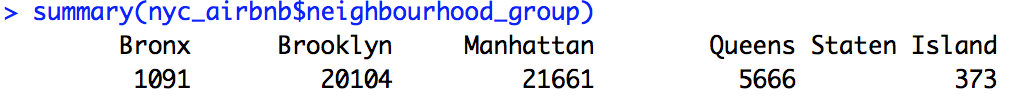
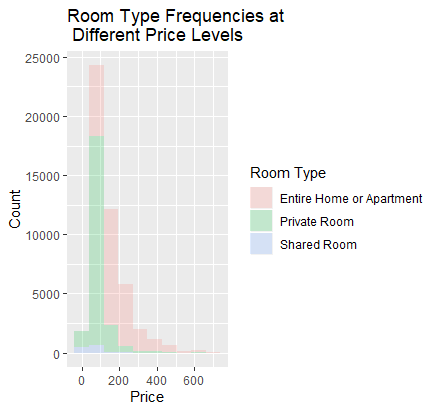
Ivan Wang (zw4fc)

STAT 5330 Project: Final Report

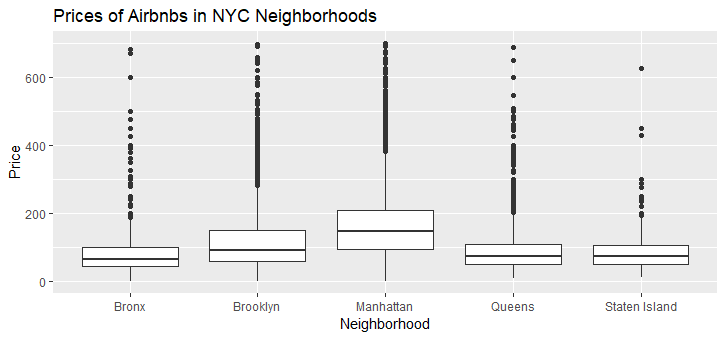
1. **Executive Summary**
   1. The two questions of interest
      1. What is the fair value market price for an Airbnb based on the neighborhood, room type, number of reviews (popularity) and other factors?
      2. What type of housing should a prospective resident/visitor expect an Airbnb to be, whether it is a private room or an entire home/apartment, based on its neighborhood, price, availability and other factors?
   2. How do the analyses answer these two questions of interest?
      1. From the regression conducted, I concluded that the neighborhood, room type, number of reviews, and availability of the listing were all significant variables in predicting the price of an Airbnb.
      2. My classification model helped in answering the questions of interest regarding the type of an Airbnb listing by showing the effect that different predictors have on the response variable, room\_type. From this analysis, I found that as price decreases, an observation will be more likely to be classified as being a private room. In addition, a listing with a smaller number of negative reviews and listings per host is more likely to be a private room. In the context of the questions of interest, a prospective resident should expect a listing with a lower price, smaller number of reviews, smaller number of listings per host, and higher availability to be a private room as opposed to an entire home.
      3. From my decision tree analysis, I found that Bronx, Brooklyn, and Queens have cheaper listings, regardless of room type. As one might expect, an entire home/ apartment is most important in determining price, with an entire home being more expensive than a private room.
      4. Each of my analyses gave different types of information that allowed me to gain important insight on the questions of interest.
   3. Interesting Findings
      1. The fair price for an Airbnb is worth predicting because the model could provide visitors a guideline of what the market price of the Airbnb that they are looking for should be based on the property’s location, room type, popularity, etc.. This would prevent the visitors who are probably not familiar with NYC's price levels from being overcharged and give them an educated idea of what kind of housing they should look for based on their budget. Moreover, the model is useful for Airbnb owners and investors who wish to invest in certain Airbnb. It helps the Airbnb owners to estimate an appropriate price to set for its Airbnb in order to prevent overcharging as well as undercharging. It also helps investors to project expected returns in the short-term and predict their annualized ROI in the long to see if they can potentially profit on the investment.
      2. The prediction of the room type of an Airbnb based on various factors, such as neighborhood, price, availability, etc. is important as well. Based on the model, the Airbnb owner could predict which type of housing would converge to the preferences of the visitors. For instance, certain types of visitors might prefer to live in one particular neighborhood or have a relatively high budget, whereas other visitors who travel with a larger group of people might choose an entire home or apartment over a private room or prefer a specific neighborhood that may be cheaper or quieter to live in. Thus, the prediction of the choice of what type of housing should be built or invested in as an Airbnb for what price, in which neighborhood, etc. is essential.
2. **Data Processing and Cleaning**
   1. Remove columns of no interest to the model building: from observing the dataset, columns such as id, name, host\_id, and host\_name have no clear relationship with the final price of an Airbnb listing.
   2. Meanwhile, the information conveyed in variables such as Longitude and Latitude should not be used since I am not using spatial models for this analysis. Neighborhood\_group and neighborhood are functions of each other, therefore I chose to use only neighborhood\_group in my analysis. I subsetted out the 8 attributes that will have a greater impact on the price of a specific Airbnb based on intuition. Those attributes/columns are neighbourhood\_group, room\_type, price, minimum\_nights, number\_of\_reviews, reviews\_per\_month, availability\_365, calculated\_host\_listings\_count.
   3. Remove the column “reviews\_per\_month” because there are 10000 NAs
   4. Remove outliers in terms of the airbnb that has overly expansive or overly cheap price, subset the price that is between $0 and $700 per night.
   5. Neighborhood\_group variables with value equal to “Staten Island” were removed from the dataset, since the number of instances are so small in comparison to the other 4 factors, as shown below:



