MIS-64060 Assignment 1

Steve Spence

9/4/2019

Assignment #1 Objectives:

1. Download a dataset from the web. You may use any source, but specify the source in your code. Also ensure that the data has a mix of quantitative and qualitative (categorical) variables.
2. Import the dataset into R
3. Print out descriptive statistics for a selection of quantitative and categorical variables.
4. Transform at least one variable. It doesn't matter what the transformation is.
5. Plot at least one quantitative variable, and one scatterplot
6. Upload your R program, and any associated datafiles to your git account. Remember to create a separate repository for this class.
7. Paste the address to your repository in the answer box here in BB.

Objective #1:

Downloaded CSV data set from Kaggle website. Source claims to have captured all AirBnb data related to New York city listings since 2008. Link to data set is shown below:

https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/downloads/new-york-city-airbnb-open-data.zip/3

Objective #2:

```
# Read CSV file in from directory

raw_data <- read.csv("~/Desktop/M.S. Business Analytics/3. Fall - 2019/MIS-64060 Fundamentals of Machine Learning/Assignment 1/Assignment 1 Data -- AB_NYC_2019.csv")
```

The result of this objective was a comma separate value (csv) dataset being imported into our R environment as a data frame named "raw_data" for further analysis.

Objective #3:

```
# Returns the structure of the data frame and all variables captured in the data set.
```

```r
str(raw_data)
```

```
## 'data.frame':    48895 obs. of  16 variables:
##  $ id                            : int  2539 2595 3647 3831 5022 5099 5121
5178 5203 5238 ...
##  $ name                          : Factor w/ 47906 levels "","  1 Bed Apt i
n Utopic Williamsburg ",..: 12573 38016 45018 15591 19219 24849 8257 24896 15
486 17573 ...
##  $ host_id                       : int  2787 2845 4632 4869 7192 7322 7356
8967 7490 7549 ...
##  $ host_name                     : Factor w/ 11453 levels "","  Valéria",..
: 4997 4791 2913 6210 5929 1938 3549 9649 6880 1235 ...
##  $ 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 62 138 14 96 203 36 ...
##  $ 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 1749 1124 1751 1048 1736 ...
##  $ reviews_per_month             : num  0.21 0.38 NA 4.64 0.1 0.59 0.4 3.4
7 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
# Return the descriptive statistics for all the variables within the dataset

summary(raw_data)
```

```
##        id                                             name
##  Min.   :    2539   Hillside Hotel                    :   18
##  1st Qu.: 9471945   Home away from home               :   17
##  Median :19677284                                     :   16
##  Mean   :19017143   New york Multi-unit building      :   16
##  3rd Qu.:29152178   Brooklyn Apartment                :   12
##  Max.   :36487245   Loft Suite @ The Box House Hotel:   11
##                     (Other)                           :48805
##     host_id                 host_name        neighbourhood_group
##  Min.   :     2438   Michael     :  417   Bronx        : 1091
##  1st Qu.:  7822033   David       :  403   Brooklyn     :20104
##  Median : 30793816   Sonder (NYC):  327   Manhattan    :21661
##  Mean   : 67620011   John        :  294   Queens       : 5666
```

```
##  3rd Qu.:107434423   Alex        :  279   Staten Island:  373
##  Max.   :274321313   Blueground  :  232
##                      (Other)     :46943
##             neighbourhood      latitude       longitude
##  Williamsburg      : 3920   Min.   :40.50   Min.   :-74.24
##  Bedford-Stuyvesant: 3714   1st Qu.:40.69   1st Qu.:-73.98
##  Harlem            : 2658   Median :40.72   Median :-73.96
##  Bushwick          : 2465   Mean   :40.73   Mean   :-73.95
##  Upper West Side   : 1971   3rd Qu.:40.76   3rd Qu.:-73.94
##  Hell's Kitchen    : 1958   Max.   :40.91   Max.   :-73.71
##  (Other)           :32209
##            room_type         price         minimum_nights
##  Entire home/apt:25409   Min.   :    0.0   Min.   :   1.00
##  Private room   :22326   1st Qu.:   69.0   1st Qu.:   1.00
##  Shared room    : 1160   Median :  106.0   Median :   3.00
##                          Mean   :  152.7   Mean   :   7.03
##                          3rd Qu.:  175.0   3rd Qu.:   5.00
##                          Max.   :10000.0   Max.   :1250.00
##
##  number_of_reviews     last_review    reviews_per_month
##  Min.   :  0.00              :10052   Min.   : 0.010
##  1st Qu.:  1.00    2019-06-23: 1413   1st Qu.: 0.190
##  Median :  5.00    2019-07-01: 1359   Median : 0.720
##  Mean   : 23.27    2019-06-30: 1341   Mean   : 1.373
##  3rd Qu.: 24.00    2019-06-24:  875   3rd Qu.: 2.020
##  Max.   :629.00    2019-07-07:  718   Max.   :58.500
##                    (Other)   :33137   NA's   :10052
##  calculated_host_listings_count availability_365
##  Min.   :  1.000                Min.   :  0.0
##  1st Qu.:  1.000                1st Qu.:  0.0
##  Median :  1.000                Median : 45.0
##  Mean   :  7.144                Mean   :112.8
##  3rd Qu.:  2.000                3rd Qu.:227.0
##  Max.   :327.000                Max.   :365.0
##
```

