TEAM 36 – Final Report

Where Did All the Cheap Cars Go?

Authors: Shawn Azzu, Andrew Bartels, Ngoc Nguyen , Matt Palmer, Jason Young

Introduction and Background:

Problem Statement

"A devastating shortage of microchips — which are necessary for all manner of critical electrical components — is slowing car production worldwide, choking the supply of new models and driving their prices skyward. High dealer markups and a lack of options are forcing more buyers to shop secondhand, chipping away at used-car inventories" - citation

Objective

Our company, Eagle Fang Data Consulting, has been hired by Penske Motor group, LLC to investigate potential arbitrage opportunities in the used car market. The objective of this project is to create a model to predict automobile prices in different regions of the country and build an app that allows used car managers to quickly compare prices. To supplement our data set and improve the accuracy of our model, Penske Motor Group has also asked us to use regional economic data to strengthen our predictions.

Data, Preparation, & Exploration:

Datasets

We have selected three datasets to train our model. Our primary dataset consists of used car data that was scraped from Craigslist during the months of April and May in 2021. This dataset contains 426,880 observations and 22 explanatory variables including craigslist id, image URL, year, make, model, region, state, odometer, cylinders, fuel type, title status, transmission type, drive type, condition, paint color, description, latitude, longitude, VIN, posting date, car type, as well as our dependent variable, price.

The complementary datasets that we combined with our car table include economic data pulled from the 2020 census as well as outdoor recreation data by state from 2021. The Census Data contains 3 useful variables including median family income, median non-family income, and ZCTA code which we can map to cities and use for joining. Finally, the outdoor recreation dataset contains two useful variables including total outdoor recreation employment and total outdoor compensation which we can map using the state variable in our car's dataset. Furthermore, we also created new variables with our existing ones to explore new and/or different relationships to implement into our models, such as car age.

Data Preparation and Cleaning

Data wrangling is the process of cleaning and transforming data for usability, modeling, and analysis.

Figure 1. Data Cleaning

For the first step of the cleaning process, we wrote the tables to an SQLite3 database for lightweight storage and efficient use between team members. Next, we created a preprocessing notebook to begin cleaning up the car dataset which contained 1,688,231 NA values and several variables that we do not intend to use. The variables we decided to immediately drop from the analysis included id, image URL, county, latitude, and longitude. Craigslist allows for a lot of flexibility in how a user inputs data when creating a vehicle listing and this presented a significant challenge for our data preparation. One of the most difficult challenges we faced when preprocessing the data was figuring out how to deal with all the missing values for the manufacturer and model variables. To solve this problem, we created a regex pattern of all known manufacturers and used a string replacement function to extract vehicle manufacturer names from the description variable which allowed us to recover 11,975 datapoints that would have otherwise been dropped from the dataset.

Here is a summary of NA values we were able to recover:

Variable	# NA Values	% of Observations	Amt. Recovered
Year	1,205	28%	1,184
Manufacturer	17,646	4.13%	11,975
Paint color	130,203	30.5%	37,758

Variable	# NA Values	% of Observations	Amt. Recovered
Type	92,858	21.75%	23,214
Drive	130,567	30.59%	12,998

Figure 2. Summary of NA values before and after recovery

For the remaining NA values, we have decided to drop them from the dataset. The reason for this decision is because many of the observations contain 5 or more variables with NA values and we do not believe using methods of imputation like mean or mode will improve the accuracy of our model. With that said, if we were looking for ways to improve the accuracy of our model in future analysis, this might be an area that we could revisit.

Another interesting finding from the initial cleanup of the craigslist car dataset was how many vehicles had extremely low odometer readings relative to their age. As an example, there were 8,599 vehicles older than 2019 that had odometer readings less than or equal to 50. Observations like these can create unwanted leverage points that will impact our model's ability to make accurate predictions, so we have created a custom function that can help us quickly filter out this data. In addition, other steps we have taken to clean

this dataset include casting datatypes, stripping punctuation, and creating factor levels for categorical variables.

For the Census Data, we were able to extract the zip code from the ZCTA code variable and map the zip codes to another table of city names. We also removed unwanted punctuation characters and dropped unwanted variables. The two variables that we decided to merge with our cars dataset are Median Family Income and Median Non-Family Income at the city level.

The Outdoor Economic Dataset required only minimal cleaning. We changed datatypes and converted the State variable to uppercase so we can join with the other tables.

APPROACH & METHODOLOGY:

Our company sets out to compare a variety of multiple variable regression models to measure the amount of influence each independent variable has on used cars' prices and analyze major determinants of resale value in used cars. As discussed in the articles (Citation), there are many segments that affect a car's value, and some are beyond your control:

- **Vehicle age:** The older the age of the car, the more depreciated it is, the less it is worth.
- **Odometer:** A used car's value mostly hinges on how many unused miles it has left.
- **Brand Name**: Slow depreciation for brands that have long-standing for reliability and durability.
- **Luxury vehicles**: Luxury brands charge more for their vehicles, but that does not necessarily mean they will hold a greater percentage of their value than a lower-priced vehicle.
- **Condition**: The cleaner and better functionality a used vehicle is, the less reconditioning it will require – hence, the more it is worth.
- **Service records**: Proof of regular maintenance, i.e., receipts of oil changes, tire rotations, fluid flushes can also boost resale value
- **Accidents or damage**: A vehicle that has never been in an accident will be worth more than one that has.
- **Number of owners**: The fewer owners a vehicle has had, the less it will depreciate
- **Fleet vehicles**: A one-owner, privately owned vehicle tends to take care of one's car than a rental car. Thus, a used car's value will be worth more
- **Features and current technology**: Needless to say, the more technologies and features, the more it is worth.

Based on the article summary above, we conducted Exploratory Data Analysis (EDA) for quantitative variables, i.e., year, odometer, cylinders, etc. and other qualitative variables, such as manufacturer and market segments, which can be measured using dummy variables. For example, for quantitative variables, we added the "age" column to calculate vehicle age, i.e., 2021-year, since the dataset is originally collected up to 2021.

