AirBnB & Zillow Data Challenge

Avinash Vashishtha

March 22, 2019

1.Introduction

We are helping a real estate company that has a niche in purchasing properties to rent out short-term as part of their business model specifically within New York City. They want us to build a data product and provide conclusions to help them understand which zip codes would generate the most profit on short term rentals within New York City.

1.1.Data Available

- **Cost data**: Zillow provides us an estimate of value for two-bedroom properties
- **Revenue data**: AirBnB is the medium through which the investor plans to lease out their investment property

1.2.Assumptions

- The investor will pay for the property in cash (i.e. no mortgage/interest rate will need to be accounted for).
- The time value of money discount rate is 0% (i.e. $1 today is worth the same 100 years from now).
- All properties and all square feet within each locale can be 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.)

2.Data Quality checks

2.1 Steps performed in Data Quality checks

- **1.Data Loading**- Revenue and Cost file was loaded with the same name
- **2.Missing value check**- Data is checked for missing values and result is displayed
- **3.Relevant Column Names**- Column names of both Datasets were checked and relevant columns are filtered from the dataset.List of relevant columns includes-neighbourhood_group_cleansed,zipcode,bedrooms,square_fe 06(Latest cost of 2 bedroom apartment property)
- **4.Missing values in columns**-Check count of missing values in relevant columns. Zipcode has 1.5% and bedrooms has 0.16% missing values.
- **5.Data Cleaning**- Making zip to 5 digits. Dropping the dollar sign and changing the format to numeric format
- **6.Merging the file**-Merging with the revenue file to get property cost of 2 bedroom apartments on the zip
- **7.Filtering for NY Zips**-Filtering for NY zip codes that will be considered for the anaysis

Code for Loading Tables

```r
datasets<-c("listings.csv.gz","Zip_Zhvi_2bedroom.csv")
for (i in seq_along(datasets))
{

  file_path=paste0(datasets[i])
  if(file.exists(file_path))
  {
    df<-read_csv(file_path)
    assign(datasets[i],df)
  } else
    print("No such file exits")
}
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   id = col_double(),
##   scrape_id = col_double(),
##   last_scraped = col_date(format = ""),
##   host_id = col_double(),
##   host_since = col_date(format = ""),
##   host_is_superhost = col_logical(),
##   host_listings_count = col_double(),
##   host_total_listings_count = col_double(),
##   host_has_profile_pic = col_logical(),
##   host_identity_verified = col_logical(),
##   latitude = col_double(),
##   longitude = col_double(),
##   is_location_exact = col_logical(),
##   accommodates = col_double(),
##   bathrooms = col_double(),
##   bedrooms = col_double(),
##   beds = col_double(),
##   square_feet = col_double(),
##   guests_included = col_double(),
##   minimum_nights = col_double()
##   # ... with 24 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   RegionName = col_character(),
##   City = col_character(),
##   State = col_character(),
##   Metro = col_character(),
##   CountyName = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

Missing Values -

```
for (i in seq_along(datasets))
{
missing_values<-any(is.na(get(datasets[i])))
if (missing_values)
{

  print(paste(datasets[i],"has Missing values"))
} else
  print(paste(datasets[i],"has no Missing values"))
}
```

```
## [1] "listings.csv.gz has Missing values"
## [1] "Zip_Zhvi_2bedroom.csv has Missing values"
```

Column Names

Percentage of missing values Zip,bedrooms,availability_30,number_of_reviews,price

```
## [1] 0.01499276
```

```
## [1] 0.001693127
```

```
## [1] 0
```

```
## [1] 0
```

```
## [1] 0
```

```
## [1] 0
```

Cleaning Data

- Changing zip to 5 digit format
- Extracting the integer from character format for price column
- you can check number of rows in listings(raw file) and after filtering for NY city.Reduced from ~41k to ~9k when filtered for NY city

```
## [1] 40753    97
```

```
## [1] 8940  102
```

3.Data Analysis

Inputs

Following inputs were used which can be changed to see results for different values-

cutoff_review-Only properties with reviews greater than cutoff_review were considered in the analysis

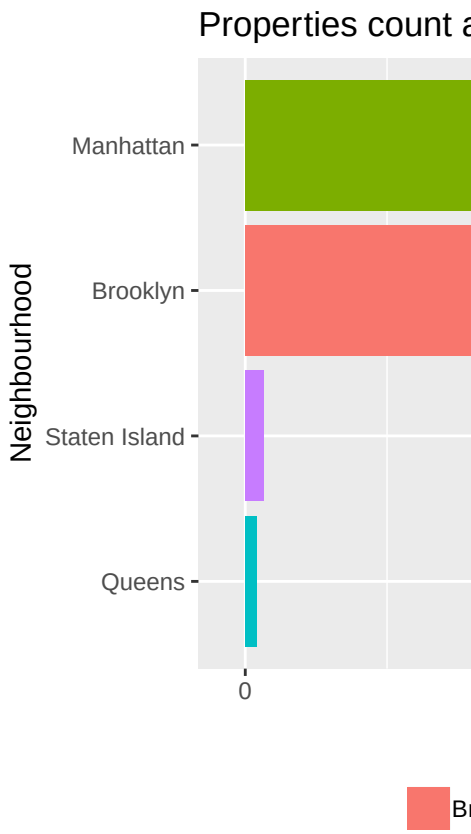
price_min_cutoff/price_max_cutoff-These values were used in deciding which properties should be included based on price