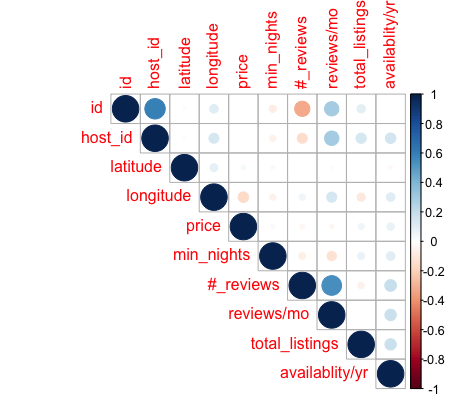
1. **Exploratory Data Analysis**
   1. Figure 1: Room type frequencies at different price levels



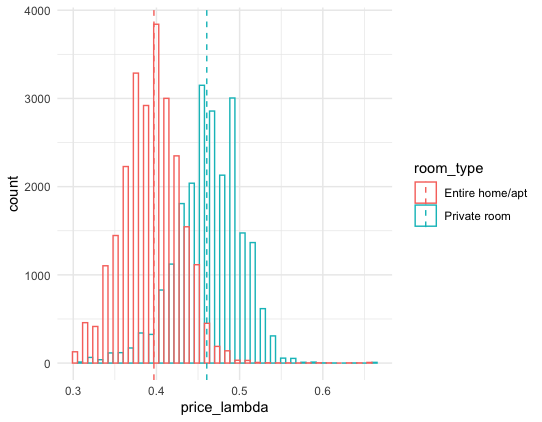
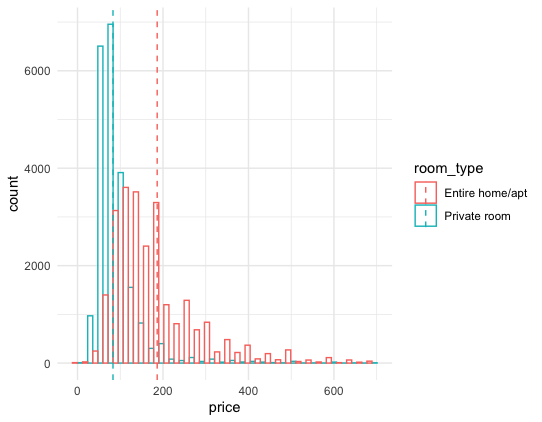
* 1. Figure 2: Prices of airbnbs based on different neighborhoods



