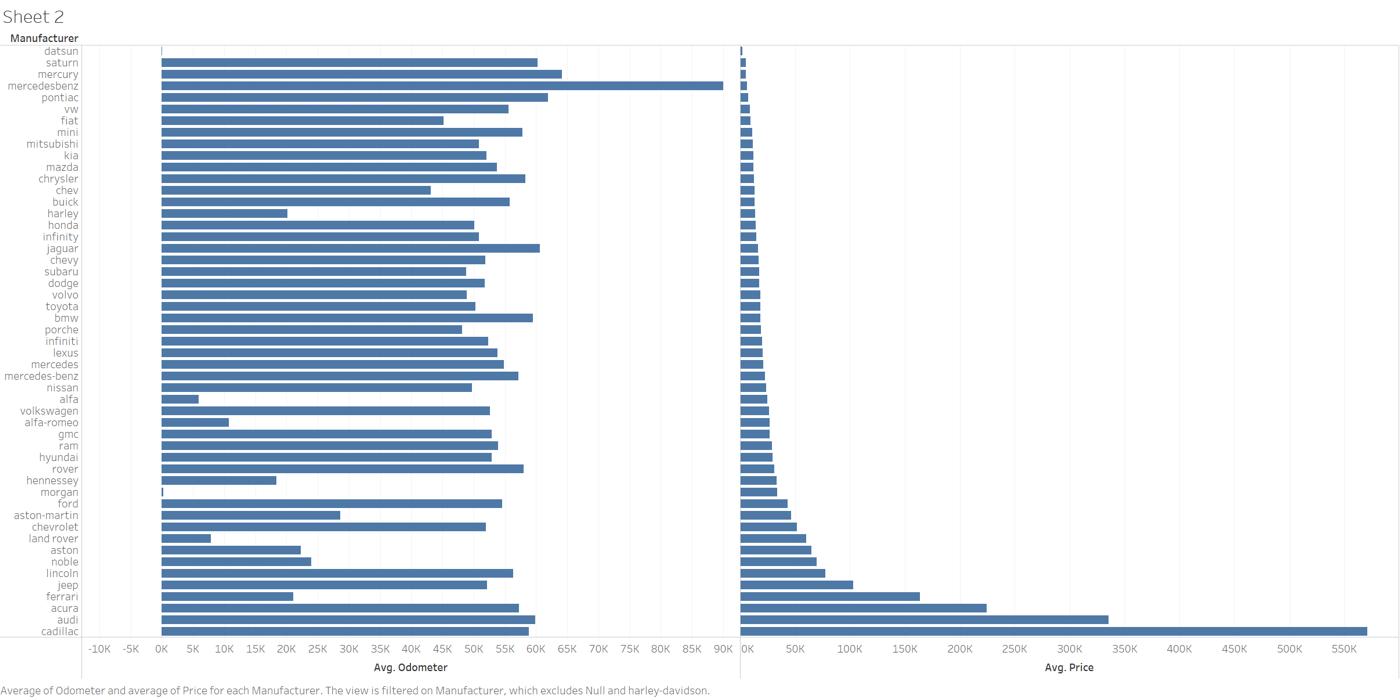
* **Size**



Odometer for each type of vehicle

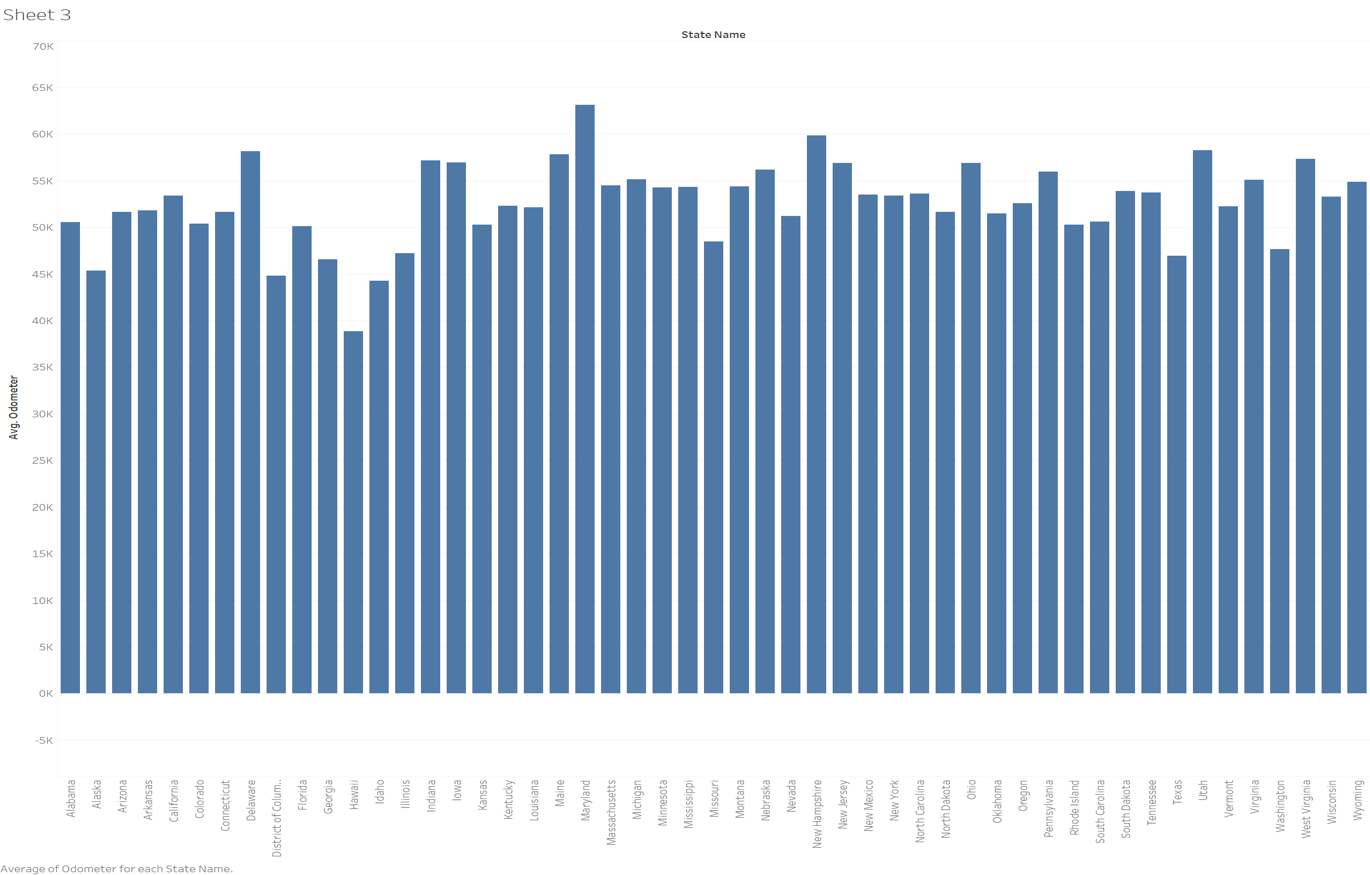


**Odometer by manufacturer**