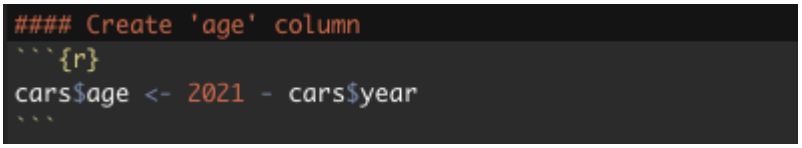

Figure 3. Quantitative variables

For qualitative variables, such as state, manufacturer, model, fuel, transmission, paint color, drive, title status, etc., we converted them to the factor data structure for multiple regression model use purpose.

```r
cars$state <- cars$state %>% factor()
cars$manufacturer <- cars$manufacturer %>% factor()
cars$fuel <- cars$fuel %>% factor()
cars$transmission <- cars$transmission %>% factor()
cars$type <- cars$type %>% factor()
cars$paint_color <- cars$paint_color %>% factor()
cars$drive <- cars$drive %>% factor()
cars$title_status <- cars$title_status %>% factor()
```

Figure 4. Qualitative variables

By utilizing correlation matrix, we want to study relationships and confirm there is no collinearity existing amongst these variables (See Figure 5 below)

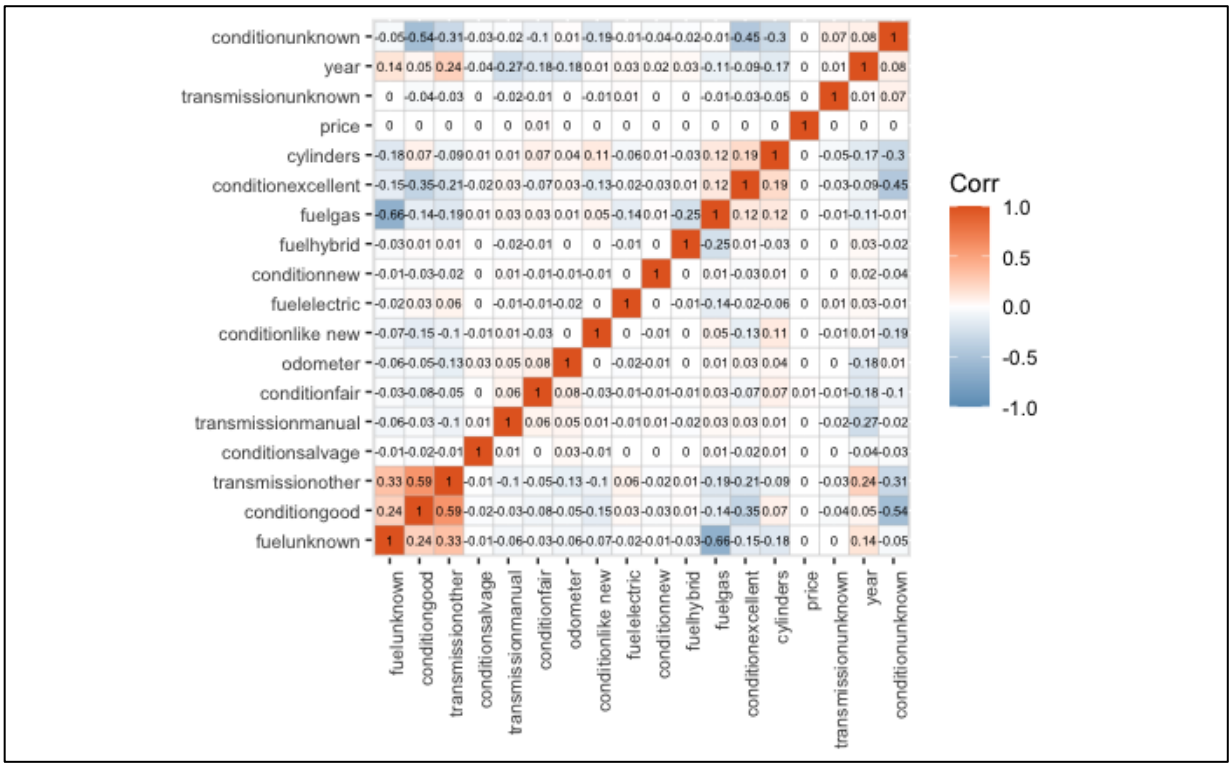

Figure 5. Initial Correlation Matrix on entire original Cars Dataset

To continue EDA, we decided to use location variables, i.e., region, city, state, zip code to check where and how the used car is available for sale in the USA. According to an article we found (Citation), different weather varies throughout the US regions and has a direct impact on your vehicle's value, i.e., sunlight will rob you of your car's shine, rain initiates rust, and the presence of snow can lead to corrosion from road salt. Thus, we expect to see price distribution of various car makes/models/trims differ across regions.

According to this article (Citation), transportation expenses count as the second biggest expense in US households, running as much as **10 to 20%** of the household incomes. Both globally and in the United States specifically, the relative strength of the economy, i.e., unemployment rate, state GDP (Gross

Domestic Product), or consumer spending habits, indirectly influence cars' price value and the overall state of the used-car market. Thus, we are joining Outdoor Economic and vehicles dataset together to study this relationship.

