Capital One – Airbnb and Zillow Data Challenge

New York Property Analysis

Introduction:

This data product is built for a real estate firm to understand and gain conclusions to invest in zip codes that will generate the maximum profit on Two-bedroom short term rentals in New York.

For this analysis publicly available data sets have been used from Airbnb and Zillow

1. Cost data: Zillow provides us and estimate of value for two-bedroom properties

2. Revenue data: Airbnb is the medium through which the investor plans to lease out their invested property.

Assumptions:

1. Occupancy rate of 75%

2. Investor will pay upfront in cash (i.e. no mortgage/interest rate will need to be accounted for).

3. Time value of money discount rate is 0% (i.e. $1 today is worth the same 100 years from now)

4. All properties and all square feet within each locale is assumed to be homogeneous (i.e. a 1000 square foot property in a locale such as Bronx or Manhattan generates twice the revenue and costs twice as much as any other 500 square foot property within that same locale.)

Data Preparation:

Software: **R studio**

The following packages are used for this project

data.table #Extension of `data.frame`

kableExtra #build common complex tables and manipulate table styles.

GGally #GGally' extends 'ggplot2'

naniar #For visually exploring missing data structures

tidyverse #dplyr, ggplot2, readr, tidyr etc---collection of R packages designed for data science---

Rmisc #many functions useful for data analysis and utility operations, I used it for multiplot

plotly #Plotly's R graphing library makes interactive, publication-quality graphs---

```
# loading data
airbnb <- read.csv("C:\\Users\\Mallikarjuna\\Desktop\\Data Challenge\\listings.csv")
zillow <- read.csv("C:\\Users\\Mallikarjuna\\Desktop\\Data
Challenge\\Zip_Zhvi_2bedroom.csv")
```

Quality Check, Data Munging and Exploratory Data Analysis

1. Dimensions of raw data

```
> dim(airbnb)
[1] 48895      106

> dim(zillow)
[1] 8946    262
```

2. Removed rows that are not needed from Airbnb and Zillow data – we need only NY state two-bedroom data.

```
#replaced the words to have unique name for NY state
airbnb$state <- (gsub("New York","NY",airbnb$state))
airbnb$state <- (gsub("ny","NY",airbnb$state))
airbnbfiltered <- airbnb[which(airbnb$state=="NY" & airbnb$bedrooms == 2),]
zillowfiltered <- zillow[which(zillow$State =="NY"),]
```

3. Merged the Airbnb and Zillow data using zip codes

```
#region name is zip code in data dictionary so converting it to zip code
colnames(zillowfiltered)[2] <- "zipcode"
# convert zipcode to char in zillow to not lose data
zillowfiltered$zipcode <- as.character(zillowfiltered$zipcode)
airbnbfiltered$zipcode <- as.character(airbnbfiltered$zipcode)

# merge two datasets by zipcode
mergedata <- merge(airbnbfiltered, zillowfiltered , by = "zipcode" )
```

4. checking the summary of data and names of columns and retaining only important variables

