Exploratory Data Analysis

Art Tay

Appendix - Code

```
# Libraries
library(tidyverse)
library(VIM)
library(mice)
```

```
# Load in Data
data_full <- read.csv("AB_NYC_2019.csv", stringsAsFactors = T, header = T)
dim(data_full)
```

```
## [1] 48895    16
```

```
colnames(data_full)
```

```
##  [1] "id"                             "name"
##  [3] "host_id"                        "host_name"
##  [5] "neighbourhood_group"            "neighbourhood"
##  [7] "latitude"                       "longitude"
##  [9] "room_type"                      "price"
## [11] "minimum_nights"                 "number_of_reviews"
## [13] "last_review"                    "reviews_per_month"
## [15] "calculated_host_listings_count" "availability_365"
```

```
str(data_full)
```

```
## 'data.frame':    48895 obs. of  16 variables:
##  $ id                            : int  2539 2595 3647 3831 5022 5099 5121 5178 5203 5238 ...
##  $ name                          : Factor w/ 47906 levels "","'Fan'tastic",..: 12573 38017 45019 155
##  $ host_id                       : int  2787 2845 4632 4869 7192 7322 7356 8967 7490 7549 ...
##  $ host_name                     : Factor w/ 11453 levels "","'Cil","-TheQueensCornerLot",..: 4997 4
##  $ neighbourhood_group           : Factor w/ 5 levels "Bronx","Brooklyn",..: 2 3 3 2 3 3 2 3 3 3 ...
##  $ neighbourhood                 : Factor w/ 221 levels "Allerton","Arden Heights",..: 109 128 95 42
##  $ latitude                      : num  40.6 40.8 40.8 40.7 40.8 ...
##  $ longitude                     : num  -74 -74 -73.9 -74 -73.9 ...
##  $ room_type                     : Factor w/ 3 levels "Entire home/apt",..: 2 1 2 1 1 1 2 2 2 1 ...
##  $ price                         : int  149 225 150 89 80 200 60 79 79 150 ...
##  $ minimum_nights                : int  1 1 3 1 10 3 45 2 2 1 ...
##  $ number_of_reviews             : int  9 45 0 270 9 74 49 430 118 160 ...
##  $ last_review                   : Factor w/ 1765 levels "","2011-03-28",..: 1503 1717 1 1762 1534 1
##  $ reviews_per_month             : num  0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...
##  $ calculated_host_listings_count: int  6 2 1 1 1 1 1 1 1 4 ...
##  $ availability_365              : int  365 355 365 194 0 129 0 220 0 188 ...
```