* 1. Figure 3: Correlation matrix

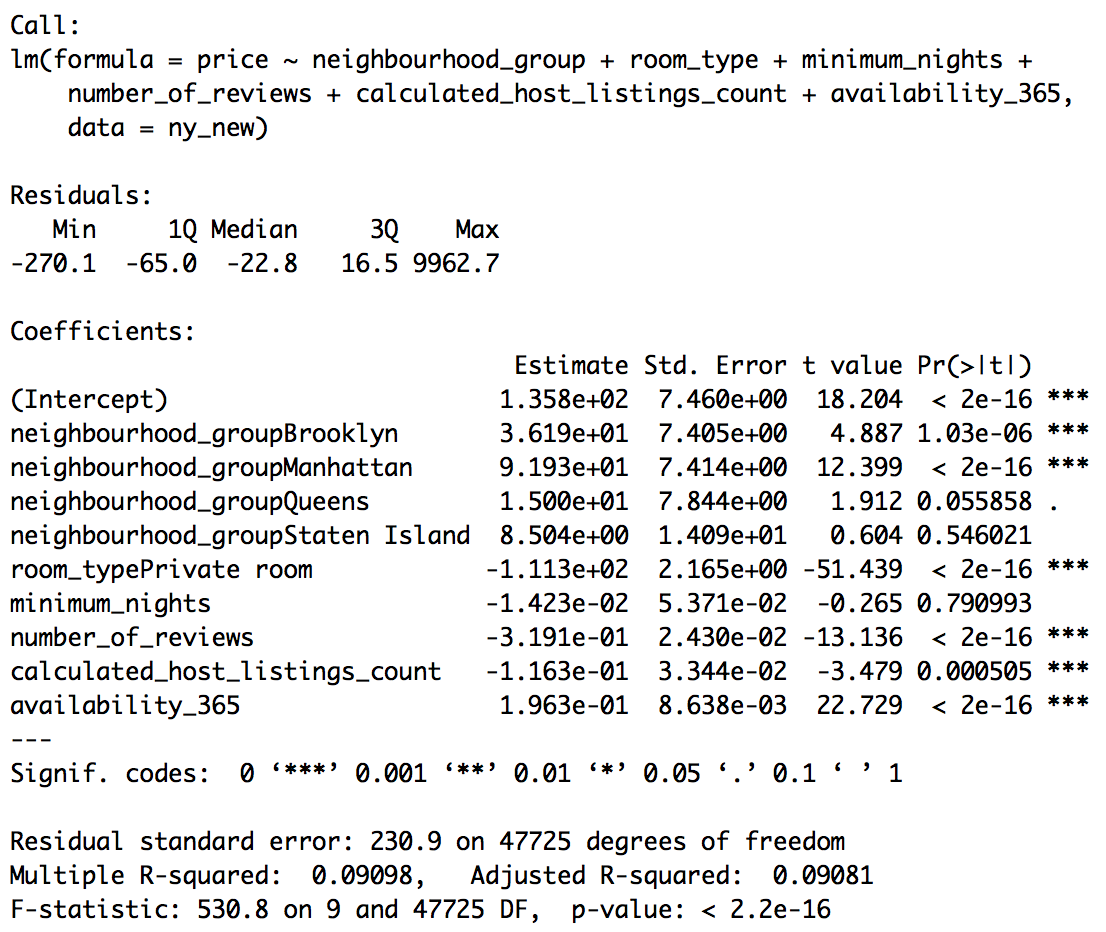


* 1. Figure 4: Histogram of Price across the two room types

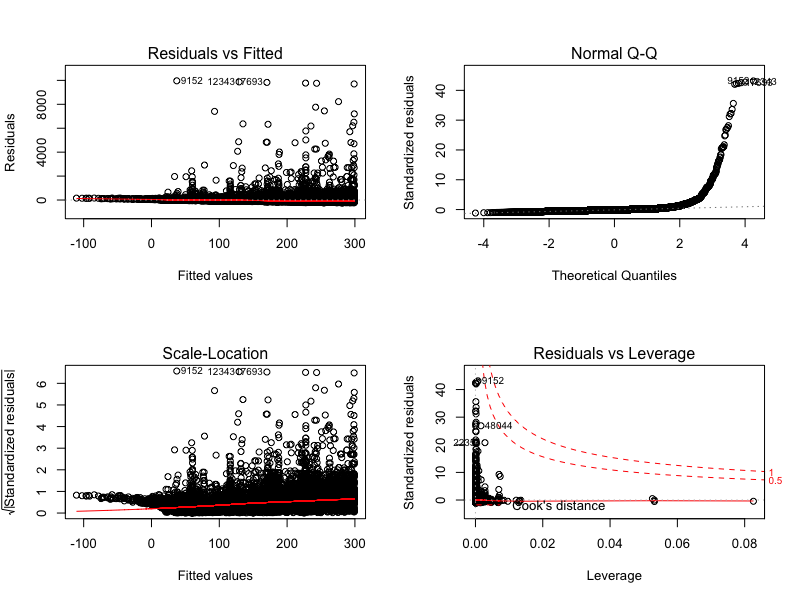


* 1. I explored these graphical summaries because they are relevant to the questions of interest.
     1. In Figure 1, I explore the relationship between the frequency of room type and varying price levels. This may be helpful in distinguishing what types of listings tend to be priced higher.
     2. In Figure 2, I created boxplots of prices of Airbnbs in different neighborhoods. This was helpful in discovering a possible relationship between price and neighborhood.
     3. Figure 3 is a correlation plot of the variables in the data set. This was helpful in deciding which predictors are correlated which may cause issues in the analysis.
     4. Figure 4 shows two histograms of price vs room\_type as well as price\_lambda respectively. Distribution of price on the left shows that the division between the two groups is not super clear because of the positive (right) skewness. As a result of that, I tried the same experiment on the price\_lambda (lambda = -0.30).

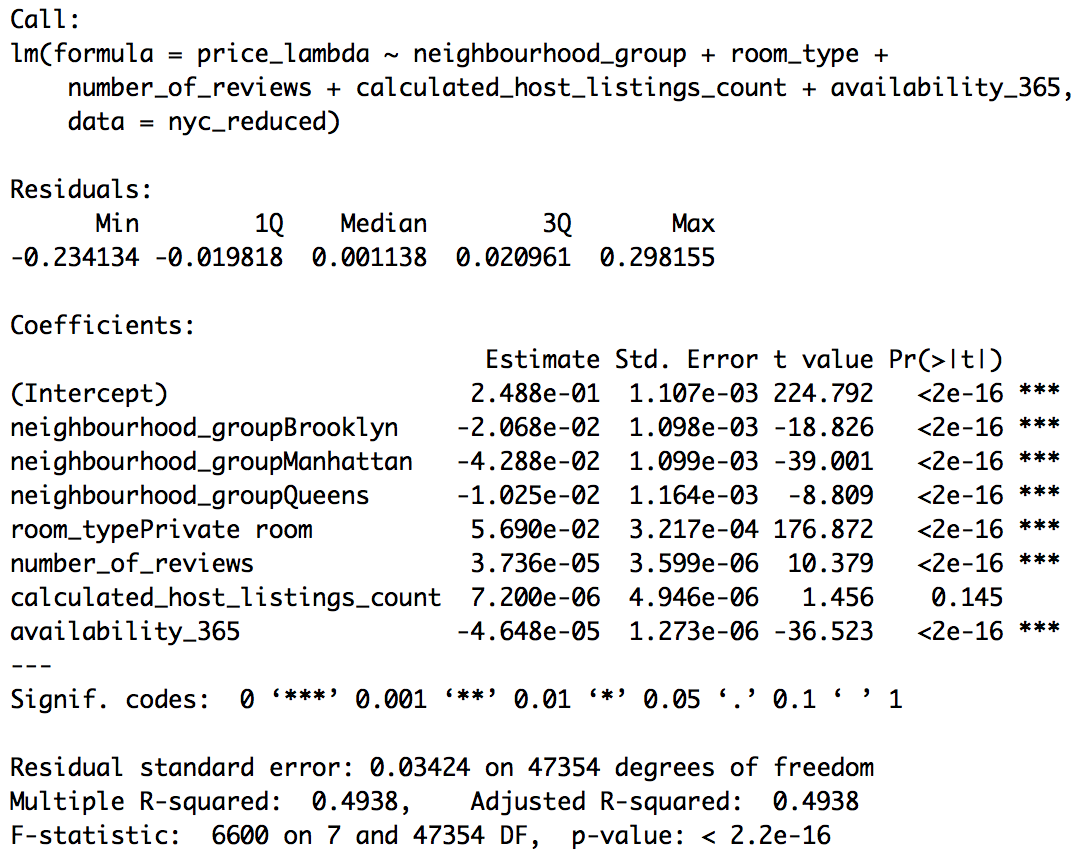
1. **Analysis from Regression, Classification, Trees**
   1. **Linear Regression Models:** 
      1. Initial Model:



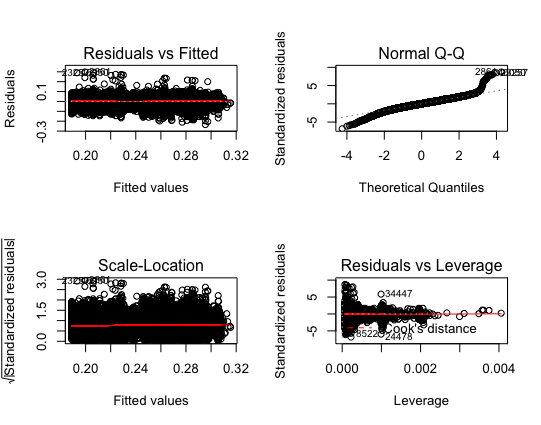
* + - 1. Of the 5 neighborhood groups, Queens and Staten Island has the least statistically significant impact on the predicted price of an Airbnb.
      2. Minimum\_nights has little to no impact on the overall price.
      3. The residual standard error is huge in this context, as $230 is a lot in terms of the prediction error of a property’s price.
      4. The linear model is only accounting for 9% of the total variation
    1. First Regression Diagnostic:



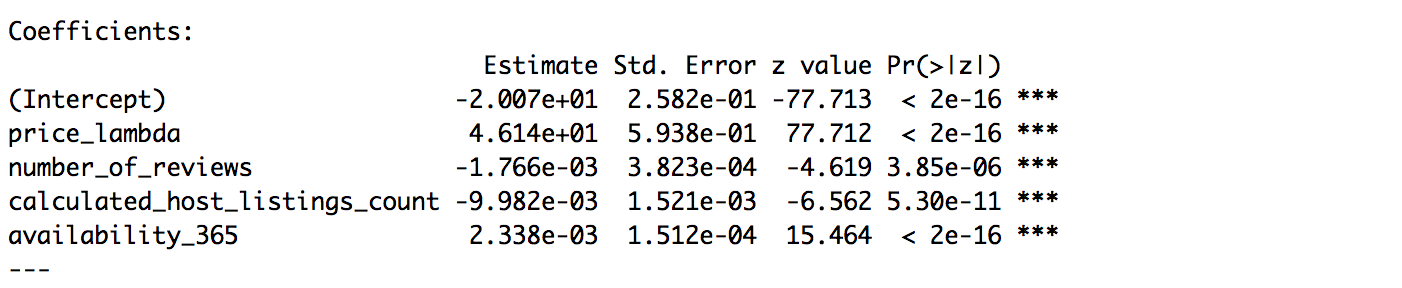
* + - 1. The model I have currently does not meet the assumption for linearity and normality. The residual vs fitted value plot shows nonconstant variance and non-linearity. The Normal Q-Q plot is severely right-skewed, meaning that most of the data is distributed on the left side with a long “tail” of data
         1. I dropped rows with neighborhood\_group = Staten island (only about 500 instances belong to this region, thus sample size too small) and the column of minimum\_nights since its p-value is so large.
         2. I transformed the response variable, price, with exponent lambda using the box cox transformation. The optimal lambda is -0.3030303.
    1. Improved Model:

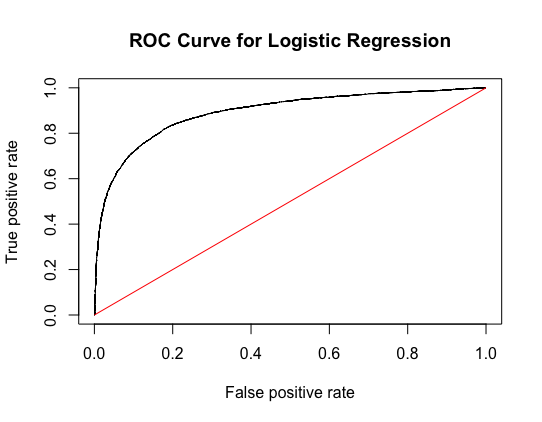


* + - 1. New model is performing significantly better as the residual standard error is reduced and r-squared is improved.
    1. Improved Model Diagnostic:

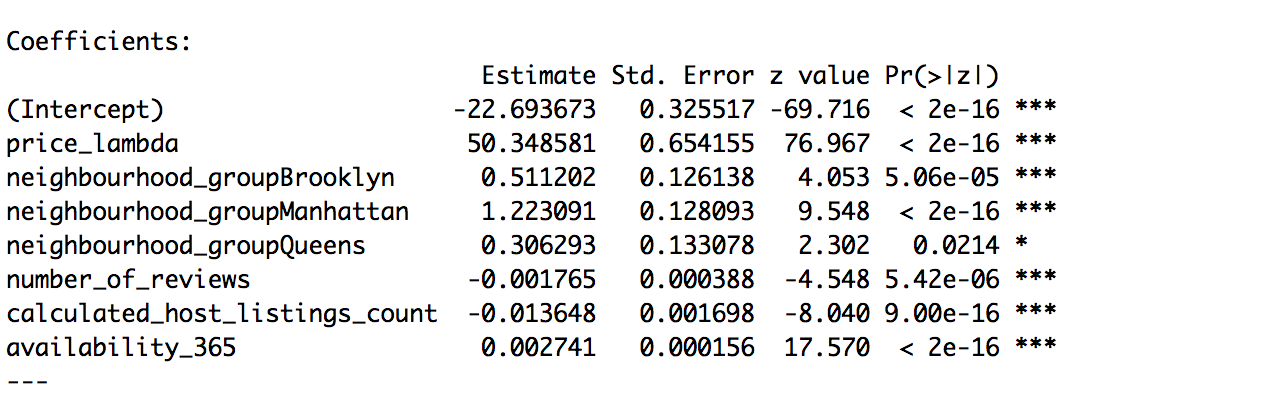


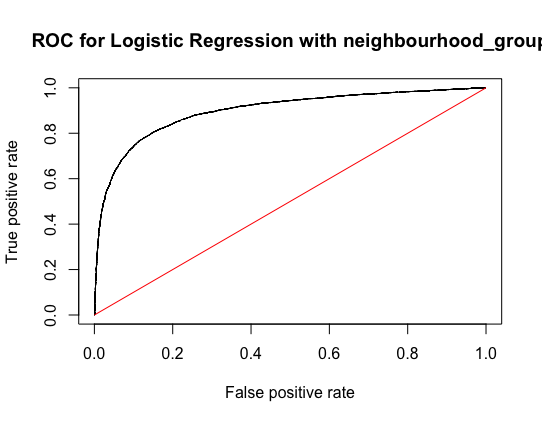
* + - 1. The first plot (residual plot) has the red line along the x-axis without any apparent curvature, indicating that the form of the model is reasonable.The normal Q-Q plot shows that the residuals mostly fall along the 45-degree line, meaning that the residuals are normal.The third plot validated that the variance are constant.
      2. However, there is still a sizable portion of influential outliers in the model.
    1. Regression Model Key Improvements:
       1. The first model had a very low value for R-squared 0.09. With the second model, I get an R-squared value of .49. The second model explains the variance in the data much better than the first.
       2. The first model included all the independent variables that I hypothesized could be significant when determining the independent variable price. The p-values of variable neighborhood\_group = Staten island and variable minimum\_nights are very statistically insignificant (p-value > 0.5). Thus, I dropped those variables in the improved models, which resulted in a higher R-squared.
       3. After transforming the response variable with boxcox in the new model, the new model held up for the assumptions of mean zero, normality, and constant variance, with minimal outliers.
  1. **Logistic Regression Model**
     1. First model without neighborhood\_group:

****

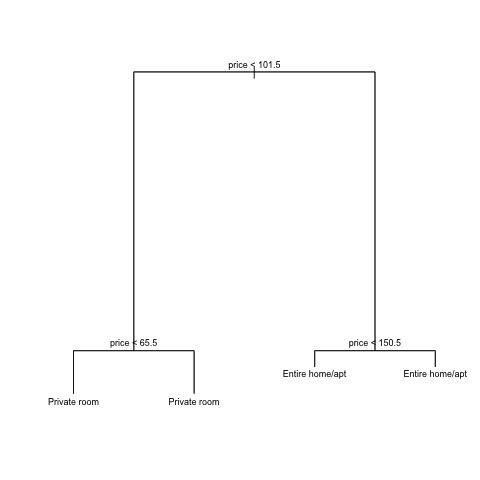
 AUC = 0.8925323

* + 1. Improved model with neighborhood\_group:

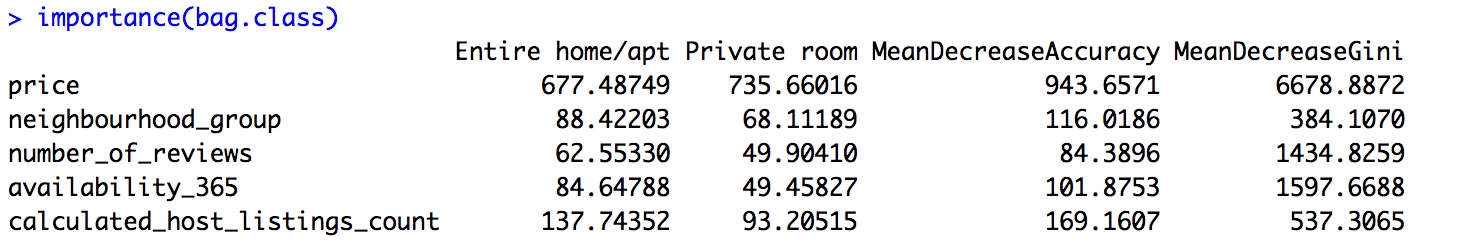
****

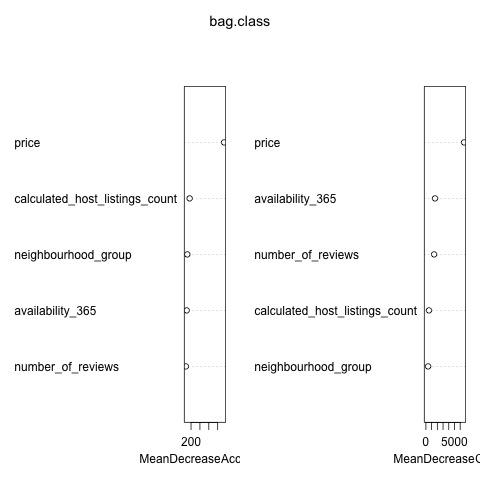
 AUC = 0.8978974

* + 1. Output Analysis
       1. The output from the two logistic regression models can be seen above. In this analysis, all of the predictors were significant in both models. These results were not surprising.
       2. By adding a categorical variable to the original logistic regression, I was able to improve the model by a little bit.
       3. The coefficient for price\_lambda is not surprising because the sign of the coefficient is positive, which means the higher the price lambda, the higher the probability of having a private room. Since price\_lambda has an inverse relationship with the actual price, the higher price would lead to higher probability of having an entire home/apt. This makes sense because an entire home/apt must be more expensive than a private room.
       4. For example, the categorical predictor neighbourhood shows some very interesting results. The base level is Bronx by default as it carries a coefficient of 0, whereas the coefficients for all three other neighbourhoods are positive as shown on the output below. Since the Bronx has a coefficient of 0, it means that when the property is in the Bronx, there is a much higher likelihood that the room type is entire home/apt rather than private room given the price and other predictors’ levels. As the neighbourhood moves from Queens to Brooklyn to Manhattan, their respective coefficients increases from 0.30 to 0.51 to 1.22, indicating that the probability for any given airbnb listing to be classified as a private room versus an entire home/apt would also get larger. This is very intuitive as there are more houses and entire homes in the Queens and Bronx districts, but since that whole apartments are much rarer and more expensive in Brooklyn and Manhattan, more private rooms in those areas are listed on Airbnb.
  1. **Decision Trees**
     1. Classification Tree: predicting room type

