Determining the Price of San Francisco AirBnB Listings

The Outliers: Hailey han, Joshua Minwoo Kang, Jihong Lee

```
#### Initial Setup

# read in dataset
setwd("/Users/jihonglee/Dropbox/Spring-2020/STAT 346/Final Project")
listings <- read.csv("listings.csv")
dim(listings)
```

```
## [1] 8111  106
```

```
# choose subset with variables of interest
listings <- subset(listings, select=c("price", "host_response_rate", "host_is_superhost", "neighbourhood",
                                      "latitude", "longitude", "property_type", "room_type", "accommodates",
                                      "bathrooms", "bedrooms", "number_of_reviews", "review_scores_rating"))

# organize into variables for later ease
listings$price <- as.numeric(gsub("[$]", "", listings$price))
listings$host_response_rate <- as.numeric(gsub("[%]", "", listings$host_response_rate))

# clean of invalid observations
listings <- listings[complete.cases(listings), ]
dim(listings)
```

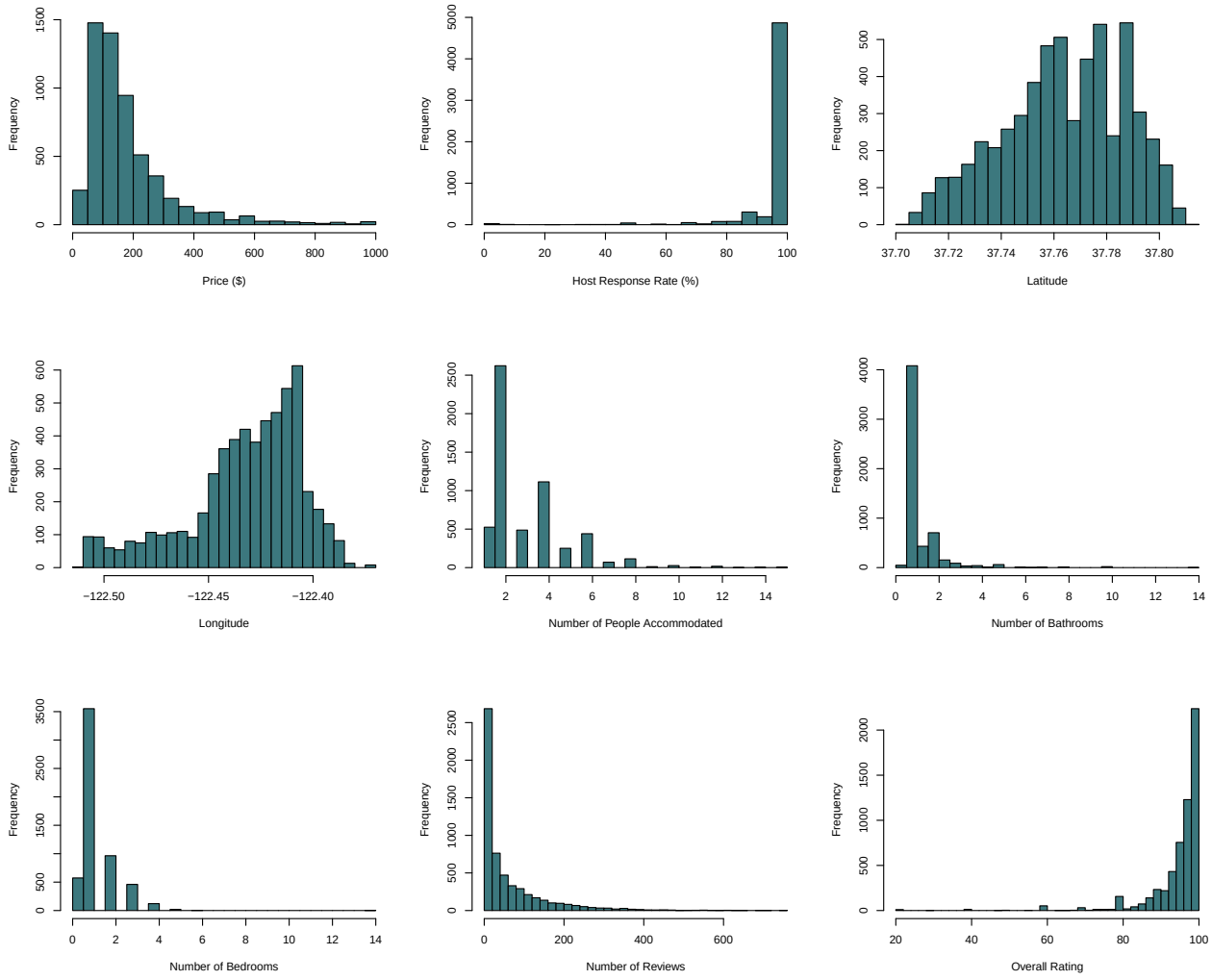
```
## [1] 5692   13
```

```
#### Exploratory Data Analysis

# quantitative variables in dataset: price, host response rate, latitude, longitude, accommodates, bathrooms, bedrooms, number of reviews, review scores rating
listings_quant <- subset(listings, select=c("price", "host_response_rate", "latitude", "longitude", "accommodates",
                                            "bathrooms", "bedrooms", "number_of_reviews", "review_scores_rating"))
# categorical variables in dataset: host is superhost, neighborhood, property type, room type
listings_cat <- subset(listings, select=c("host_is_superhost", "neighbourhood", "property_type", "room_type"))

## Univariate EDA

# statistical display of quantitative variables
l_mean <- apply(listings_quant, 2, mean)
l_sd <- apply(listings_quant, 2, sd)
l_med <- apply(listings_quant, 2, median)
l_iqr <- apply(listings_quant, 2, IQR, na.rm=T)
l_min <- apply(listings_quant, 2, min)
l_max <- apply(listings_quant, 2, max)

statsum <- data.frame(l_mean, l_sd, l_med, l_iqr, l_min, l_max)
rownames(statsum) <- c('Price ($)', 'Host Response Rate (%)', 'Latitude', 'Longitude', 'Number of People', 'Bathrooms', 'Bedrooms', 'Number of Reviews', 'Review Scores Rating')
colnames(statsum) <- c("Mean", "Standard Deviation", "Median", "IQR", "Minimum", "Maximum")
statsum
```

```
##                                    Mean Standard Deviation     Median
```

```