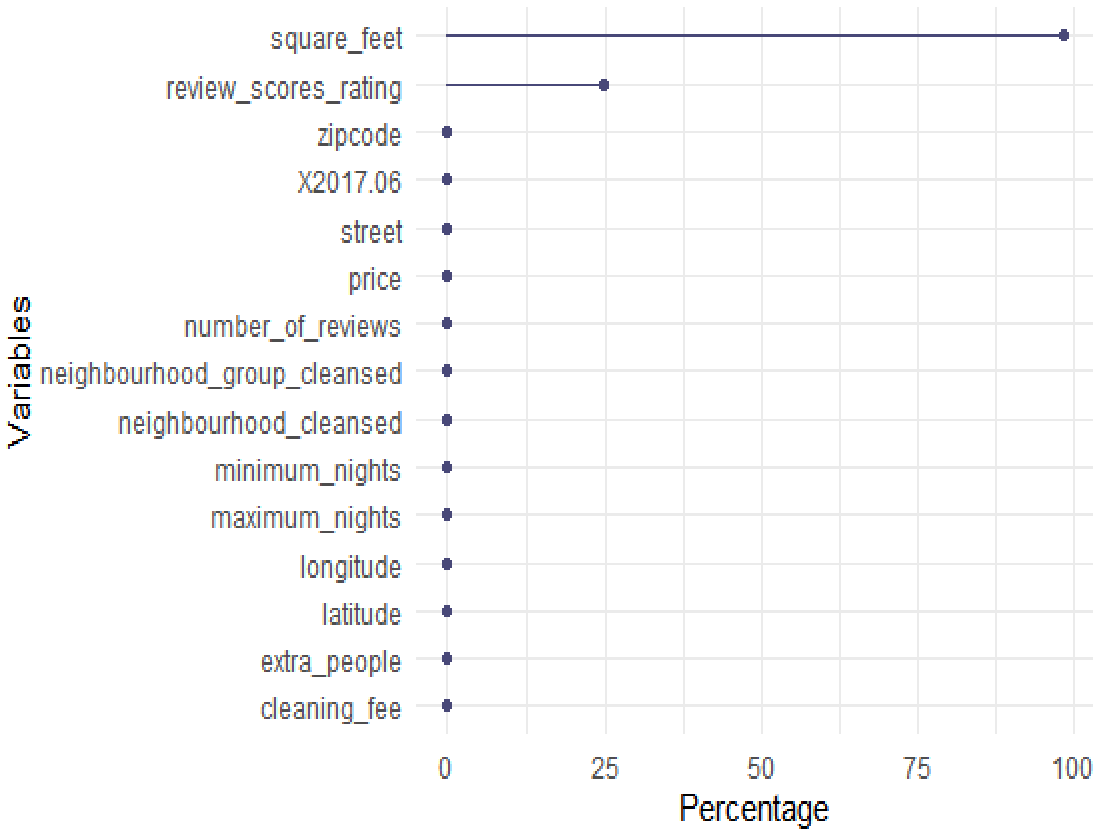
```
  mergedatafil <- mergedata
mergedatafil <- mergedatafil[,c(1,39,41,42,49,50,60,61,65,67,68,69,83,87,367)]
```