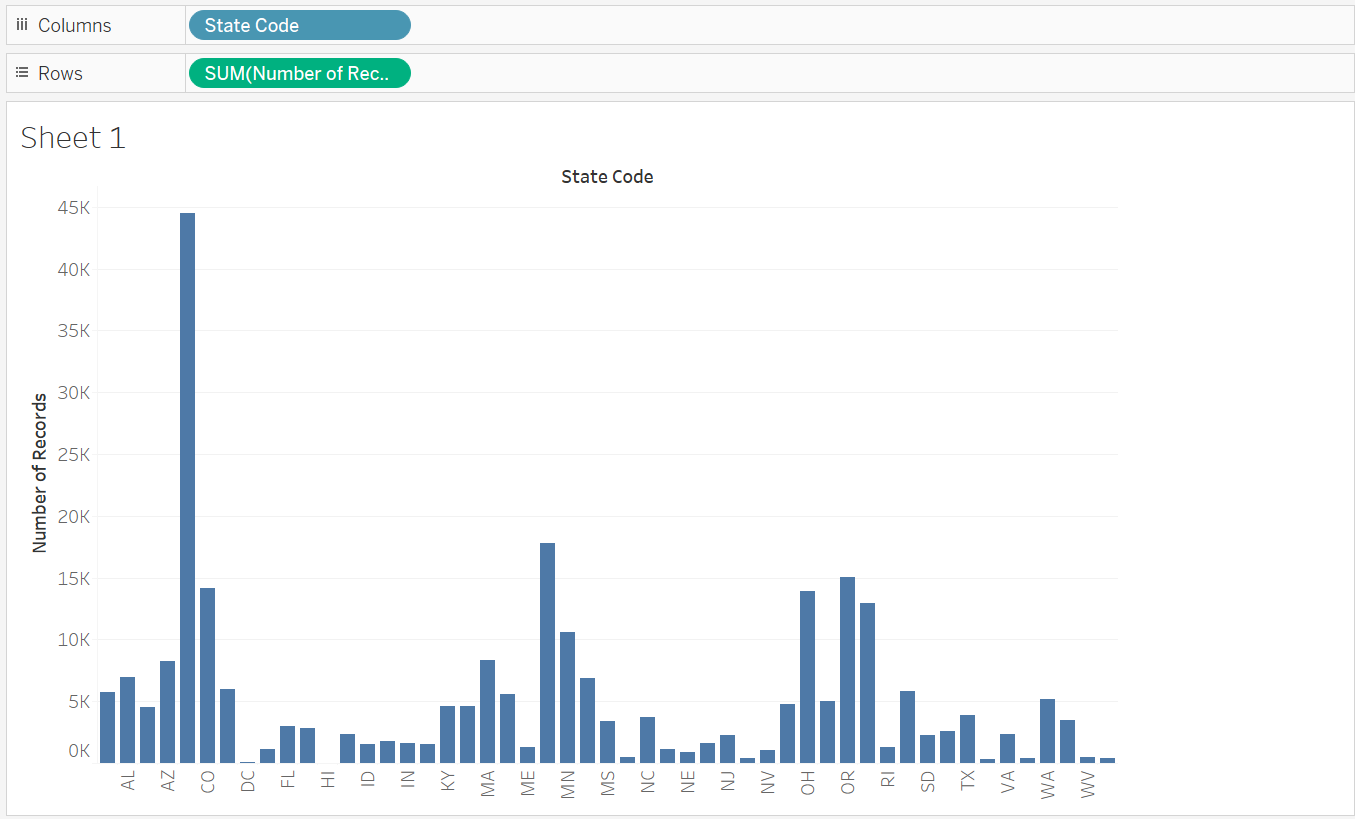
People are more probable to list brands like Acura, Audi and Cadillac with higher odometer readings at a premium price.

