**Analytical Process:**

The process underlying the Kaggle project was a very challenging, demanding, and testing endeavor which has really expanded my understanding of applying analytics to a real-world dataset. During the project, I had gotten a much better hands- on understanding of interpreting a dataset, identifying predicting variables, iteratively trying various approaches, and gaining greater comfort with the R language.

I started the project by looking at the variables in the dataset primarily with the str and summary functions. I had identified about 8 to 10 variables I had a strong inclination were going to be significant. Amongst these were the following:

* *Neighborhood*: I started with making the variable into a factor and seeing the list of possible results from with the levels function. I then lined up all of the levels in a more easily viewable CSV file and manually identified which neighborhoods stood out to me as being a significant factor in a price of an apartment being higher. I then created a logical variable with this list of neighborhoods for my model. I would later find a much more refined way of achieving my goal with this variable.
* *Reviews*: I believed that reviews were going to be a strong predictor of price, therefore I took a look at each review based variable individually. The most helpful way to discern which ones I should use was with the corrplot function. This showed me that the three most correlated review variables with price were: review\_scores\_cleanliness, review\_scores\_location review\_scores\_rating. I created a variable that summed these three values into on variable. While it was not my most significant variable, it was part of the final model I submitted.
* *Room type:* I found that this was a very important variable and is used in the final submitted model. With this variable I also made a separate variable to just identify the ‘Entire Home/ Apt’ and ‘Private room’. I would later find that this variable performs better just being left as a factor by itself.

In these beginning stages, I had not done several things which would have been useful. First, instead of picking out a handful (about 5 to 8) variables which I had a sense would be useful, I should have separated as many variables as I thought could be relevant into one dataset. There were certain variables which were apparent that they weren’t going to be significant (for example, character variables listing\_url, last\_scraped, thumbnail\_url), so starting with as many as possible and dwindling to the key ones from there would have been helpful. Second, I did not do a good job of identifying the amount of NA values in different variables, as well as discovering some of the outliers within the various variables. Overall, this more granular approach to the beginning of my project had made it more difficult for me to compartmentalize which are the collection of most important variables that I should use. Within the first 2 to 3 weeks of the project, the lowest RMSE that the best model achieved was around 71.

As the project went on, I was able to better organize and come to a collection of variables which would be my base for the rest of the project. The simplest but most common measure I used was correlation and corrplots to identify how certain variables were going to work with price. In addition, the following methods I relied on heavily:

**Lasso:** This was a very useful technique that helped me identify a significant number of variables which would ultimately be the most important in my model:

* *Zipcode*: By using the lasso model to run with just the zipcode variable, I identified which zip codes were most significant importance to price. I then tested performance of all non-0 coefficient zip codes provided by the lasso model to identify which coefficients had the highest performance. I identified that the variables above a 20 coefficient were most significant, therefore took all zipcodes with a higher coefficient than 20 and made them into a separate logical variable (as a numeric data type).
* *Cleaning\_fee:* When running a lasso model on several testing iterations, it identified cleaning\_fee as a very important variable in the model. This variable was a very significant one with the only complication being that there were NA’s within the variables. I made these 0 as many apartments did not have an associated cleaning fee (generally, for NA’s, I tried to input the median or mean value of the variable).

**Mutate:** this function through dplyr along with the Lookup function through QdapTools were very useful in putting wanted calculations to new variables in the dataset.

* Host\_neighborhood: In the beginning stages I mentioned manually identifying what I felt were important variables and then putting them into a logical variable. When I gained more experience through the mutate function it allowed me to take a much more effective approach to have neighborhoods be judged by numerical coefficients that is tied to the mean price of apartments (I had found that the mean had worked better in predicting final price than medium). After creating a new variable (titled ‘neighclean’) in the training and test set I created a lookup object through the aggregate function to align pricing coefficients from testing variables to the scoring data. This was a significant variable for me and this process of creating it is something I will definitely rely on in future projects.
  + I had also tried this approach with zip code, but interestingly it did not perform as well as the neighborhood function.

**Random forest:** Through the majority of the project, I found the best results with random forest model. The formula was extremely flexible with various data types and consistently provided amongst the best performance results. In order to get a better visual understanding of a tree structure, I would use my highest performing variables, create a model and plot decision trees using rpart. I would often times take the information gained through lasso feature selection or through corrplots and incorporate them within random forest. Random forest models with the same variables performed approximately a half an RMSE point better than models generally considered to be more powerful such as boosting. One limitation which I came across first using Random Forest was the computational limitation of my workstation. I would generally test the performance of a new set of variables with 100 trees, and then run to get optimal results with 1000 trees, as anything above that level would not be able to be computed by the resources provided in the Lenovo Yoga i5 2 core laptop that I utilized.