Data Cleaning

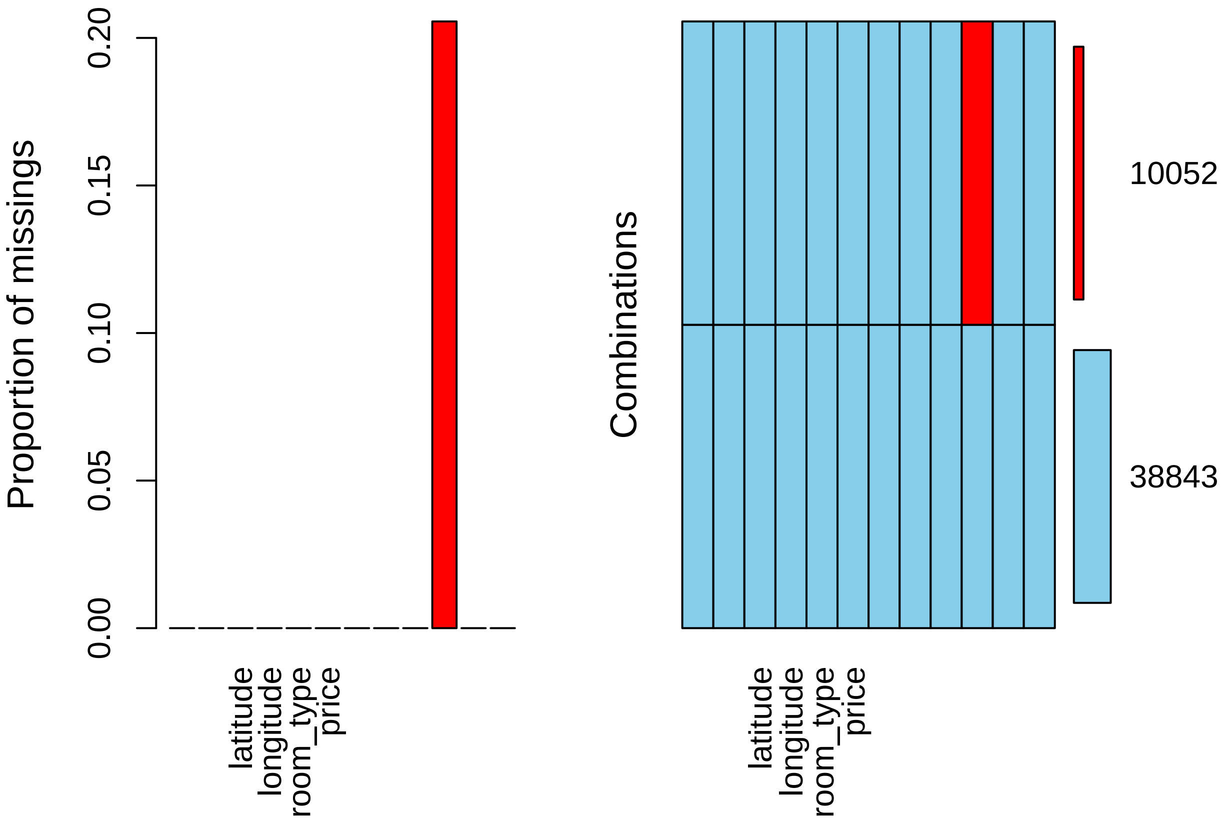
```r
# Data cleaning

# Removing uninformative variables (names).
data_quant <- data_full %>% select(-c(id, host_id, name, host_name))
str(data_quant)
```

```
## ’data.frame’:    48895 obs. of  12 variables:
##  $ neighbourhood_group           : Factor w/ 5 levels "Bronx","Brooklyn",..: 2 3 3 2 3 3 2 3 3 3 ...
##  $ neighbourhood                 : Factor w/ 221 levels "Allerton","Arden Heights",..: 109 128 95 42
##  $ latitude                      : num  40.6 40.8 40.8 40.7 40.8 ...
##  $ longitude                     : num  -74 -74 -73.9 -74 -73.9 ...
##  $ room_type                     : Factor w/ 3 levels "Entire home/apt",..: 2 1 2 1 1 1 2 2 2 1 ...
##  $ price                         : int  149 225 150 89 80 200 60 79 79 150 ...
##  $ minimum_nights                : int  1 1 3 1 10 3 45 2 2 1 ...
##  $ number_of_reviews             : int  9 45 0 270 9 74 49 430 118 160 ...
##  $ last_review                   : Factor w/ 1765 levels "","2011-03-28",..: 1503 1717 1 1762 1534 17
##  $ reviews_per_month             : num  0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...
##  $ calculated_host_listings_count: int  6 2 1 1 1 1 1 1 1 4 ...
##  $ availability_365              : int  365 355 365 194 0 129 0 220 0 188 ...
```

