Capstone 2: Milestone Report 1

Introduction:

Vehicles are an important part of many people's lives. Whether the vehicle is for transporting ones family, commuting to work, or a vehicle for work, a vehicle is an important investment that is practically mandatory in many parts of America. With that being said, I was interested in analyzing some type of vehicular data for one of my capstone projects, and fortunately for this second capstone project I was able to do so. Using data gathered by Kaggle user CooperUnion, I wanted to analyze factors that could potentially affect the MSRP that a manufacturer chooses for a vehicle. Many factors could affect MSRP, from manufacturer biases about how they perceive the reputation of their products to cutting edge technology that is expensive to produce and sell. This specific dataset describes many of the most common features of vehicles that are evaluated by a customer while shopping for a vehicle so I believed it to be a good data set for a preliminary assessment on what may affect some manufacturers choices when suggesting a price for their vehicles. In this document I will be describing how I cleaned this data for analysis and exploratory data analysis that I performed with this data before making any predictions with models.

Data Description and Cleaning:

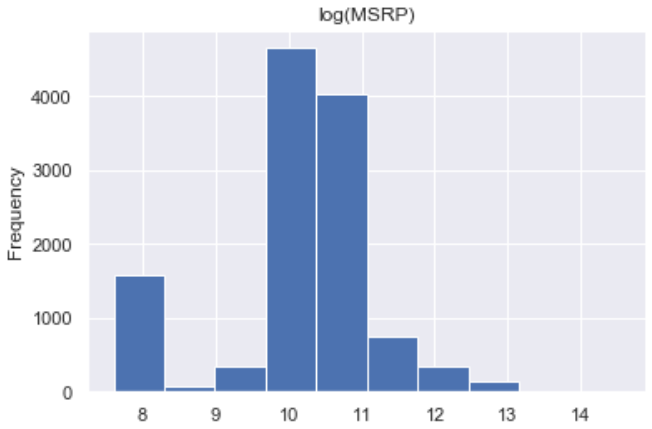
The data source for this project is linked previously, but I will take some time to describe the data and the features contained within it. The data originally contained around 12000 rows with 16 columns. The columns are:

- Vehicle identification columns:
    - Make, model, year.
- Vehicle description columns:
    - Fuel type, horsepower, number of cylinders, transmission type, drive type, number of doors, market category, vehicle size, vehicle style, highway mpg, city mpg
- A column titled "popularity" which seems to be a count of number of tweets about the vehicle in question(this feature isn't documented well in the data description).
- A column with the MSRP of each vehicle.

This dataset originally appeared to be very clean, so the cleaning process was fairly standard, at least initially. The first thing I did was to rename the columns to be shorter, more standardized, and a bit more descriptive of what they actually contained in order to ease using the data while programming. After changing the names, I took the time to look closely at each feature to try to find incorrect or missing data. The make, model, and year columns did not seem to contain any erroneous data, as all of the manufacturers were spelled correctly and I was unable to find any text errors in the

model column. The other features of the data such as number of cylinders and number of doors, were similarly clean. The first hiccup in the data was in the highway MPG column, which had a single anomalously high datapoint which I was able to look up and fix.