**Average Odometer readings in each state**

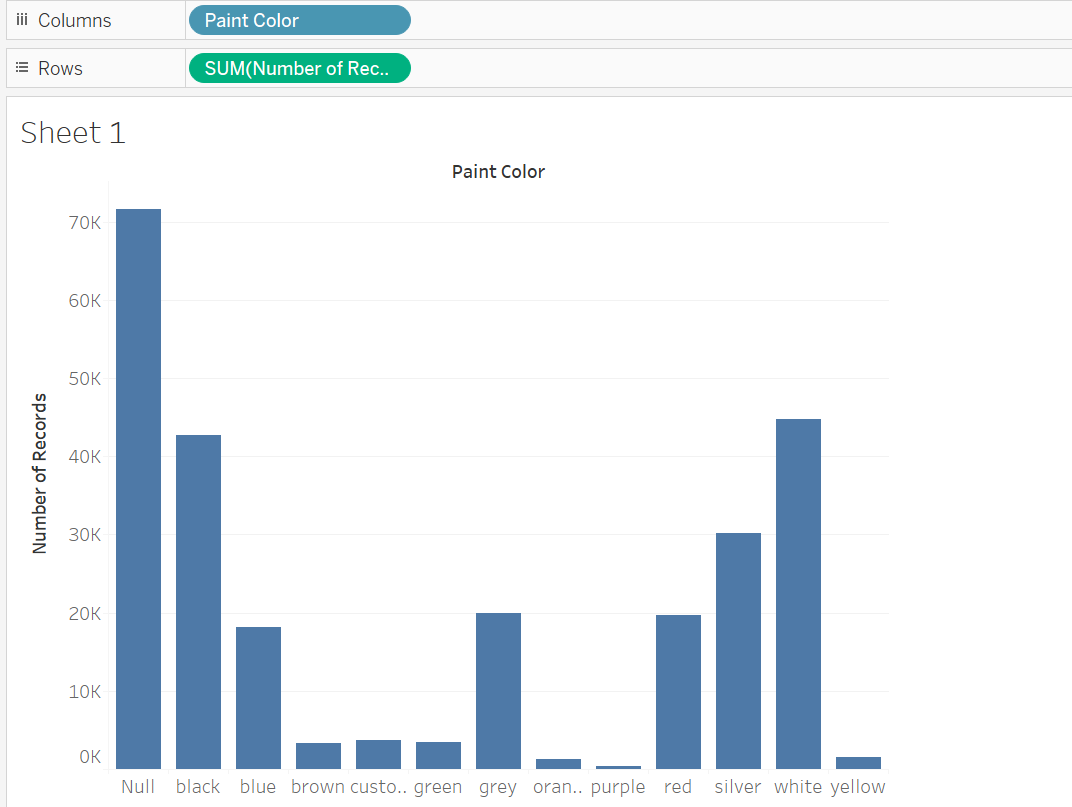


Maryland, New Hampshire, Utah and Delaware have the highest average odometer readings for cars bought post – 2000.

