Name: **Muley, Tushar**
Assignment: **DSC 680 - Week 3 Milestone 2 White Paper**
Date: **February 27, 2022**

Craigslists Used Vehicle Analysis

Draft White Paper – Milestone 2

Table of Contents

Introduction: 2
Data Set/Research Questions: 3
Assumption/Limitation/Challenges: 3
Ethical Considerations: 4
Method: 4
Exploratory Data Analysis: 5
Test/Training the Model: 9
Model Evolution: 12
Conclusion: 13
Reference: 14
Appendix: 15
Questions: 15

Introduction:

The used car market has been a little out of control since the pandemic has introduced us to a new way of doing everything it seems. In the last year used car prices have gone up by almost 40.5% between January 2021 and January 2022 [6]. This has been due to a lack of new cars being produced due to a storage of silicon chips. This storage as caused a increase in used car prices. Individuals and dealers are getting in on the pandemic fueled frenzy to purchase new cars to avoid public transits or to have a car to get away from being stuck at home. There are a lot of web sites that allow individuals and dealers to post their vehicles for sale.

Where does an individual go to sell cars without selling them to a dealer or some other third? There are great websites for selling your car on your own terms. These sites include Facebook under the Facebook Market Place banner, Auto Trader (the original car selling publication), Cars.com, eBay Motors and Car.com to name a few. Most people will turn to Craigslist. Craigslist is a low-cost solution to selling your unwanted or unneeded car. While it might not be fasts it is the most cost effective unless you have a lot of friends on Facebook that can help you market your car quickly. Craigslist was also the original site to go for selling or buying stuff before Facebook Market Place.

Craigslist.org created by Craig Newmark back in 1995 is a classified advertisement website that has various section devoted to jobs, vehicle for sale, items wanted, services and other sections. It is based in San Francisco, California. The service is currently available in 70

different countries. The Craigslist site has 49.4 million unique monthly visitors [2]. All those views rank Craigslist as 72nd most viewed site worldwide and 11th most viewed site in the United States.

In this study I will perform Exploratory Data Analysis, answer research questions about Craigslist vehicle selling site. Then I will split the data into test and training to perform a predictive analysis on sale prices of cars on Craigslist. From there will provide the best model for the analysis with resulting data.

Data Set/Research Questions:

The data set was obtained from Kaggle.com which was surprising. Another individual doing a project for school created a scrapper to pull prices from Craigslist.org for all vehicles that have been listed for sale in the United States between the month of March and April 2021. The data contains 426,881 rows and 26 columns of data. It is the most current version of the data which is version 10. The appendix section contains column names and description of the attributes for the single table. I did remove data points and columns contain nulls values. A data dictionary is provided in the appendix section of this paper for the table used in the analysis.

Data obtained from: https://www.kaggle.com/austinreese/craigslist-carstrucks-data/version/10

I am planning to resolve the below research questions from the data.

1. Which model years are most frequently come up for sale?
2. What is the average odometer reading?
3. What is the most common condition listed?
4. What states are seeing the largest used car sales?
5. Which make of car makes up the majority of vehicles listed?
6. Which manufactures are seeing highest average sale prices by state?
7. Predicting the sale price of similar vehicles?

This next question is a little bit of a stretch since it will be difficult to determine which are dealer advertisement compared to sale by owner listings. I am going to list it as an optional question to be completed if the above six questions and the predictive model are complete in time.

- Determine which listings are from auto dealers compared to sale by owner as the seller of the vehicle?

Assumption/Limitation/Challenges:

I am actually expecting a lot of challenges with this data. It appears a good job has been done to scrape a lot of good data from the site. It is also missing data fields like number of images, posting removal date and seller type. I don't believe this information is needed but

would provide additional information for the analysis. If the data was available we could determine days on market, if image count was available it could be a feature in degerming sale timelines.

The other challenge is the data is limited to two months. More data would be useful and is available but comes with issues. As new versions where added fields have been dropped or added. This becomes a challenge when trying to combine different version to increase the amount of data. I still believe I can use the data to predict future days prices or weekly price changes. While car's prices in the past have not been as volatile as they are in the pandemic car market.

Ethical Considerations:

This data is publicly available due to the nature of the advertising. The data does not contain any personally identifiable information. As Craigslist does try its best to warn posters and setup email address to mask actual email address of the posters. Phone numbers being displayed in posts are frowned upon and warns are given by Craigslist. But that sometimes does not prevent poster from providing some personal data in hopes the item can be sold quickly. When I provide the data for further analysis I will try to remove phone numbers in the description field or remove that column once I have pulled information I need for the analysis.

If more attributes could be pulled like posting removal date or scraping the seller type that would help with the analysis quite a bit. Craigslist is pretty popular compared to other sites as some charge 50 dollars or more to post on their website. Depending the state dealership like to post on Craigslist for additional exposure, even if they have their own sites. This show how popular Craigslist can be.

I believe the data can be used for future price predictions even if limited to two months. This would help seller and buyers alike. On the seller side they could determine better range to sell their vehicle for the most amount of money. On the buyer side this would assist in tempering expectation of what they can pay for a make or model they are interested in.

Dealership could also use this information to predict what range they could offer their customers when they are trading in a vehicle. While dealerships have greater resource to determine this in the current market sites like KBB, commonly known as Kelley Blue Book and North American Dealer Association or NADA price predictors are not able to determine expected vehicle prices accurately.

Method:

Once the table was loaded into Python I had to do some data cleansing as there where a lot of null values. I dropped unused parameters like region_url, url, VIN, image_url, country, lat, long, size, type, and paint color as these are not need for the analysis. I updated the nan values

to zeros as I wanted to keep other information. I transformed the date time stamp parameter to just date. Since there was no sale timestamp having the time did not make sense. During the EDA process I discover anomalies in the data.