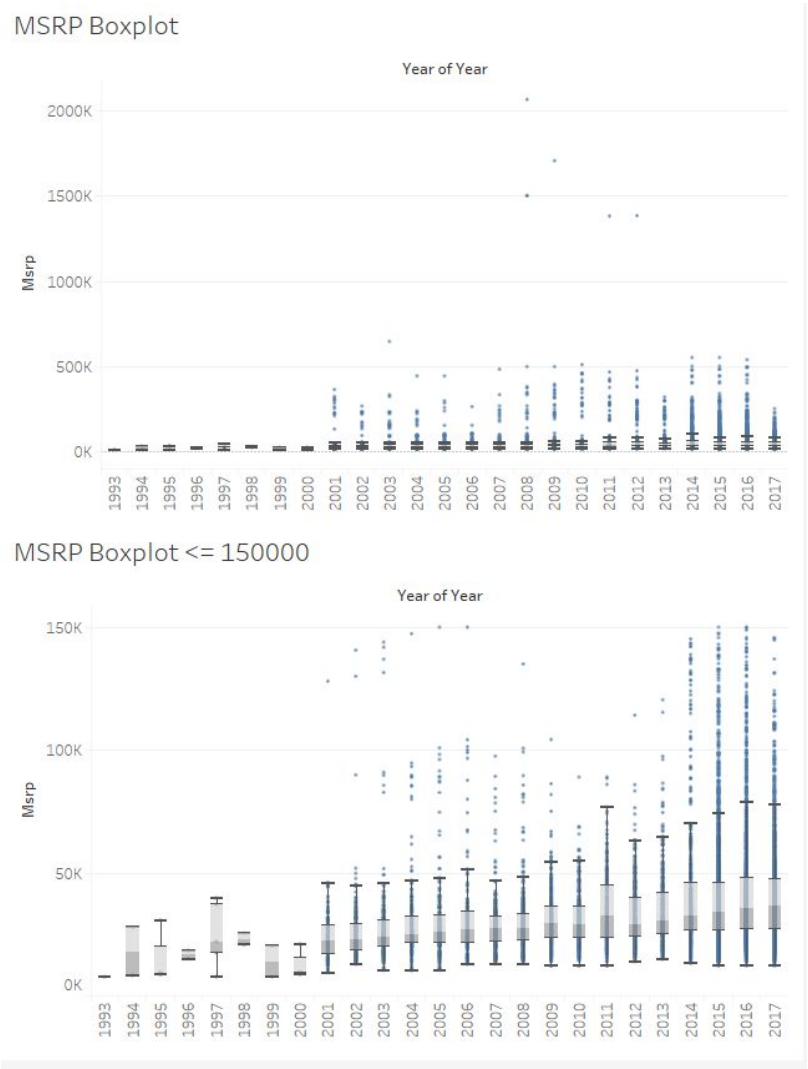
After fixing the MPG data, I went about putting the category column into a more usable form. The category column was originally a list of words that can be descriptive of vehicles such as “hatchback” or “exotic”. I went about splitting this column into multiple columns each column containing a single descriptive word from the category column. Some vehicles had up to five words in the category column. After looking at the words in these new column after splitting I came to two conclusions. The first conclusion was that for vehicles that were missing a value in the category column; they were missing a value because they did not fit any of the special niches that were described by the categorical words that were in the dataset. Since this was the case, I decided to replace the missing values in the first column with the word “Standard” since the vehicles that were missing this value simply didn’t fall into any of the categorical words initially in the data set, and did not seem to have any special defining features beyond being the standard versions of those vehicles.