## Price ($)                       188.361033       147.67957394  145.00000
## Host Response Rate (%)           97.093992        10.27417767  100.00000
## Latitude                         37.763858         0.02303580   37.76452
## Longitude                      -122.432058         0.02754039 -122.42649
## Number of People Accommodated     3.115952         1.83675125    2.00000
## Number of Bathrooms               1.369290         0.94945646    1.00000
## Number of Bedrooms                1.311490         0.88295209    1.00000
## Number of Reviews                60.459768        86.79206402   24.00000
## Overall Rating                   95.237175         7.74988807   98.00000
##                                        IQR    Minimum    Maximum
## Price ($)                      126.0000000    0.00000  999.00000
## Host Response Rate (%)           0.0000000    0.00000  100.00000
## Latitude                         0.0337875   37.70474   37.81031
## Longitude                        0.0337400 -122.51306 -122.37043
## Number of People Accommodated    2.0000000    1.00000   15.00000
## Number of Bathrooms              0.5000000    0.00000   14.00000
## Number of Bedrooms               1.0000000    0.00000   14.00000
## Number of Reviews               77.0000000    1.00000  757.00000
## Overall Rating                   6.0000000   20.00000  100.00000
```

```r
# graphical display of quantitative variables
par(mfrow=c(3,3))
quantvars <- with(listings, cbind(price, host_response_rate, latitude, longitude, accommodates, bathroo
quantnames <- c('Price ($)', 'Host Response Rate (%)', 'Latitude', 'Longitude', 'Number of People Accom
for(i in 1:quantnum) {
  hist(quantvars[, i], main="",
       xlab=quantnames[i], ylab="Frequency",
       breaks=30, col="#3C787E")
}
```




```
# statistical display of categorical variables
superhost_tab <- table(listings$host_is_superhost)
neighborhood_tab <- table(listings$neighbourhood)
property_tab <- table(listings$property_type)
room_tab <- table(listings$room_type)

superhost_tab
```

```
##
##    f    t
## 2643 3049
```

```
neighborhood_tab
```

```
##
##          Alamo Square        Balboa Terrace               Bayview
##                    37                    43                   155
##        Bernal Heights             Chinatown         Civic Center
##                   297                    60                    14
##           Cole Valley           Cow Hollow       Crocker Amazon
```

```