**State**



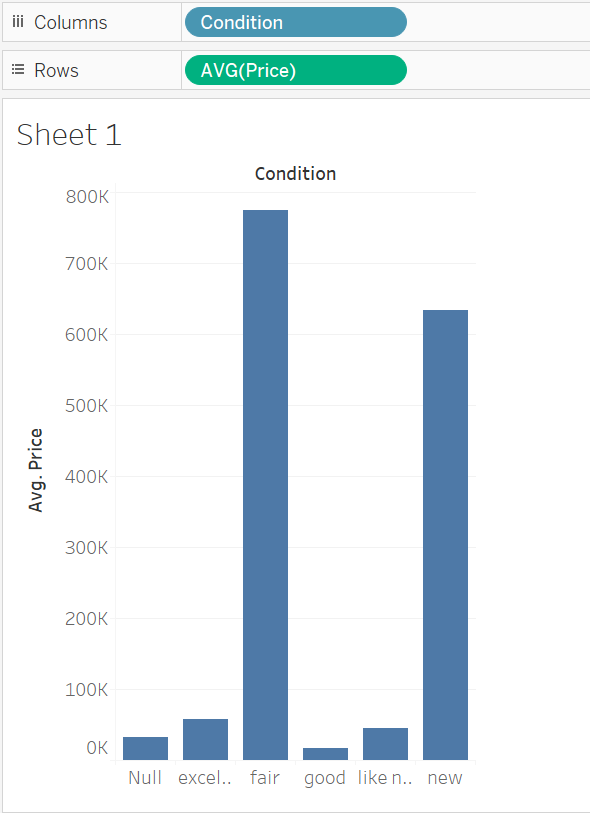
**Color**



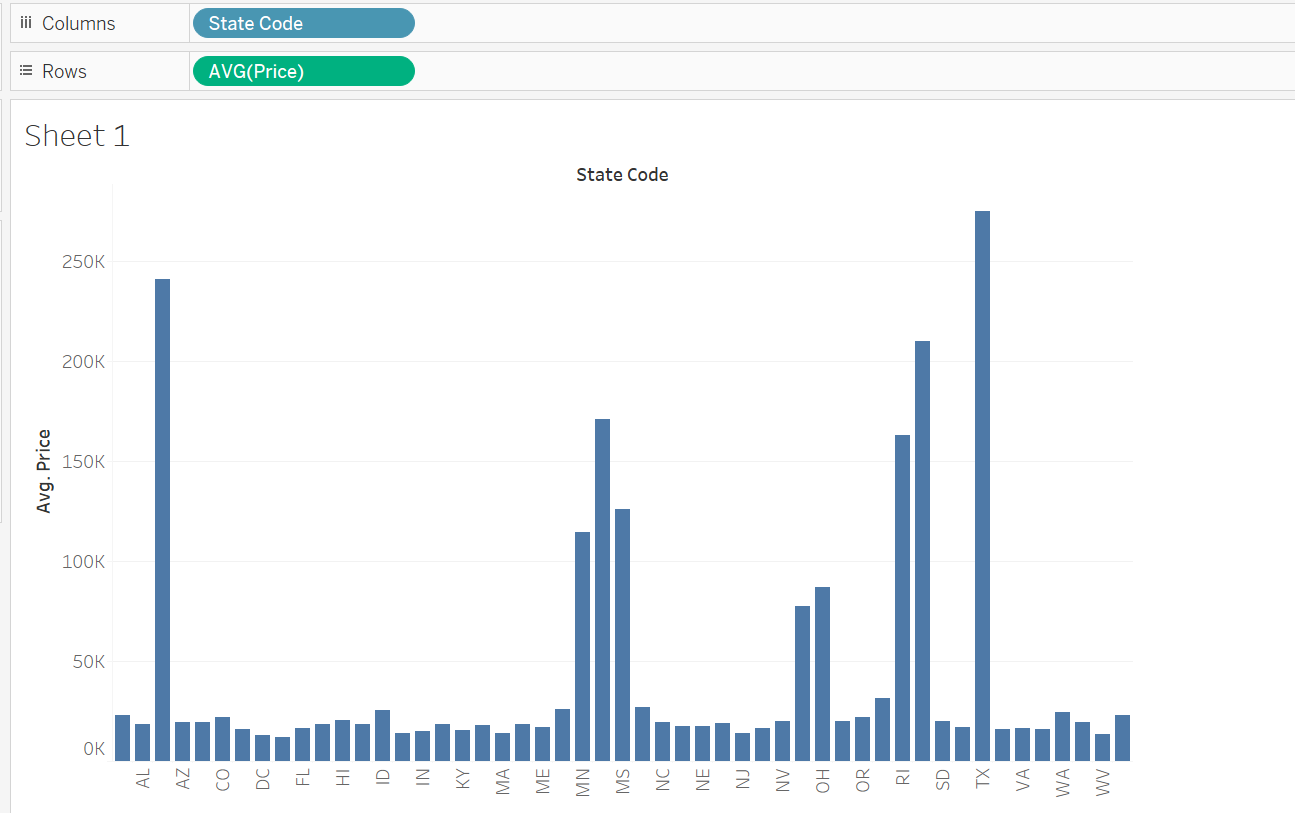
According to this graph, black, white and silver cars are the most popular