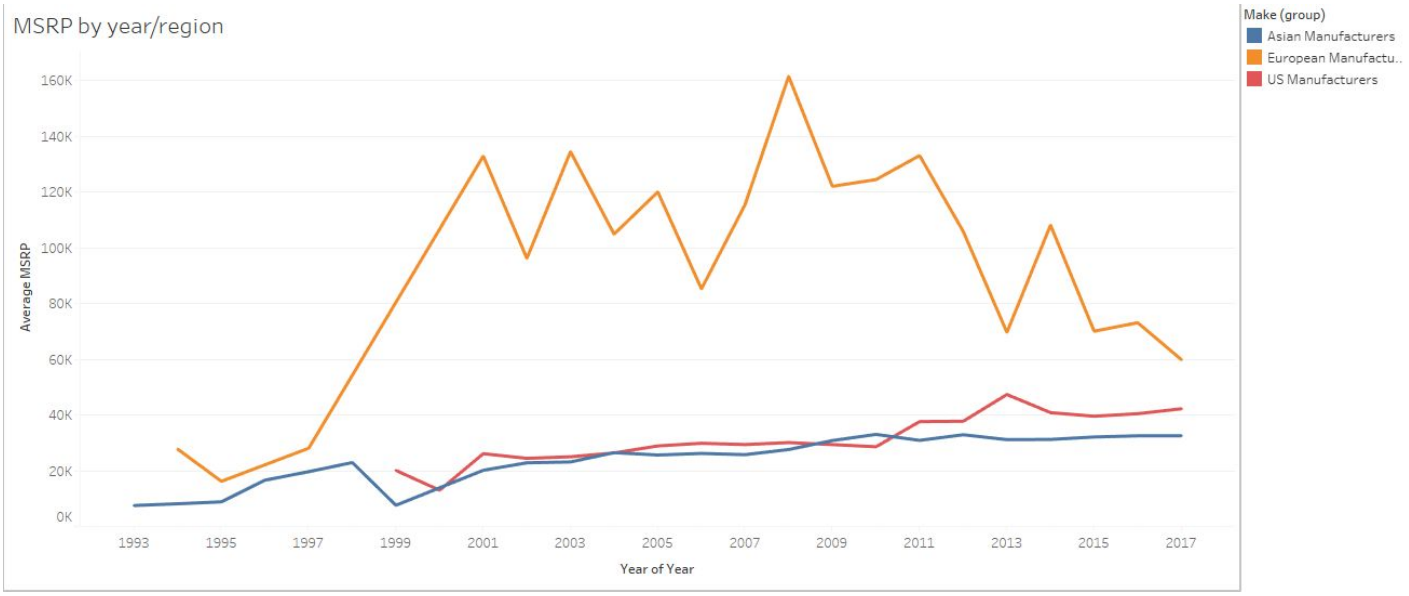


As I was looking into the numerical features of the data in more depth, I discovered something very strange occuring in the distribution of the MSRP values of the dataset, which can be observed in this histogram. The MSRP data was exponentially distributed, so I looked at the distribution of the log values of the data and found that there was an interestingly high number of relatively small values in the dataset. When I looked into these small values, they all seemed to be coming from the earliest years of the dataset, and none of them were simply off by a 0. There seemed to be a default value of 2000 that was placed into the data for cars that may not have had an MSRP when the data was scraped. There is no way to confirm why the data is erroneous but it seems that most of the data from before the year 2000 had some error that resulted in a very low MSRP being put in the place of the actual MSRP of the vehicles. Not wanting to filter out meaningful data but also not wanting to have junk data with a bad target variable to build my model with, I decided to try filtering the data by removing all of the vehicles with an MSRP less than $7,500 from my data to have more accurate MSRP values. This resulted in most of the data from before the year 2000 being dropped from the dataset, but the dataset still contained 10,278 rows of data.

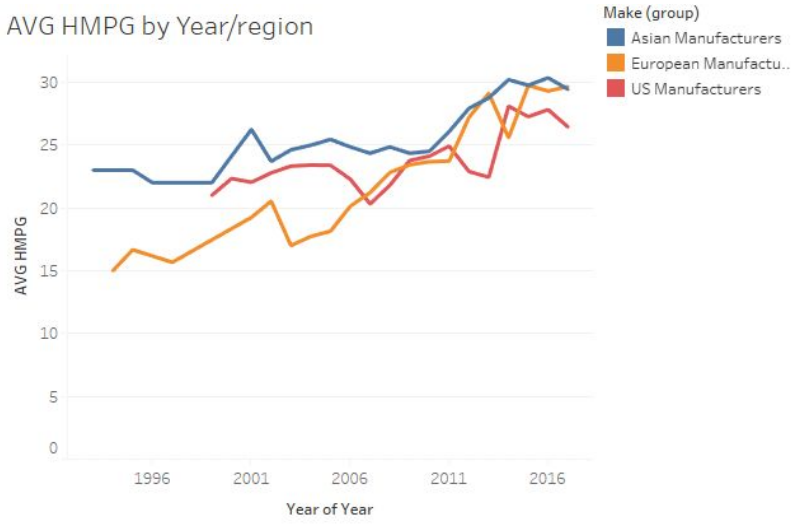
Exploratory Data Analysis:

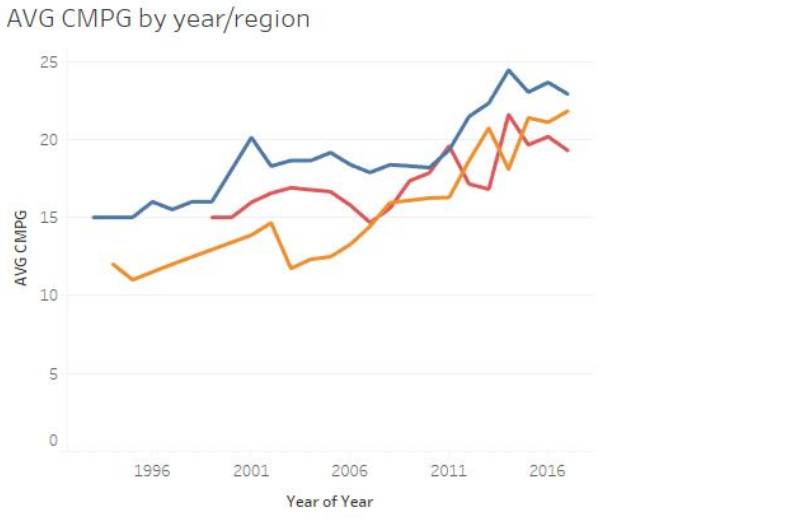
The primary objective for the EDA of this project is to attempt to find relationships between MSRP and the other features of the dataset. As can be observed by the boxplots on the right, it seems that MSRP has slowly increased over time. With median values of MSRP not varying too significantly while slowly increasing. An interesting quirk of the MSRP values in this dataset is the ranges between years can vary wildly depending on whether or not an incredibly expensive exotic sports car was manufactured during that year. Another quirk of the data to make note of is the limited number of data remaining from the year 2000 and earlier, which was addressed earlier. Many of the visualizations moving forward may appear a bit skewed because of this fact. Below you can see a graph that plots MSRP over time by year and split by region. This specific region set was chosen because it is popular to distinguish automobile manufacturers by their general location





