STAT 346: Statistical Analysis of San Francisco AirBnB Listing Prices and Review Ratings: EDA

The Outliers 1

Due Monday, April 27

EDA

The dataset that we will use for our project is `listings.csv`, which includes 8,111 unique listings of AirBnBs in the San Francisco area. The dataset was obtained through the website https://www.kaggle.com/jeploretizo/san-francisco-airbnb-listings. As the second phase in conducting our project, we first ran univariate EDAs for each of the variables that we wanted to look at in the dataset (price, overall rating, latitude, longitude, and neighborhood) and bivariate EDAs between each of the variable pairs. Price is our response variable that we want to predict given different values of the other variables as predictors.

Here are our two selected EDA results: (1) univariate EDA of rating and (2) bivariate EDA of relationship between price and rating.

```
# Read in data set
data = read.csv("listings.csv")
subset <- data[,c("price", "review_scores_rating", "latitude", "longitude")]

# Make variables
price <- as.numeric(data$price)
rating <- data$review_scores_rating
latitude <- data$latitude
longitude <- data$longitude
neighborhood <- data$neighbourhood
```

Univariate EDA of Rating

```
ratingmean <- mean(rating, na.rm=TRUE)
ratingvar <- var(rating, na.rm=TRUE)
ratingmed <- median(rating, na.rm=TRUE)
ratingiqr <- IQR(rating, na.rm=TRUE)

summary.stat <- data.frame(ratingmean, ratingvar, ratingmed, ratingiqr)
rownames(summary.stat) = c("Rating")
colnames(summary.stat) = c("Mean", "Variance", "Median", "IQR")

library(xtable)
options(xtable.floating = FALSE)
options(xtable.timestamp = "")
print(xtable(summary.stat), comment=FALSE)
```

	Mean	Variance	Median	IQR
Rating	95.42	57.32	98.00	5.00

```
hist(rating, main="Distribution of Rating of Listed AirBnB",
     xlab="Rating of Listed AirBnB",
```

```
     ylab="Frequency",
     breaks=40, col="#256788")
```

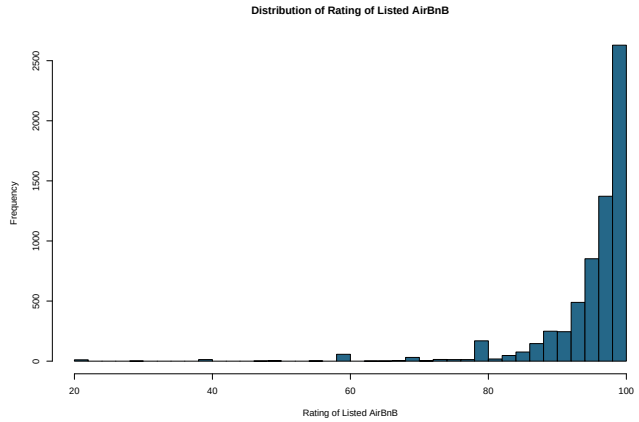

Figure 1: Univariate EDA of Rating in the AirBnB Listings Dataset with breaks = 40

As seen above in Figure 1, the distribution of ratings in the AirBnB listings dataset is unimodal with a single mode around a 100 rating units. The distribution is extremely left-skewed as shown in the histogram and this fact is also supported by the summary statistics table where the mean is much less than the median. There are outliers in the lower range of ratings around 20 rating units and a slight increase in frequency around 80 rating units.

Bivariate EDA of Relationship Between Price and Rating

```
plot(rating, price, pch=16, cex=1)
```

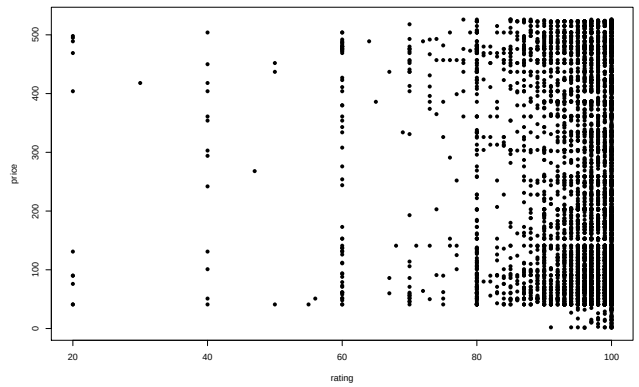

Figure 2: Bivarite EDA of Relationship Between Price and Rating in the AirBnB Listings Dataset

As seen above in Figure 2, the bivariate relationship between price (y-axis) and rating (x-axis) of AirBnB listings in San Francisco is shown in a scatterplot. The presence of a linear relationship is not clear from the scatterplot and there seems to be a weak relationship between the two variables. There is a high concentration of high values in rating for all values of price. There are noticeable points at lower ratings, once again both in the lower price range as well as the higher price range.