**Northeastern University**

**DS 5110 Project - Spring 2019**

**A Report On**

Airbnb Insights

And

Predictions

Authors

Anjita Shrestha

Rahul Tyagi

Nick Tyler

Zhaoxi Zhang

1. Summary

This report gives an overview of the Airbnb insights in the Boston neighbourhoods. For the project, we have used the dataset which is publicly available at the Airbnb website,<http://insideairbnb.com/get-the-data.html> . The Dataset comprises of mainly 3 tables:

* **Listings.csv**: Listings table consists of detail data of all the listings in the file. It contains 6,155 columns and 106 rows. Some of the variables that we have used in our analysis are, *listing\_id (categorical), price(continuous), longitude(continuous), latitude(continuous), room\_type(categorical), neighbourhood(categorical), neighbourhood\_cleansed(categorical), review\_scores\_rating (continuous), beds(categorical)* among others.
* **Reviews.csv**: Reviews table consists of detail reviews of the listings given by the users. It contains 199,106 columns and 6 rows. Key attributes include *date (date), listing\_id (categorical) and comments (character).*
* **Calendar.csv**: Calendar table consists of detail dates about bookings and property listings. It contains 2,280,155 columns and 7 rows. Key attributes include *date (date), listing\_id (categorical) and price(continuous).*

**1.1 Objectives**

The aim of this project is to provide an insight to customers to make it easy to choose an Airbnb as per their budget and requirements. To fulfill the objectives, we did some analysis based on the following questions:

* What are the important variables that directly affect the pricing of the listings?
* What is the correlation among different types of variables and what are their effects on our model for predicting the price?
* What features are to be considered while searching for an Airbnb listing.
* Which neighborhood is most costly?
* Is there any impact of number of listing on price?

**1.2 Outcomes**

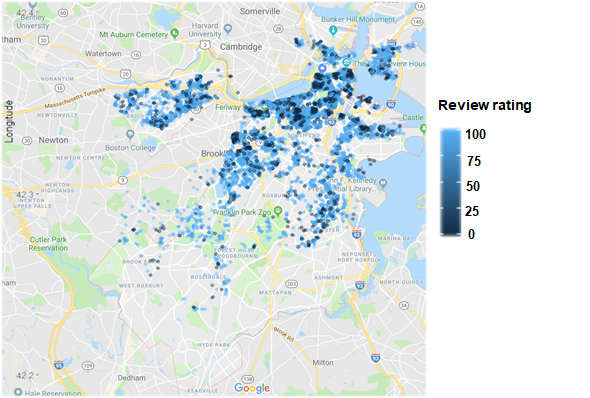
Overall, we found a number of interesting features that seemed to affect the pricing of a listing described in detail below. We found that factors such as neighbourhood and day of the week can result in either a higher or lower price. We also found that neighbourhoods with a high amount of listings tend to have a lower price per listing. In addition to our discoveries during the EDA portion of our project, we were also able to fit a successful model of price using three dataset features: *accommodates, neighbourhood, and room\_type*. From these features alone we were able to generate a model with an RMSE of .749. The residuals from this model were normally distributed around 0 when compared with the explanatory variables.

2. Methodology

We performed various data transformation and data manipulation techniques and methods such as :

* Replacing all the null values with 0
* Converting categorical variables into numerical variables for our model.
* Made user defined functions to deal with datetime and pricing variables in order to convert them into proper format.
* Used log transformation for our model in order to get a more linear fit.
* Used subsets of the data in our exploratory data analysis
* Used stopwords and “bing” sentiment in our sentimental analysis