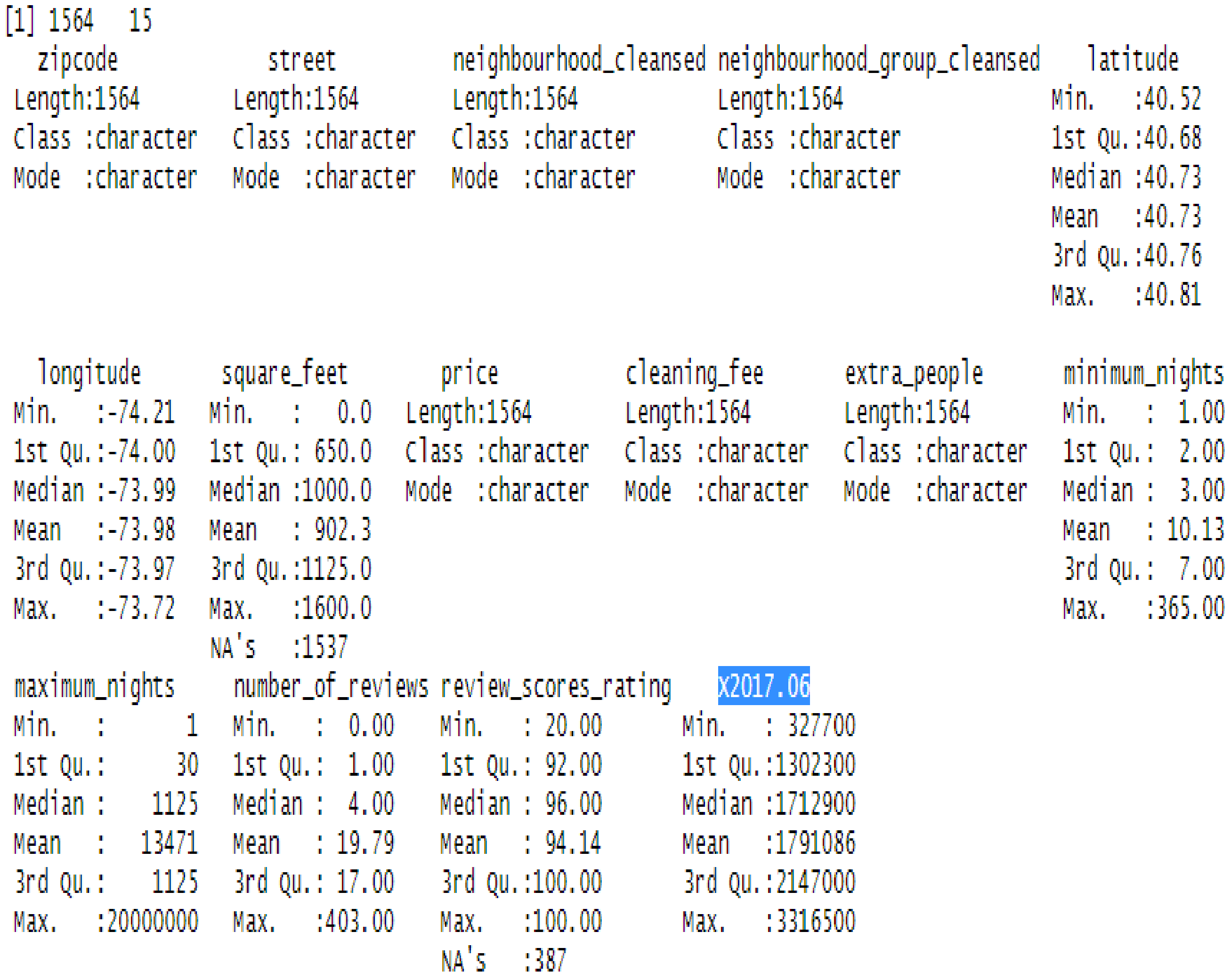
Data dictionary for the filtered merge data

Column number selected	Field	Description
1	Zipcode	Zip code where the property is located.
39	street	Street address where the property is located
41	neighbourhood_cleansed	Verified neighborhood name where the property is located.
42	neighbourhood_group_cleansed	Name of the area where the property is located.
49	latitude	The angular distance of a place north or south of the earth's equator, expressed in degrees and minutes.
50	longitude	The angular distance of a place east or west of the meridian at Greenwich, England, expressed in degrees and minutes.
60	square_feet	Square footage of the property or space for rent.
61	price	Price the host is charging to stay per night.
65	cleaning_fee	Price the host is charging to clean up after your stay.
67	extra_people	Additional charge per additional guests you bring.
68	minimum_nights	Minimum amount of nights the host is willing to rent out the property.

69	maximum_nights	Maximum amount of nights the host is willing to rent out the property.
83	number_of_reviews	Number of reviews received for the property for its entire existence within AirBnB.
87	review_scores_rating	Overall score given based on accuracy, cleanliness, check-in, communication, location, and value.
367	X2017.06	Indicates the historical median price within that area