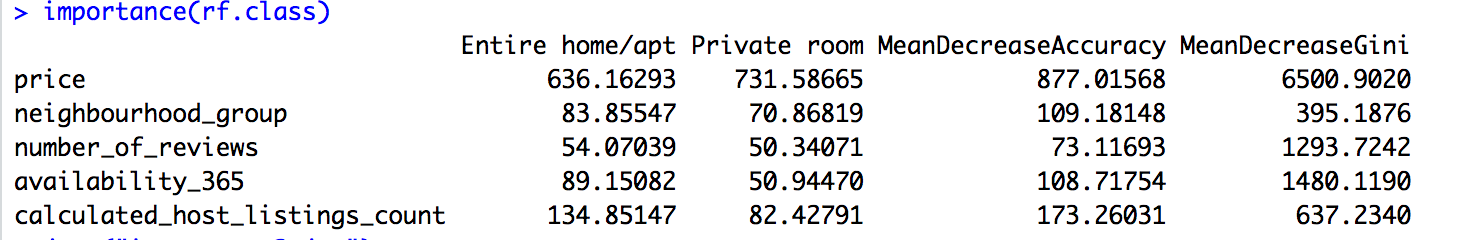


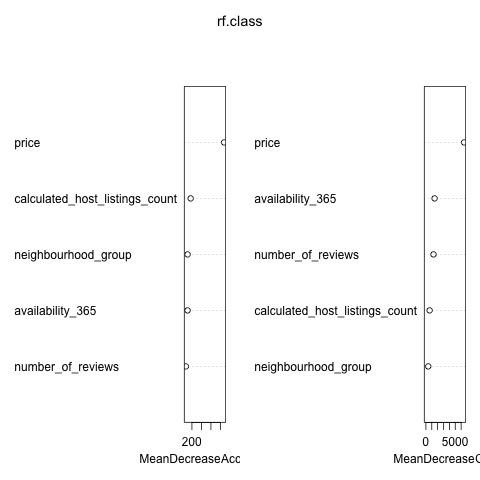
* + - 1. Price is the only useful predictor in distinguishing between a private room or an entire room or apartment.
      2. A price greater than 101.5 for a private room would be considered expensive. And a price less than 101.5 for an entire home or apartment would be relatively inexpensive.
      3. I can expect that the variance in price is pretty high within the categories of private rooms or entire home/apartments because there is an extra split in price during the recursive binary splitting process for both private rooms and entire home/apartments
      4. Based on the tree, if the price of a listing is less than $101.50, it is most likely a private room. If the price of a listing is greater than $101.50, it is most likely an entire home/apartment. Therefore, if a prospective visitor is paying less than $101.50 for a listing, they should expect it to be a private room.
    1. Importance (Bagging and Random Forests) for Classification Tree