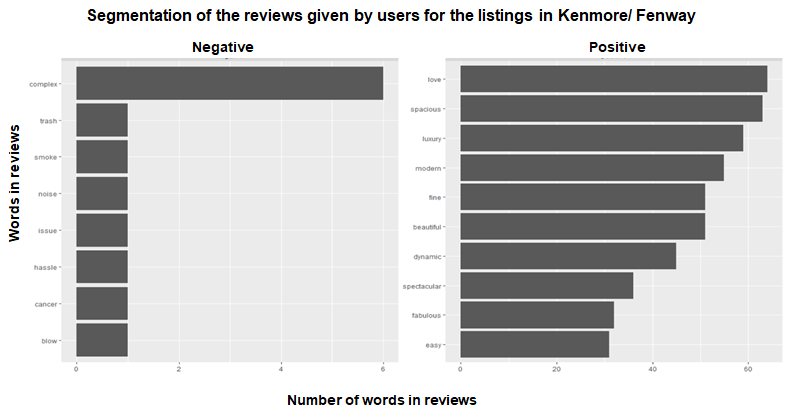
3. Results

**3.1 Choosing the neighbourhood**

First, we used the “overall\_score\_rating” to map out the density of the number of airbnb listings around great boston area. From the graph above we can see that the most dense area provides us the most options, the further we are out of the city, the fewer options there are.The review rating is presented by the color, the users can spot the areas which have high review ratings, and have the overall idea where has more options for the airbnb. Though, review rating does not have a strong correlation with the price.

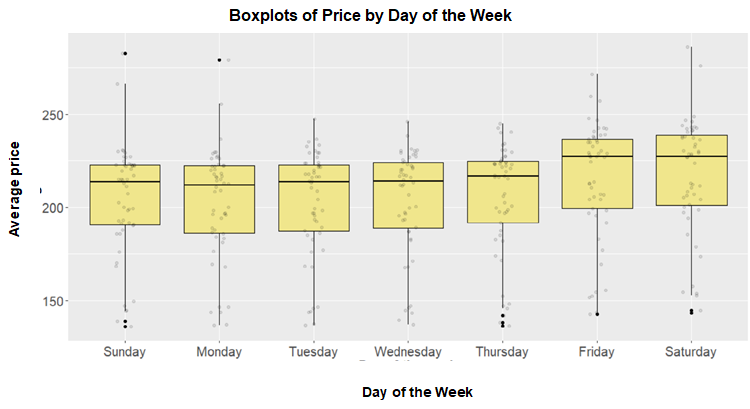


Then, we wanted to see what type of airbnb can we get with our certain budget and what is the availability of the desired room type. We divided the price into 4 different price range, $0-74, $75-149, $150-249,$250 to higher. The graph shows the trend of room type in different price range. As we can see, the number of private room listings keep decreasing as the price increases, conversely, the number of entire room listings have dramatic increase with the price. The users can easily find their desired type of room within their certain budget by looking at this graph.

The graph on the right side shows the sentiment analysis of reviews given by the users for the listings in fenway/kenmore area. This is one of the interesting finding which we got when doing our EDA on the reviews. From this graph a user can get a good idea about the overall experience of other users who had previously used airbnb in that particular area and can compare reviews based on Positive and Negative sentiments. An overall more positive reviews for the listings in an area means that customers are satisfied and a user can find good listings in that area. An overall more negative reviews means customers are not at all satisfied and by looking at the main words used in the reviews a user can find out what are the major drawbacks and what kind of problems are there in that area. This graph can help users in refining their search for listings and can make their work easy.

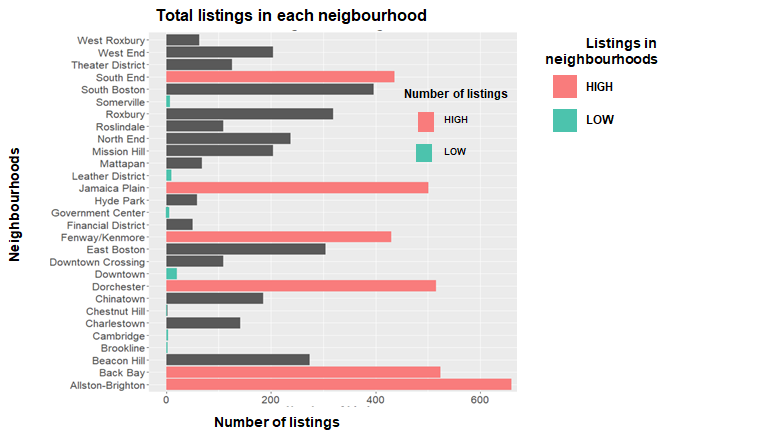
**3.2 Analysing the factors affecting price**