**Boosting/ XGBoost:** While different boosting methods did not give me the best RMSE, I had gotten a sense of the effectiveness of these models, and these approaches is what I want to spend most time on learning more from what was covered this semester. One experience that shocked me was a gbm boosting model on 10,000 trees took more than 12 hours to complete.

When I tried XGBoost, the train and test data results were all promising, with train data always being in the low 40’s for RMSE and on the test it was consistently in the 50’s. From knowing the general theory around how boosting works (i.e that it sequentially runs forest models and improves on the error of the previous), boosting models should show an improvement off of random forest, therefore I would like to continue familiarizing my understanding of various hyperparameters applied.

To end this section outlining the process taken on this project, I wanted to summarize the key missteps I had as I believe these are critical in achieving a better score in a similar competition next time. First, my initial approach to the data was too granular in scope. I selected only a few variables and made models on them instead of picking out a big set of variables I thought were important and dwindling down to the more significant ones. Second, my focus was lacking in the areas which should take the most time in data modeling, tidying and cleaning up of individual variables. Due to not spending an adequate amount of time on this process, I found myself doing data tidying well into the model making part of my analysis, often times after adding new variables later in the process instead of having analysis ready variables prepared (through cleaning data and a more top down approach). Third, a better understanding of hyperparameters and their impact on models. Boosting methods which I utilized did not provide an improvement on my model that I had hoped, and my understanding of utilizing various hyperparameters was the most glaring reason for that.

**Implementation of other process’ and approaches:**

**Linear Regression:** Linear regression was a starting point for many of the models. I often used linear regression models in the beginning process of my analysis, where I was testing different combinations of variables. The RMSE for my models had never gotten below 66 with a linear regression model.

**Feature Selection:** While I went over my reliance on Lasso, I had found other measures such as ridge and principal component analysis to be less helpful for this project. Ridge did not outline the least performing variables as well as lasso, while principle component analysis was not very interpretable even when the score provided was good.

**Logistic Regression:** The problem that the project was modelling for was inherently not a logistic regression problem. Therefore, it was very early on that I saw logistic solutions perform very poorly for this prediction.

**What I would like to do differently in the future:**

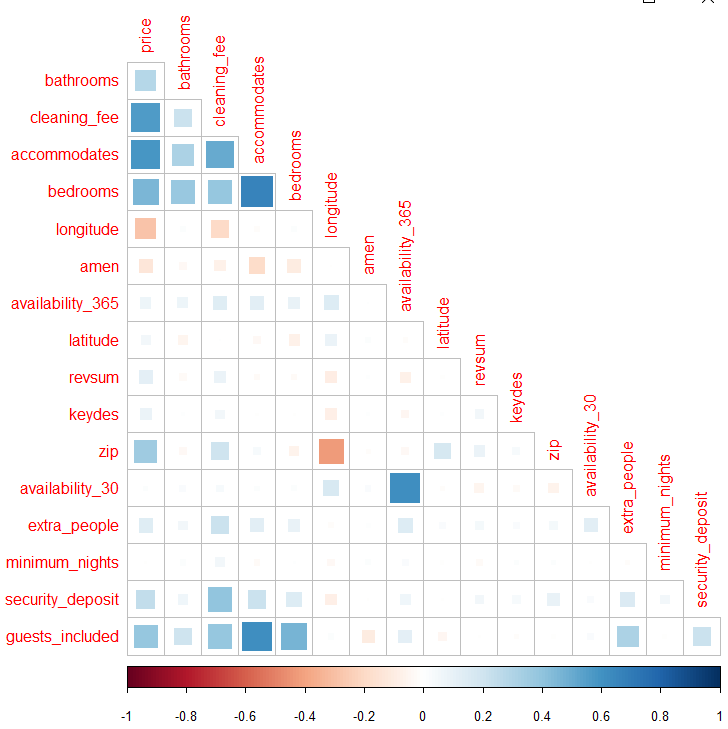
First and foremost, as outlined earlier, I will take a more top down approach when it comes to my initial analysis of a data set. Instead of picking certain variables based on my instinct and rushing to see how a model would perform on them, I will take significant grouping of variables and identify their significance with the proper feature selection approaches. In addition, in order to improve on my ability to build successful models, I need to get more comfortable with tools that analyze significance of model variables. My primary analysis point on the significance of individual variables was correlation to the price while for collection of variables it was Lasso model. I would like to expand the toolkit I am comfortable with when it comes to variable selection.