From the “str” command we get an idea of the structure of the data set. There are sixteen total variables in the data frame with the following data types:

Seven (7) integer variables: id, host_id, price, minimum_nights, number_of_reviews, calculcated_host_listings_count, and availability_365.

Six (6) factor variables: name, host_name, neighbourhood_group, neighbourhood, room_type, and last_review

Three (3) numeric variables: latitude, longitude, and reviews_per_month

From the “summary” command we get a list of descriptive statistics for the variables in the data set. In this case we can see the values typically displayed on a boxplot (quartiles,

median, min, and max) plus the mean value of numeric variables and count for catergorical variables.

Objective #4:

For the transformation objective, I decided to transform the "price" variable by normalizing it via the z-transformation, and storing those values in a new variable named "price_normalized".

```
# Transform the "price" variable by normalizing values via the z-transformati
on method and storing in a new variable named "price_normalized".

raw_data$price_normalized <- (raw_data$price - mean(raw_data$price)) / sd(raw
_data$price)

summary(raw_data$price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    69.0   106.0   152.7   175.0 10000.0

summary(raw_data$price_normalized)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.63593 -0.34861 -0.19454  0.00000  0.09277 41.00399
```

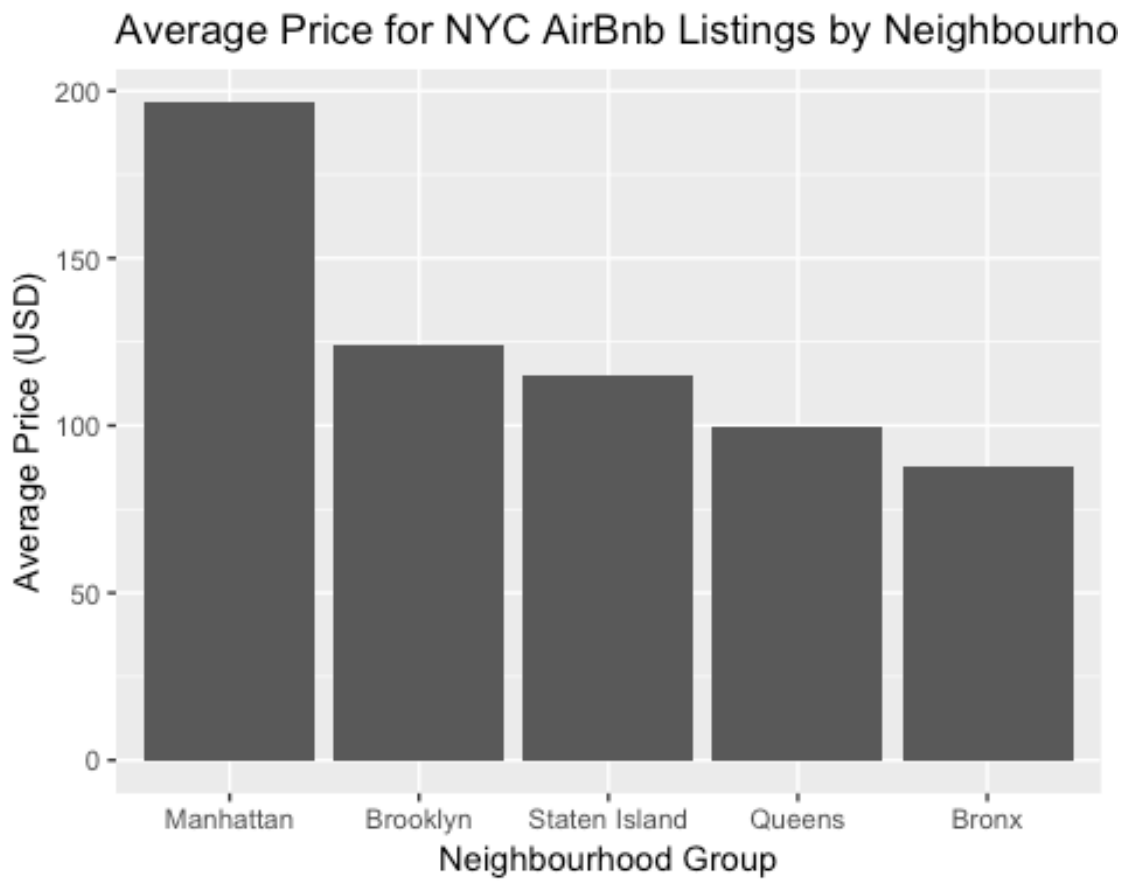
From the comparison of the two variables summary statistics, we can see that the range has now changed from 0 to 10,000 (original data) to -0.63593 to 41.00399 (z-transformation) based on the calculation in the previously shown R script.