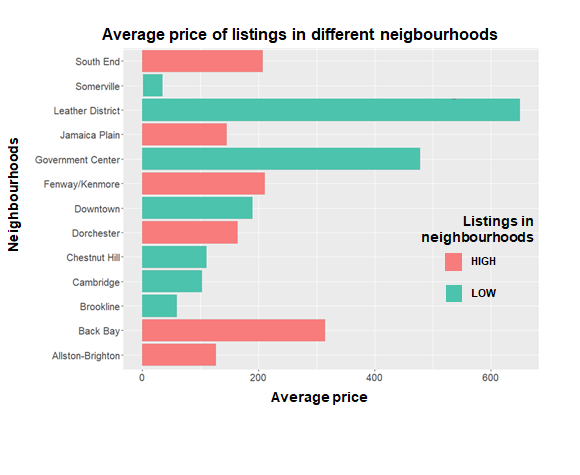
In this section, we have conducted some analysis to know the factors that are affecting the price of Airbnb listings in Boston. Considering these factors, will be helpful to book Airbnb, if there is any budget constraint.

One of the major factor that affects the pricing of the listings is the duration of your stay. It plays a major role in pricing. That is what this graph is depicting. We analyzed and found out that the mean price is almost same on the weekdays and there is not much variance in pricing but on weekends including friday we can see that there is sudden increase in the mean price of the listings which is meant to be obvious because family and friends often come in town to meet their relatives or to attend some parties or functions and there is sudden increase in demand which leads to increase in price on the weekends. 



After that, using data from “Listings” table, we analyzed how the number of beds is affecting the price. First, we converted price in categorical variable which was continuous before. As we can see, in the bar plot on left side, as the number of bed increases, number of listings for lower price range seems to decrease, but the listings for higher price range seems to increase. From this plot it can be concluded that there is direct correlation between the number of beds and price.

Next, we wanted to see if there is any impact of number of listings in price or not. So, first we figured out total number of listings in each neighbourhoods using “neighbourhood” column instead of “neighbourhood\_cleansed”.Because, we didn’t want to eliminate neighbourhood with very low number of listings. In the bar plot on right side, we can see that some neighbourhoods have very high number of listings and some have very low number of listings. As per the findings, we assumed that higher number of listings could be higher demand for Airbnb. So, higher demand could affect the price. 

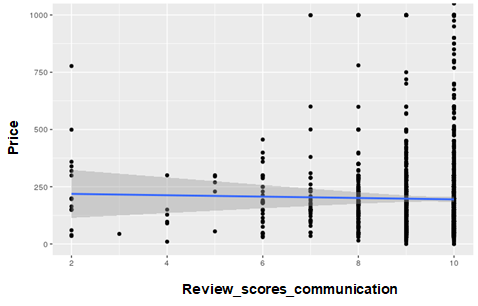
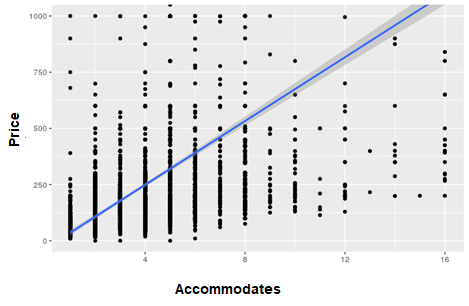


Before coming to any conclusion, we performed another EDA, for which we chose only neighbourhoods with very high and very low number of listings and calculated their average price. In the bar plot on left side, as we can see, even if the listings are high in the suburb neighbourhoods like Allston-Brighton, Dorchester, their average price are relatively lower than the neighbourhoods like Leather District, Government Center which are located inside the city and have very low number of listings, perhaps due to close proximity of the neighbourhoods to the Boston and with the place of attraction.

So, it can be concluded that the location of neighbourhood has more impact on price rather than the number of listings.

