Queens Apartment Price Prediction

Final project for Math 390 Data Science at Queens College

May 24, 2020

By:Tyron Samaroo

**Abstract-- Predicting the apartment price phenomenon seems like a simple task given the amount of data that can be collected. This is solved by finding those significant features that can truly estimate the predictive outcome. Once these features have been found. there is a need to correct any errors and handle any missingness. After doing this, different models like Random Forest, Regression Trees and Linear Model(OLS) are used and fine tuned to provide a useful model that can perform accurate real life predictions.**

**I. Introduction**

The phenomenon apartment sales price will be the response variable which we will call . One problem we have when predicting a phenomenon is that we do not know the true causal inputs and the true relationship between them. We can express this as a formula below.

The next best thing we can do is approximate the with measurable data which we can call and the relationship between them . Now the phenomenon can be expressed as follows.

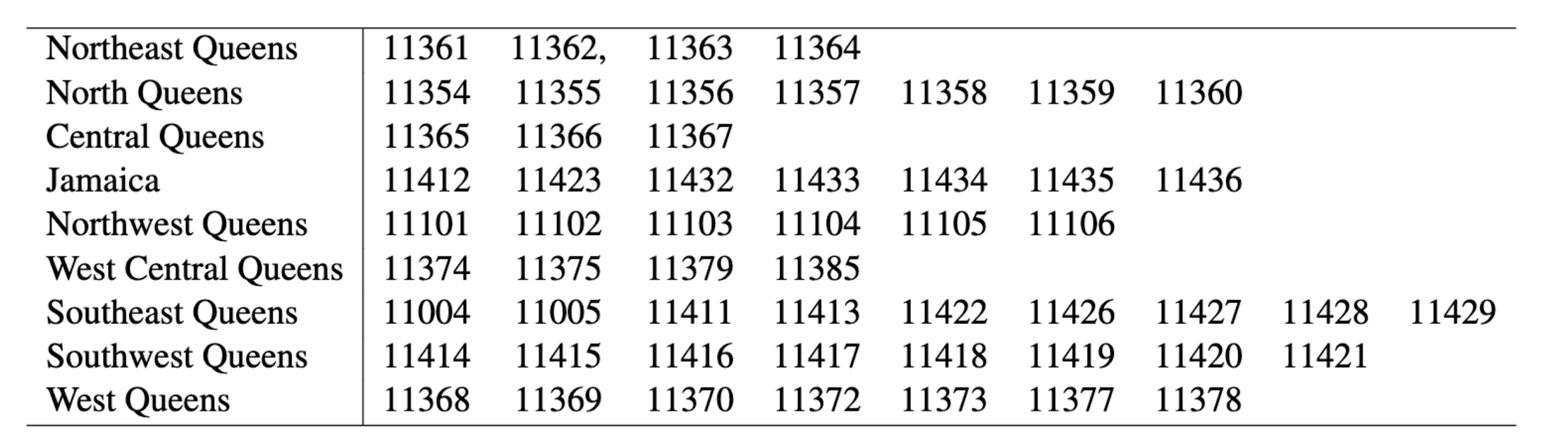
We can only approximate the phenomenon so we can include to be the error due to ignorance which we have no control over. This will be the predictive model.

A predictive model's job is to perform real world approximation on a given reality or a phenomenon..It is a model that is trying to find meaning to the phenomenon it is tasked with, by finding meaningful observable features. In this case predicting Queens apartment prices is the phenomenon that is being predicted based on observable features ) that is provided from the 2016 and 2017 apartment sales data in Queens, New York. But be wary a famous statistician George Box once stated that “All models are wrong but some are useful.” (Box, 1987). Geroge is warming many to not expect much from models, they only provide a way for humans to have a better understanding of a problem that is out of the scope of their knowledge.

In order to solve this problem three different modeling techniques such as linear modeling, regression trees modeling, and random forest modeling are used. Linear Model is tasked with finding meaning between each of the observed features and its influence on the apartment sales price. Regression Tree Model tries to find the best possible way to split the dataset D and construct a binary decision tree based on the given split. Random Forest extends bagging that performs a bootstrap sampling of the data. Random Forest Model computes many regression trees.While it computes these trees at each of their node construction, it splits on a subset of features in order to reduce correlation which also reduces variance.

**II**. **THE DATA**

The data used is Queens, New York housing data from 2016 to 2017 collected from MLSI. This dataset contains information from mainland Queens.