Percentage of missing values in each parameter





```
[1] 1564   15
   zipcode             street          neighbourhood_cleansed neighbourhood_group_cleansed    latitude    
 Length:1564        Length:1564        Length:1564            Length:1564                  Min.   :40.52  
 Class :character   Class :character   Class :character       Class :character             1st Qu.:40.68  
 Mode  :character   Mode  :character   Mode  :character       Mode  :character             Median :40.73  
                                                                                           Mean   :40.73  
                                                                                           3rd Qu.:40.76  
                                                                                           Max.   :40.81  

   longitude       square_feet        price           cleaning_fee       extra_people       minimum_nights  
 Min.   :-74.21   Min.   :   0.0   Length:1564        Length:1564        Length:1564        Min.   :  1.00  
 1st Qu.:-74.00   1st Qu.: 650.0   Class :character   Class :character   Class :character   1st Qu.:  2.00  
 Median :-73.99   Median :1000.0   Mode  :character   Mode  :character   Mode  :character   Median :  3.00  
 Mean   :-73.98   Mean   : 902.3                                                            Mean   : 10.13  
 3rd Qu.:-73.97   3rd Qu.:1125.0                                                            3rd Qu.:  7.00  
 Max.   :-73.72   Max.   :1600.0                                                            Max.   :365.00  
                  NA's   :1537                                                                              
 maximum_nights     number_of_reviews review_scores_rating    X2017.06      
 Min.   :       1   Min.   :  0.00    Min.   : 20.00       Min.   : 327700  
 1st Qu.:      30   1st Qu.:  1.00    1st Qu.: 92.00       1st Qu.:1302300  
 Median :    1125   Median :  4.00    Median : 96.00       Median :1712900  
 Mean   :   13471   Mean   : 19.79    Mean   : 94.14       Mean   :1791086  
 3rd Qu.:    1125   3rd Qu.: 17.00    3rd Qu.:100.00       3rd Qu.:2147000  
 Max.   :20000000   Max.   :403.00    Max.   :100.00       Max.   :3316500  
                                      NA's   :387                           
```

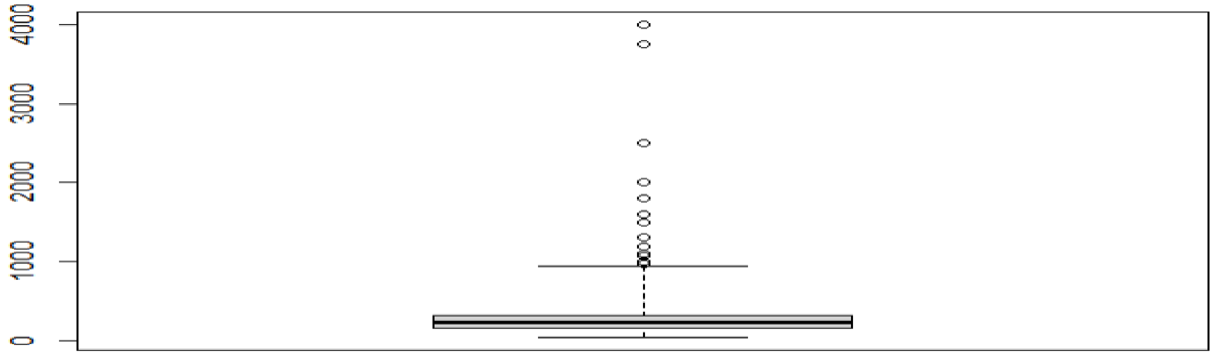
5. Cleaning Data and EDA: Price, cleaning fee and extra people were character data types and should be converted to numeric by removing the $ and other characters in them.

6. Square feet has more than 90 percent of missing values and removing will cause incorrect analysis. Sq feet value can’t be zero so zero values are converted to na values and then the median value which is around 1000 sq ft is considered to fill the data.

