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)
#colnames(data_full)
#str(data_full)
```

Data Cleaning

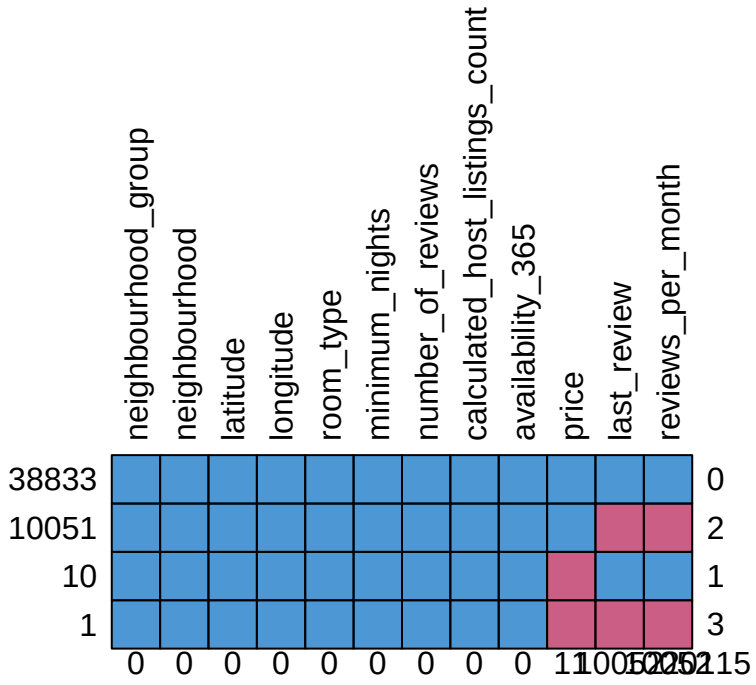
```
# Data cleaning

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

```
# Missing data.

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

# 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(ifelse(x == " " | x == "", NA, x))
}

data_quant_mis <- apply(data_quant_mis, MARGIN = 2, FUN = check_empty_string)

data_quant_mis <- as.data.frame(data_quant_mis)

colnames(data_quant_mis) <- colnames(data_quant)
```

```
# Plot the percentage and patterns of missing values.
missing_percent <- apply(data_quant_mis, MARGIN = 2,
    FUN = function(x){sum(is.na(x)) / length(x)})

# Filter out non-missing variables
missing_percent <- as.data.frame(missing_percent) %>%
                   filter(missing_percent > 0)

# Add variable names to the data frame.
missing_percent$Variable <- c("price", "last review date", "reviews per month")

# Round and change proportion to percentages.
missing_percent$missing_percent <- round(
        missing_percent$missing_percent * 100, 2)

plot_1 <- missing_percent %>%
          ggplot(aes(x = reorder(Variable, missing_percent),
            y = missing_percent, fill = Variable)) +
          geom_bar(stat = "identity") +
          geom_text(aes(label = missing_percent), vjust = 1.6) +
          theme_bw() + theme(legend.position = "none") +
          ggtitle("Percentages of Missing Values by Variable") +
          ylab("Percent Missing") + xlab("")
```

```
# Plot the pattern of missing values
md.pattern(data_quant_mis, rotate.names = T)
```




```
##       neighbourhood_group neighbourhood latitude longitude room_type
## 38833                   1             1        1         1         1
## 10051                   1             1        1         1         1
## 10                      1             1        1         1         1
## 1                       1             1        1         1         1
##                         0             0        0         0         0
##       minimum_nights number_of_reviews calculated_host_listings_count
## 38833              1                1                               1
## 10051              1                1                               1
## 10                 1                1                               1
## 1                  1                1                               1
##                    0                0                               0
##       availability_365 price last_review reviews_per_month
## 38833                1     1           1                1     0
## 10051                1     1           0                0     2
## 10                   1     0           1                1     1
## 1                    1     0           0                0     3
##                      0    11       10052            10052 20115
```

```
# boxplot by neighborhood
plot_2 <- data_quant %>%
    ggplot(aes(x = neighbourhood_group, y = log(price))) +
    geom_boxplot()
```