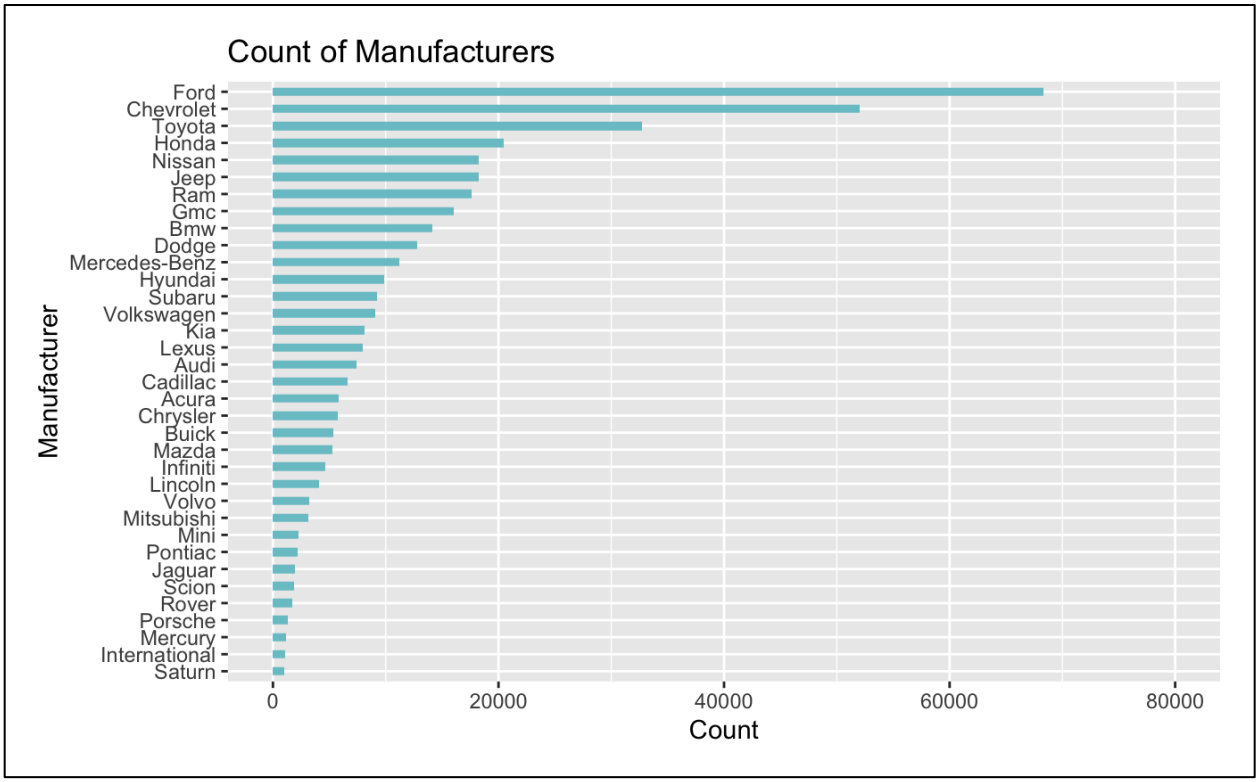

Figure 6. Market Share of Different Car Manufacturers

Through our exploration of the data, we realized that creating an accurate price prediction without using the vehicle make and model variables was going to be difficult. We decided to focus our efforts on building models using the following manufacturers: Ford, Chevrolet, Toyota, Honda, and Ram. Thus, we are focusing on these top manufacturers for machine learning and analysis where we want to train a model that can give us the estimated price given the specifications.

For our analysis, we decided to focus on the manufacturer "Ford" to keep consistency and simplicity of interpretation:

```
> summary(ford)
     city              state           manufacturer          model              year      
 Length:43170       Length:43170       Length:43170       Length:43170       Min.   :1915  
 Class :character   Class :character   Class :character   Class :character   1st Qu.:2009  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :2014  
                                                                             Mean   :2012  
                                                                             3rd Qu.:2017  
                                                                             Max.   :2022  
      age            condition          cylinders             fuel             odometer       
 Min.   : -1.000   Length:43170       Length:43170       Length:43170       Min.   :       0  
 1st Qu.:  4.000   Class :character   Class :character   Class :character   1st Qu.:   48182  
 Median :  7.000   Mode  :character   Mode  :character   Mode  :character   Median :   97835  
 Mean   :  9.336                                                           Mean   :  107275  
 3rd Qu.: 12.000                                                           3rd Qu.:  143567  
 Max.   :106.000                                                           Max.   :10000000  
 title_status       transmission         drive              type           paint_color       
 Length:43170       Length:43170       Length:43170       Length:43170       Length:43170       
 Class :character   Class :character   Class :character   Class :character   Class :character   
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   


     price       med_family_income med_non_family_income
 Min.   :    0   Min.   : 40218    Min.   :16445        
 1st Qu.: 6980   1st Qu.: 61986    1st Qu.:33129        
 Median :15999   Median : 71336    Median :37179        
 Mean   :20068   Mean   : 73344    Mean   :38572        
 3rd Qu.:29999   3rd Qu.: 80263    3rd Qu.:41995        
 Max.   :99700   Max.   :137864    Max.   :78362        
```

Figure 7. Ford Dataset Summary

As a part of focusing on the Ford Data, we also thought it would be worthwhile to explore the relationship between our predictor (Price) and our independent variables once again as the correlation matrix above (see figure 5) surprised us quite a bit. We initially expected parameters, such as odometer and year, to be highly correlated to the price of the cars, but that was not shown to be the case when examining the entire dataset. As we can see in correlation matrix of narrower dataset which is Ford (Figure 8), a few linear relations are clearly visible like car price and the year when the car was manufactured. This finding, in fact, makes sense. For the same given model/brand, a car manufactured in a later year is expected to be in better condition and could have more technologically built-in features than the ones manufactured in previous years, its price would be higher as a result.