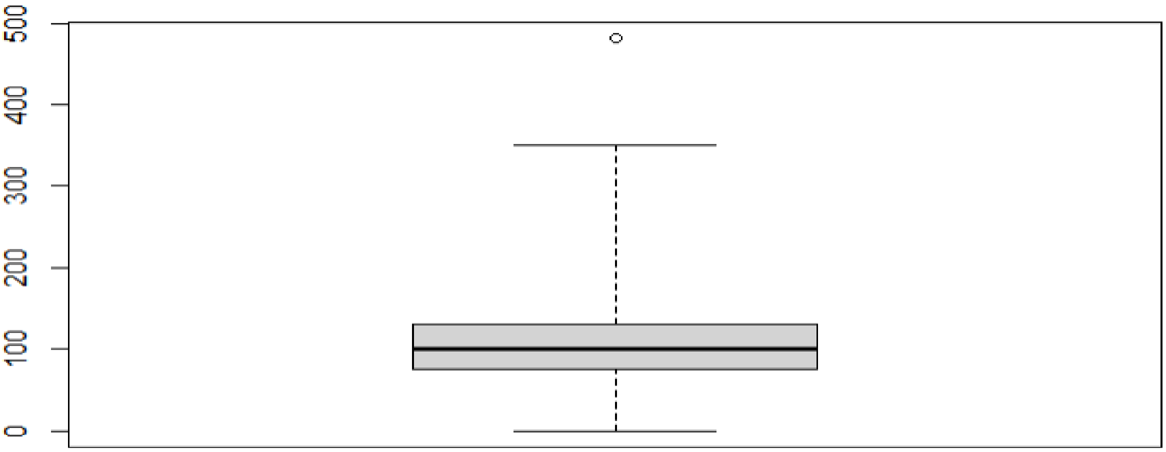
7. Cleaning fee and extra person fee are calculated perday and the cost of the property X2017.06 cannot be compared directly.

8. Displaying and removing the extreme values from price, cleaning fee and extra people by using boxplot.

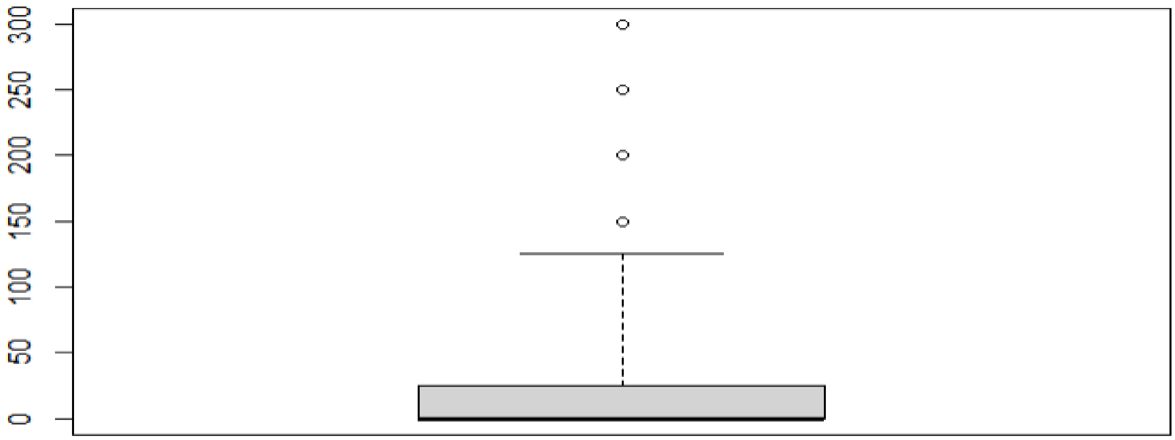
```
k = boxplot(mergedatafilclean$price, range = 4)
extremes <- which(mergedatafilclean$price %in% k$out)
mergedatafilclean <- mergedatafilclean[-c(extremes),]
```