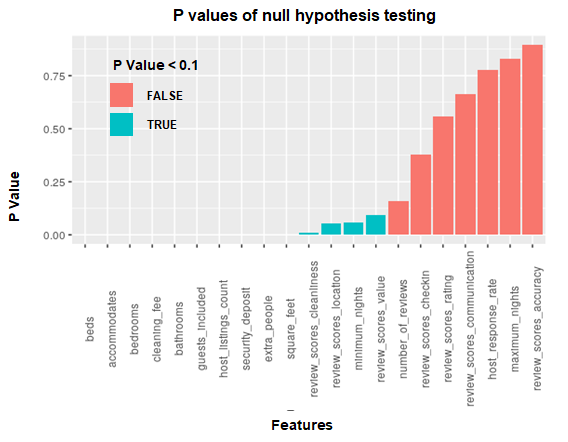
**3.3 Selecting features for a price model**

Our feature selection process followed three primary steps: visual feature analysis, hypothesis testing, and correlation analysis. During each step of the process, we found variables that we believed to be good candidates based on the analysis, and then made an overall decision based on how the variables performed across all three steps. We began first by doing visual analysis on our dataset’s numeric variables, creating scatter plots of our variables against price, and overlaying a line of best fit. We looked for variables that showed possible or likely correlation in any direction, and selected those to be candidates for our model features.



Comparing the two scatter plots above, we were looking for graphs and trends similar to what is seen in the left figure. The left graph presents a clear positive relationship between the variables. The graph on the right represents an example of a variable performing poorly, where the relationship between the candidate variable and the price is either unclear or not present.

We performed a similar visual analysis on our categorical variables, using boxplots to show mean and distribution of each categorical value. For each categorical variable in the dataset we looked to see if there were noticeable differences in the medians and quartile barriers of price distribution. If we noticed clear differences, the variable was considered to be a good candidate for our model.



From the visualisation portion of our analysis, we found the following features to be good candidates for a model of price: *neighbourhood, property\_type, room\_type, accommodates, square\_feet, bedrooms, extra\_people, guests\_included, bathrooms, beds*

Next, we moved to the hypothesis testing portion of our analysis. We tested each variable against our null hypothesis that the variable did not have a relationship with price. We then did p-value analysis at a significance level of .1. While strong evidence against the null hypothesis does not imply that there is a predictive relationship between the two variables, we used the variables with strong evidence against the null hypothesis during testing as candidate variables to be considered holistically including the results from our visual analysis and correlation analysis. Below is a chart outlining the variables we found to have significant evidence against our null hypothesis, with those variables being colored in blue.

Lastly, we performed a correlation analysis against our variables to ensure that we were not including variables that were highly correlated with each other, and therefore unlikely to contribute to a model where both are included. During our analysis, we considered any two variables with an R2 value of .7 or higher to be highly correlated.

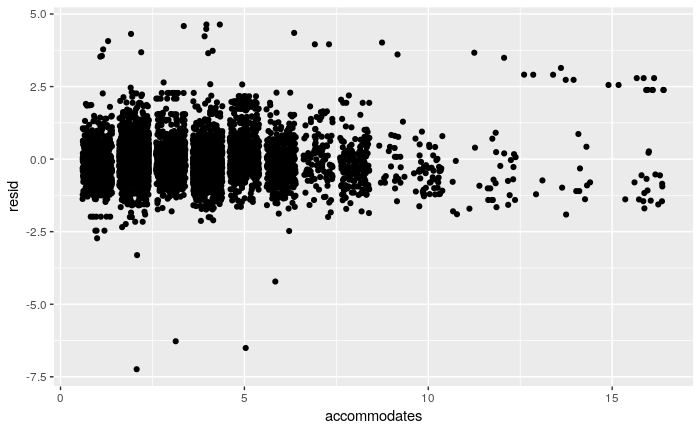
From our analysis, we identified two distinct groups of highly correlated variables. We found accommodates, beds, bedrooms, and square\_feet to all be highly correlated. We also found all of the variables related to the different review score ratings to be highly correlated. From this analysis we decided that even if we found multiple variables in a certain group to be predictive from our previous steps, that we should be careful about including multiple variables from one group in our model.