```
cutoff_review<-4
price_min_cutoff<-0.05
price_max_cutoff<-0.95
```

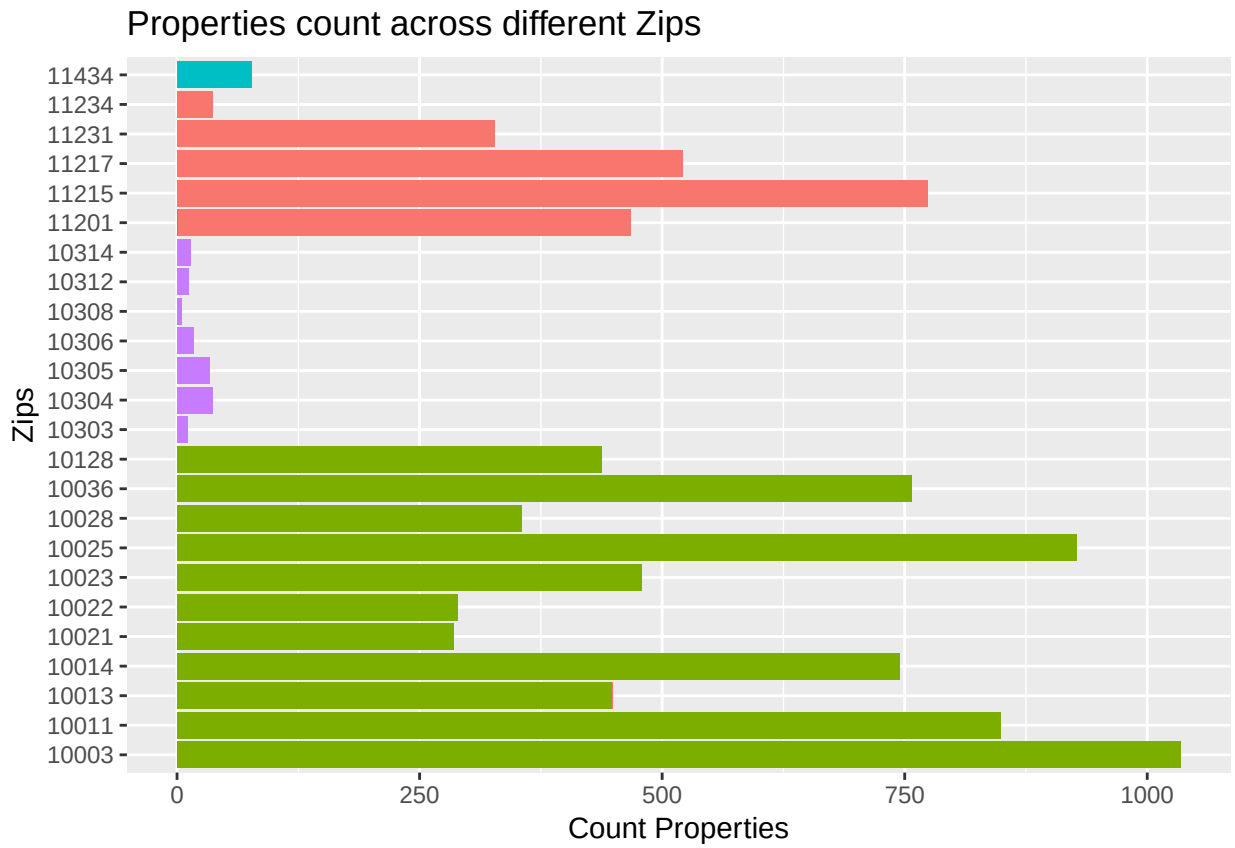
3.1 Methodology

- ***Step 1-EDA (3.2)***- Checking count across Zips and Neighbourhoods
- ***Step 2-Cost Factor (3.3)***- Based on average square feet area across bedroom count, Cost factor was calculated. This was used later to calculate property cost
- ***Step 3-Calculating Occupancy Rate (3.4)*** -Calculating Occupancy rate across review_score_location and same was used later to calculate occupancy for each property. Occupancy rate was calculated using available_30(Indicates the number of days the property is available for rent within 30 days)
- ***Step 4-Outlier Removal on Pricing (3.5)*** -Price/day for properties across neighbourhood were checked to identify outliers and only properties which lie in the middle 90%(5%-95%) were considered in the analysis. This was done to ensure that outliers(Properties with very high price/day) don't influence our results