The EDA process was a little longer than expected. Based on some odd numbers for mean analysis on prices I dropped records of cars below 500 dollars. The reason for this is a well know practice of offering cars for one dollar to get more views. This meant the actual asking price were inaccurate. Removing the outliers at the under 500 dollars and cars above 70,000 brought the mean to a level of common listings on Craigslist. Figure 1 show some of the odd mean prices. Note 'nj' and 'or' have average mean that is 350,000 and 250,000.

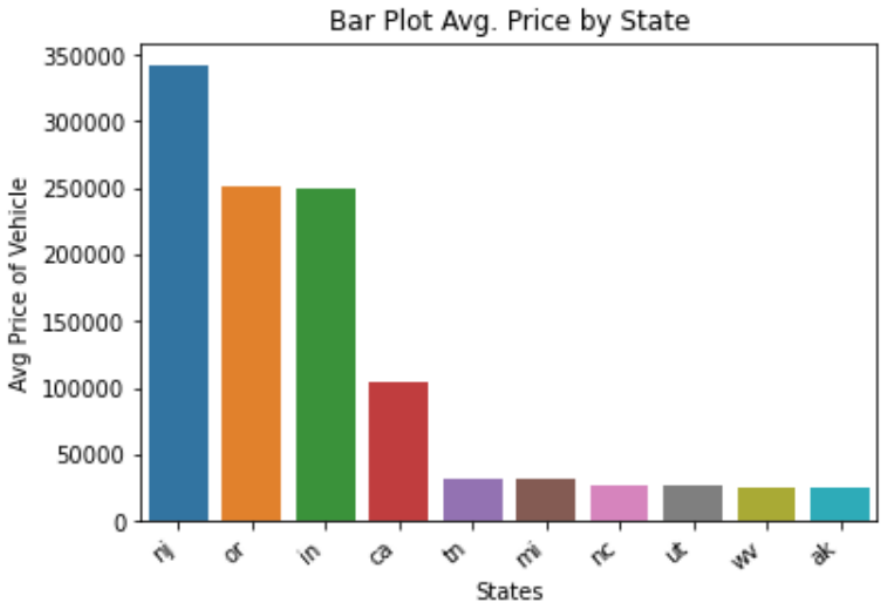

Figure 1: Bar chart of Average Used Car Prices by State. Note 'nj', 'or','in' and other state being over 100,000

In review of manufacture count and cylinder I made the decision to remove Harley-Davison since we are looking at car prices not motor cycles. I also dropped 'other' cylinder counts. This left only cars.

Exploratory Data Analysis:

I did exploratory analysis on the data and transformed some of the features along the way. As part of my research questions, which model years are most frequently come up for sale? Below is a pie chart showing the model year and percentage vehicle's for sale.

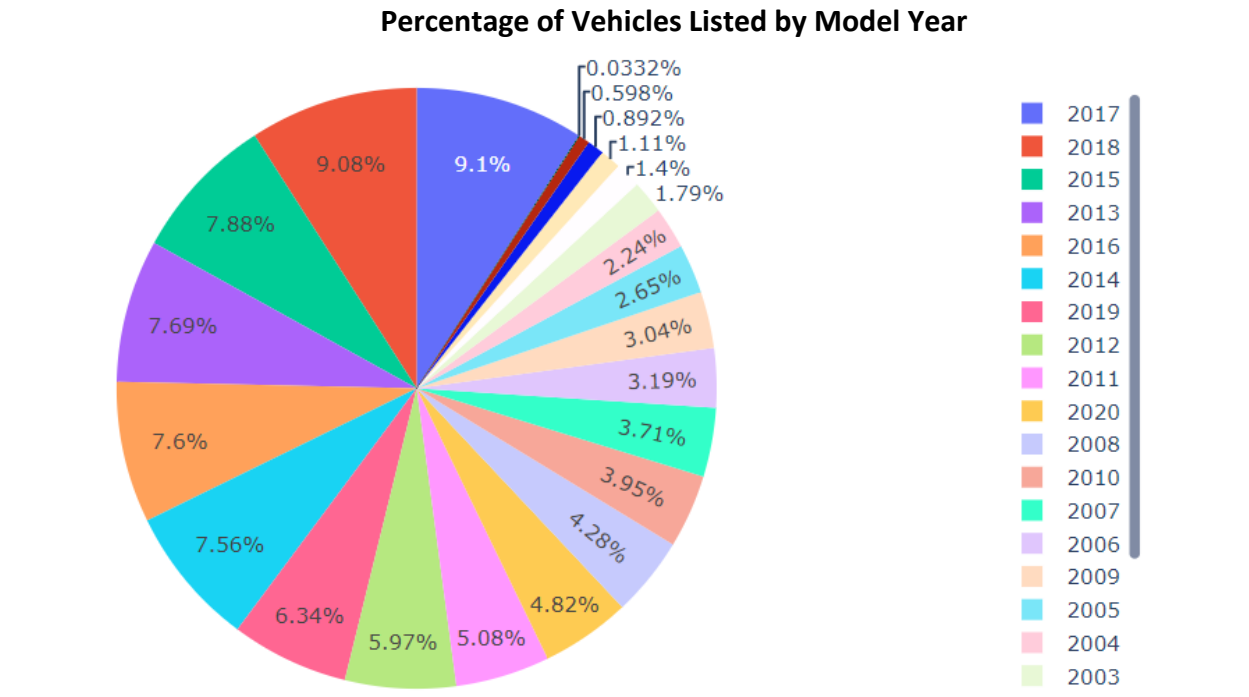

Figure 2: Percentage of Vehicles Listed by Model Year

Vehicles models years range from 2013 to more common 2017 model year vehicles. Vehicles that are a year or older only making up 4.82% for 2020 and 6.34% for model year 2019. Majority of the vehicles are in the 4-to-6-year range (2017 – 2015).

