PSTAT 131 Project Proposal

Matthew Coleman 5267398
Nick Reyes 4089215
Jeff Pittman 5865159
Austin Mac 3065273

Airbnb NYC
https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

**Research Question** What is the research question? Why is it important? What is already known?

Possible research questions include:

- Can we predict anything about the host based on listings, prices, reviews, etc.?
- Can we predict anything about room type?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?
- What can be learned from listings with 0 reviews?
- Does the number of listings in a neighborhood affect the prices of those listings?
- Which hosts are running a business with multiple listings and where are they?

These questions are important because Airbnb has become a major tool for tech-savvy tourists all over the world. Airbnb is upending the hotel industry while simultaneously raising concerns over gentrification. We already have a decent amount of information from Airbnb as it makes publicly available information on host listings, prices, reviews, etc.

**Data** Describe the data source(s) you plan to use. Include verification of data use. – Which is your response variable? – How many predictor variables are there? How many of each type (e.g. numeric, categorical, binary).

- From the dataset page, https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data, we see: license is CC0: Public Domain
- 16 variables - 10 numeric, 6 categorical (date is currently a string but will be converted)
- Possible response variables are price, number of reviews, reviews/month depending on what research question we wish to pursue more

**Analysis Plan** What do you plan to do? – Descriptive statistics: what basic summaries (e.g. mean, median, frequencies) do you plan to report? On which variables? – What exploratory graphics do you plan to include? – What type(s) of statistical machine learning (supervised or unsupervised) are you planning? – Model building: how do you plan to train your classifier? e.g. cross-validation. – Model validation: how many hold-out observations are planned?

- Supervised Learning: Plot heatmap of lat vs. long and classify using tree, knn, lda…

- Summary statistics of price by neighborhood, ex: min, max, median price of properties in Kensington, Midtown, Harlem
- Plot frequencies of properties by neighborhood
- Our initial classifier was trained with 5 fold CV
- There are ~8000 observations

**References**

http://insideairbnb.com/about.html
https://medium.com/datadriveninvestor/airbnb-listings-analysis-in-toronto-october-2018-2a5358bae007
https://rstudio-pubs-static.s3.amazonaws.com/365075_ec9ebe4da4cc465ba9beaef25cda6bad.html

**Data Overview** Write R code to complete the following and include as an appendix in your Rmarkdown file. – Read the dataset(s) in. – Report the dimensions of each. – Summarize the missingness in the data. – Split the data into training, validation (optional), and test sets. Report dimensions. – Using only your training set: * Summarize the response variable using numerical or graphical summaries. * Fit a very basic model to the training set. e.g. a logistic regression with 1 or 2 explanatory v

# Data Overview

*Matthew Coleman*

*2/12/2020*

Data Overview Write R code to complete the following and include as an appendix in your Rmarkdown file. – Read the dataset(s) in. – Report the dimensions of each. – Summarize the missingness in the data. – Split the data into training, validation (optional), and test sets. Report dimensions. – Using only your training set: * Summarize the response variable using numerical or graphical summaries. * Fit a very basic model to the training set. e.g. a logistic regression with 1 or 2 explanatory v

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidy
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.2     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
bnb <- read.csv('AB_NYC_2019.csv')
dims <- dim(bnb)