- ***Step 5-Creation Of Metrics (3.6)***-Annual Revenue, Property cost and Breakeven Years were calculated for properties
- ***Step 6-Analyzing results across zips and neighbourhood (3.7)*** - Results were analysed after rolling up numbers at Zip and neighbourhood level.Top performing zips were pulled based on breakeven period

3.2 Exploratory data analysis



Initially, we just checked the count of properties across buroughs and zips



3.3 Calculating cost factor

Analysis of Bedroom count

- Majority of the properties are 1 bedroom apartments
- Bedrooms count was checked as this would be an important factor in deciding the cost of the apartment
- We only have cost of 2 bedroom apartment in a particular zip
- This factor will be used in deciding the property cost based on the bedroom count
- 0 bedrooms based on average_sq_feet value were assumed to be studio apartments similar to 1 bedroom apartment

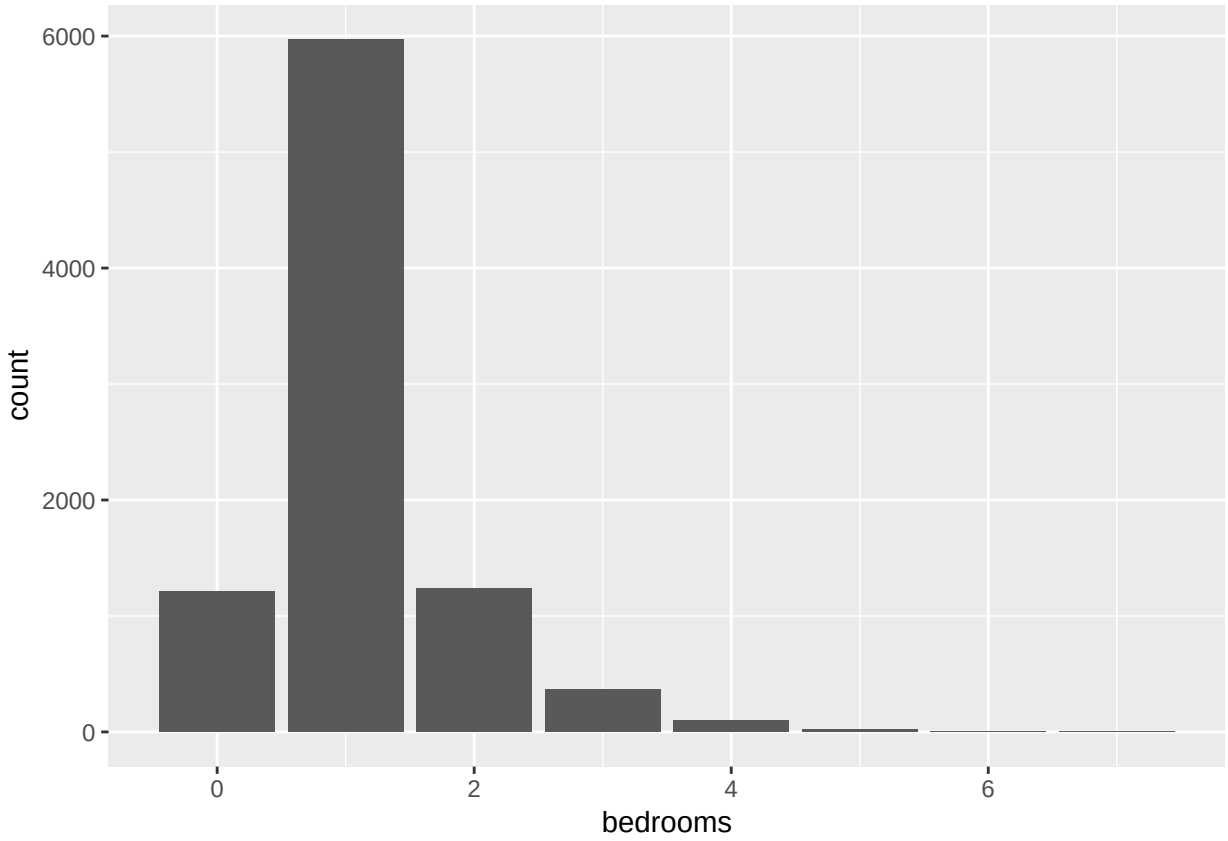


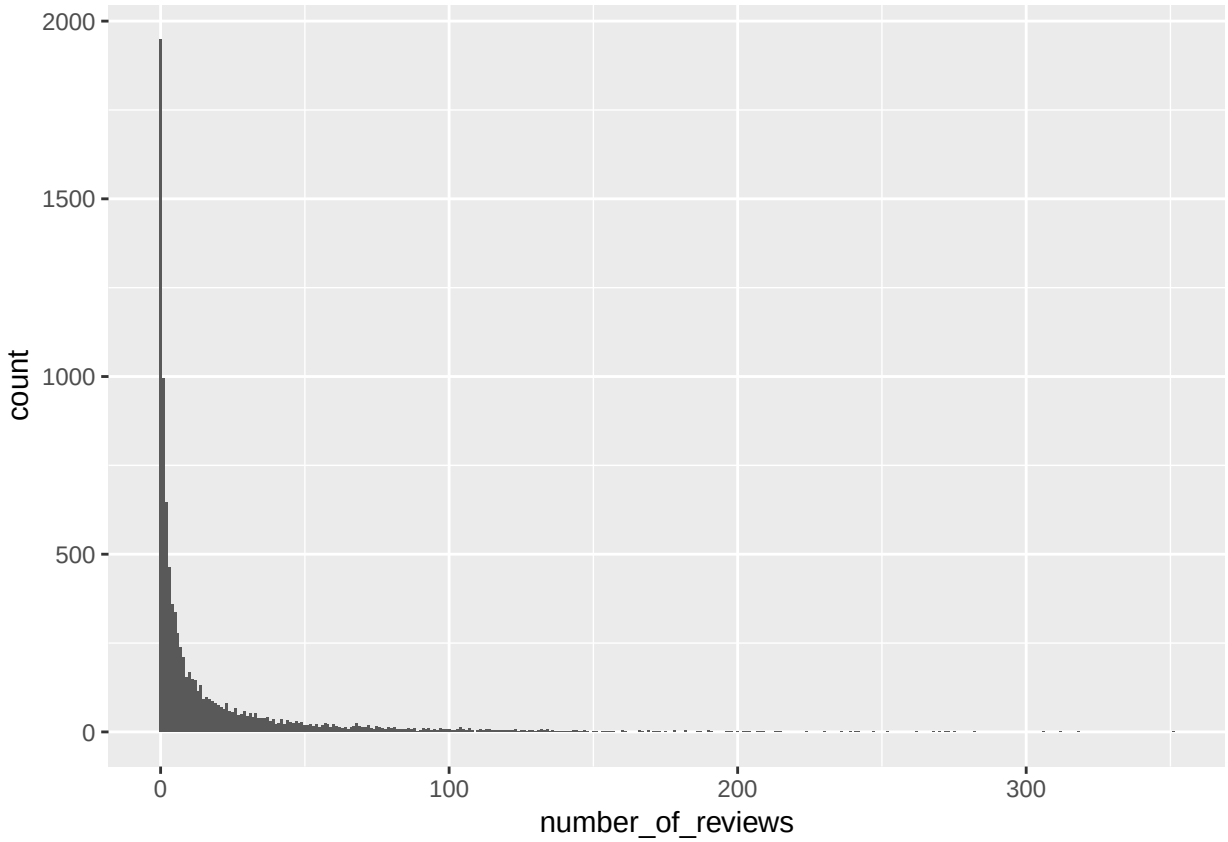
Table 1 : Table showing cost factor that will be used in the analysis