Figure 3 shows the average odometer reading of these vehicles listed by state.

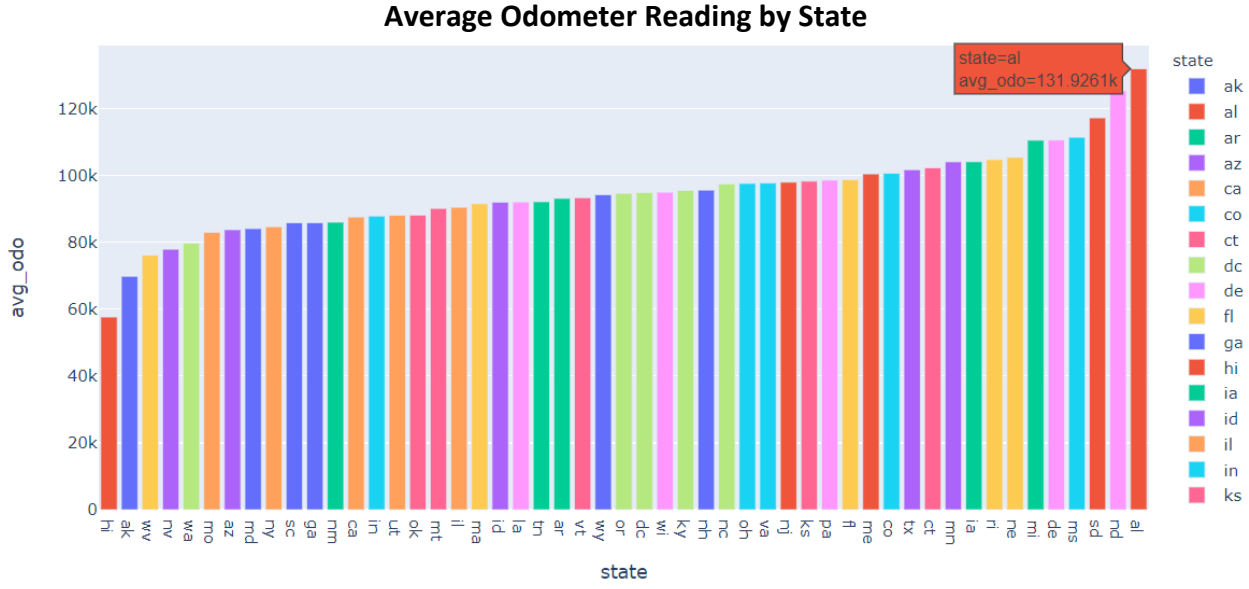

Figure 3: Average Odometer Reading by State in Ascending Order

As you can see smaller states like Hawaii have lower average odometer readings at 57,000 miles compared to any of the mainland state. Alabama has the highest average odometer reading at 131,000 miles.

What is the most common condition listed for these cars? The unique bubble chart shows three most common conditions that bring good prices for used cars.

Count of Condition Types of Listed Vehicles

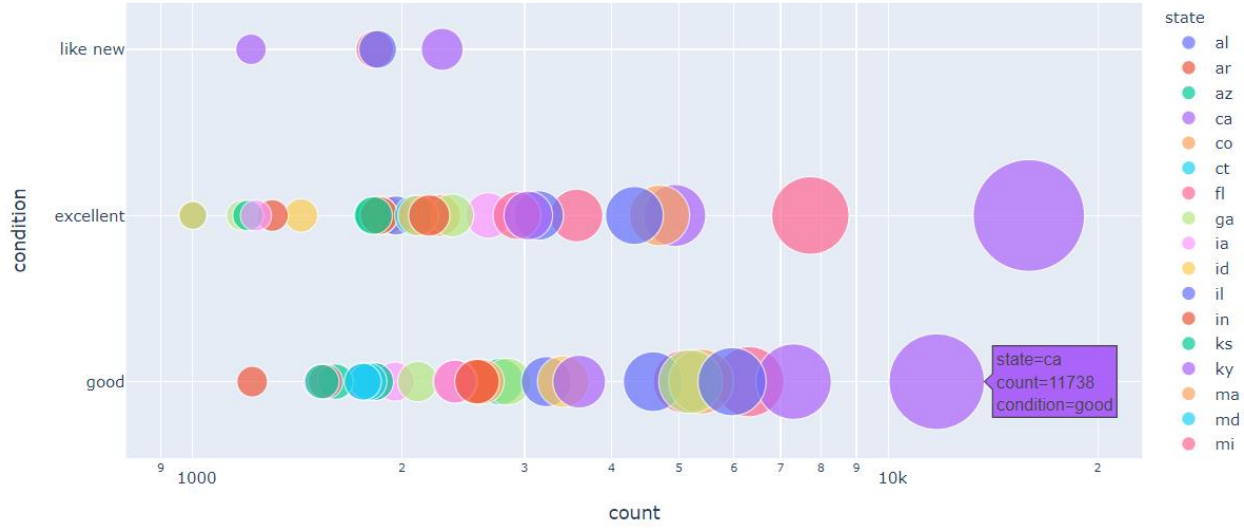

Figure 4: Count of Condition Type of Listed Vehicles

Californian appear to have the opinion they take care of their cars better than any other states owners. This might be related to firm belief in the California car culture. As you can see by the chart no other state comes even close to Californian's ratings of their cars.

Based on the above understanding the below bar chart should not be a surprise.