##                   81                  37                  77
##            Daly City      Diamond Heights             Dogpatch
##                    8                  17                  21
##             Downtown      Duboce Triangle            Excelsior
##                  232                  70                 110
##   Financial District    Fisherman's Wharf          Forest Hill
##                   28                  29                  10
##            Glen Park       Haight-Ashbury         Hayes Valley
##                   60                 176                  77
##            Ingleside         Inner Sunset            Japantown
##                   47                 117                   4
##            Lakeshore         Lower Haight               Marina
##                   39                  71                  70
##          Mission Bay     Mission District      Mission Terrace
##                   12                 523                  65
##             Nob Hill          Noe Valley          North Beach
##                  217                 270                  36
##            Oceanview         Outer Sunset      Pacific Heights
##                   40                 320                 121
##             Parkside             Portola         Potrero Hill
##                   38                  54                 163
##             Presidio     Presidio Heights    Richmond District
##                    1                  15                 322
##          Russian Hill          Sea Cliff                 SoMa
##                   63                   1                 329
##           South Beach          Sunnyside       Telegraph Hill
##                   42                  83                  95
##            Tenderloin         The Castro           Twin Peaks
##                   85                 262                  58
##          Union Square    Visitacion Valley          West Portal
##                  112                  60                  10
## Western Addition/NOPA
##                  308
```

```
property_tab
```

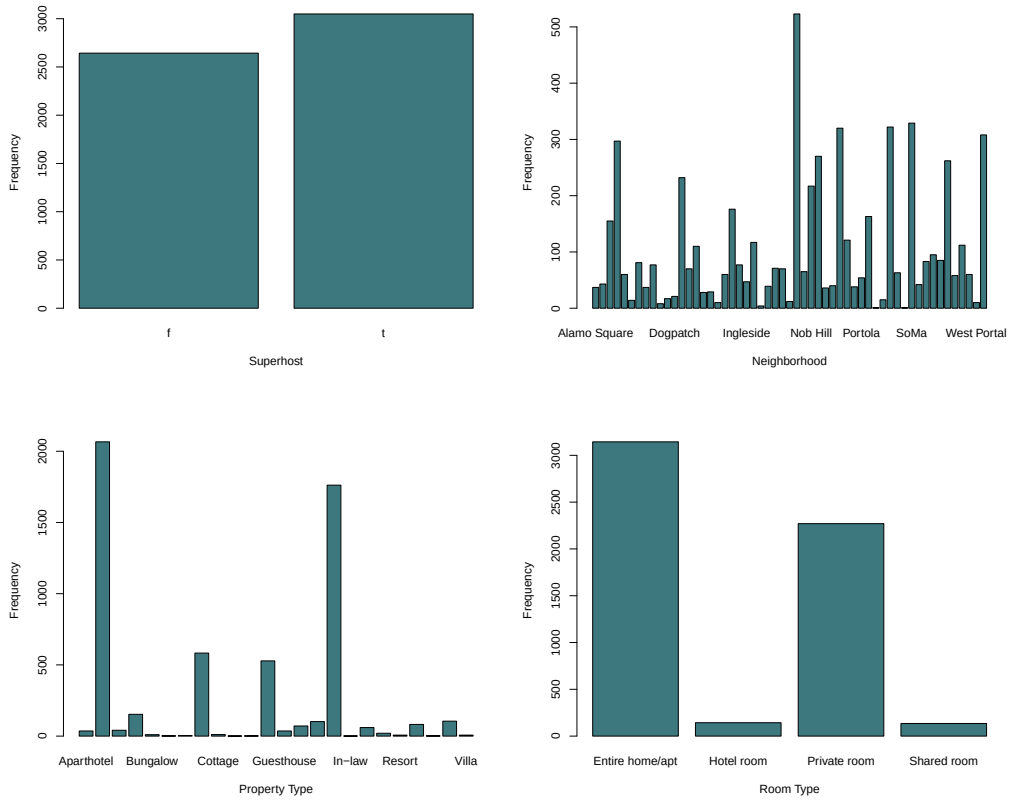
```