```
## # A tibble: 9 x 3
##   bedrooms mean_sq_feet mean_factor_2bedroom
##      <dbl>        <dbl>                <dbl>
## 1        0         408.                0.442
## 2        1         585.                0.634
## 3        2         924.                1.00
## 4        3        1333.                1.44
## 5        4        3350                 3.63
## 6        5        4000                 4.33
## 7        6        2300                 2.49
## 8        7        3700                 4.00
## 9       NA         400                 0.433
```

3.4 Calculating Occupancy Rate

Analysis of Reviews

- Next, we looked at reviews count as properties with fewer reviews suggests that either those properties are new or there is not enough data to make an accurate judgement based on review score
- We will use review score to decide on occupancy rate. Properties with lower location review score are likely to have lower occupancy rate
- We are looking at ~50% of the property **if we take a default cutoff value(This can be changed above) of 5 reviews in our anaysis**. So, only properties where we have 5 or more reviews will be considered for the analysis

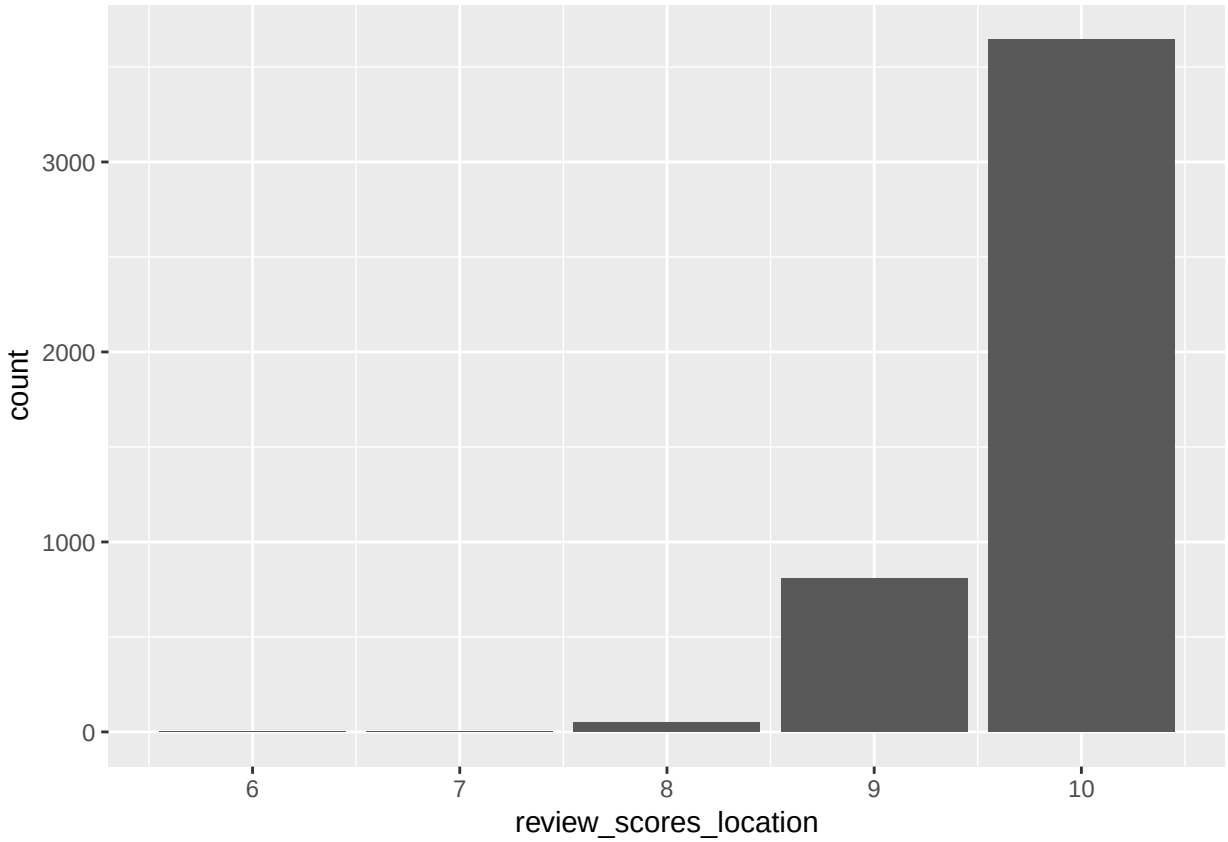


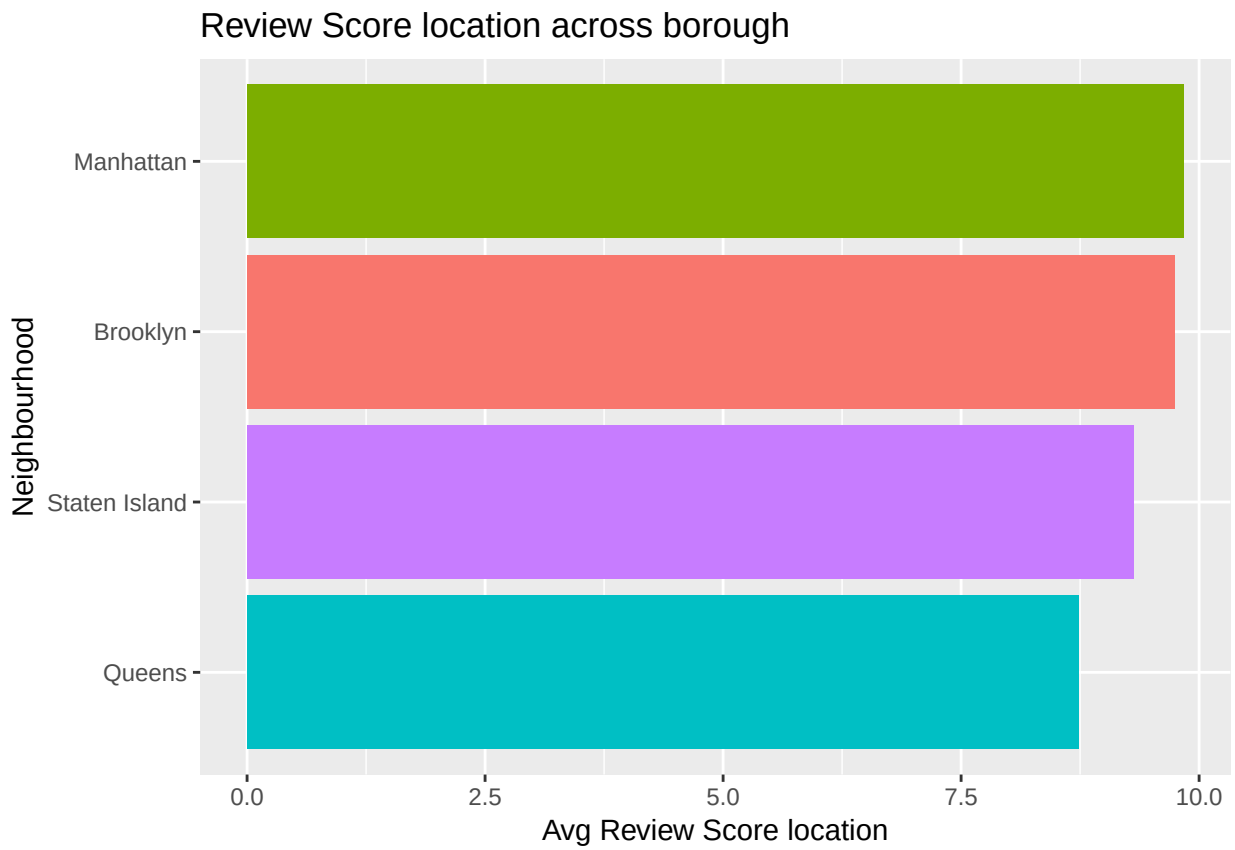
We are only considering 50% of the properties with 5 or more than 5 review. This value is dynamic and can be changed in the anaysis (Input at the top Input 1)