Our initial selection of features included the following six features: accommodates, bathrooms, neighbourhood, property\_type, room\_type, square\_feet

However, after beginning to create our model, we found a number of our original features to be somewhat redundant, and eliminated them from our model to create a simpler model which was no less accurate. The final features which we built our model off of were: **accommodates, neighbourhood, and room\_type.**

4. Modelling

Our final model was against the log2 transformation of price. We used the features accommodates, neighbourhood, and room type as predictors for the price variable.

Given that the variables neighbourhood and room type were both categorical, we elected to transform these variables into numeric variables to make them simpler to include in our model. For each of these two variables, we substituted the variable for the mean price of listings with that value. I.e, we replaced the value “alliston” for neighbourhood with the average price of all listings with the neighbourhood value “alliston.” This was done for both neighbourhood and room type.

Our final model resulted in a 5-fold cross validated RMSE of .749. After getting this value, we analysed the residuals of our model against our explanatory variables. For each of our three explanatory variables, we found graphs similar to the figure above. For all three of our variables, the residual graph showed no obvious relationship, and the residuals appeared to be normally distributed around 0. This gave us confidence that the assumptions we had made during our linear modeling held true.

4. Discussion

Overall we feel that the analysis we have done, both during the EDA portion as well as during the modeling portion, can be impactful for a person on either side of an AirBNB transaction. We feel that the EDA portion of our analysis demonstrated several ways to analyze the price of a listing, in addition to several factors that affect the price such as location and day of the week. We also feel that the model itself could be helpful to either party. Those listing their property on AirBNB may be able to get a more accurate valuation of their listing. Those searching for a listing can use the model to better understand what goes into the price of a listing, and tailor their search based on that information, and the features which are important to them but may not affect the price.

We also feel like there are several ways we could improve on the work we did during this project. We feel that while our model performed well, there was still a decent amount of variation in the price of a property that was unexplained by the model. Our belief is that this is likely due to features that either are not present in our dataset, or that we did not use. We believe that something like “year built” could help to explain some of the variation, but this variable was not available to us. We also believe that some of the text based features that were available to us, such as “description” may explain the variance.

5. Statement of contributions

**Anjita Shrestha:** Contributed substantially in, data transformation, plotting, analyzing and interpreting EDAs, organizing Rmd files and PPT with all the group members. Contributed in formatting plots in a presentable manner.

**Rahul Tyagi:** Plotted and analyzed graphs for the Exploratory data analysis and cleaned the data for the modelling process.Contributed in feature extraction and modelling along with other members.

**Nick Tyler:** Performed feature analysis and hypothesis testing for our price model. Did initial cleaning and standardisation of our dataset, including making sure columns were properly formatted. Performed residual analysis of the price model. Contributed to PPT and report with other group members.

**Zhaoxi Zhang:** Plotted and analyzed EDAs, and its transformation and interpretation. Organized Rmd files and PPT with all the group members.

References

<http://insideairbnb.com/get-the-data.html>

<https://towardsdatascience.com>

<https://www.google.com/maps>

Appendix

**Pre-work:**

**1.1**

HousingData %>%

filter(stripDollars(price) < 30)

# Some useful helper functions

stripDollars <- function(dollarValue) {

as.numeric(gsub("[\\$,]", "", dollarValue))

}

stripDollars("$10.50")

stripDollars("$1,500.24")

unPercent <- function(percentValue) {

as.numeric(gsub("[\\%]", "", percentValue)) / 100

}

unPercent("20%")

unPercent("33.33%")

**1.2**

ModelingData <- transmute(HousingData,

id,

host\_id,

host\_response\_rate=unPercent(host\_response\_rate),

host\_acceptance\_rate=unPercent(host\_acceptance\_rate),

host\_is\_superhost=as.integer(host\_is\_superhost),

host\_listings\_count,

host\_has\_profile\_pic=as.integer(host\_has\_profile\_pic),

host\_identity\_verified=as.integer(host\_identity\_verified),

neighbourhood\_cleansed,

is\_location\_exact=as.integer(is\_location\_exact),

neighbourhood,

property\_type,

room\_type,

accommodates,

bathrooms,

bedrooms,

summary,

beds,

square\_feet,

price=stripDollars(price),

weekly\_price=stripDollars(weekly\_price),

monthly\_price=stripDollars(monthly\_price),

security\_deposit=stripDollars(security\_deposit),

cleaning\_fee=stripDollars(cleaning\_fee),

guests\_included,

extra\_people=stripDollars(extra\_people),

minimum\_nights,

maximum\_nights,

number\_of\_reviews,

review\_scores\_rating,

review\_scores\_accuracy,

review\_scores\_cleanliness,

review\_scores\_checkin,

review\_scores\_communication,

review\_scores\_location,

review\_scores\_value,

lat=latitude,

lon=longitude,

city=city,

requires\_license=as.integer(requires\_license),

instant\_bookable=as.integer(instant\_bookable))