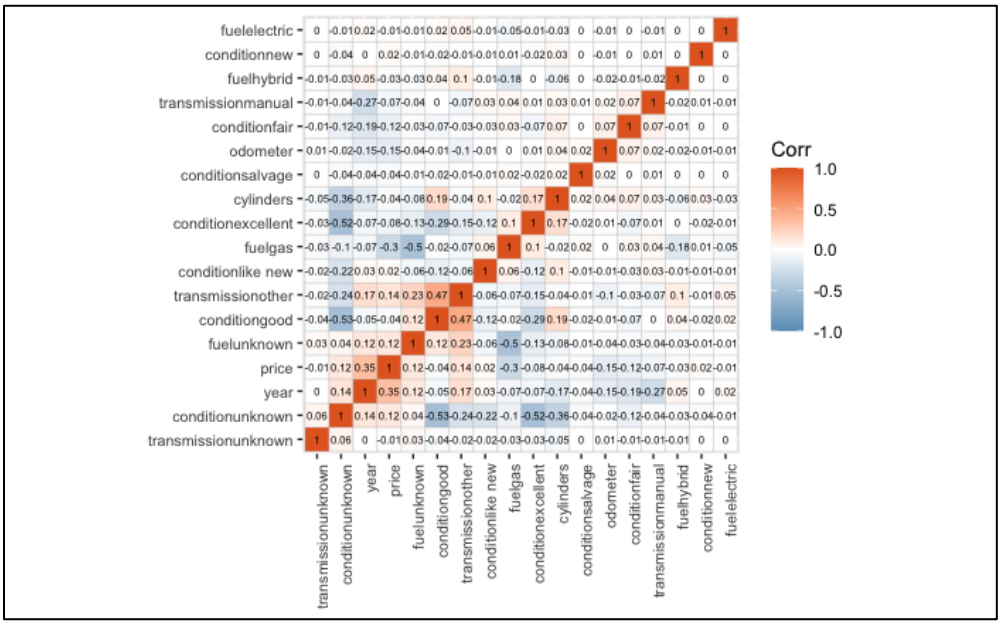

Figure 8. Correlation Matrix on Ford Dataset

Application

Eagle Fang Data Consulting set out on a mission to make an easy-to-use and reliable tool for Penske Motor Group, LLC that would allow them to visualize the data they need, discover rare and arbitrage opportunities in the market, and see fair value for their car purchases. Our team made this tool in the form of an R Shiny Application.

This tool will first allow our clients to inspect the data they desire:

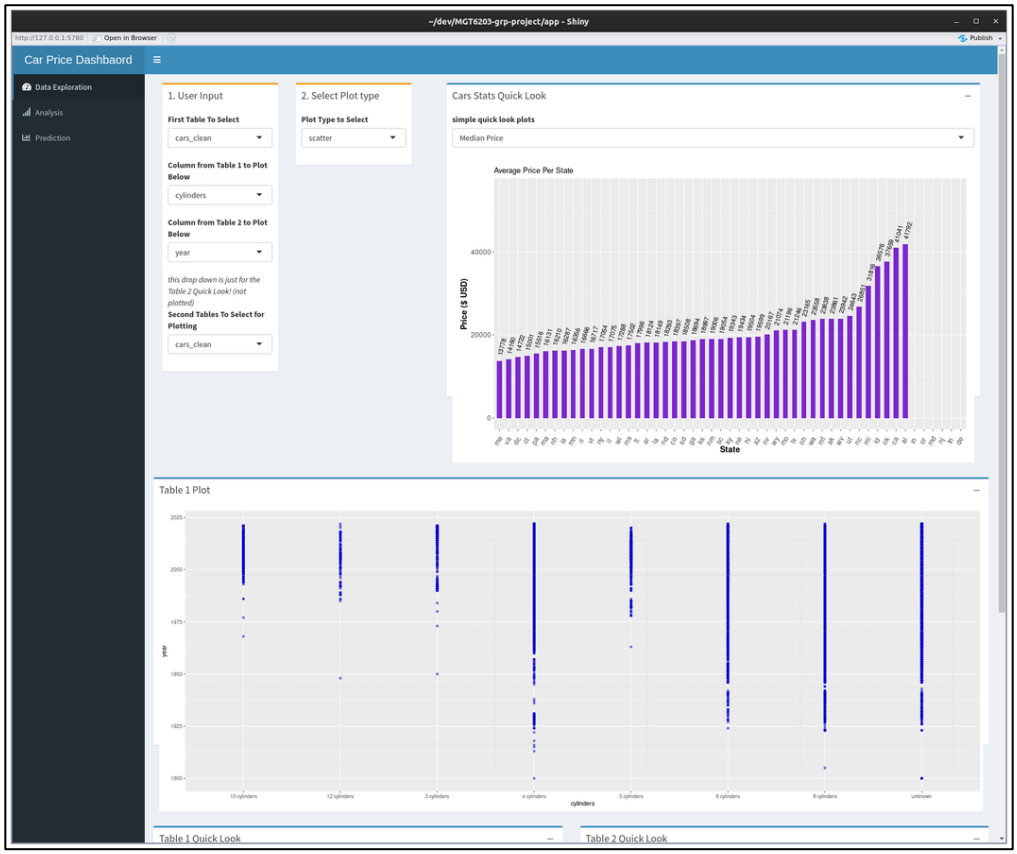

Figure 9. Data Exploration Tab

Here, the user can select their desired input, starting with the desired data. Above, we display the user selecting the table cars_clean, but this can also be the Ford, Chevy, Toyota, or other manufacturers. Further, they select the desired parameters they want to compare against the price. This results in the Average Price Per State Table shown above. They also can select a column from the table to compare against the first parameter, here we show the year number of Cylinders against the year; they choose which Plot Type they would like. This results in the “Table 1 Plot”, a Scatter Plot showing outliers and general quartiles the data lies upon.

Other interesting and useful plots they can see here in this first tab are the Count of Cars Per Manufacturer (see figure 6), the Average Price Per Manufacturer, and the Box Plot of desired variables.