```
## [1] 0.5063758
```

Using review score location to define Occupancy rate

- After filtering for properties with 5 or more than 5 reviews, we checked their review score rating and review scores location to understand the distribution
- Since, we are more interested in location, we calculated occupancy rate based on Location review score and same will be used to caluculate revenue
- We can see that occupancy of the property goes down as review score location rating decreases
- Manhattan has the highest location review rating followed by brooklyn, staten Island and Queens


3000
2000
1000
0
count
6
7
8
9
10
review_scores_location



This graph shows that higher review scores lcation result in higher occupancy rate

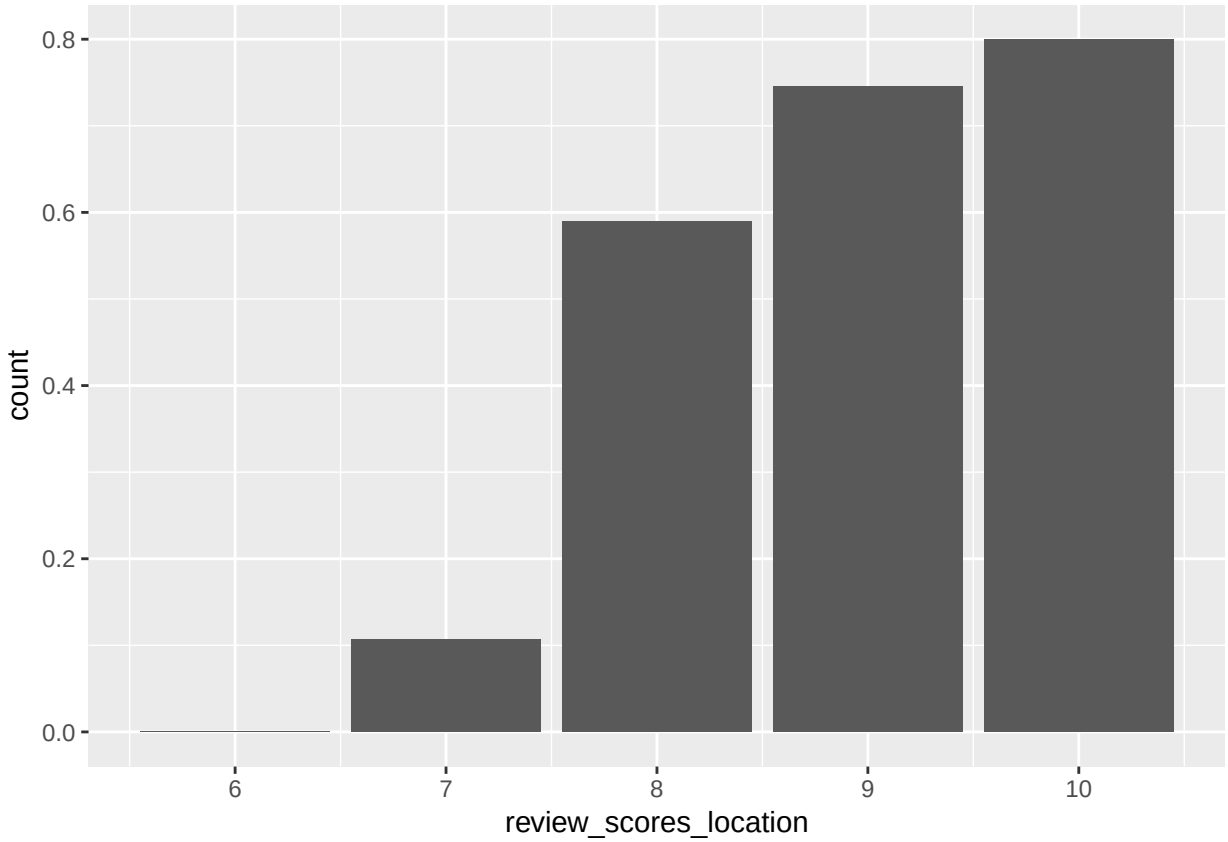


Table 2 : This table shows the occupancy rate across review_scores_location values

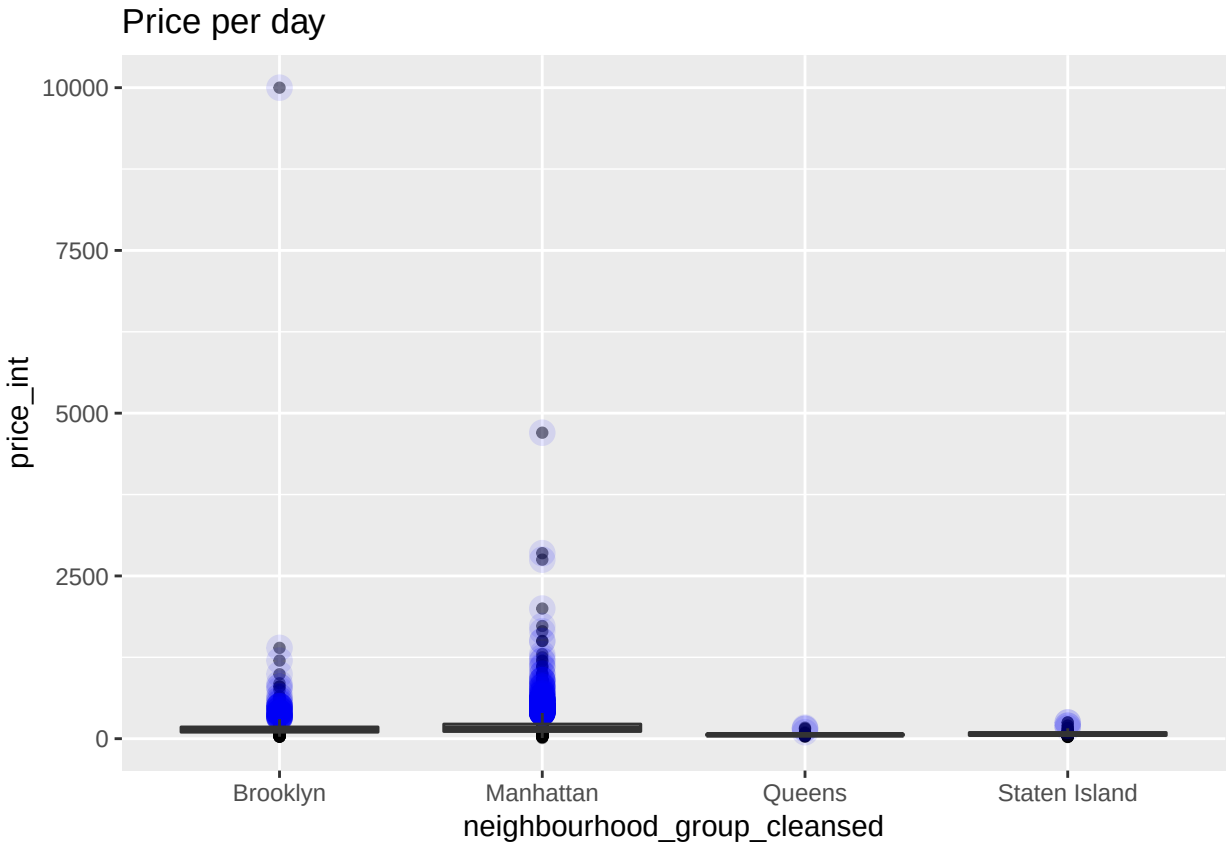
```