The importance of processing power is another major topic I would like to consider in the future. While I did not utilize AWS, I have a better understanding of the significance in computing power with larger datasets. Understanding this aspect of analytics (integration and processing power) is something that interests me, and I hope to get further exposure of this area in the field as I continue the program.

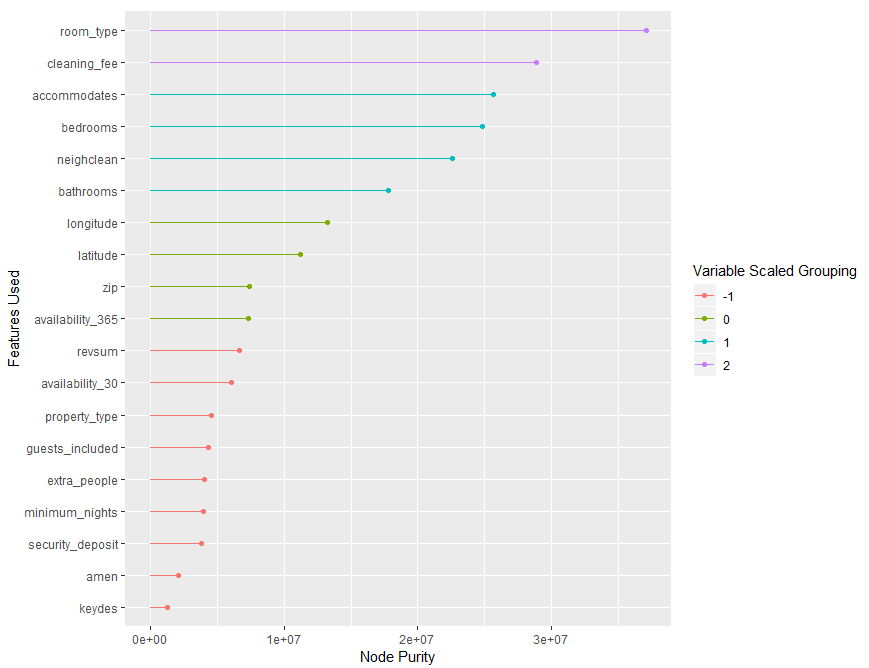
The experience of this project has left me more excited than ever to be in the field. While I am ultimately disappointed in the final RMSE figure which I had achieved on the scoring data (56.39 public, 58.7 private), this process showed me the way to start and approach a data science project, something I found rather intimidating before with a lack of knowledge of how to start. However, the project has also shown me the breadth of information that I do not understand and need to gain more comfort in order to get where I ultimately want to be. This was a very important first step in getting there.

**Charts**:

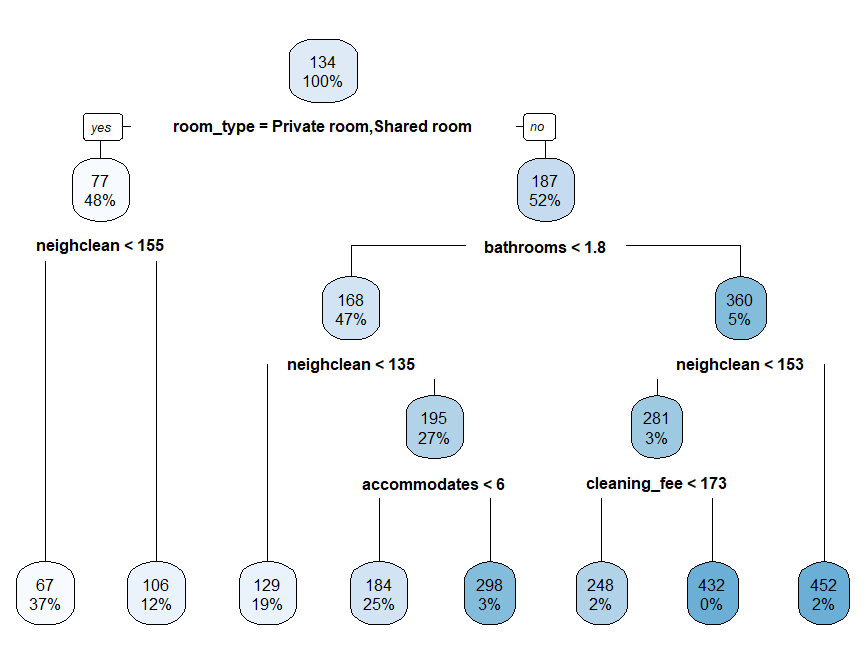
        Corrplot



         Importance Variables



         Rpart.plot tree



* Lasso Model Best Lambda