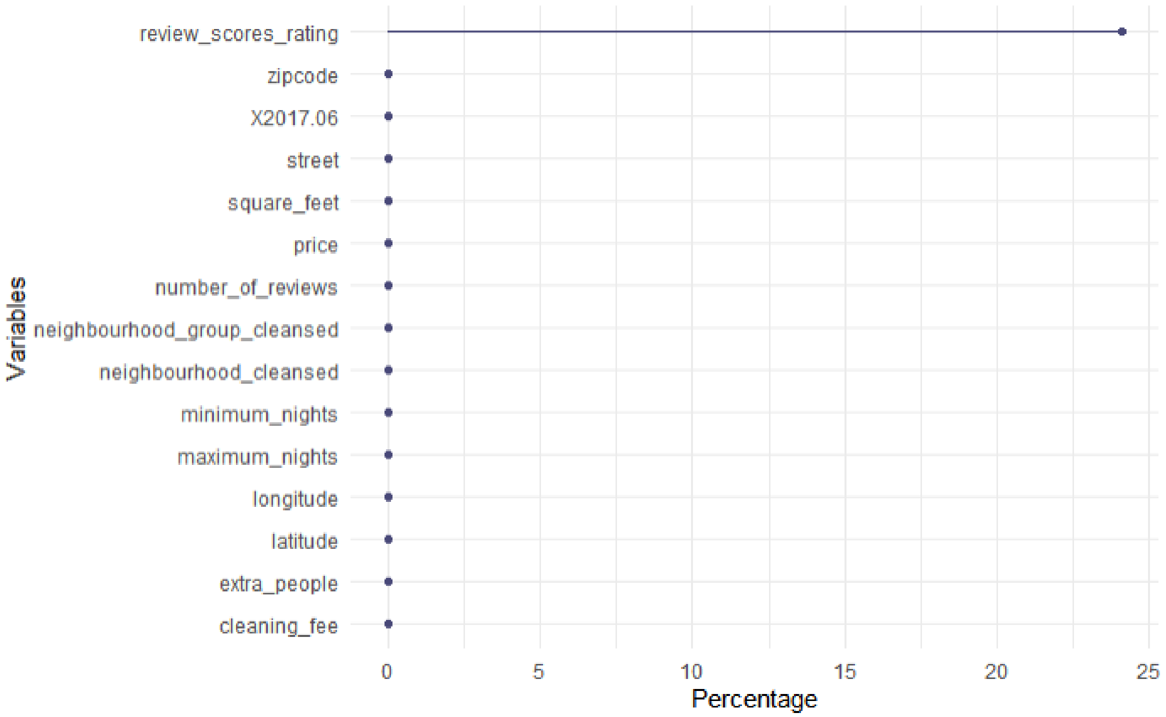
```
q = boxplot(mergedatafilclean$cleaning_fee, range = 4)

extremes1 <- which(mergedatafilclean$cleaning_fee %in% q$out)

mergedatafilclean <- mergedatafilclean[-c(extremes1),]
```



```
j = boxplot(mergedatafilclean$extra_people, range = 4)

extremes2 <- which(mergedatafilclean$extra_people %in% j$out)

mergedatafilclean <- mergedatafilclean[-c(extremes2),]
```



Percentage of missing values in each parameter



9. After updating the median of each square feet, cleaning data and extra people respectively into their missing values, the above graph shows the amount of missing variable in data set.

10. Missing Review scores ratings are left without updating as the median ratings scores are high and fluctuating. If they are updated with median value, they will tell a different story if they are wrongly updated.

Key Assumptions:

1. The price of the property is taken based on the recent available data which is

2017-06

2. Occupancy rate is assumed as 75 percent so when calculating total annual income of property.

3. we assumed that 1 in 5 times there might be an extra guest.

4. we might have to use 50 percent of cleaning fee charged when room is occupied which is almost 40 percent.