**1.3**

ModelingDataSubset <- transmute(ModelingData,

price,

host\_has\_profile\_pic,

neighbourhood\_cleansed,

property\_type,

room\_type,

accommodates,

square\_feet,

bedrooms,

security\_deposit,

summary,

extra\_people,

cleaning\_fee,

guests\_included,

bathrooms,

beds)

**Output:**

**2.1**

Reviewrating<- mutate(ModelingData,review\_scores\_rating=ifelse(is.na(review\_scores\_rating),0,review\_scores\_rating ))%>%

mutate(Satisfaction=ifelse(review\_scores\_rating>=93,"High",

ifelse(review\_scores\_rating<=93| review\_scores\_rating > 87,"Medium","Low"))) %>%

mutate(price\_ca=cut(price, breaks = c(0, 75, 150, 250, 5000),

labels = c('$0-74', '$75-149', '$150-249', '$250-5000'),

include.lowest = TRUE,

right = FALSE))

Massmap<- ggplot2:: map\_data("state",region = "massachusetts")

ggplot(data=subset(Reviewrating,!is.na(review\_scores\_rating))) +

geom\_polygon(data = Massmap,aes(x=long,

y=lat,group=group), fill="cornsilk",color="red")+

geom\_point(data = Reviewrating,

mapping = aes(x=Reviewrating$lon,

y=Reviewrating$lat,

color=`review\_scores\_rating`),alpha=0.4)+

coord\_fixed(xlim = c(-71.3,-70.9),ylim =c (42.2,42.48))+

theme(axis.text=element\_text(size=15),

axis.title=element\_text(size=13,face="bold"),

legend.text = element\_text(size=12),

legend.title = element\_text(size = 14))+

labs(title = 'Mass Airbnb overall review ratings',

x = 'Latitude',

y = 'Longitude')

**2.2**

price\_list <-transmute(ModelingData,

room\_type,

beds,

price,

accommodates,

neighbourhood)

price\_list1<- price\_list %>%

mutate(price\_range=cut(price, breaks = c(0, 75, 150, 250, 1000),

labels = c('$0-74', '$75-149', '$150-249', '$250-1000'),

include.lowest = TRUE,

right = FALSE,na.rm=TRUE))%>%

filter(!is.na(price\_range),!is.na(beds))%>%

filter(!is.na(neighbourhood))

price\_list1%>%

ggplot(aes(x=price\_range))+ geom\_bar(aes(fill=room\_type), position="dodge")+

ggtitle("Availibility of room types with price")+

theme(axis.text=element\_text(size=15),

axis.title=element\_text(size=13,face="bold"),

legend.text = element\_text(size=12),

legend.title = element\_text(size = 14))+

theme(plot.title = element\_text(hjust =0.5))+

xlab("Price Range")+

ylab("Number of listings")

**2.3**

knitr::opts\_chunk$set(echo = TRUE)

## Setimaental Analysis of User Reviews in Fenway

ModelingDataSubset[is.na(ModelingDataSubset)]<- 0

sum(is.na(ModelingDataSubset))

Reviews\_Fen<-ModelingDataSubset%>%

filter(ModelingDataSubset$neighbourhood\_cleansed=='Fenway')