**Price correlation**

* Price/ Condition
* This graph shows that cars that are in better condition sell for higher prices than cars that are not in as good condition

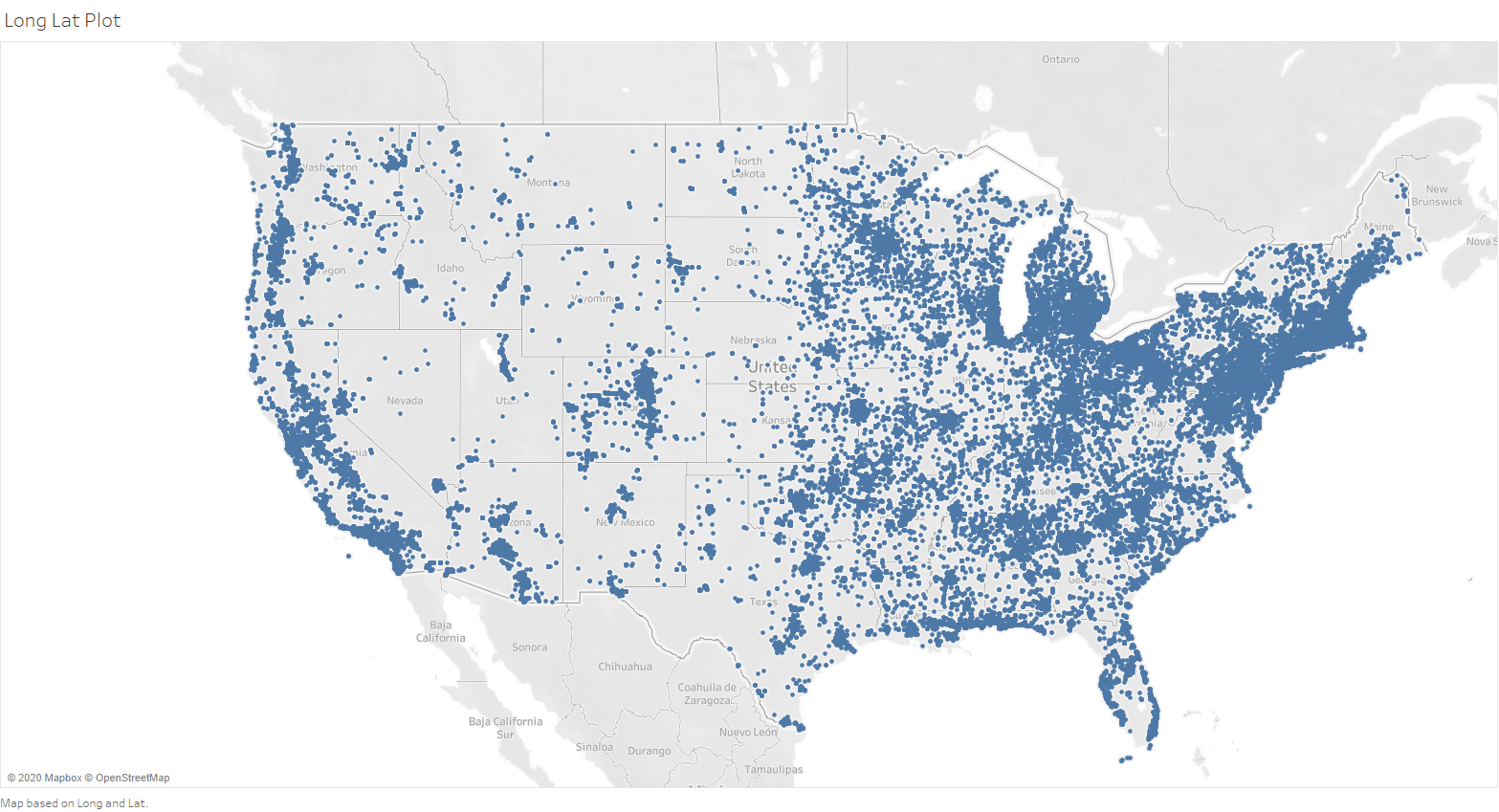


* Price/ State



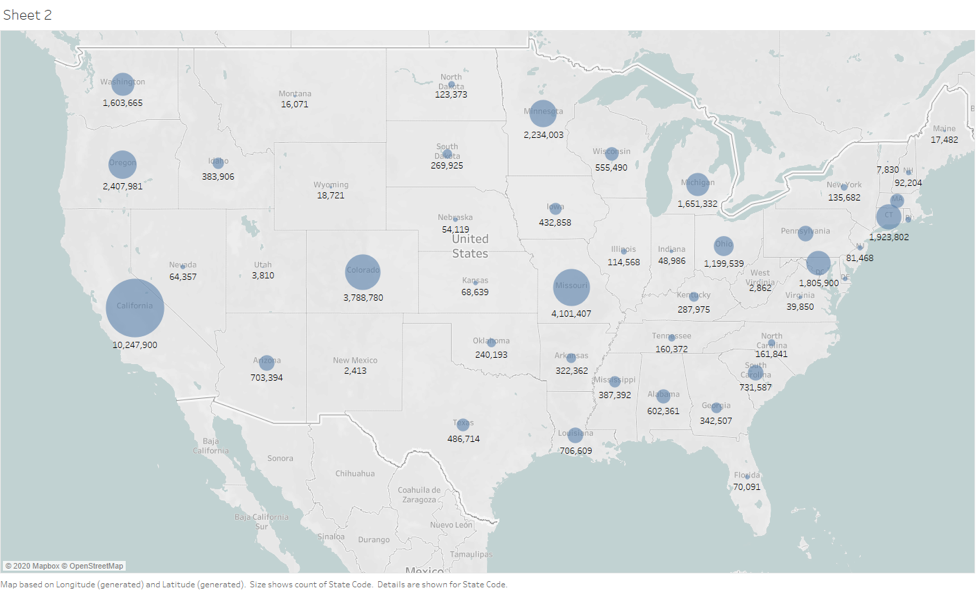
**Map**

* **Heat map of the USA on eligible cars:** We can see the majority of the listings located in the Eastern part of the states, and we see a very dense aggregation of listings in the east coast, the west coast and the Illinois, Indiana and Ohio area