So We calculated annual income of property as price * (0.75 * 365) + cleaning_fee * (0.40 * 365) + extra_people * (0.2 * 365)

We have added 4 new parameters to our final data set for our analysis

1. total annual income

2.price per square feet

3.Years to start profiting

4.revenue in ten years

Summary of final data set after adding above parameters

```
neighbourhood_group_cleansed     latitude          longitude
  Length:1526                  Min.    :40.52    Min.    :-74.21
  Class :character             1st Qu.:40.68    1st Qu.:-74.00
  Mode  :character             Median :40.73    Median :-73.99
                               Mean    :40.73    Mean    :-73.98
                               3rd Qu.:40.76    3rd Qu.:-73.97
                               Max.    :40.81    Max.    :-73.72
    square_feet          price          cleaning_fee     extra_people
  Min.   :   3.0   Min.   : 50.0   Min.   :   0.0   Min.   :   0.00
  1st Qu.:1000.0   1st Qu.:162.0   1st Qu.: 75.0   1st Qu.:   0.00
  Median :1000.0   Median :225.0   Median :100.0   Median :   0.00
  Mean   : 998.6   Mean   :257.5   Mean   :107.4   Mean   : 14.93
  3rd Qu.:1000.0   3rd Qu.:300.0   3rd Qu.:130.0   3rd Qu.: 25.00
  Max.   :1600.0   Max.   :950.0   Max.   :350.0   Max.   :125.00

  minimum_nights   maximum_nights      number_of_reviews
  Min.   :  1.0   Min.   :         1   Min.   :  0.00
  1st Qu.:  2.0   1st Qu.:        30   1st Qu.:  1.00
  Median :  3.0   Median :      1125   Median :  4.00
  Mean   : 10.2   Mean   :     13786   Mean   : 20.04
  3rd Qu.:  7.0   3rd Qu.:      1125   3rd Qu.: 17.00
  Max.   :365.0   Max.   :20000000   Max.   :403.00

  review_scores_rating     X2017.06          totalannualincome
  Min.   : 20.00         Min.   : 327700   Min.   : 15148
  1st Qu.: 92.00         1st Qu.:1302300   1st Qu.: 58400
  Median : 96.00         Median :1712900   Median : 77928
  Mean   : 94.08         Mean   :1780431   Mean   : 87254
  3rd Qu.:100.00         3rd Qu.:2147000   3rd Qu.:104641
  Max.   :100.00         Max.   :3316500   Max.   :281963
  NA's   :368
  pricepersqrft        Yearstostartprofiting revenue_in_ten_years
  Min.   :  327.7   Min.   :  4.604         Min.   :-2978875
```

```
1st Qu.:    1302.3     1st Qu.: 15.450             1st Qu.:-1284356
Median :    1712.9     Median : 21.169             Median : -823225
Mean   :    2022.6     Mean   : 23.410             Mean   : -907890
3rd Qu.:    2147.0     3rd Qu.: 27.810             3rd Qu.: -502313
Max.   :356933.3       Max.   :122.444             Max.   : 1315625
```

Metadata of final data set

index	Field	Description
1	Zipcode	Zip code where the property is located.
2	street	Street address where the property is located
3	neighbourhood_cleansed	Verified neighborhood name where the property is located.
4	neighbourhood_group_cleansed	Name of the area where the property is located.
5	latitude	The angular distance of a place north or south of the earth's equator, expressed in degrees and minutes.
6	longitude	The angular distance of a place east or west of the meridian at Greenwich, England, expressed in degrees and minutes.
7	square_feet	Square footage of the property or space for rent.
8	price	Price the host is charging to stay per night.

9	cleaning_fee	Price the host is charging to clean up after your stay.
10	extra_people	Additional charge per additional guests you bring.
11	minimum_nights	Minimum amount of nights the host is willing to rent out the property.
12	maximum_nights	Maximum amount of nights the host is willing to rent out the property.
13	number_of_reviews	Number of reviews received for the property for its entire existence within AirBnB.
14	review_scores_rating	Overall score given based on accuracy, cleanliness, check-in, communication, location, and value.
15	X2017.06	Indicates the historical median price within that area
16	totalannualincome	Total annual income of the property
17	pricepersqrft	Price per each square feet
18	Yearstostartprofiting	Years the property takes to break even amount it is purchased
19	revenue_in_ten_years	Revenue generated in 10 years.