Fen<-tibble(line=1:length(Reviews\_Fen$summary), text=Reviews\_Fen$summary)

bigrams <- Fen%>%

unnest\_tokens(bigram, text, token="ngrams", n=2) %>%

separate(bigram, c("word1", "word2"), sep = " ")

bigrams %>%

filter(!word1 %in% c("not", "no", "never", "without")) %>%

filter(!word2 %in% c(stop\_words$word)) %>%

inner\_join(get\_sentiments("bing"), by=c("word2"="word")) %>%

count(word2, sentiment, sort=TRUE) %>%

mutate(word2 = reorder(word2, n)) %>%

group\_by(sentiment) %>%

top\_n(10) %>%

ggplot(aes(x=word2, y=n)) +

geom\_col(show.legend=FALSE) +

facet\_wrap(~sentiment, ncol=2, scales="free") +

coord\_flip()

**2.4**

combinedCalendar <-calendar

combinedCalendar$price <- as.numeric(gsub(",", "", substring(combinedCalendar$price, 2)))

groupedCalendarAll <- combinedCalendar %>% group\_by(date = date) %>%

summarise(averagePrice = mean(price, na.rm = TRUE)) %>% mutate(year = year(date), commonYear = paste("2019",substring(date, 6),sep="-"))

groupedCalendarAll <- groupedCalendarAll %>% mutate(day = strftime(date,'%A'))

groupedCalendarAll$day <- factor(groupedCalendarAll$day, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"), labels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

ggplot(groupedCalendarAll, aes(x = factor(day),

y = averagePrice)) +

geom\_boxplot(fill = "khaki", color = "Black") +

geom\_jitter(alpha = 0.1, width = 0.1, color = "black") +

ggtitle("Is Airbnb expensive on weekends?",

subtitle = "Boxplots of Price by Day of the Week") +

labs(x = "Day of the week", y = "Average Price") +

theme(plot.title = element\_text(face = "bold")) +

theme(plot.subtitle = element\_text(face = "bold", color = "Black", size = 15)) +

theme(plot.caption = element\_text(color = "grey68"))+

theme(plot.title = element\_text(hjust =0.5))+

theme(plot.subtitle = element\_text(hjust =0.5))+

theme(axis.text=element\_text(size=13),

axis.title=element\_text(size=15,face="bold"))

**2.5**

price\_list1%>%

filter(beds<=5)%>%

ggplot(aes(x= beds ,fill=price\_range))+

geom\_bar(position = "dodge")+

ggtitle("Number of listings vs. Number beds in different Price range")+

theme(axis.text=element\_text(size=15),

axis.title=element\_text(size=13,face="bold"),

legend.text = element\_text(size=12),

legend.title = element\_text(size = 14))+

xlab("Number of beds")+

ylab("Number of Listings")

**2.6**

avg\_listing <- ModelingData%>%

filter(!is.na(neighbourhood))%>%

filter(neighbourhood !="Allston-Brighton", neighbourhood != "Back Bay" ,

neighbourhood != "Dorchester", neighbourhood != "Jamaica Plain",

neighbourhood !="Fenway/Kenmore", neighbourhood !="South Boston",

neighbourhood !="Somerville", neighbourhood != "Leather District" ,

neighbourhood != "Government Center", neighbourhood != "Downtown",

neighbourhood !="Chestnut Hill", neighbourhood !="Cambridge",

neighbourhood !="Brookline")%>%

group\_by(neighbourhood) %>%

summarise(average\_price = mean(price))

ModelingData%>%

filter(!is.na(neighbourhood))%>%

ggplot( aes(x=neighbourhood))+

geom\_bar() +

labs(x = "Neighbourhood", y = "Number of Listings")+

coord\_flip()+

theme(axis.text=element\_text(size=15),

axis.title=element\_text(size=16,face="bold"),

legend.text = element\_text(size=12),

legend.title = element\_text(size = 16))+

ggtitle("Total listings in each neighbourhood")+

theme(plot.title = element\_text(face = "bold", color = "Black", size = 17))+

theme(plot.title = element\_text(hjust =0.5))

**2.7**

ModelingData%>%

filter(!is.na(neighbourhood))%>%

ggplot( aes(x=neighbourhood))+

geom\_bar() +

labs(x = "Neighbourhood", y = "Number of Listings")+

coord\_flip()+

theme(axis.text=element\_text(size=15),

axis.title=element\_text(size=16,face="bold"),

legend.text = element\_text(size=12),

legend.title = element\_text(size = 16))+

ggtitle("Total listings in each neighbourhood")+

theme(plot.title = element\_text(face = "bold", color = "Black", size = 17))+

theme(plot.title = element\_text(hjust =0.5))