Figure 1. Zip codes from houses in the dataset

The dataset contains 2,230 numbers of rows and 55 columns. The rows represent the number of observations collected and the columns represent features about each observation. Using the skim() method from Skimr R Package, 36 factor variables, 6 logical variables, and 14 numeric variables were found. There is also a lot of data that is missing especially for sales price our response . Many of the observations can be concluded to be unnecessary “junk” that is not representative in predicting sales price. After evaluating the data the number of columns/features was reduced to 15 that are believed to provide a meaningful prediction. The features includes categorical variables such as (fuel\_type, coop\_condo, dining\_ro om\_type, dogs\_allowed, cats\_allowed, garage \_exist) and continuous variables such as (sq\_footage, walk\_score, common\_charges, num\_full\_bathrooms, num\_floors\_in\_buildin g, num\_full\_bathrooms, maintenance\_cost, num\_total\_rooms, num\_bedrooms, appox\_ye ar\_built)

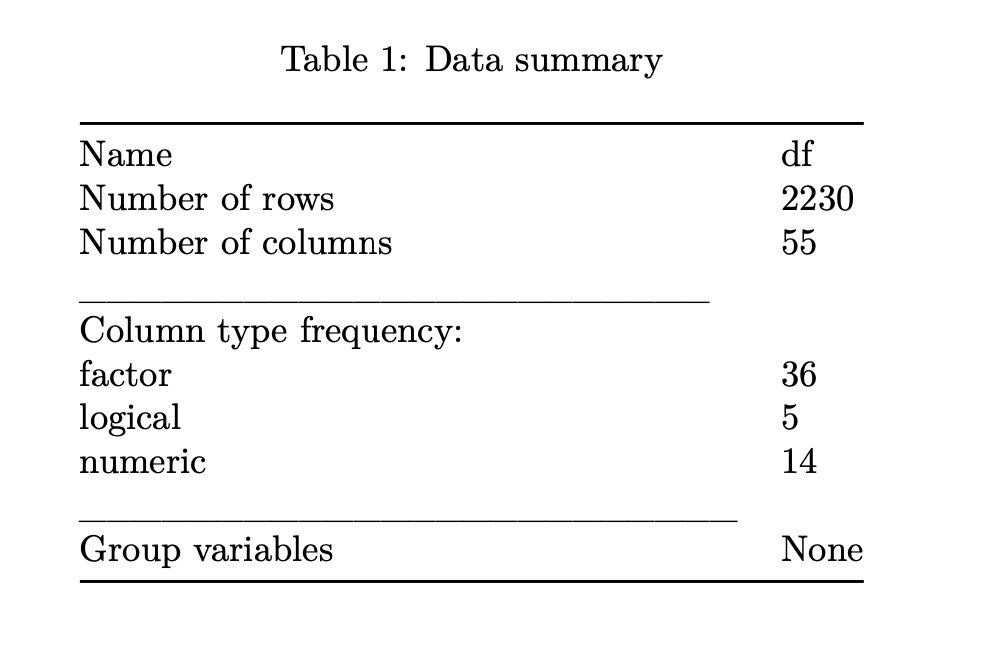
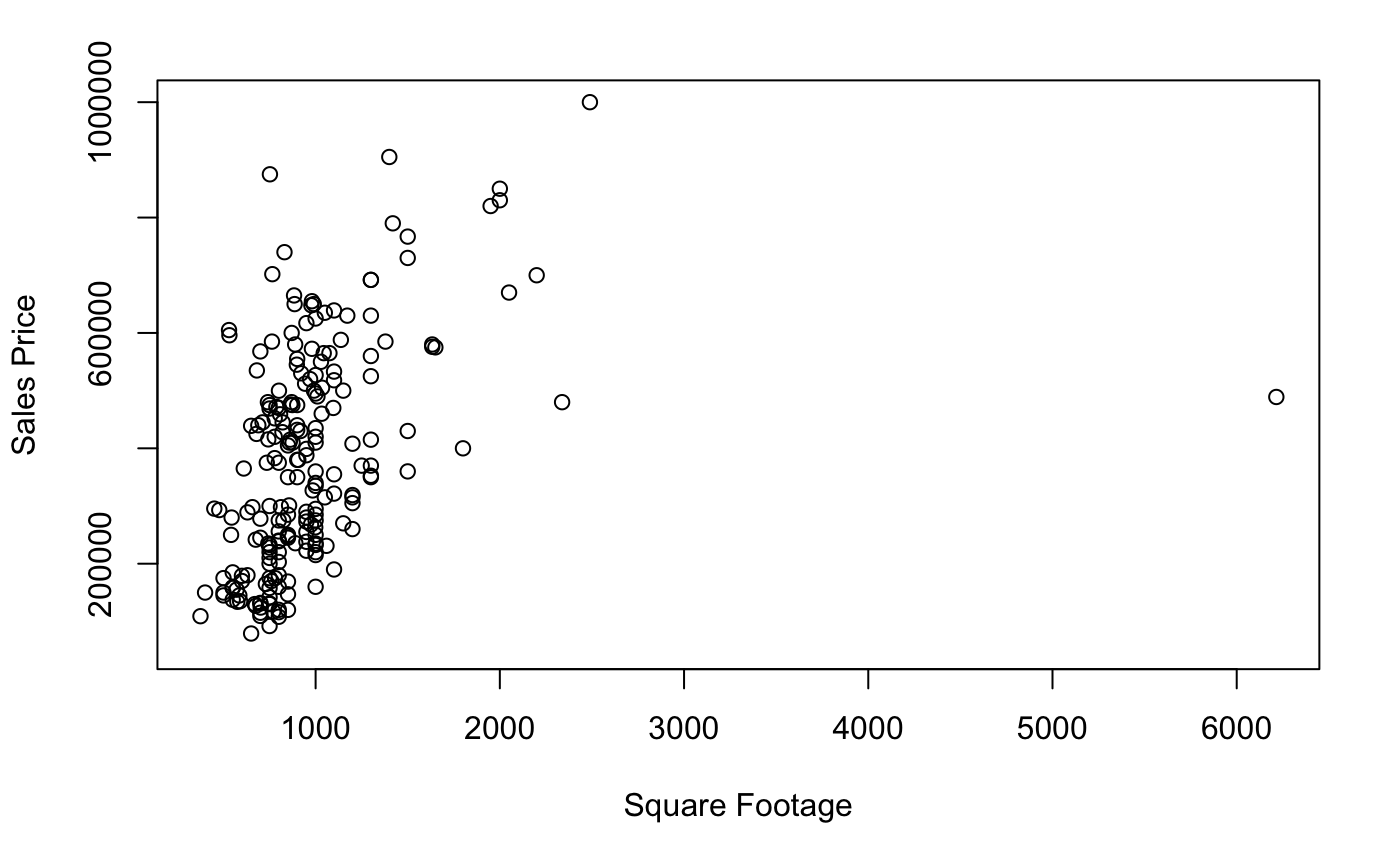