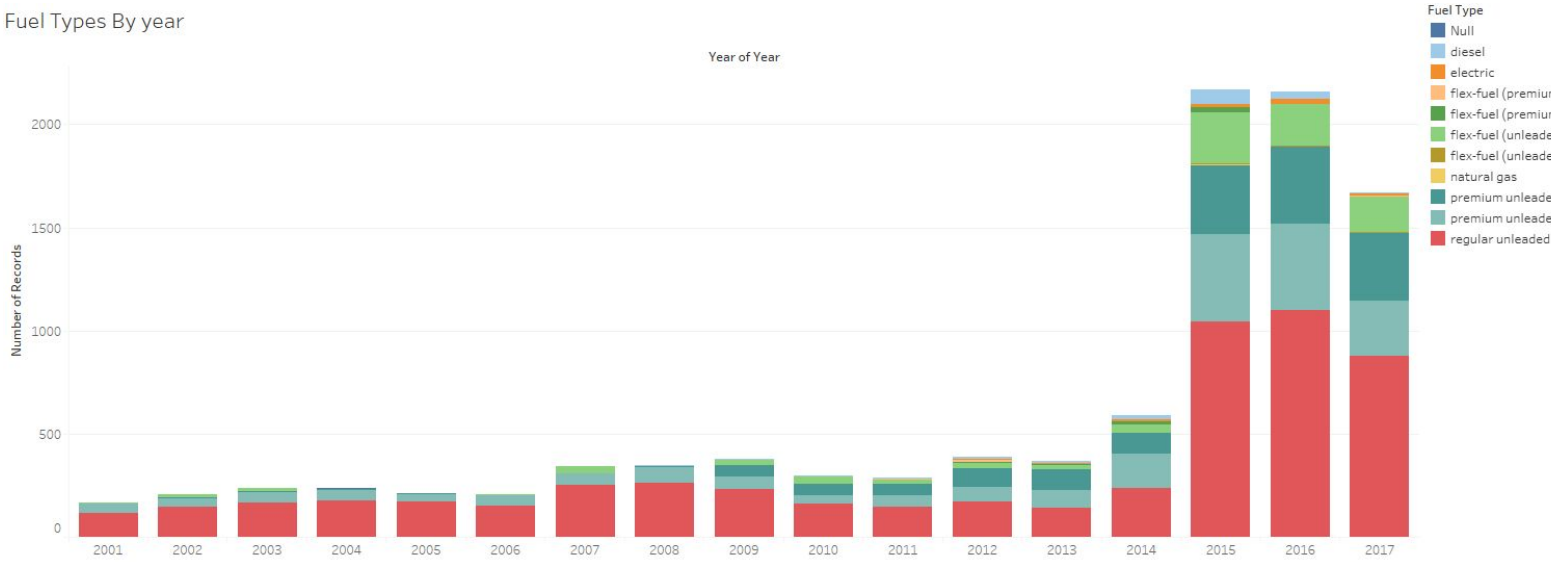
as an American, European, or Asian automobile manufacturer. This specific plot shows average MSRP by region over time. The plot shows that average MSRP increases relatively slowly over time for both US and Asian auto manufacturers, while the average MSRP for European manufacturers is higher than the other two regions consistently, and varies wildly across years. The European average is so much higher for one primary reason; which is the fact that most 'Exotic' or 'Luxury' manufacturers, such as Ferrari, Lamborghini, and Bentley, are classified as European manufacturers. Manufacturers in the US and Asia have much more of a focus on producing economy cars or utility vehicles that do not carry the hefty price tag that something like a Bentley or Ferrari has.