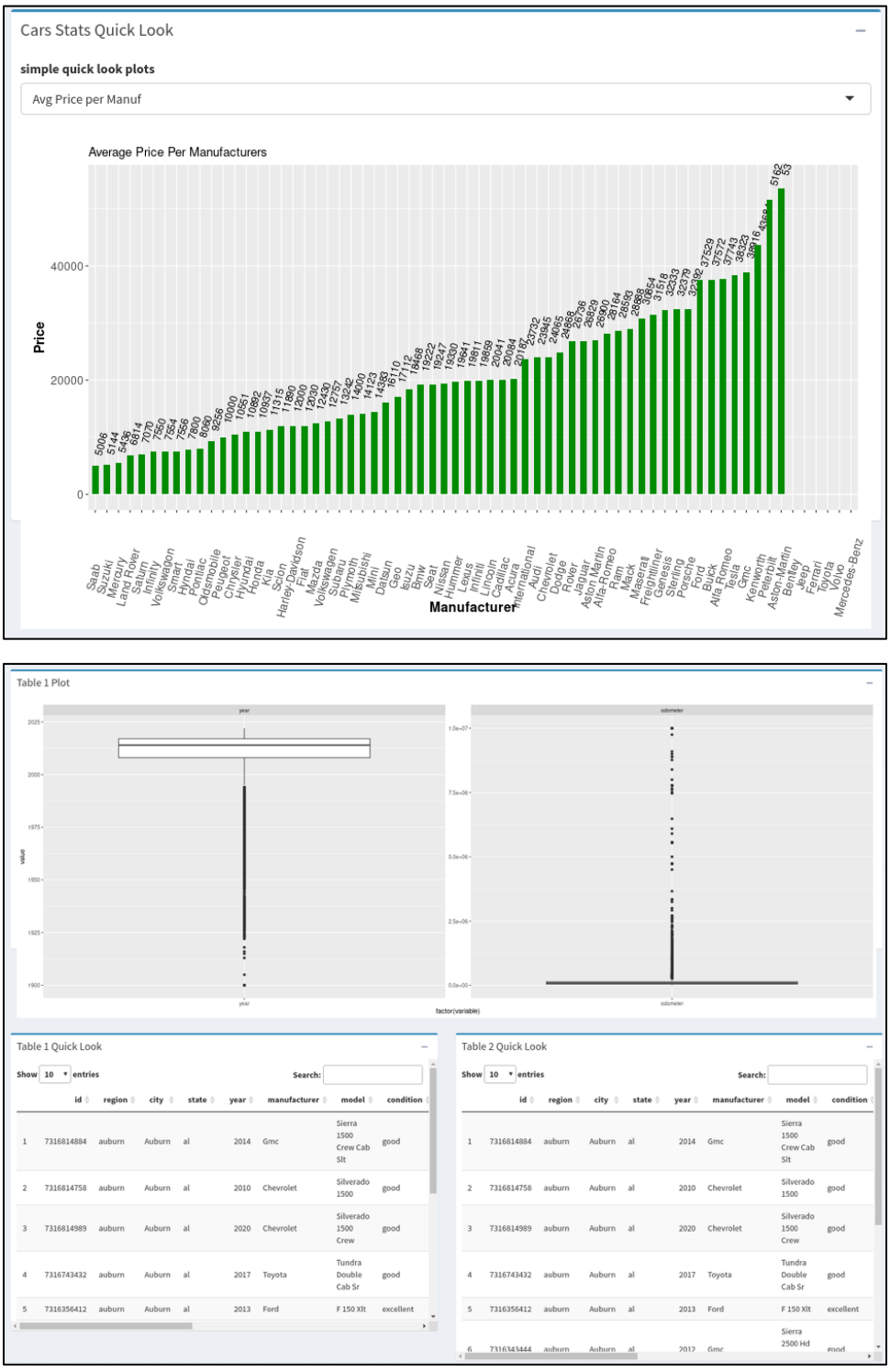

Figure 10. Data Exploration Tab Other Charts

This data exploration tab dovetails nicely with our Analysis Tab. Here the user can dive deeper into the data, selecting a specific manufacturer to see their model lineup and see medians, quartiles (25, 50, 100), and positive and negative outliers all against price in a colorful and vibrant Box Plot, see figure below.

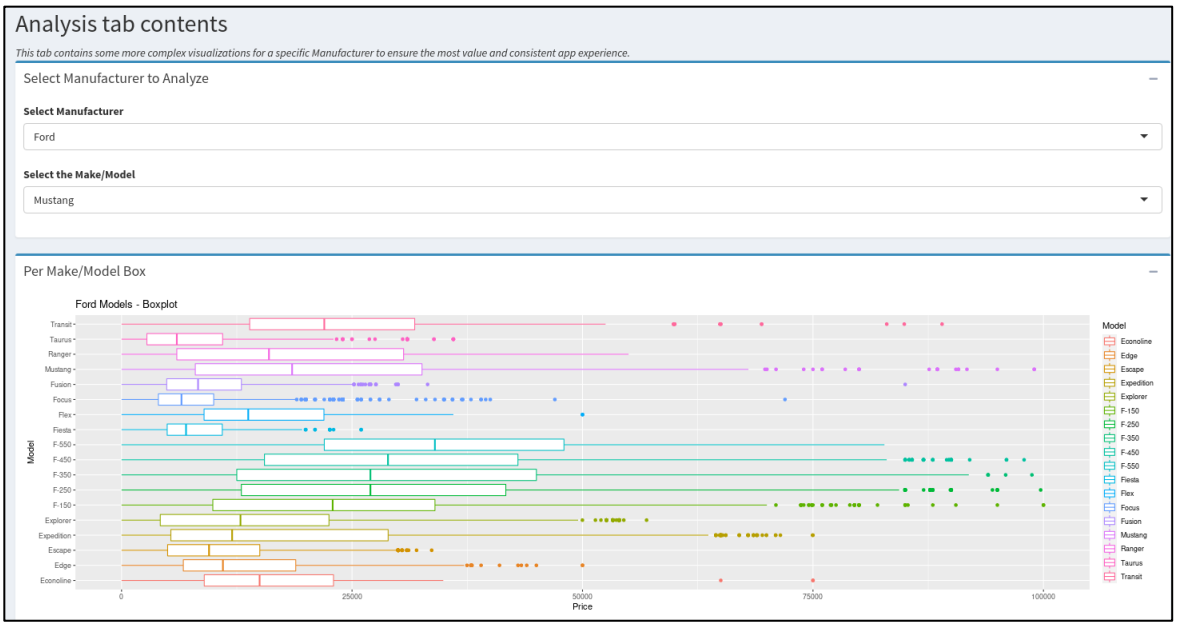

Figure 11. Analysis Page, Box Plot

Furthermore, it is here in the analysis Tab that the User can get their first taste of the power behind our application. We allow the user to Select their Manufacturer, the Make/Model, and the year to generate a radar chart displaying price discrepancies based on region.