**2.8**

NumericData <- transmute(ModelingData,

price,

host\_response\_rate,

host\_acceptance\_rate,

host\_listings\_count,

accommodates,

bathrooms,

bedrooms,

beds,

square\_feet,

security\_deposit,

cleaning\_fee,

guests\_included,

extra\_people,

minimum\_nights,

maximum\_nights,

number\_of\_reviews,

review\_scores\_rating,

review\_scores\_accuracy,

review\_scores\_cleanliness,

review\_scores\_checkin,

review\_scores\_communication,

review\_scores\_location,

review\_scores\_value)

for (col in names(NumericData)) {

if (col != "price") {

plt <- ggplot(NumericData, aes(x=NumericData[[col]], y=price)) +

geom\_point() +

geom\_smooth(method='lm',formula=y~x) +

coord\_cartesian(ylim = c(0, 1000)) + # Look at the important part of data, not outliers

xlab(col)

print(plt)}}

**2.9**

pv\_frame <- data.frame(column=character(),

p\_value=double(),

stringsAsFactors = FALSE)

for (col in names(NumericData)) {

if (col != "price") {

fit <- NumericData %>%

lm(as.formula(paste("price ~ ", col)), data=.)

row <- data.frame(column=col, p\_value=(glance(fit)$p.value))

colnames(row) <- c("column", "p\_value")

pv\_frame <- rbind(pv\_frame, row)

}

}

ggplot(pv\_frame, aes(x=reorder(column, p\_value))) +

geom\_bar(aes(y=p\_value, fill=(p\_value < .1)), stat = "identity") +

theme(axis.text.x = element\_text(angle=90)) +

xlab("Feature") +

ylab("P Value")

ggplot(pv\_frame, aes(x=reorder(column, p\_value))) +

geom\_bar(aes(y=p\_value, fill=(p\_value < .1)), stat = "identity") +

scale\_y\_continuous(trans = "log10") +

theme(axis.text.x = element\_text(angle=90))

**3.0**

# Doing Predictions: Basic Data tidying and Data Cleaing. Replacing missing values with 0

ModelingDataSubset[is.na(ModelingDataSubset)]<- 0

sum(is.na(ModelingDataSubset))

# From this plot we can see that there is somekind of relationship between the independent and dependent variables

ggplot(ModelingDataSubset,aes(x=log2(price), y=log2(ModelingDataSubset$bedrooms+ModelingDataSubset$accommodates+ModelingDataSubset$bathrooms+ModelingDataSubset$square\_feet))) + geom\_point()

## Now we will try to do modelling with different predictors and try to fit the model.

## Predictions:

predict1<-lm((log2(1+price)) ~((square\_feet+property\_type)),data=ModelingDataSubset)

predict2<-lm(log2(1+price) ~log2(1+bathrooms),data=ModelingDataSubset)

predict3<-lm(log2(1+price) ~ ((square\_feet+bathrooms+accommodates+neighbourhood\_cleansed+room\_type+property\_type)),data=ModelingDataSubset)

predict4<-lm((log(1+price)) ~((square\_feet+property\_type)),data=ModelingDataSubset)

predict5<-lm((log(1+price)) ~((square\_feet\*property\_type\*bedrooms\*accommodates)),data=ModelingDataSubset)

Predict6<-lm(log(1+price) ~((bathrooms+accommodates+bedrooms+room\_type+property\_type+neighbourhood\_cleansed)),data=ModelingDataSubset)

summary(predict1)

summary(predict2)

summary(predict3)

summary(predict4)

summary(predict5)

summary(Predict6)

predict3<-lm(log2(1 + price) ~ accommodates+avg\_neighbourhood\_prc+avg\_room\_type\_prc,data=ModelingDataSubset)

rmse(predict3, ModelingDataSubset)

ModelingDataSubset %>%

add\_residuals(predict3) %>%

ggplot(aes(x=accommodates, y=resid)) + geom\_point(stat='identity', position='jitter')