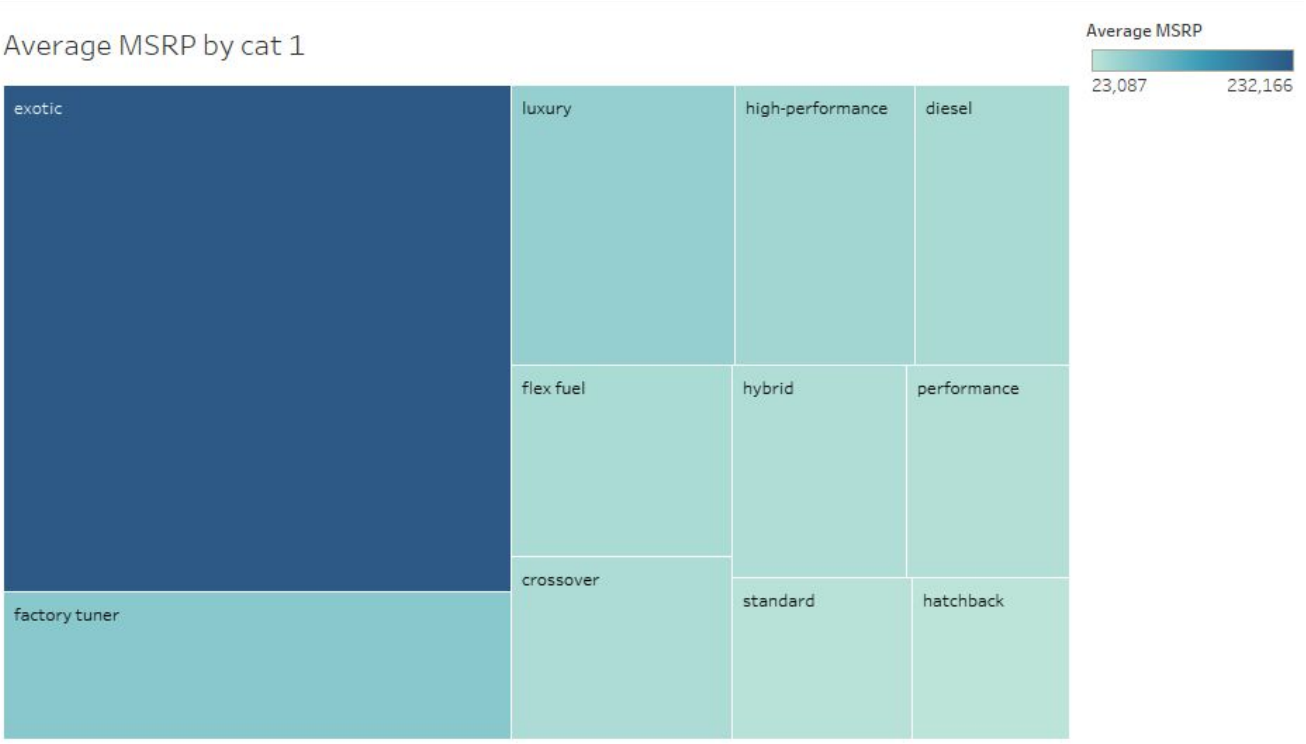
Taking a look at some of the other features of the dataset, it can be seen that each region has made some progress in increasing average city and highway fuel efficiency of their vehicles over time, with Asian manufacturers leading the way and US manufacturers lagging slightly behind the other two regions. As time has gone on, different types of fuel have been developed for vehicles such as E85, and the number of records by fuel type can be observed in the chart on the next page. It seems prudent to note that the vast majority of vehicles being manufactured today still seem to require 'normal' gasoline, or the premium unleaded equivalent, but proportionally there are many more flex-fuel and E85 vehicles in the 2015-2017 range than in the past in this data.

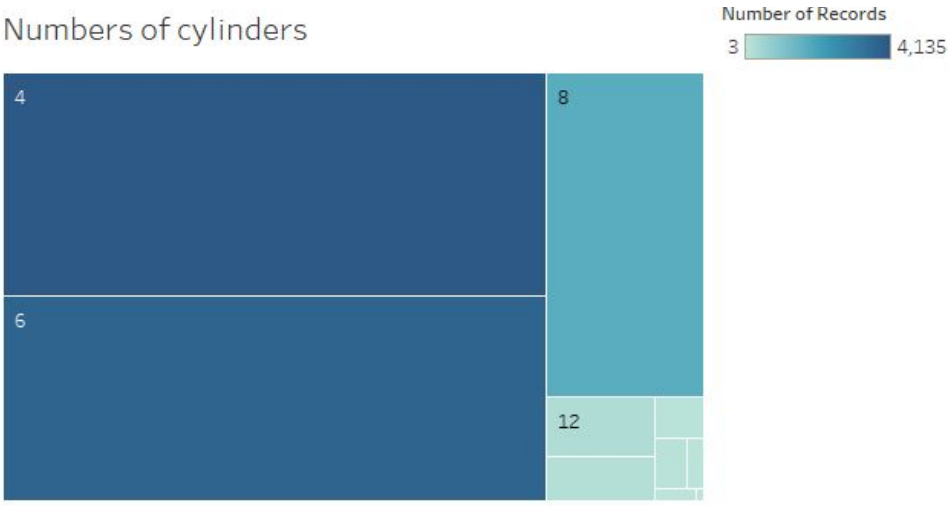
The 'style' a vehicle has the potential to be one of the more important aspects in regards to predicting MSRP for a vehicle. On the next page you can see a chart that shows relative average MSRP by year separated by the styles that are included in the data set. The chart reveals that most styles' average MSRP's do not



vary much over time, and average MSRP does not seem to vary significantly between

styles, with the two exceptions to this being a few years in the ‘convertible’ and ‘coupe’ columns. It does seem that those two styles are the most expensive of the columns in terms of average MSRP.