## # A tibble: 6 x 3
##   review_scores_location count_reviews mean_occupancy
##                    <dbl>         <int>          <dbl>
## 1                      6             1          0
## 2                      7             5          0.107
## 3                      8            50          0.589
## 4                      9           809          0.745
## 5                     10          3646          0.800
## 6                     NA            16          0.744
```

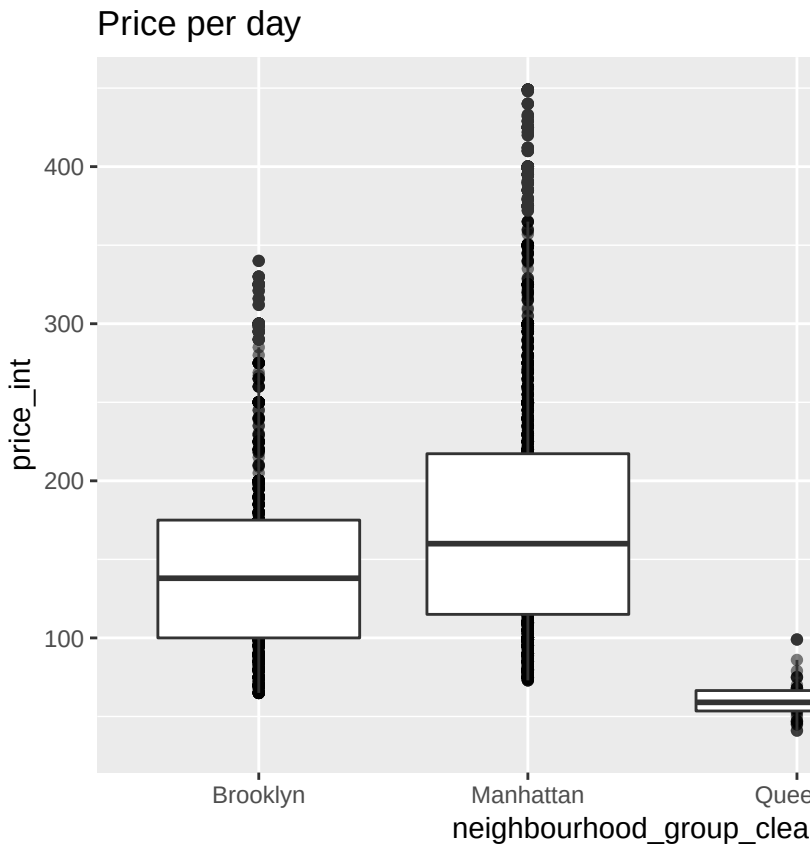
This table will be used to calculate occupancy rate for properties based on their location review score

3.5 Outlier Removal on Pricing

Analysing Price to look for outliers

- We checked price across neighbourhood groups and identified several outliers as shown in the graph below
- We used default values (0.05, 0.95) to remove outliers from both sides to get more robust results ***This values can be changed at the top-Input values***