Figure 2. Original Data Summary from Skimr.

**III** **FEATURIZATION**

The original dataset looks like it has plenty of observations with many features but a closer look at the data shows that the *sale\_price* that is missing 1702 observations, that's about 76% of our data missing its response *sale\_price.* All observations with missing *sale\_price* were completely dropped, leaving a total of 528 observations. When observing the 55 features availability many of them don't really help with explaining *sale\_price.* Features such as (Title, Description, CreationTime, WorkerId, Expiration) were removed. On the other hand, features like *sq\_footage* were kept because they represent the size of the land. The average square footage was around 965, majority of the square footage was under 3,000 except for one that was 6,000. Since this was the only occurrence it was treated as an outlier and was removed from the data. Figure 3 is a scatter plot that visualizes sales price and square footage where you can see the distribution of the data. It ranges from 375 square feet to 6,215 square feet and has a standard deviation of about 364 square feet. Clearly this provides some insight about sales price as some sort of correlation can be seen.

 Figure 3. Sales price according to Square Footage

Another feature *approx\_year\_built* ranged from being built in 1915 to 2016 with a standard deviation of 20 years. There isn't a clear visual relationship as when compared to sales price(Figure 4) it seems like there is almost no correlation but there is at least something that the model can learn from.

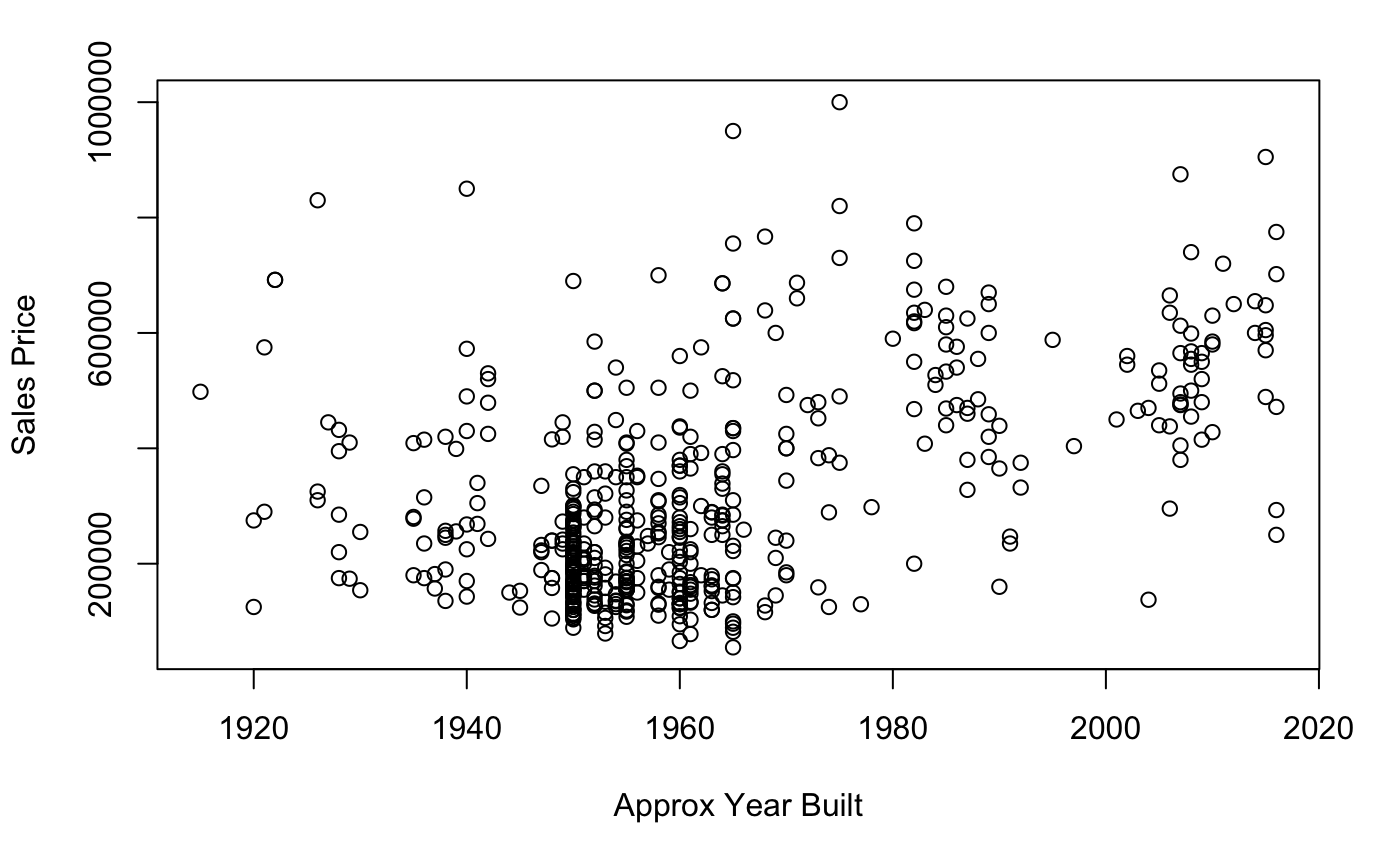
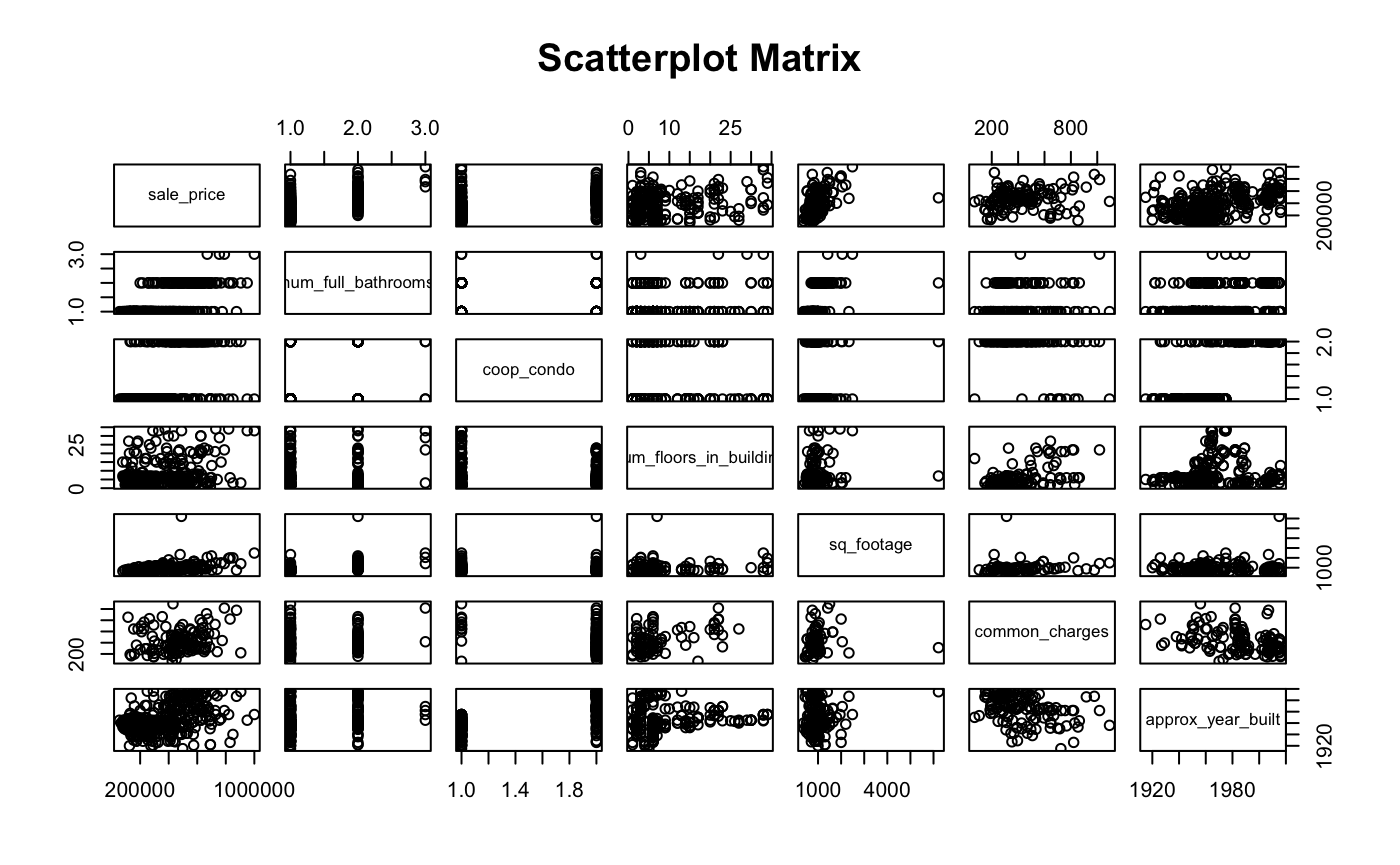