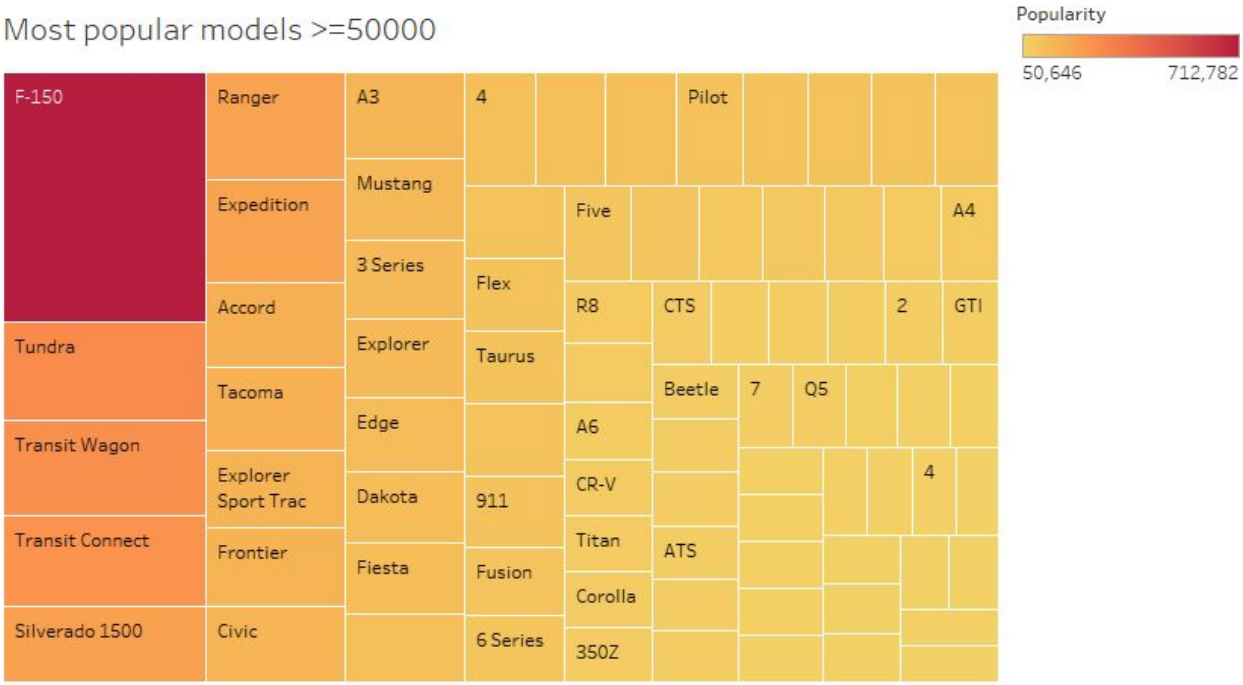




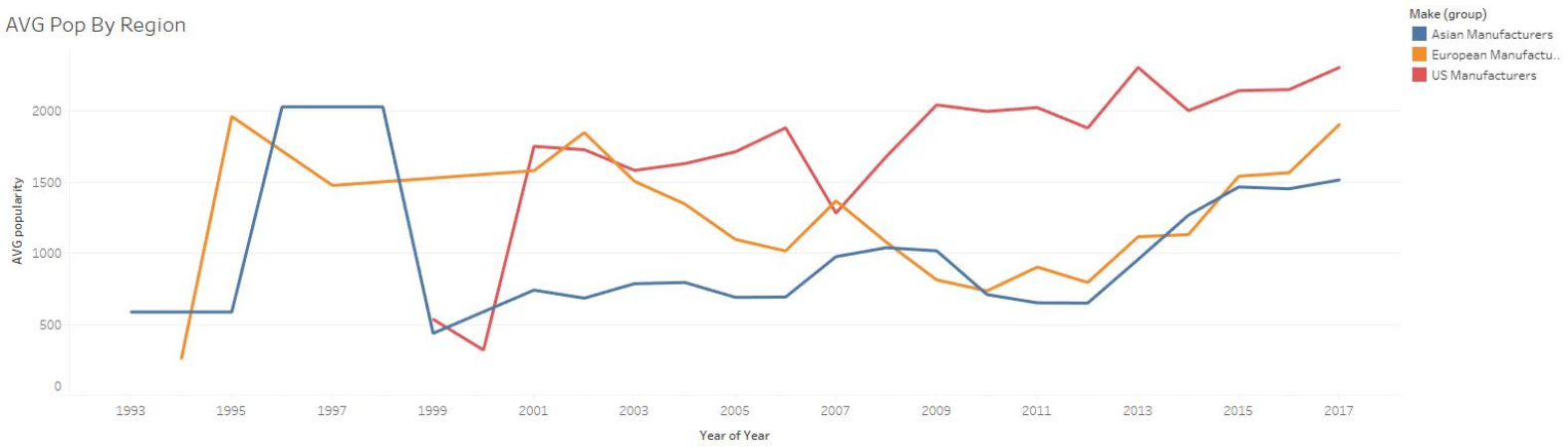
There are just a few more aspects of this data that I would like to explore and show before moving on to the next parts of this project. The chart above gives a relative sense of average price by category for every vehicle in the data. The information here is not surprising but it does show that categories one would expect to be more expensive such as ‘exotic’ and ‘factory tuner’ are indeed more expensive than models one would expect to be less expensive such as ‘standard’ and ‘hatchback’. We can also observe that most of the vehicles in the dataset have four-cylinder and six-cylinder engines, with eight-cylinder engines being the next

most common. This information is not surprising, however it may be that any models that are created to predict MSRP may be better at predicting MSRP for the engine types that occur much more often than the others.

Finally, I would like to take a look at what is going on with the popularity data in this dataset. Since we can not actually be sure how exactly the popularity value was generated, it seems important to take some time to see what these values seem like they will be able to tell us. To be quite honest, the validity and usefulness of these values seems quite questionable, as some models have identical popularities for multiple years, which seems quite impossible if this was counting tweets for each model and each year, however, at the advice of my mentor I have not dropped the data, and will now present some of the information that the popularity data can provide.