Number of Vehicles Listed for Sale by State

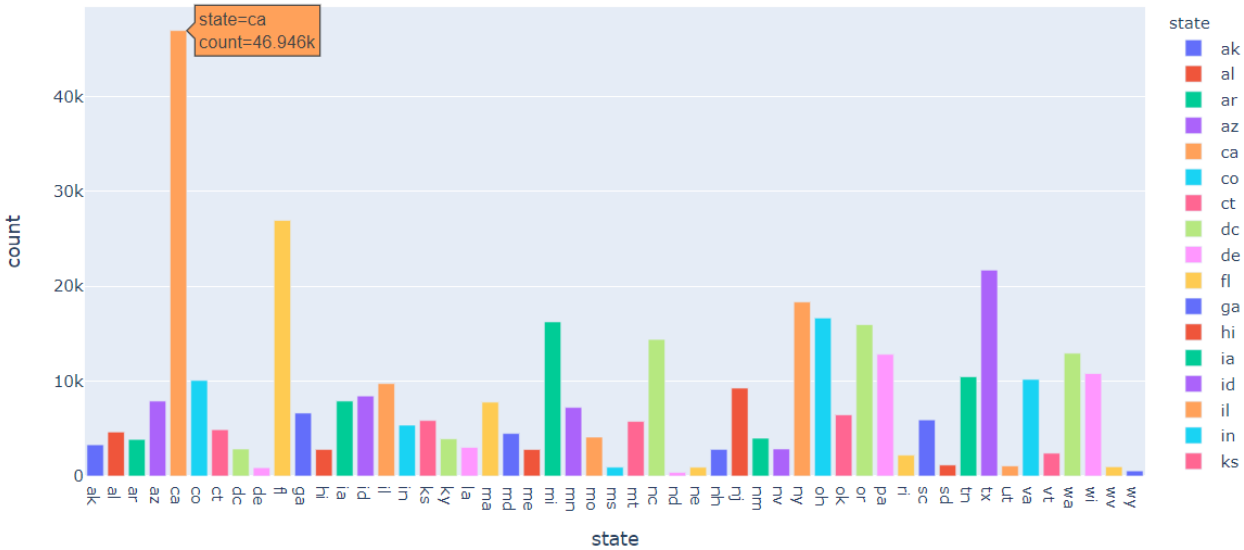

Figure 5: Number of Vehicles Listed for Sale by State

California sees more cars listed for sale compared to any other state. California sees 46,000 plus vehicles. The next largest is Florida at 26,934 and Texas in distant third at 21,704 cars listed. This might speak more to how popular Craigslist might be in those states compared to other states.

Based on claims made by manufactures like Ford who have the best-selling new trucks on the market the bar chart listed below show something very similar.

Count of Vehicles Listed by Model

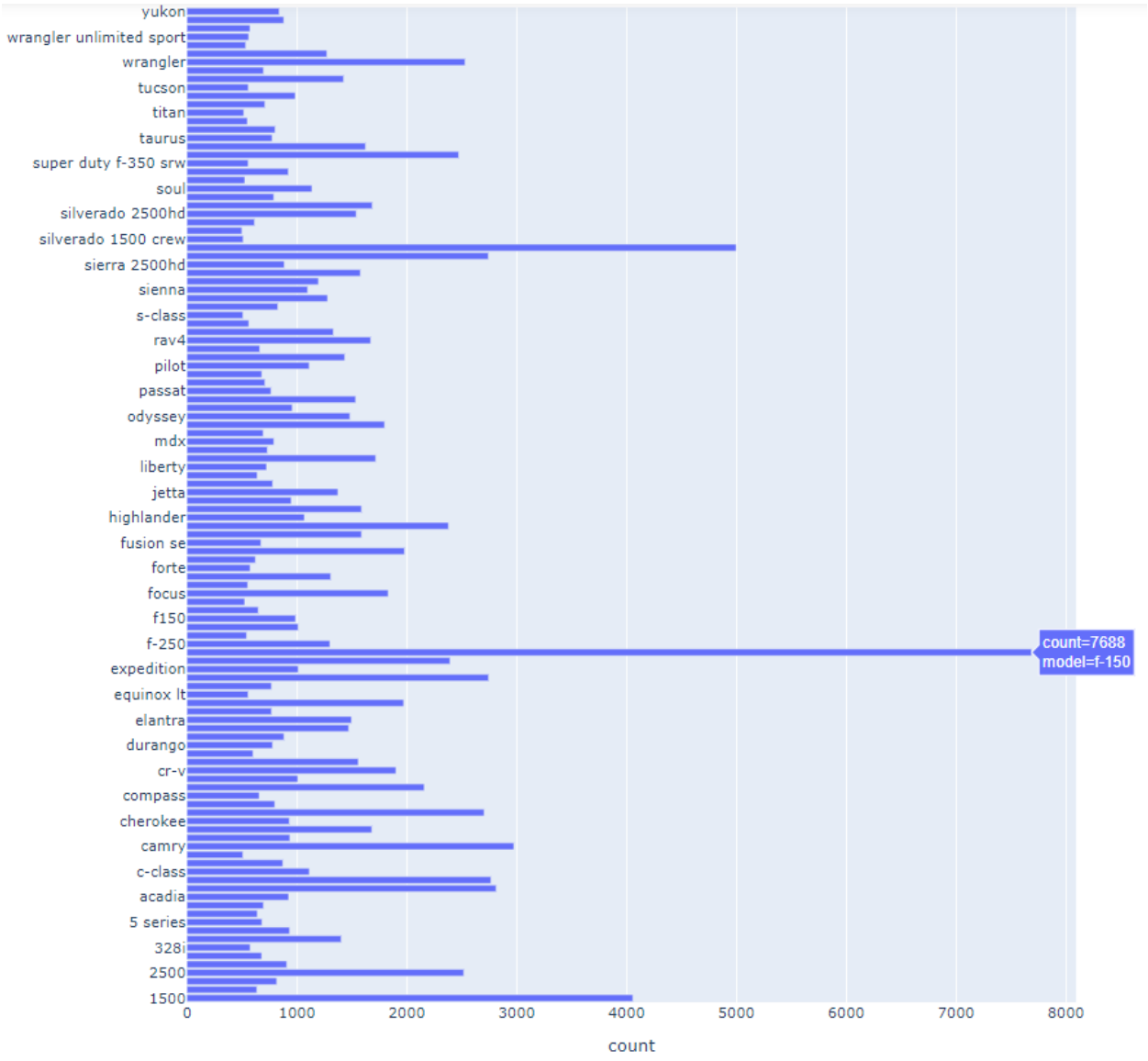

Figure 6: Number of vehicles by model

As the horizontal bar graph shows F-150s are the largest number of listed vehicles on Craigslist. Majority of the vehicles on the graph are version of a truck or sport utility vehicle. Note the lower number of luxury vehicles like BMWs or Mercedes.