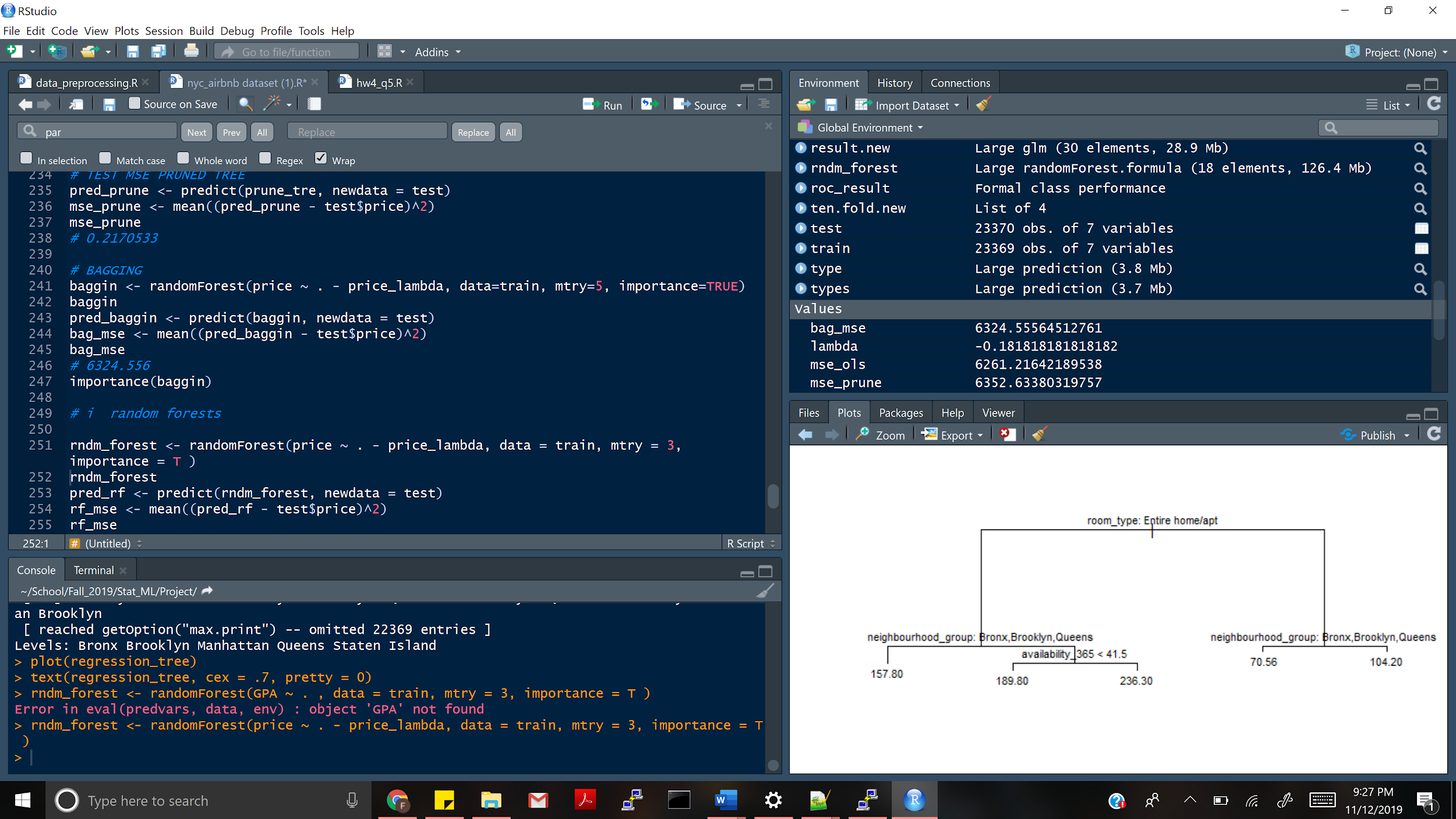


1. The predictors that were found to be most important in bagging were price and calculated\_host\_listing\_count.

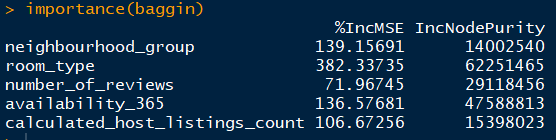




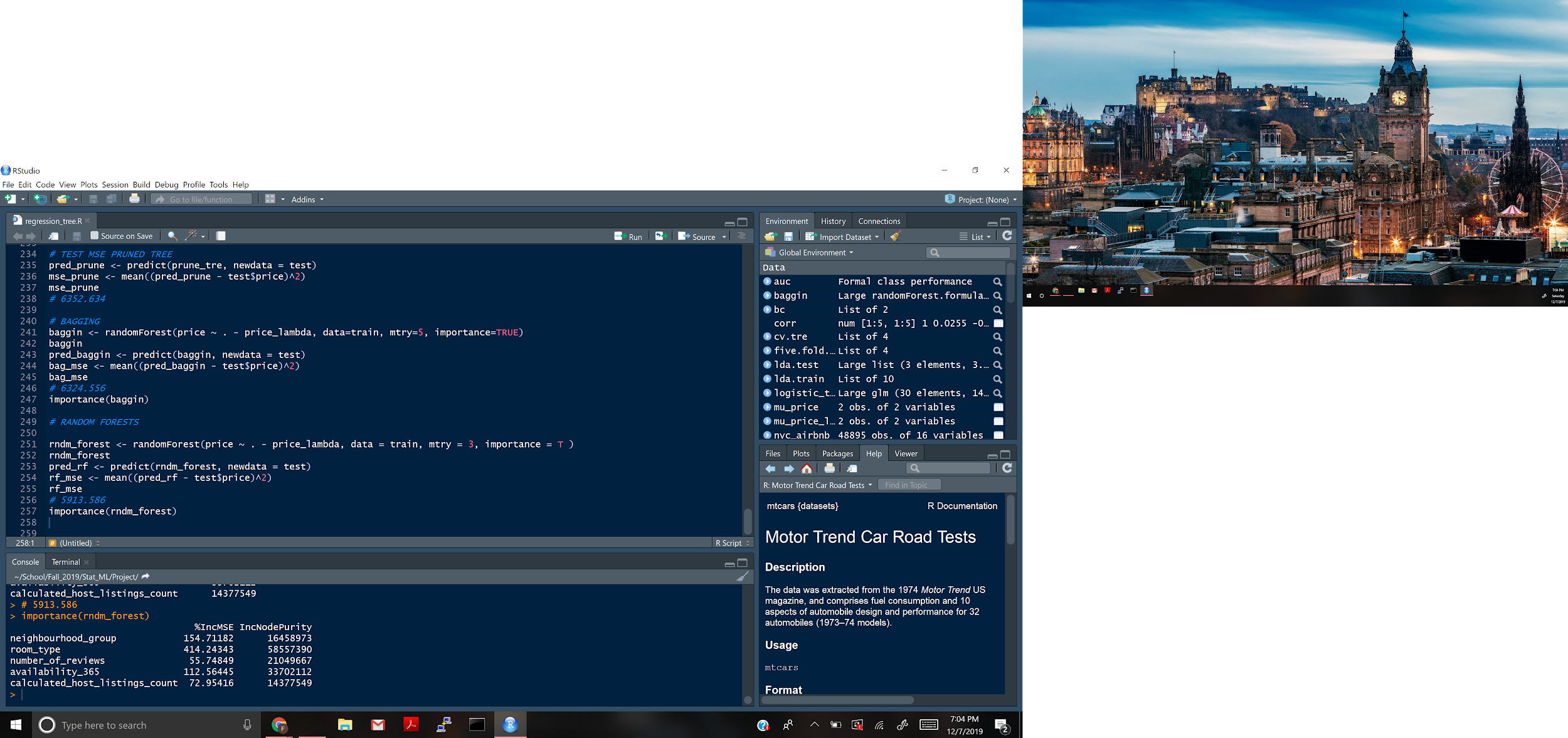
1. The predictors that were found to be most important in random forests were also price and calculated\_host\_listings\_count.
   * 1. Regression Tree: predicting room price