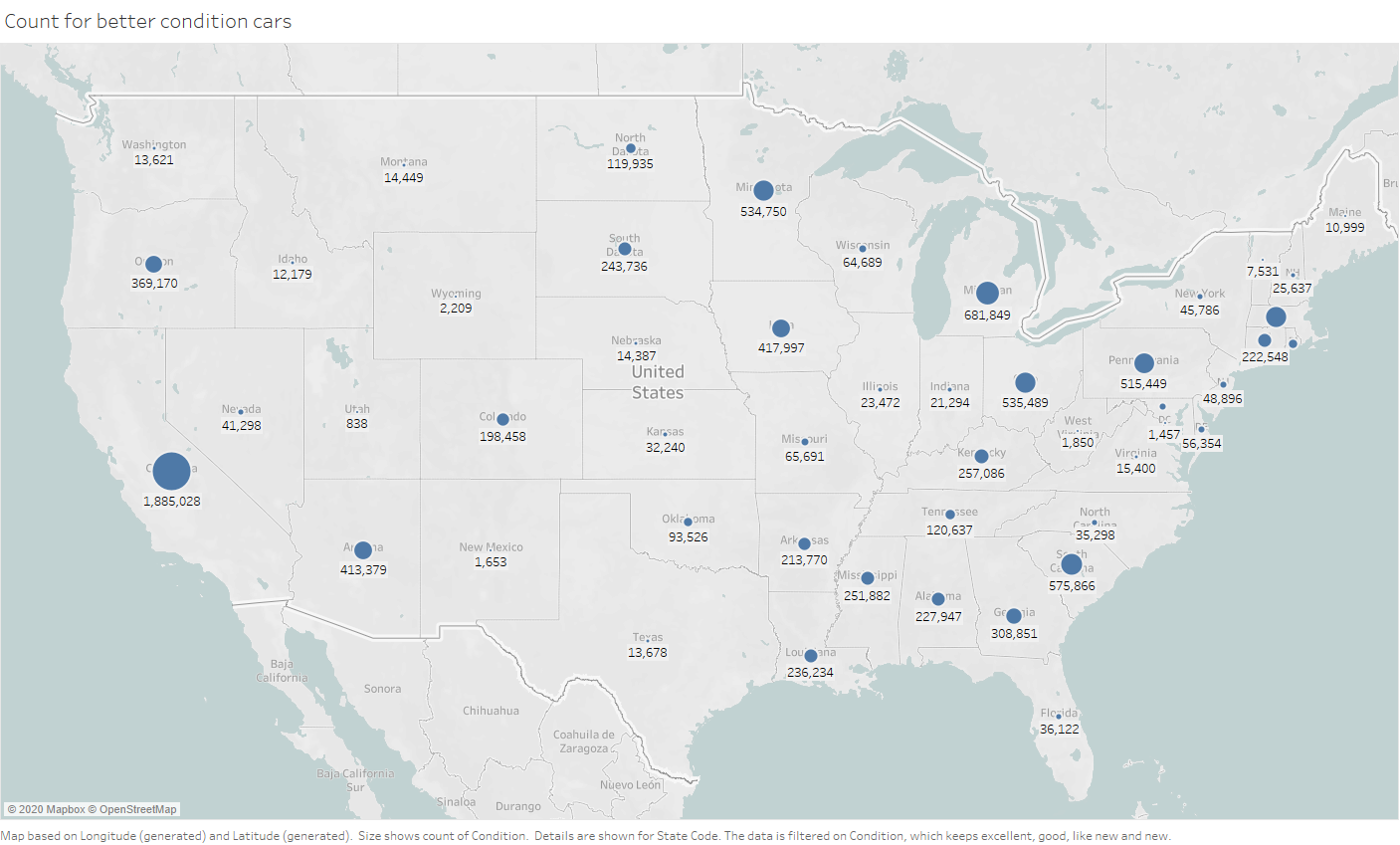
**Heat map of state**

* The following states have above 1 million listings: CA, CO, MO, MD, MN, OH, WA, OR MI



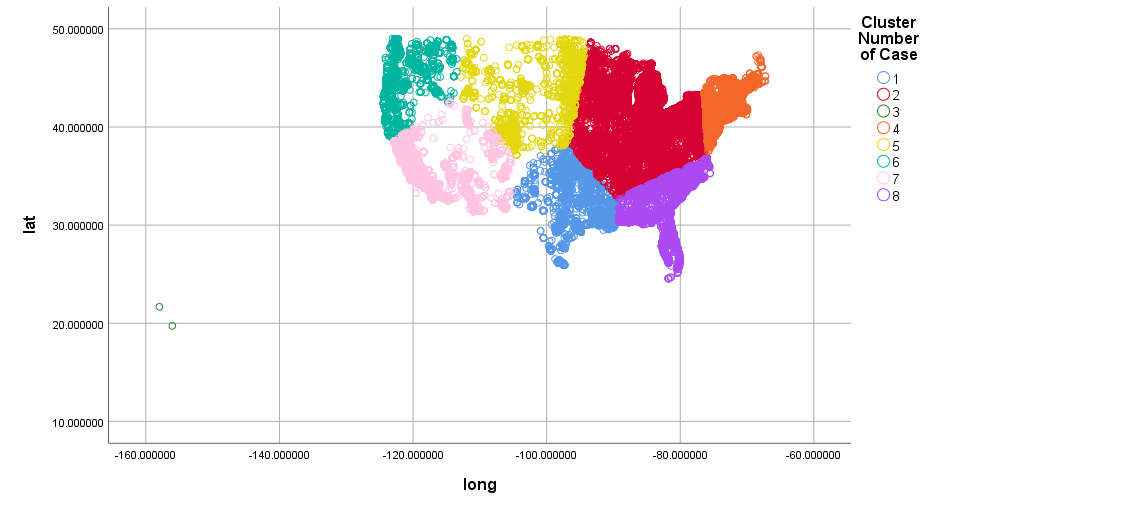
**Heat map of state with better conditioned cars (excellent, good, like new, new)**

* State: CA, MI, SC, MA, OH, MN, PA have more car in better condition overall



**Cluster analysis**

* K means cluster with a k of 8 is used to split the county into 8 districts based on an euclidean distance algorism

In order for Carvana to adapt a new method of purchasing cars from individuals, Carvana will develop an operation plan as well as an advertising plan, which will be the 2 aspect that we will be focusing on for our recommendations.

**Operations recommendations:**

* Having an experienced mechanic inspects the car before purchasing is essential to mitigate risk of used car purchases. Therefore, we believe it is very necessary for Carvana to work with local dealership or mechanic as cars from individual seller come in all kinds of conditions. According to the prescriptive analysis done on manufactures: Ford, Chevy, Nissan, Toyota and Jeep are 5 most popular manufacturers. Therefore, it would be wise to work with dealership or mechanic that specialized in inspecting domestic and Japanese cars. For selling only customer, it could allow customer to drop their car off at a dealership and leave with instant cash offer in hand. For trade in customer, it could enable a very smooth drop off- inspect – go car shopping procedure.
* Since major costs of Carvana are vehicle transportation cost and warehousing costs. To minimize transportation costs, we used a k-mean cluster analysis with Euclidean distance algorism to split the country into 7 districts. Each district will have a centralized warehouse and vehicles purchased from each district could be directed transported to the district warehouse instead of the one centralized location. Also, the company could allocate warehouse and staff base on the percent of vehicles distributed in the district.