Figure 4. Sales price according to Approx Year Built

Another interesting feature was coop\_condo it categorical data that has about 76% co-ops and 24% condos. This feature basically classifies an apartment as being part of a cooperative ownership that's owned by multiple people or condo that is individually owned.

Figure 5. Scatterplot Matrix of some of the features used

**IV ERROR AND MISSINGNESS**

While exploring the dataset there were a couple of errors. For the categorical feature *cats\_allowed* there were some *y* labels with the existing *yes* and *no* labels. It did not make sense to have *yes* and *y* since it makes sense to have only two choices when asked if cats are allowed. So all observations that have a *y* label were changed to *yes* label. Another categorical feature *fuel\_type* also had two levels *other*  and *Other*, it seems like it was a mistake so I combine the two factors to just be *other.* Another issue was *common\_charges* and *maintance\_cost*, they were both factors but were changed to numeric as it makes more sense since these are continuous variables. Lastly the response *y* which is the *sales\_price* was also a factor that was also changed to numeric type.

From the beginning there was a huge problem with the given dataset. There are 2,230 observations with their 55 features. The problem is that for the response *sales\_price* there were a total of 1702 observations missing their sales price, that's about 76% of the data missing the feature that will be predicted on. Given that *sales\_price* is being predicted it will not make sense to impute it so, it was removed from the dataset which was left with now 528 observations. After doing some featurization that was mentioned beforehand a total of 15 features were left to make a prediction on *y.* There was still some more missingness in *dining\_room\_type, sq\_footage, common\_charges, maintence\_cost* which were imputed using two methods, missingness vector and missForest. Data that was missing was encoded in a new feature as 0 or 1 depending if it was missing or not. Then the data that was missing was imputed using missForest that uses the random forest algorithm to impute all missing values. Now the dataset has no missing values and now has 22 features because missingness vector was used.

**V MODELING**

The data is now fully prepared for modeling. There are many different types of models that can be used in the machine learning world. We will only discuss three different types, a linear model, regression tree model, and random forest model. All three use different approaches to make accurate predictions. Each with different statistical interpretation and real life interpretation.

**VI Regression Tree Modeling**

Regression trees