Objective #5:

The first plot for a quantitative variable is a barplot via ggplot2. This barplot takes the average listing price for each of the five "neighbourhood_group" locations called out in the data set. Once it has the average value for each group, it will display the values in descending order by average price.

```
# Ensure ggplot2 package is installed prior to executing code.

require(ggplot2)

## Loading required package: ggplot2

# Create a bar plot of "neighbourhood_group" and "price" using the average pr
ice for each group and returning the x-axis in descending order by average pr
ice.

ggplot(raw_data, aes(x = reorder(neighbourhood_group, -price), price)) +
```

```
  geom_bar(stat = "summary", fun.y = "mean") +
  ggtitle("Average Price for NYC AirBnb Listings by Neighbourhood Group") +
  xlab("Neighbourhood Group") +
  ylab("Average Price (USD)")
```



As shown in the barplot, it is clear that Manhattan has a significantly higher average price per listing than the other four groups. This makes sense in real-world verification, since it is the more sought after area of New York City; however, the data will need further cleaning and scaling to give a more accurate representation.

The second plot for a scatter plot will plot the "number_of_reviews" variable versus the "price" to see if we can define a correlation between quantity of review and price per night of the listing.

```
# Ensure ggplot2 package is installed prior to executing code.

require(ggplot2)

# Create a scatter plot of "number_of_reviews" and "price" via ggplot2

ggplot(raw_data, aes(number_of_reviews, price)) +
  geom_point() +
  ggtitle("Number of Reviews for NYC AirBnb Listings Versus Price per Night")
```

```
+
  xlab("Number of Reviews") +
  ylab("Price (USD)")
```

From the scatter plot, we can see that there is a correlation between number of reviews and price per night for that listing. The larger the quantity of reviews for a listing, the cheaper the listing costs per night from the collected data. This makes sense in a reality check, because a cheaper listing is more likely to get a lot more visitors since it is more accessible to the population. Alternatively, it can be seen that expensive listings have much fewer reviews, because a smaller amount of people can afford to stay there. However, there are probably other factors that need additional review to see their impact as well (such as length of time of listing, average review, etc.). Therefore, additional cleaning/review of the data is needed to get a more complete picture.

Objective 6:

The R Markdown file will be uploaded to my Github account, as well as the original csv data set, for review.

Objective 7:

The address to my Github repository will be included in the Blackboard submission for Assignment 1.

https://github.com/sspenc12/MIS-64060