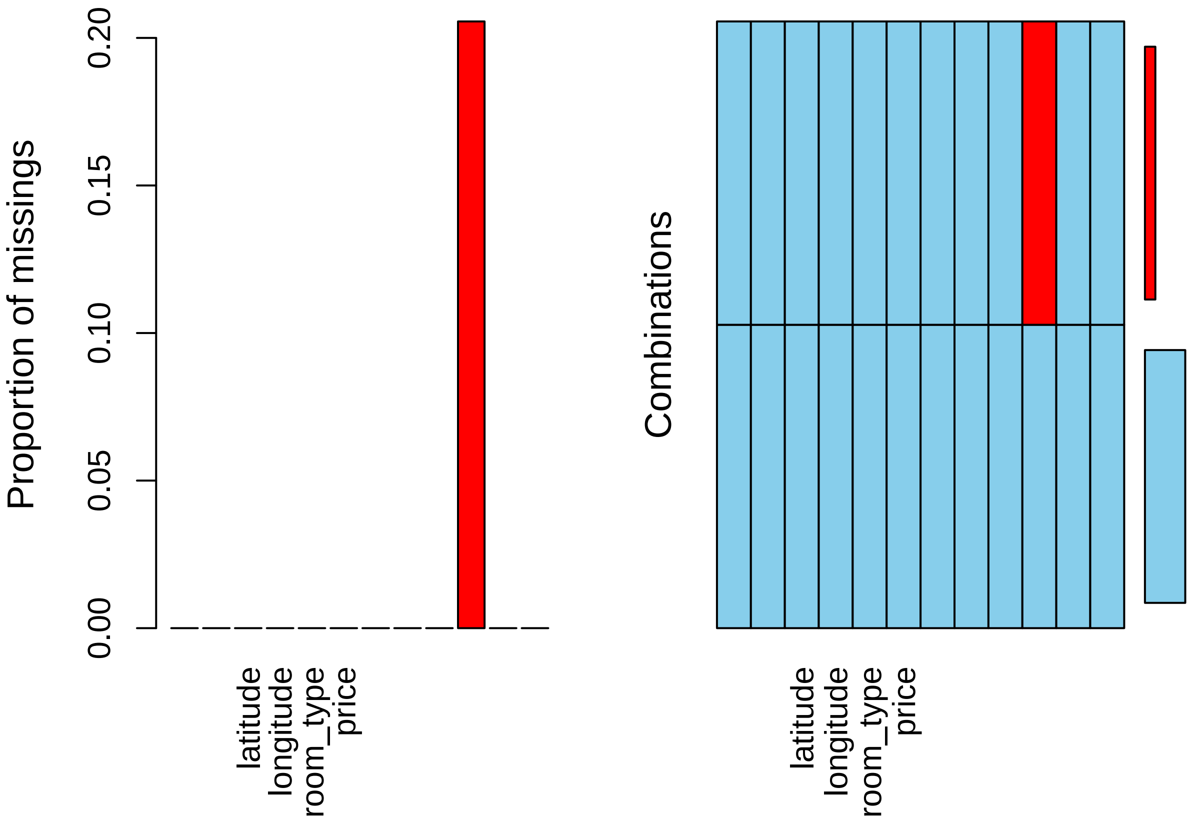
```r
# Missing data.

# Code value that might mean missing.
# price == 0 -> NA
# lattitude == 0 -> NA
# longitude == 0 -> NA
# min_night == 0 -> NA
# factors == "" or " " -> NA

# A functions that checks values of factors to
# see if they are " " or "".
# If they are the function replaces them with NA.
# Otherwise it returns the original value.
check_empty_string <- function(x){
    return(x)
}

data_quant_mis <- data_quant %>%
    mutate(price, ifelse(price == 0, NA, price)) %>%
    mutate(latitude, ifelse(latitude == 0, NA, latitude)) %>%
    mutate(longitude, ifelse(longitude == 0, NA, longitude)) %>%
    mutate(minimum_nights, ifelse(minimum_nights == 0, NA, minimum_nights))

head(data_quant_mis)
```

```
##   neighbourhood_group neighbourhood latitude longitude       room_type price
## 1            Brooklyn    Kensington 40.64749 -73.97237    Private room   149
## 2           Manhattan       Midtown 40.75362 -73.98377 Entire home/apt   225
## 3           Manhattan        Harlem 40.80902 -73.94190    Private room   150
## 4            Brooklyn  Clinton Hill 40.68514 -73.95976 Entire home/apt    89
## 5           Manhattan   East Harlem 40.79851 -73.94399 Entire home/apt    80
## 6           Manhattan   Murray Hill 40.74767 -73.97500 Entire home/apt   200
##   minimum_nights number_of_reviews last_review reviews_per_month
## 1              1                 9  2018-10-19              0.21
## 2              1                45  2019-05-21              0.38
## 3              3                 0                            NA
## 4              1               270  2019-07-05              4.64
## 5             10                 9  2018-11-19              0.10
## 6              3                74  2019-06-22              0.59
##   calculated_host_listings_count availability_365 ifelse(price == 0, NA, price)
## 1                              6              365                           149
## 2                              2              355                           225
## 3                              1              365                           150
## 4                              1              194                            89
## 5                              1                0                            80
## 6                              1              129                           200
##   ifelse(latitude == 0, NA, latitude) ifelse(longitude == 0, NA, longitude)
## 1                            40.64749                             -73.97237
## 2                            40.75362                             -73.98377
## 3                            40.80902                             -73.94190
## 4                            40.68514                             -73.95976
## 5                            40.79851                             -73.94399
## 6                            40.74767                             -73.97500
##   ifelse(minimum_nights == 0, NA, minimum_nights)
## 1                                               1
## 2                                               1
## 3                                               3
## 4                                               1
## 5                                              10
## 6                                               3
```

```
# Check for amount and types of missing data.
mis_plot <- aggr(data_quant, number = T, prop = c(T, F))
```