For my final question I wanted to show, which manufactures are seeing highest average sale prices by state? The below stacked bar graphs showing bot the top 10 and bottom 10 with respective state shows how high-priced luxury vehicle impact the overall prices.

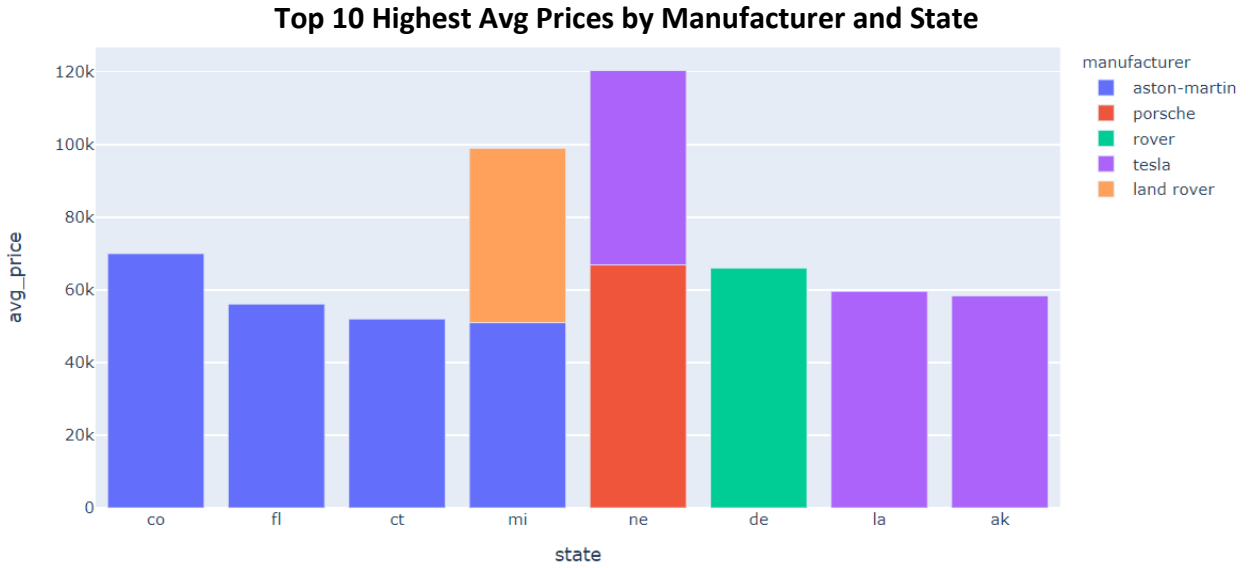

Figure 7: Top 10 Highest Average Price by Manufacturer and State

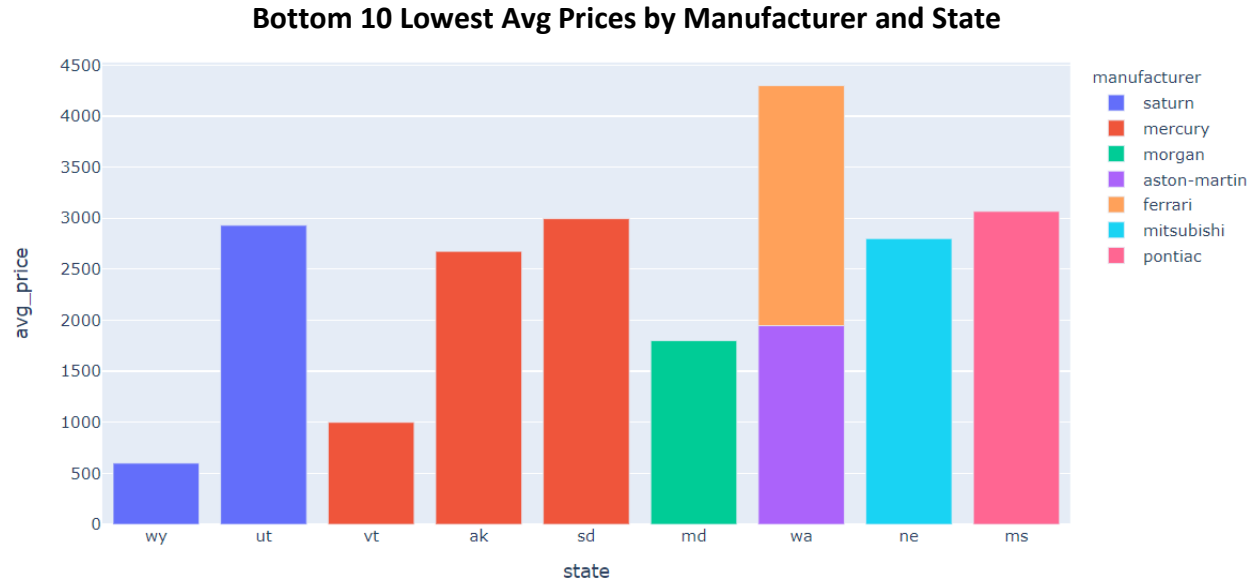

Figure 8: Bottom 10 Lowest Average Price by Manufacturer and State

Oddly enough both Top 10 and Bottom 10 contain luxury of expensive vehicles. The bottom 10 probably represent bargains that might need some fixing up like the Ferrari and Aston-Martin.