This chart shows the most popular models in the dataset by the sum of their popularity across all years. It is important to note that this specific chart seems to be very misleading, as most of these models that have the highest popularity are also some of the most common models in the dataset, and it does not seem like the person that scraped this data made an effort to account for minor differences in models in the same year and their popularity value. On the next page, I've created a visualization that shows average popularity by year and region. Excluding the early data which has few samples so averages will tend to vary wildly there anyway, it seems that since 2011 the average popularity of all vehicles in all regions has seemed to increase over time. This makes intuitive sense since the popularity data was derived from Twitter, however it may be worth considering this popularity increase simply has to do with the number of people



using Twitter as opposed to actual objective measure of popularity, since we do not know exactly how this particular value was measured. It will be interesting to see if the popularity values assist in predicting MSRP or simply add noise to any of the models that are created.

Moving Forward:

Now that we have a better picture of what is contained within the data and have been able to draw out some insight for the data, it is time to move forward with creating models. The next part of this project will be focused on using an XGBoost regression to predict MSRP as well as using some other methods to analyze feature importance to attempt to determine what aspect of the data can help best predict MSRP.

Project data source: https://www.kaggle.com/CooperUnion/cardataset
Github link for related code:
https://github.com/mbuck86/Capstone-2/blob/master/Data%20Cleaning/Data%20Cleaning%20and%20EDA.ipynb