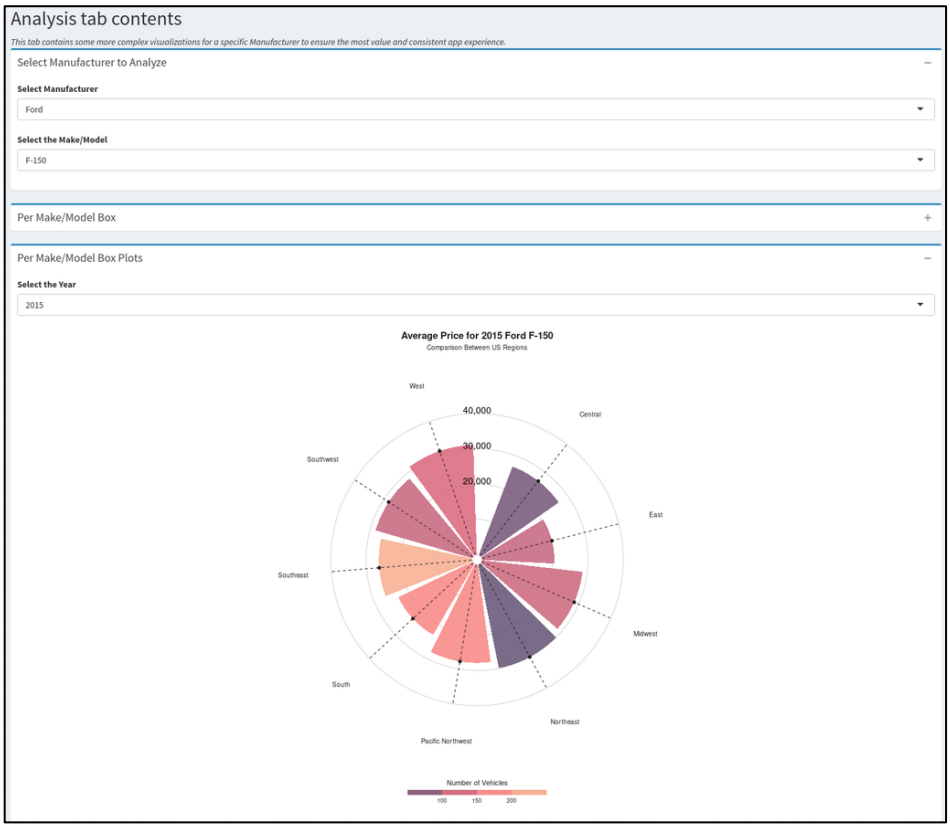

Figure 12. Analysis Page, Radar Chart

The application puts this all together in our Predictions Tab (Discussed Below).

Results

Regression Models: Linear and Linear-Log

To compare a variety of regression models, we wrote a custom function in R called Model_Comparison(), which does the following:

1) Accepts a data frame and any number of model-creation functions as arguments
2) Splits the data frame into 3 sets: training (60%), validation (20%), and testing (20%)
3) Trains each of the models on the training set
4) Uses each of the trained models to predict the price for the cars in validation set and compares the results to the actual price of those cars
5) Based on the results of 4) above, the function will indicate which of the provided models has the best accuracy, as measured by its R-squared value
6) For the selected model, the accuracy is tested again on the test set, giving a final measure of the quality of the model
7) Returns all results in a neat table format as shown below:

Sample output of Model_Comparison() function

	Model <chr>	R^2 on Validation Data <chr>	Best Model? <chr>	R^2 on Test Data <chr>
1	Model 1	0.272363884414753	0	*NA*
2	Model 2	0.369859962701729	0	*NA*
3	Model 3	0.383760232034281	1	0.389529506010875

With the Model_Comparison() function in place, it was easy to test a variety of regression models and compare their accuracy as measured by R-squared.

The table below will show the results for the following 4 models:

1) Standard LM (car info only) - This model serves as the base case. Prior to adding any location data or economic data to our fields, we create a model that only uses information about the vehicles themselves. We would expect this model to be the least accurate
2) Standard LM plus local economic data – This model adds our unique data sets, including state median incomes from the census dataset and outdoor recreation economic data from the U.S. Bureau of Economic Analysis. However, this model remains simple, with no transformations.
3) Log Age (All Columns) - Keeping the same prediction variables, we next log-transform each car's age. The non-linearity of the impact of a car's age on price was discovered through research articles and confirmed in our dataset as shown below:

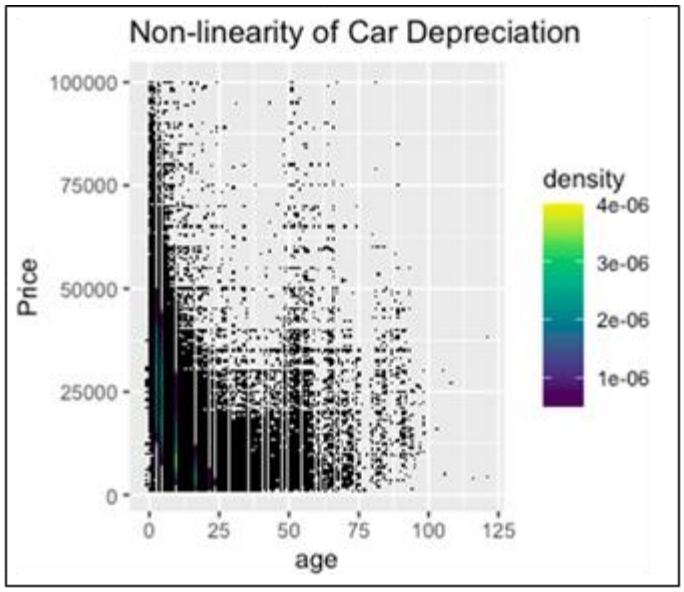

Figure 3. Non-linear Car Depreciation Scatter Plot

4) Log Age Log Odometer (All Columns) - Inspired by the success of the log-transformation in model 3), we also log-transformed the odometer field, since this predictor is likely to follow a similar influence on Price

After training, validating, and testing the above 4 models with the Model_Comparison() function, we get the following results:

	Model <chr>	R^2 on Validation Data <chr>	Best Model? <chr>	R^2 on Test Data <chr>
1	Standard LM (car info only)	0.423	0	*NA*
2	Standard LM plus local economic data	0.459	0	*NA*
3	Log Age (All Columns)	0.537	0	*NA*
4	Log Age Log Odometer (All Columns)	0.549	1	0.548

This progressive improvement in R-squared values above confirms our intuitions for improving the model incrementally at each step. Additionally, the selected best model performs nearly as well on the test set as it did the validation set.

K-nearest Neighbors (KNN) Regression Model

Beside multivariable regression model, we would like to study and analyze the result of K-nearest neighbors regression model where a data point's response is estimated based on the responses of the k nearest data points with known response (See Figure 11 for example). Once the neighbors are found, the algorithm will calculate the average value of all the neighbors, for our example below, the new data point's predicted house price would be sum of all the neighbors points divided by 5 equal $986K as the result. As you can see from the example, KNN regression model could be very intuitive algorithm and easy to explain how the predictions were made.