sprintf('Our dataset has %d observations and %d attributes', dims[1],dims[2])
```

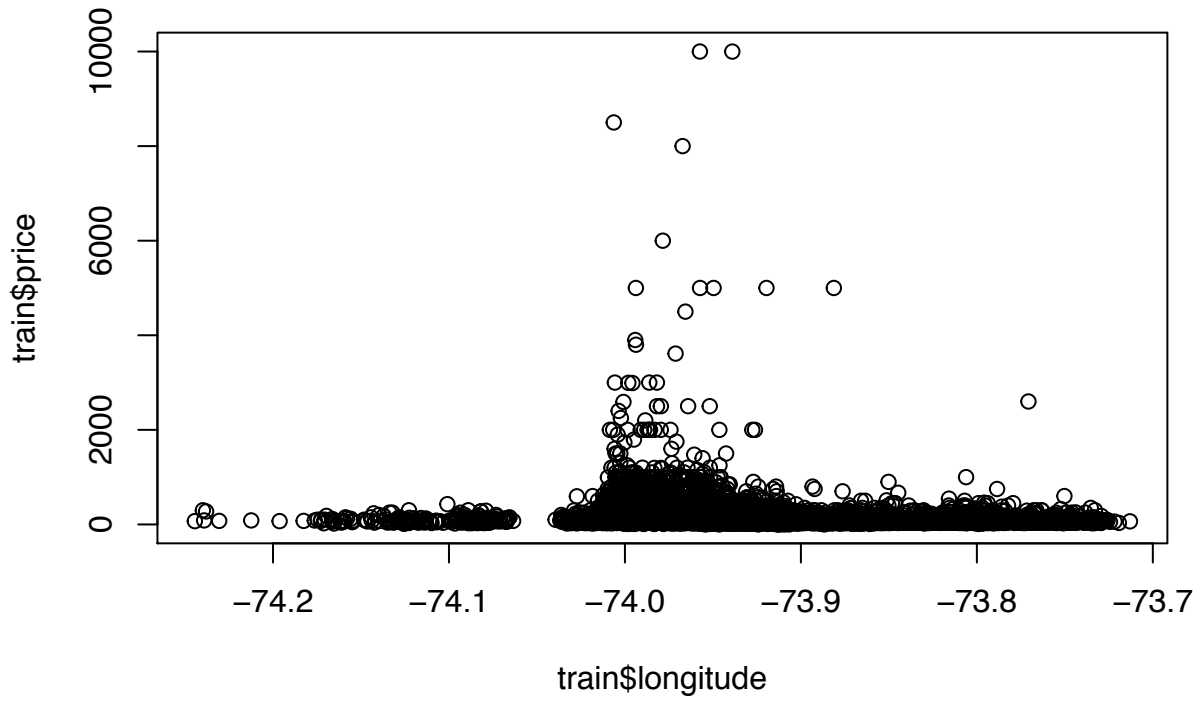
```
## [1] "Our dataset has 48895 observations and 16 attributes"
```

```
num_com <- sum(complete.cases(bnb))

per_com <- sum(complete.cases(bnb))/nrow(bnb) * 100

cat(sprintf('there are %d number of complete cases, which amounts to %2f percent of the dataset',
            num_com, per_com ))
```

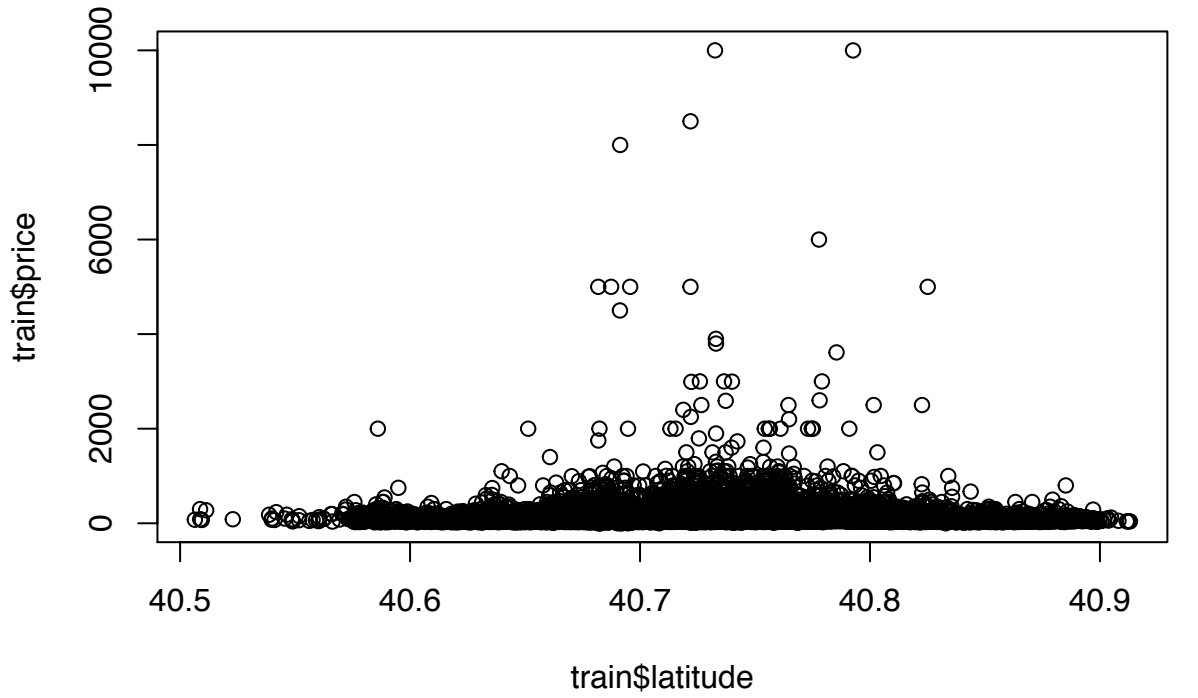
```
## there are 38843 number of complete cases, which amounts to 79.441661 percent of the dataset
```