##
##         Aparthotel          Apartment  Bed and breakfast     Boutique hotel
##                 36               2066                 41                153
##           Bungalow              Cabin             Castle        Condominium
##                 10                  2                  4                583
##            Cottage         Dome house        Earth house        Guest suite
##                 11                  1                  2                528
##         Guesthouse             Hostel              Hotel              House
##                 36                 71                102               1762
##             In-law              Loft              Other             Resort
##                  1                 60                 20                  7
## Serviced apartment         Tiny house          Townhouse              Villa
##                 82                  2                105                  7
```

```
room_tab
```

```
##
```

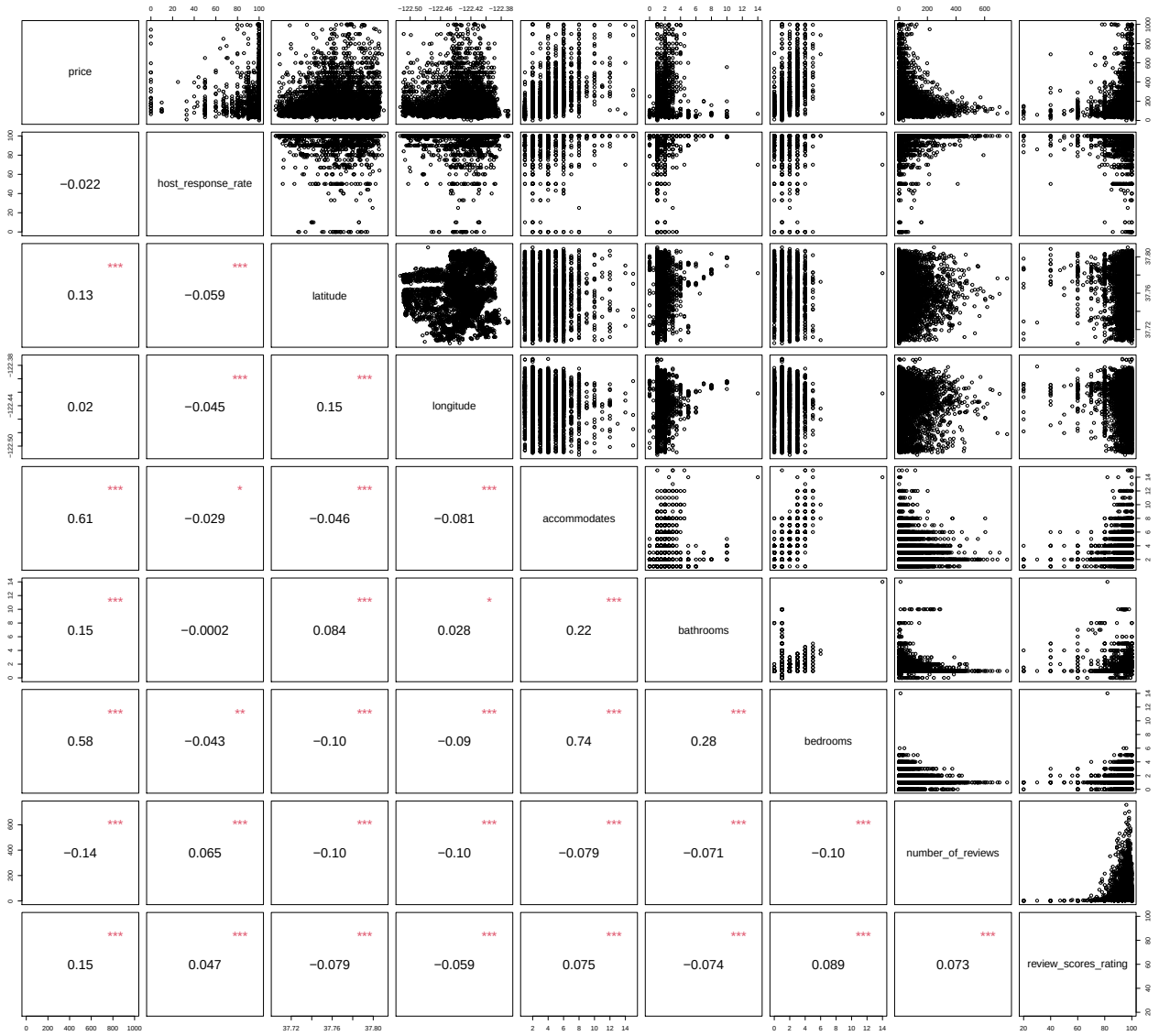
```
## Entire home/apt      Hotel room    Private room     Shared room
##            3144             143            2270             135
```

```r
# graphical display of categorical variables
par(mfrow=c(2,2))
barplot(superhost_tab, main="",
          xlab="Superhost", ylab="Frequency",
          col="#3C787E")
barplot(neighborhood_tab, main="",
          xlab="Neighborhood", ylab="Frequency",
          col="#3C787E")
barplot(property_tab, main="",
          xlab="Property Type", ylab="Frequency",
          col="#3C787E")
barplot(room_tab, main="",
          xlab="Room Type", ylab="Frequency",
          col="#3C787E")
```




```r
## Multivariate EDA

# define panel.cor function to use as personalized pairs plot
panel.cor <- function(x, y, digits=2, prefix="", cex.cor) {
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- cor(x, y)
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
```

```
  if(missing(cex.cor)) cex <- 2
  test <- cor.test(x, y)
  signif <- symnum(test$p.value, corr=FALSE, na=FALSE,
                   cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                   symbols = c("***", "**", "*", ".", " "))
  text(0.5, 0.5, txt, cex=cex)
  text(0.8, 0.8, signif, cex=cex, col=2)
}

# relationship between quantitative variables
pairs(listings_quant, lower.panel=panel.cor)
```




```
# relationship between categorical variables
par(mfrow=c(2,2))
catvars <- with(listings, cbind(host_is_superhost, neighbourhood, property_type, room_type)); catnum <-
catnames <- c('Superhost', 'Neighborhood', 'Property Type', 'Room Type')
for(i in 1:catnum) {
```

```
  boxplot(listings$price ~ catvars[, i], main="",
          ylab="Price of AirBnB Listing ($)",
          xlab=catnames[i])
}
```