1. Neighborhoods in the Bronx, Brooklyn, or Queens have cheaper airbnbs, regardless of room\_type than Neighborhoods in New York outside of these regions.
2. Rooms that are rented out less often are in general cheaper.
3. As one might expect, an entire home/ apartment is most important in determining price, with an entire apartment being more expensive.
4. The two other variables, host\_listings and number\_of\_reviews, were found to have low importance.
5. And a high number of airbnbs rented out by a landowner, is probably unhelpful in determining a higher or lower price.
   * 1. Importance (Bagging) for Regression Tree



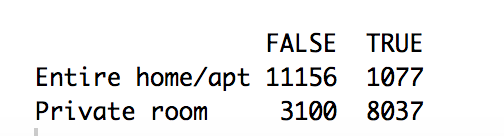
* + - 1. The most important predictors found in bagging for the regression tree was room\_type, nneighborhood\_group, and availability.
    1. Importance (Random Forests) for Regression Tree



* + - 1. The most important predictors found in random forests were room\_type and neighborhood group.
  1. **Compare and contrast the results from different types of analysis.**
     1. Comparison of Quantitative Predictions (Regression vs Regression Tree)

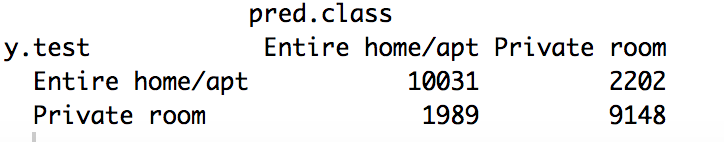
This regression tree is much easier to interpret than the output of ordinary least squares regression. A fair price estimate can be found fairly quickly by looking at this tree. The coefficients on the OLS regression are also harder to interpret because the response variable is price to the power of lambda, where lambda was optimized with a box cox transformation.

* + 1. Comparison of Classification Methods (Logistic Regression vs Classification Tree)
       1. Confusion Matrix Output of Logistic Regression:



In logistic regression all of the predictors were found to be significant and useful in predicting price. For the classification tree, only the variable price was used during the recursive binary splitting process.

* + - 1. Confusion Matrix Output of Classification Tree:



* + - 1. Comparison of LDA and Logistic Regression

In the context of this problem, logistic regression model outperforms LDA in terms of the overall classification accuracy, and that indicates that the assumptions of the LDA are not properly met by the dataset, thus resulting in a weak approximation of the Bayes Rule for classifiers since not all features’ density are normally distributed.

* 1. **Interesting Findings and Insights** 
     1. For the regression models, the variable price is extremely skewed with many high outliers. This is why the original model had a horrible fit. In order to create a good model to predict price, I would likely have to consider methods with greater flexibility.
     2. I hypothesized that the minimum nights stayed at Airbnb would be a significant independent variable in the first place because I assumed the longer the stay, the cheaper the price would be. However, it turned out that the minimum\_nights has a large p-value of 0.79 and is statistically insignificant. I did some research online and found that Airbnb actually does not allow hosts to adjust the price per night according to the length of the stay. Only a lower cleaning fee/day would apply, which is probably a relatively small change in the amount to influence the overall model. This explains the reason that the variable minimum\_nights is insignificant. Moreover, due to a large amount of requests for adjustability in the price/day from hosts, Airbnb started to test this new feature on a few housings in mid 2018. Thus, the significance of minimum nights of stay would probably change in future observations.
     3. By adding a categorical variable to the original logistic regression, I was able to improve the model by a little bit.
     4. According to the logistic regressions, the larger the number of reviews the more likely the room type is an entire house/apt. People tend to comment more when they stayed in an entire house/apt than did in a private room.
     5. According to the logistic regressions, the larger the host\_listing\_count, the more likely the room type is an entire house/apt.
     6. The logistic regression models I performed all have a similar auc so there are not big differences regarding accuracy for the models. Thus, the choice of which model performs the best could be hard and may be biased because of the way I split the dataset.
     7. By looking at the classification tree, I determined that price is the most important predictor in determining what type of housing a prospective resident/visitor should expect an Airbnb to be. I thought it was interesting that only price was used in both the classification tree and the pruned classification tree since there were other predictors used. The number of listings per host (calculated\_host\_listings\_count) was also more important than the other predictors, but not nearly as much as price.
     8. By looking at the Regression Tree, Bronx, Brooklyn and Queens have cheaper airbnbs, regardless of room\_type. Rooms that are rented out less often are in general cheaper. This can be understood in that someone that is willing to lease out their home for a short time is willing to charge less to ensure someone rents out the airbnb. As one might expect, an entire home/ apartment is most important in determining price, with an entire apartment being more expensive. The two other variables, host\_listings and number\_of\_reviews, were found to have low importance. They were not even deemed useful during the recursive binary splitting process. This makes sense because a low or high number of reviews doesn’t necessarily mean that the apartment is low quality. And a high number of airbnbs rented out by a landowner, is probably unhelpful in determining a higher or lower price.

1. **Further Work**
   * 1. If I had more time to work on this project, I would have added more variables into the analysis that may increase the accuracy of the predictions. For example, with more information I could incorporate a quantitative variable that describes the size of the listing. I could also identify whether a given listing is hosted by a superhost or not. If a listing has a superhost, this means the host has more credibility which may affect the response variables. In addition, I could add the average review rating of a listing into the model. The average review rating could help in predicting price since a higher rating may indicate a better listing and in general, better listings are associated with higher prices.
     2. If there is an API for the source of Airbnb data at different times of the months throughout multiple calendar years, I would try to incorporate the time dimension into the model as well, so that it could predict price more dynamically based on weekends, holidays, or other peak vacation seasons for the users. If this option is not available, I could incorporate categorical variables such as “IsWeekend (Friday/Saturday)” or “IsHoliday(Christmas/Thanksgiving)” to each one of the price data points.