```
bnb <- bnb[complete.cases(bnb),]
set.seed(123)

train.ind <- sample(1:nrow(bnb),nrow(bnb)*.80 )

train <- bnb[train.ind,]
test <- bnb[-train.ind,]
```

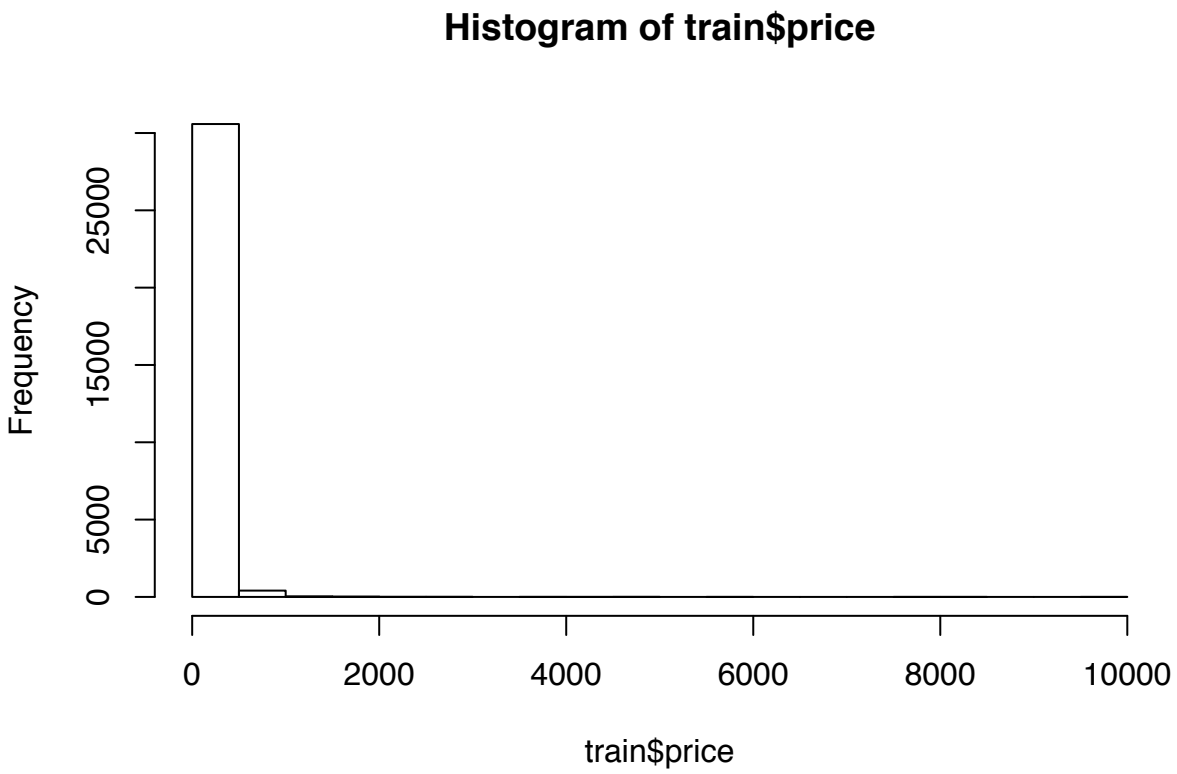
```
plot(train$longitude, train$price)
```




```
plot(train$latitude, train$price)
```




```
hist(train$price)
```




As we can see by the pair plot, there is a serious issue of overplotting, which may require that we reduce the number of observations in the dataset we analyze. We may also need log transformations on some of the variables such as the price. We will dive more in depth with the data cleaning when we begin to work more with the dataset.

```
train <- train %>% mutate(price_greater = ifelse(train$price>median(train$price),1,0))
```

```
gfit <- glm(price_greater~longitude + latitude, data = train, family = 'binomial')
summary(gfit)
```

```
##
## Call:
## glm(formula = price_greater ~ longitude + latitude, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8750  -1.1100  -0.2026   1.0506   2.8491
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1371.6715    28.2620  -48.53   <2e-16 ***
## longitude     -16.6905     0.3437  -48.56   <2e-16 ***
## latitude        3.3724     0.2219   15.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43078  on 31073  degrees of freedom
```

```
## Residual deviance: 39835  on 31071  degrees of freedom
## AIC: 39841
##
## Number of Fisher Scoring iterations: 4
```