- After removing outliers boxplot was created again
- Analysis was performed on these values only


Price per day
10000
7500
5000
2500
0
price_int
Brooklyn
Manhattan
Queens
Staten Island
neighbourhood_group_cleansed



After outlier removal the box plot looks like this

3.6 Creation Of Metrics

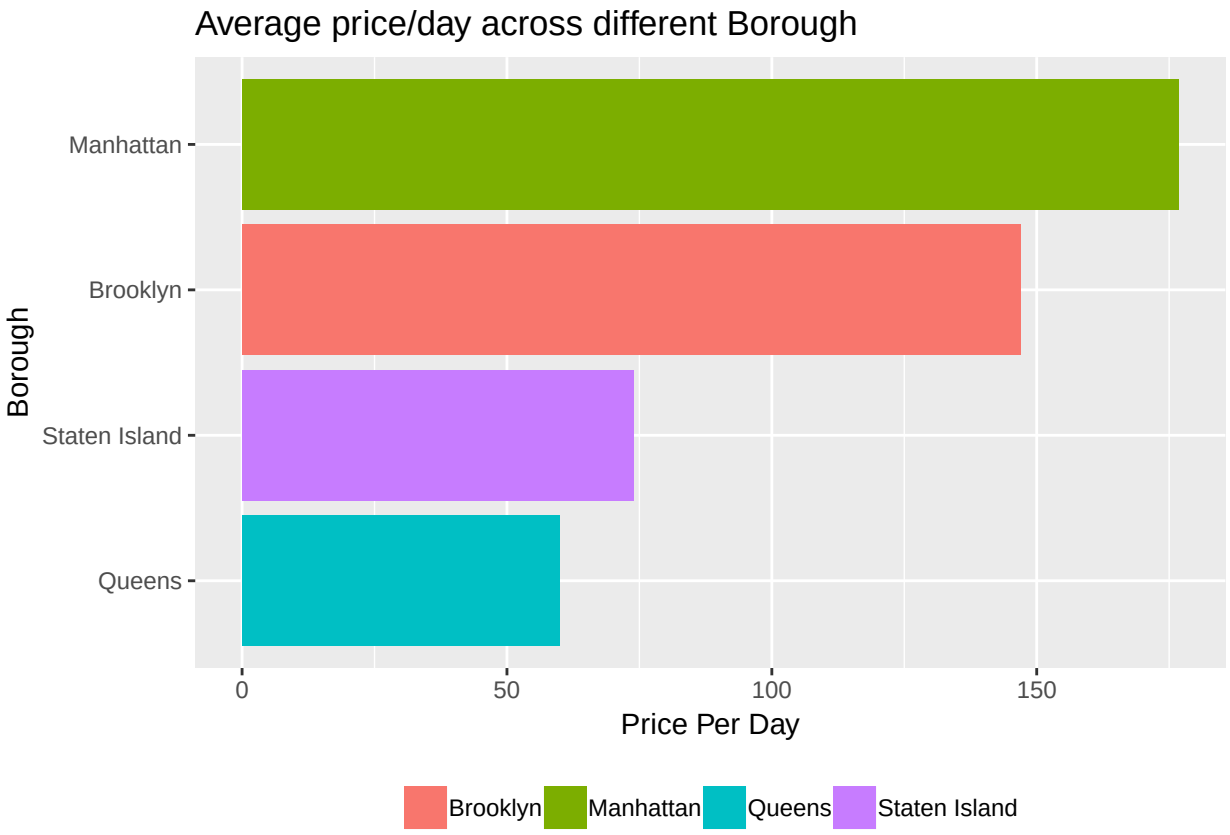
Defining Revenue, property cost and years to breakeven using the dataset created above

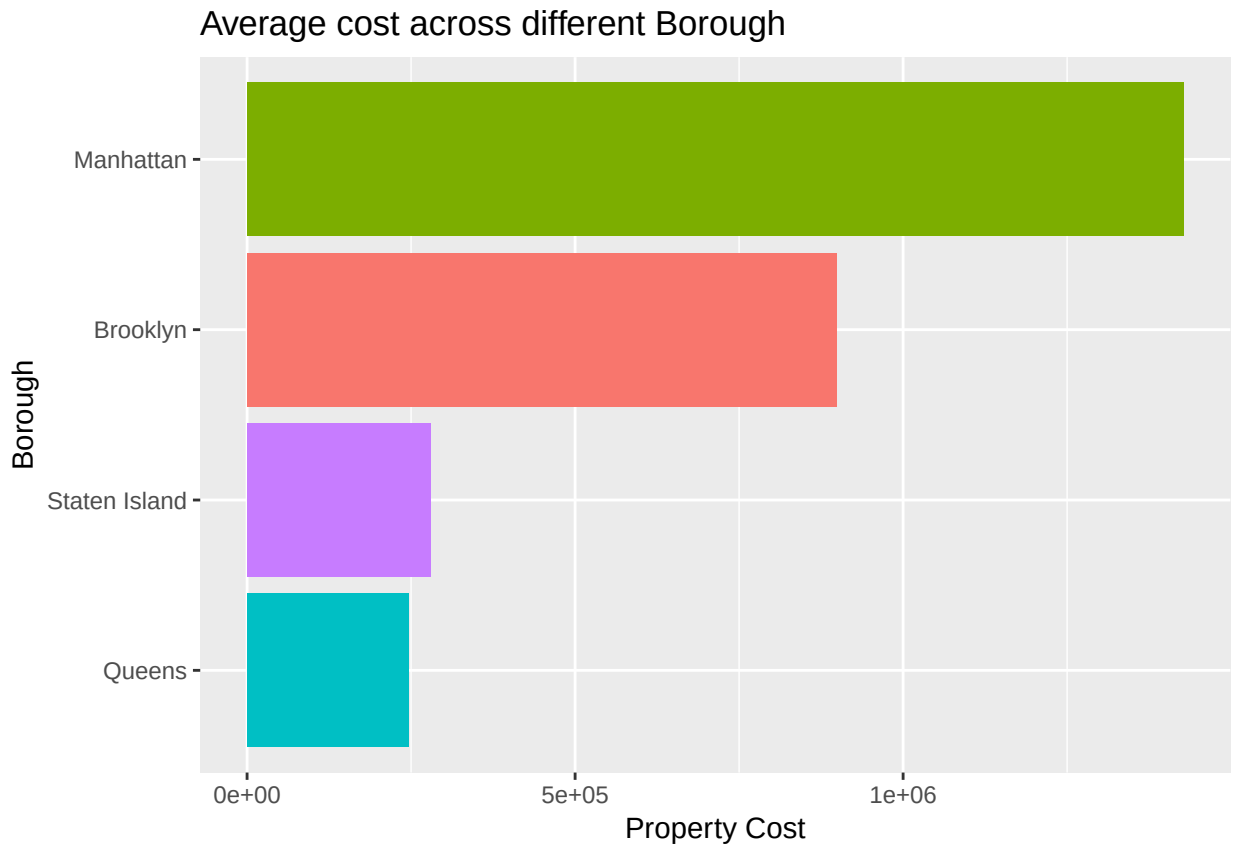
- **Annual Revenue** is defined as the **price per day X 365 X Occupancy** (Occupancy was based on the review score location as discussed earlier)
- **Property cost** is based on **Latest cost of 2 bedroom apartment X Factor calculated earlier based on bedroom count**
- **years to breakeven = Property cost/Annual Revenue**

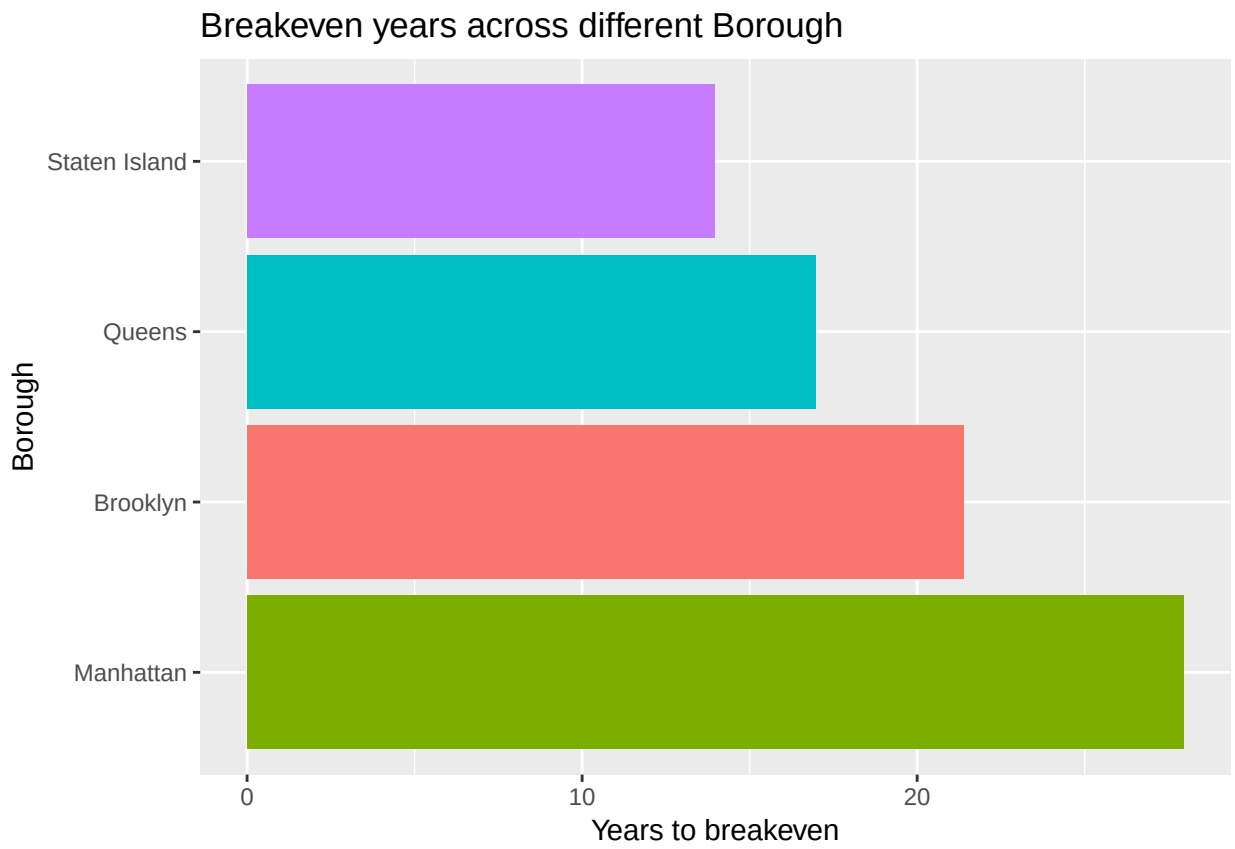
3.7 Analyzing results across zips and neighbourhood