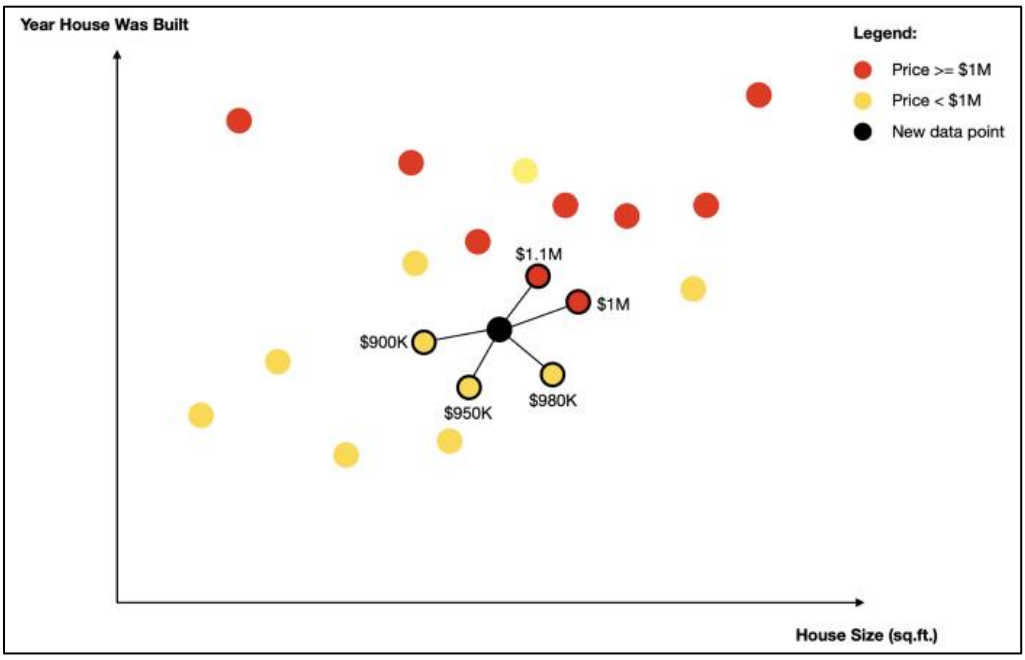

Figure 11. Example of KKNN Regression

We will use the following data fields for our initial models:

- **Predictor variables:** State, condition, age, odometer, drive, cylinders, med_family_income, med_non_family_income
- **Response variables:** Price

Our goal is to predict the price for a car given user inputs' criteria. We split the data into 80% for training and 20% for testing.

Since initial model are taking more variables and consumes a lot of time to run, it would be not efficient and could possibly overfit data. We are approaching the method of filtering the dataset by using different given user inputs' criteria before training model. By doing so, we can reduce predictor variables to only "year" and "odometer". With the narrower dataset , the benefit of this approach is to enhance running time and accuracy. However, the downside is we could underfit the data.

```
##-------------------------------------
##  Test Model_Prediction Function  --
##-------------------------------------

State_Model_Prediction_KNNReg(cars, "CA", "Sacramento", "Ford", "F-150", 2011, 160524, "good", "4wd", "8")
cars %>% filter(state=='CA',manufacturer=='Ford',model=="F-150",city=="Sacramento",condition=='good',cylinders=="8",year==2011,odometer==160524)

##
[[2]]
2-nearest neighbor regression model

[[3]]
[1] 16995

> cars %>% filter(state=='CA',manufacturer=='Ford',model=="F-150",city=="Sacramento",condition=='good',cylinders=="8",year==2011,odometer==160524)
       city state manufacturer model year age condition cylinders    fuel odometer title_status transmission drive   type paint_color posting_date price
1 Sacramento    CA         Ford F-150 2011  10      good         8 unknown   160524        clean    automatic   4WD pickup      silver   1620025200 18995
  med_family_income med_non_family_income    region
1          81053.41                46366 Southwest
>
```

Figure 12. Example of KKNN Reg Predictions when k = 2

By using this approach, the smaller k would direct us closer to the neighbors since data were already filtered to meet the criteria. Taking the same example of KNN regression prediction of 2011 Ford card, model F-158 in Sacramento, CA with good condition, 8 cylinders, 4WD drive, we can see the actual price is $18,995. With K=2, the predicted price is $16,995 (see figure 12) which gives us approximately 89% accuracy.

With K=5, the predicted price is dropped to $12,197 (see figure 13) that yields approximately 64% accuracy

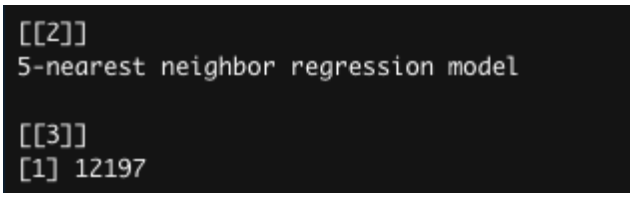

Figure 13. Example of KKNN Reg Predictions when k = 5

With K=10, the predicted price is dropped to $8,994.9 (see figure 14) that yields approximately 47% accuracy

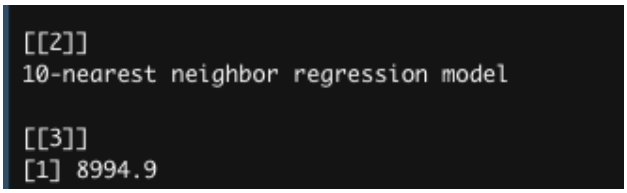

Figure 14. Example of KKNN Reg Predictions when k = 10

In general, we find it quite challenging to come up with the algorithm to detect optimal k at the time and knowing that our algorithm is not quite there yet. With that said, if we were looking for ways to improve the accuracy of our model in future analysis, this might be an area that we could revisit by collecting, studying, and researching data patterns for this model.