Test/Training the Model:

Testing and training model I ran a correlation matrix as seen below in Figure 9. Notice there are only three Features that correlated. If the pricing only depended on these it would be

simple. I decided additional Features were needed. I Feature Engineered cylinder count and vehicle condition and competed another Correlation Matrix shown in Figure 10.

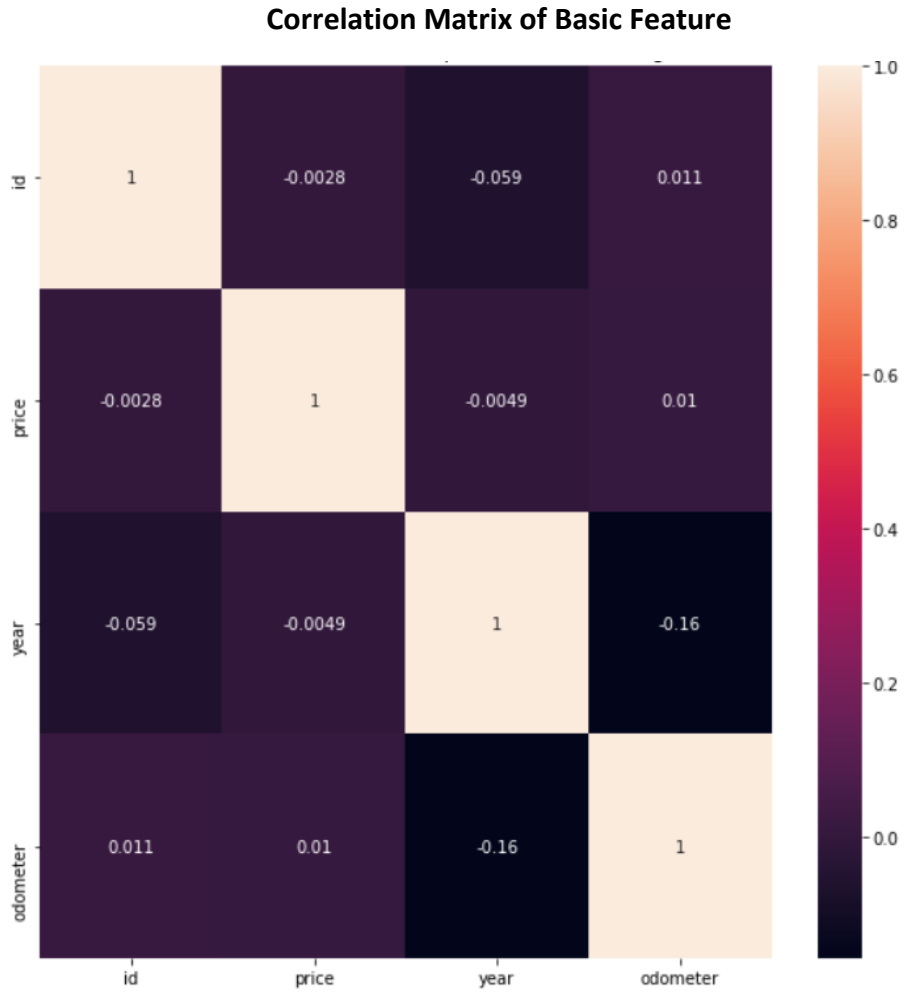

Figure 9: Correlation Matrix of basic Features

Adding cylinder counts and condition as Features of the vehicle as those good contributing factors to the value of a car based on Kelly Blue Book. This increased the number of columns in the dataframe to a 16 from four.