|  |  |  |
| --- | --- | --- |
| District # | Percent of Cars | Regions Included |
| 1 | 2% | TX, OK, LA, KS |
| 2 | 17% | Midwest, VA, West PA, South state NY |
| 3 | 12% | ME, VT, NH, MA, CT, RI, NJ, NYC and Up State NY, DE, East PA |
| 4 | 10% | MT, ND, SD, West WI, NE, CO, WY |
| 5 | 14% | WA, OR, ID, North CA |
| 6 | 34% | South CA, AZ, NM, UT |
| 7 | 12% | NC, SC, GA, AL, FL |

**Advertising recommendations**

**Online advertising**:

We could potentially approach all craigslist listings with a ‘We will buy your car’ type of invitation which will be linked to Carvana’s website’s ‘sell your vehicles’ page. Since Carvana already know its car’s details, Carvana could even give a price offer range in the email. Then, at the same time, Carvana could potentially pitch the seller a car that is listed on Carvana’s site during the selling process.

Particularly, we could reach out to the seller that is selling the following cars as they are the most popular used cars to be sold in the US. Also, these cars in these states tend are more like to have a better condition and price/ odometer ratio:

* + Manufacturer: Ford, Chevy, Toyota and Honda
  + Type: Sedan, SUV, Pickup
  + Tittle: Clean, Lien
  + Color: Black, White, Gray
  + State: CA, MI, SC, MA, OH, MN, PA

**Offline advertising**:

Currently one of Caravana's biggest offline advertising medium is the car vending machine that is located close to many highways. These giant tower like structure that are located adjacent to major highways serve as an excellent advertising medium. According to the heat map, the following states have above 1 million car listing: CA, CO, MO, MD, MN, OH, WA, OR MI. We could potentially place addition car vending machines near interstate highways to capture audience or billboard advertising in states which don’t have much purchasing activity.

**Discussion of Ethical Considerations**

There are many important ethical considerations that could be addressed about our data set. Based on our data analysis, and the conclusions we have drawn based on our analysis, we believe that our data set comes from a reliable data source recommended for this project, which is, kaggle.com. This website is reputable and specializes in data sets. We spent time as a group discussing this dataset and its reliability and we believe it is an ethical choice. Although we believe the dataset is a reliable dataset, there are some areas for potential discrimination based on our analysis.

We found that the five most popular manufacturers for cars in this dataset are Ford, Chevy, Nissan, Toyota and Jeep. All these manufacturers are either American or Japanese companies. It could also cause dealerships to discriminate against other less successful American and Japanese manufacturers. Due to the success of these manufacturers, this could cause used car dealerships to discriminate against brands from other countries, such as European cars.

Although companies should maximize profit and meet the demand for these manufacturers, they should also not discriminate against other potentially successful manufacturers. Another source of discrimination could be geographical discrimination based on sales. More sales occurred in the southern part of the United States than the northern part. This could cause dealerships to favor the south. More cars would be stored in states with higher sales and more money would be spent to target customers in these states. Again, although they should target a market that produces a lot of sales, dealerships should be careful not to discriminate against other markets. Discrimination against other areas of the country could lead to reduced overall sales in these parts of the country. There is also the potential for privacy issues because these posts have been made by individual sellers on Craigslist. If this data were breached, sellers’ personal information as well as financial information could be hacked, manipulated or stolen.

**Conclusion: -** In this study, the Tableau features have been used to demonstrate the ease of

working on a large data set in the virtual environment. An analysis has been performed on the used car data set to understand the various factors and its importance. Colors, vehicle type, Brand, model, mileage, cost, and vehicle ages are some of the significant factors we had a chance to analyze, which affect the time a used vehicle stays in the lot. In general, a customer’s preference is varied and complex. However, there are some characteristics that are highly desirable among used vehicles that Carvana can prioritize to purchase.

After the analysis we found many interesting facts, such as, customers preferring a vehicle that is Black, White or Gray. From popular manufacturers like Ford, Chevy, Toyota or Honda.

Location affects sales significantly. When purchasing or selling used vehicles in the low-volume states such as Texas or Oklahoma, Carvana should focus on other highly desirable characteristics like colors, brands, low mileage and age.

Customers always prefer to purchase vehicles that have low mileage, low cost and newer manufacturing dates when it comes to choosing between two equivalent vehicles in term of other characteristics.

In conclusion, future research is recommended to explore other factors that influence the sales period of a used vehicle. For example, the level of fuel-efficiency, whether the vehicle is electric or hybrid, discount level. After incorporating these factors in the analysis, we can improve the accuracy to choose vehicles which have a positive impact on profit.

**References**

Gustafson, S. (2018, June 8). Online car dealer Carvana opens 2 more car vending machines. Retrieved May 5, 2020, from <https://www.autoblog.com/2018/06/08/carvana-used-car-vending-machines/>

1. <https://www.kaggle.com/austinreese/craigslist-carstrucks-data> [↑](#footnote-ref-1)