Random Forests Model

Another model that we wanted to look at and potentially implement into the application is a Random Forest Model. Random Forests is a supervised learning algorithm that creates many decision trees to answer the root problem, in this case, statistical significance to the prediction variable, price. The decision tree created below (please see figure x) takes price as the predictor and creates 500 decision trees. Having a larger number of trees lends credence to a decrease in variance.

```
> model = randomForest(price ~ ., data = ford,
+                      do.trace = TRUE)
     |      Out-of-bag   |
Tree |      MSE  %Var(y) |
   1 | 1.142e+08    40.27 |
   2 | 1.076e+08    37.95 |
                                  499 | 5.166e+07    18.22 |
                                  500 | 5.166e+07    18.22 |
```

```
> actual <- ford$price
> predicted <- unname(predict(model, ford))
> R2 <- 1 - (sum((actual-predicted)^2)/sum((actual-mean(actual))^2))
> R2
[1] 0.9496845
```

Figure 14. Random Forest Model

This model was run using the randomForest library, which allows us to run 500 Regression Decision trees on the Ford dataset with price as our predictor against all other variables. As shown the R-squared was manually calculated to compare against other models and received a score of about 0.95. In the above figure, we can see how the variance and the mean-squared error (MSE) decrease as the number of trees increases.

Predictions on the App

The final piece of our application is the Prediction Tab (see figure below).

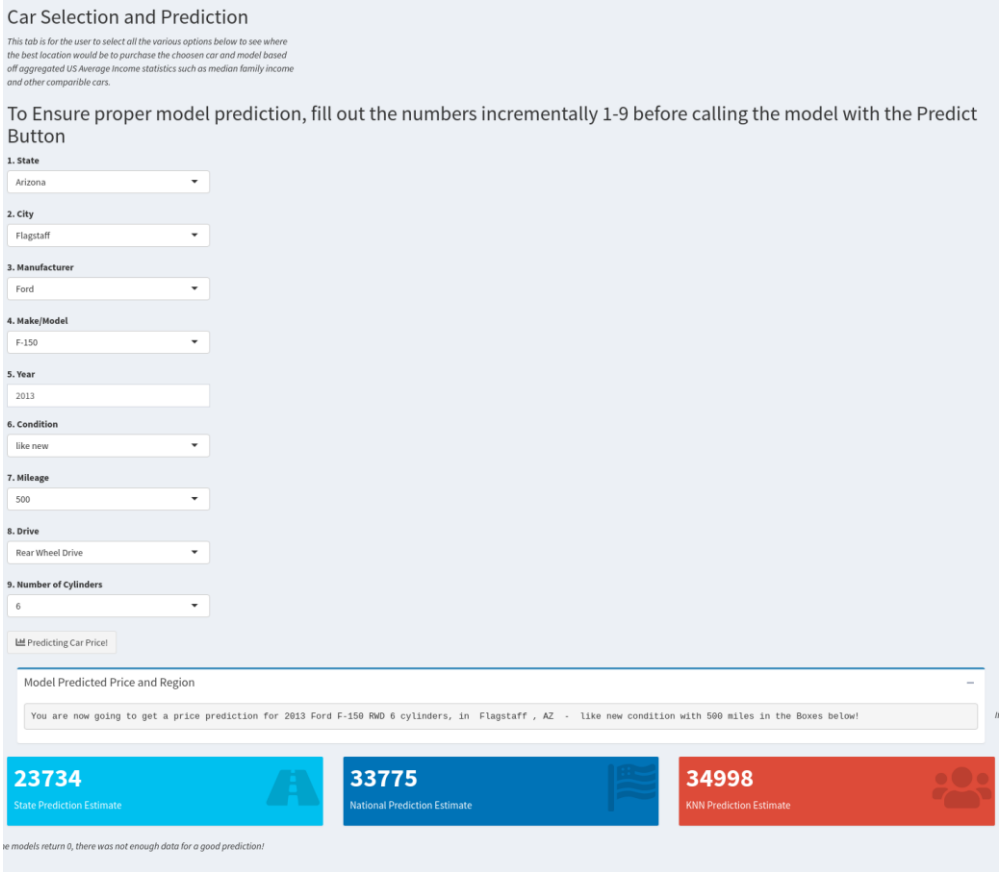

Figure 15. Sample Predictions

This tab allows for the user to make a prediction on a car's price based on a set of parameters, including State, City, Manufacturer, Make/Model, Year, Condition, Mileage, Drive, and Number of Cylinders. Upon picking the desired specs the user is prompted to press the "Predicting Car Price!" button and with the magic of analytics (our models) we display three different predictions. The first prediction utilizes our Linear Regression Model to estimate the car's price for the input state. The second prediction (middle blue) also utilizes our Linear Regression Model but instead considers data nationwide for the estimate. Finally, our third prediction (right, red) utilizes our KNN Model to make a prediction, also on data gathered nationwide.

Interpretation of Results:

Model Comparisons

As shown in our results we attempted to utilize a variety of models to find the best to use for our application. Altogether we used Linear and Linear Log Models (please see Results - Regression Models: Linear and Linear-Log section for full breakdown), KNN Regression Models, and Random Forests. Based on R-squared values alone the Random Forest came out on top with an R-squared of 0.95. However, the caveat here is that this model was the only one we were unable to train/test/validate as the processing power and compute time needed was unavailable to us, thus this model is very overfit. So, with that in mind, the next best model would be the Log Age and Log Odometer model which used all columns. This model had an R-squared of about 0.55 on the test and validation data, compare this to the Linear Regression Model which had an R-squared of about 0.42, which indicates that it would be the best model to use for making predictions. Nonetheless, the models we decided to implement into our application were the KNN Regression Model and the standard linear regression models as they were able to produce quick results and accurate predictions, keeping our application smooth, efficient, and reliable.

Interpretation of Predictions and Outcomes

Our results show that there are plenty of opportunities to take advantage of arbitrage markets for the Penske Motor Group, LLC. However, these opportunities are highly dependent on the specific/desired car make, model, and specifications. For the car market, it can be quite difficult to wrap a blanket statement saying a specific region has the best arbitrage opportunities available to take advantage of. We have found that there are many factors that go into the price of a car. We were surprised to see the great influence that economic conditions of regions could have on a car's price. In summation, the goal of producing a data exploration app for the Penske Motor Group was not just completed but exceeded every goal and metric for visualization and the ability for clients to explore these data. We have built an app that will enable our clients to easily identify and compare prices in this dataset, and future ones should more data be acquired.

-