Correlation Matrix of All Feature

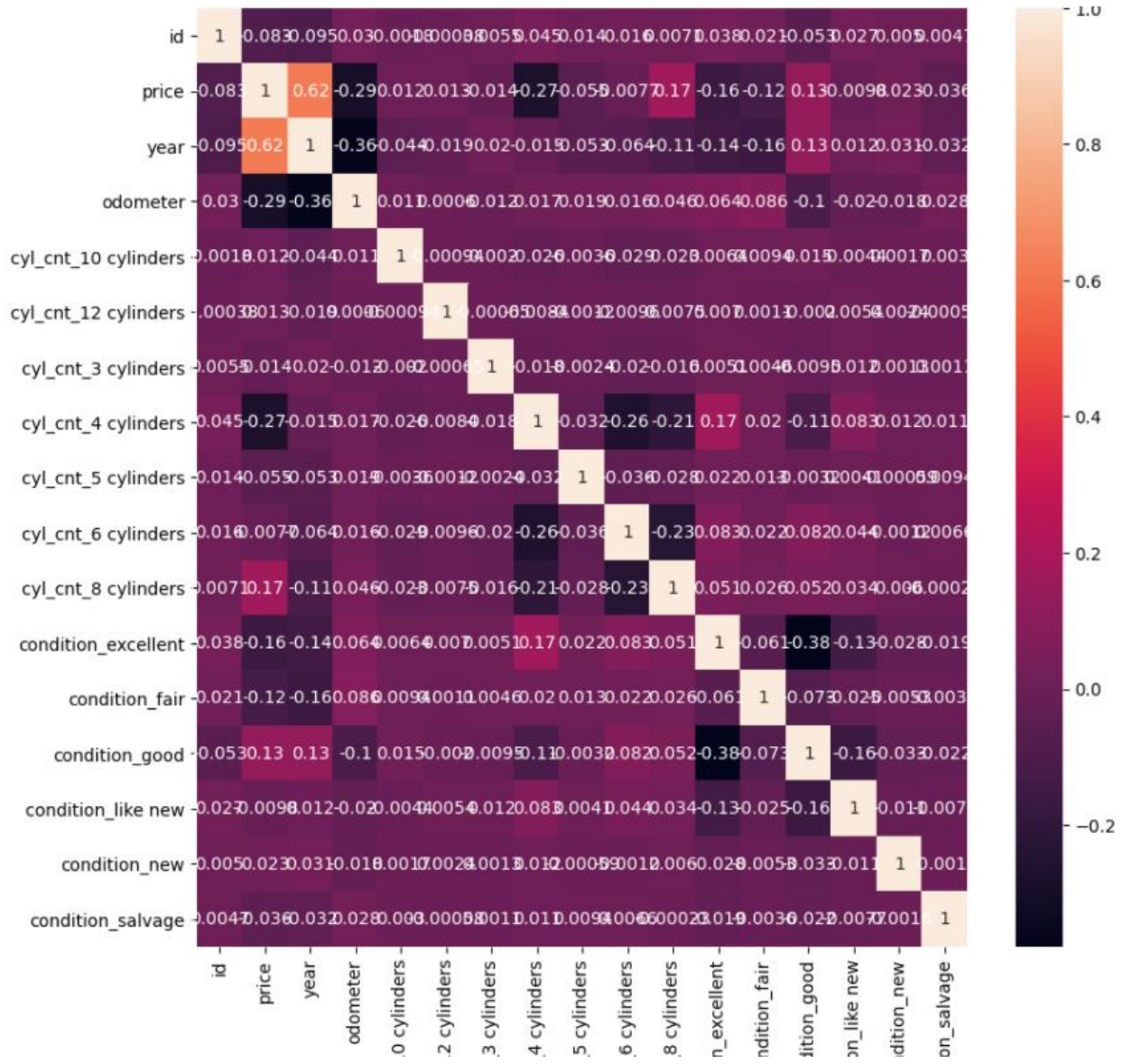

Figure 10: Correlation Matrix after adding cylinder count and condition

Next I split the data into training and test. I took a simple 80% Train and 20% Test as show in Figure 11.

```
# split out X and y

X = df_model_ready[[ 'year', 'odometer', 'cyl_cnt_10 cylinders',
       'cyl_cnt_12 cylinders', 'cyl_cnt_3 cylinders', 'cyl_cnt_4 cylinders',
       'cyl_cnt_5 cylinders', 'cyl_cnt_6 cylinders', 'cyl_cnt_8 cylinders', 'condition_excellent',
       'condition_fair', 'condition_good', 'condition_like new',
       'condition_new', 'condition_salvage']]

y = df_model_ready['price']
```

```
# split the data in training and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Figure 11: Splitting the data into Training and Testing

Model Evolution:

The dependent variable is price. We want determine how close a predictive model can come to predicting future prices of vehicles based on current values seen on Craigslist. I first ran a Linear Regression model to see what results could be seen.

The Linear Regression model returned a good respectable R^2 of 0.5012 or 50% and RMSE of 9505.

```
Test data - R-square  0.501222499073551
Test data - mean squared error  90346883.48178868
Test root mean square error (RMSE): 9505.09776287381
```

Figure 12: Evolution of Linear Regression Model

Line Graph of Linear Predictive Model Actuals to Predicted Values

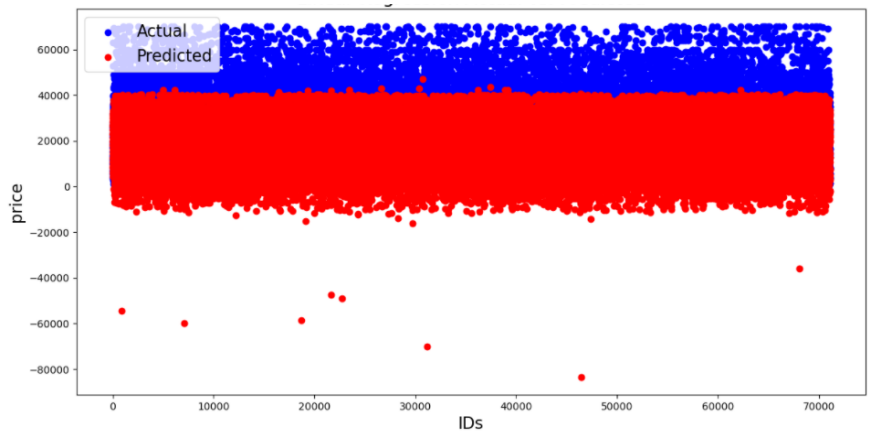

Figure 13: Linear Regression Model Line Chart

The next model I decided to evaluate was Decision Tree Regressor model. This result was better than expected. As seen in Figure 14 below.