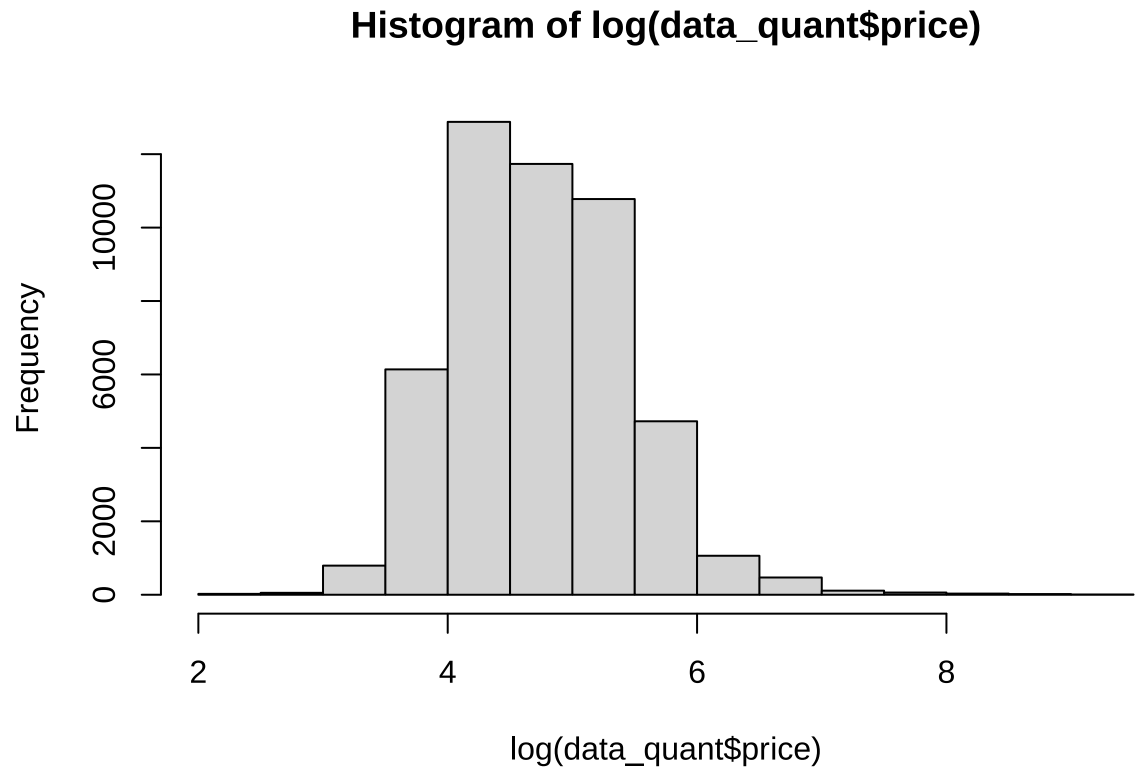
```
plot(mis_plot)
```



Feature Engineering

Visualizations

```
# Histogram of price
hist(log(data_quant$price)) # very right skewed - expected
```



```r
# boxplot by neighborhood
plot_1 <- data_quant %>%
    ggplot(aes(x = neighbourhood_group, y = log(price))) +
    geom_boxplot()
```