```
Test data - R-square  0.7216507667171639
Test data - mean squared error  50419246.45746601
Test root mean square error (RMSE): 7100.651129119498
```

Figure 14: Evolution of Decision Tree Regressor Model

An increase in R^2 from 0.5012 to 0.7216. That changed 0.2204 from the Linear model to the Decision Tree model. Below in Figure 15 is the line graph showing the outcome.

Line Graph of Decision Tree Predictive Model Actuals to Predicted Values

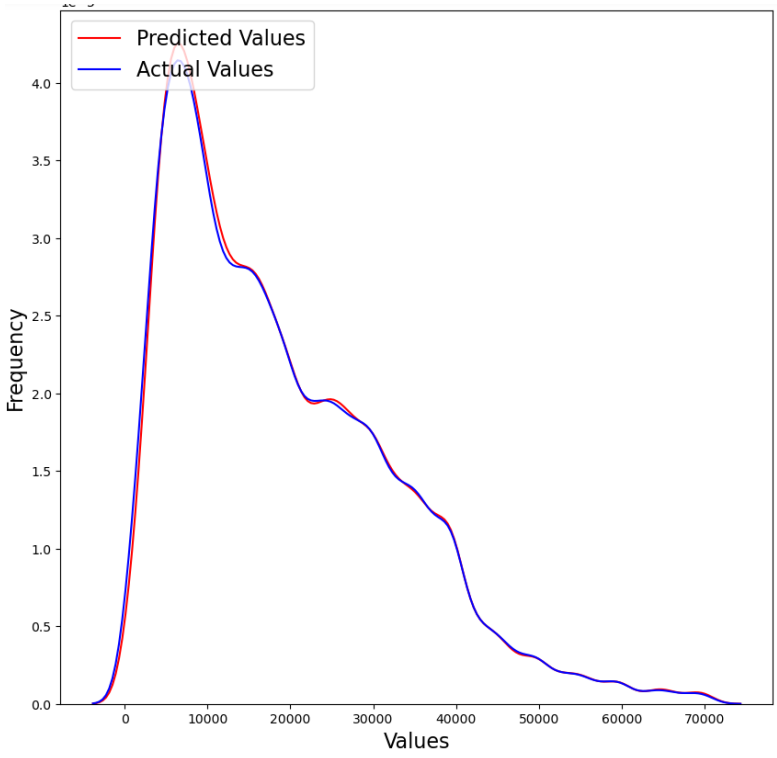

Figure 15: Linear Regression Model Line Chart

Conclusion:

The goal of this project is to predict sale price of used vehicles on Craigslist, while exploring the dataset and discovery valuable insight from the data. I started by extracting the data using Panda library. The data contained over 400,000 rows. After doing some exploratory data analysis I realized the data contained anomalies like vehicles offered for sale for a dollar. Outliers that appeared to have very high prices. I filter the data down to cars between 500 dollars and 70,000 dollars. Since the Mean was about 57,000 I felt this would get the correct amount of data and focus the analysis on the more common users of the site.

Once narrowing the data, the number of correlated features was limited to two features. I decided to perform Feature Engineering to create cylinder indicator and condition indicator. This increased the count of correlated Features. I split the data between training and test doing a 70/20 split. I ran a Linear Regression model and Decision Tree Regressor model.

The Decision Tree Regressor model showed to be the most promising. It produced a R^2 value of 72%. I evaluated each model using different evaluation methods. I got a decent 72% R-square value using Decision Tree model. I believe that can be improve by adding additional Features like Manufacturer and splitting out private party and dealership listings.

This project was pretty insightful. I recently sold one of my vehicles on Craigslist and I based my price on a few factors. Using Kelly Blue Book, offline dealership appraisals and online offers from places like CarMax, Carvana and Varoom. While I was able to get 28% better price than what the highest offer made by the online retailers. I think I would have maybe gotten a few hundred dollars more by knowing what I could expect from running my manufacturer, model and year.

Reference:

1. Reese, A. (June 2021). "Used Cars Dataset" Version 10. From Kaggle.com. https://www.kaggle.com/austinreese/craigslist-carstrucks-data/version/10
2. "Craigslist". (February 2022). From Wikipedia.org. https://en.wikipedia.org/wiki/Craigslist
3. Hodge, L (January 2022). "Welcome to Used Car Buying Hell". From Jalopnik.com https://jalopnik.com/welcome-to-used-car-buying-hell-1848381994
4. "13-Month Rolling Used-Vehicle SAAR". (January 2022). From Cox Automotive.com https://www.coxautoinc.com/market-insights/cox-automotive-13-month-rolling-used-vehicle-saar/
5. "Most Popular Sites to Buy and Sell Used Cars". (August 2021). From preownedautologistics.com. https://www.preownedautologistics.com/blog/most-popular-sites-to-buy-and-sell-used-cars/
6. Shen, M. (February 2022). "Used cars cost 40.5% more than last year as gas prices rise. New car prices also climbing". From USAToday.com. https://www.usatoday.com/story/money/cars/2022/02/13/used-cars-cost-more/6778705001/
7. "Car Values". (2022). From Kelly Blue Book (kbb.com). https://www.kbb.com/car-values/

Appendix:

Vehicle Table

Column Name	Column Description
id	A unique number for the listing
url	A unique url for the listing
region	Region in which the listing was posted. Contains city name in most cases
region_url	Region the listing was posted within
price	Asking pricing of the vehicle
year	Year of the vehicle for sale
manufacturer	Manufacturer of the vehicle
model	Type of model the vehicle is
condition	What condition the vehicle is in
cylinders	The number of cylinders of the internal combustion engine
fuel	The type of fuel required
odometer	The milage that appears on the odometer of the vehicle
title_status	The current titling of the vehicle
transmission	The type of transmission in the vehicle
VIN	The unique number call the vehicle identification number that identifies the vehicle within the DMV systems and to the manufacturer
drive	The placement of the drive wheels
size	The class size of the vehicle
type	The type of vehicle
paint_color	The color of the vehicle
image_url	URL of the images provided by seller
description	The description of the vehicle that is for sale
county	The county in which the vehicle is located
state	The state in which the vehicle is located
lat	Short for latitude coordinate of the vehicle location
long	Short for longitude coordinate of the vehicle location
posting_date	The date the vehicle was posted on Criagslist

Questions:

1. Why was the data limited to two months?
2. There a column to determine dealership listing compared to private party listings?
3. Determine what the final selling prices of the vehicle was?
4. Can this data we used to determine where used car prices are headed?
5. Can sell date be determined?
6. Is the price impacted by location or region?
7. Since Craigslist charges $5.00 to list a car how much revenue does Craigslist earn?
8. Can you determine what attributes cause a car to sell faster than another car of the same make and model?
9. What is the limitation of this type of analysis?
10. What additional insights can be gained from this data?