Bar graph showing avg price and property cost across Boroughs(Neighbourhood Group) Insights from the charts creted at neighbourhood level

- **Price per day** - Average price per day is highest for manhattan(~$190), followed by brooklyn(~$170), Staten Island(~$80) and Queens(~$60)
- **Property cost per day** - Property cost also follows the same order-Manhattan, brooklyn,Staten Island and Queens
- **Years to breakeven** - Breakeven Years is minimum for staten Island followed by Queens, Brooklyn and manhatten. ***Deep dive at zip level is needed to understand which zips are most profittable***

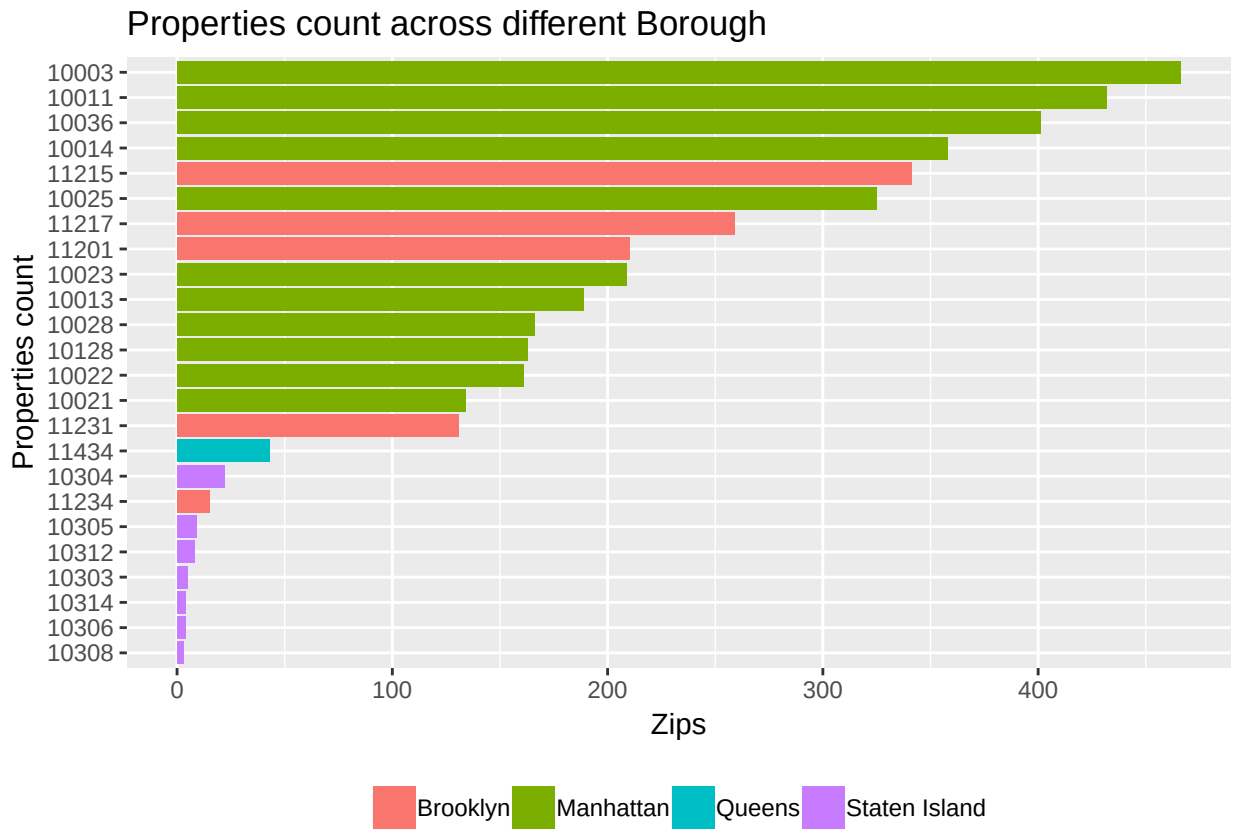

Average price/day across different Borough
Manhattan
Brooklyn
Staten Island
Queens
Borough
0
50
100
150
Price Per Day
Brooklyn
Manhattan
Queens
Staten Island


Average cost across different Borough
Manhattan
Brooklyn
Staten Island
Queens
Borough
0e+00
5e+05
1e+06
Property Cost



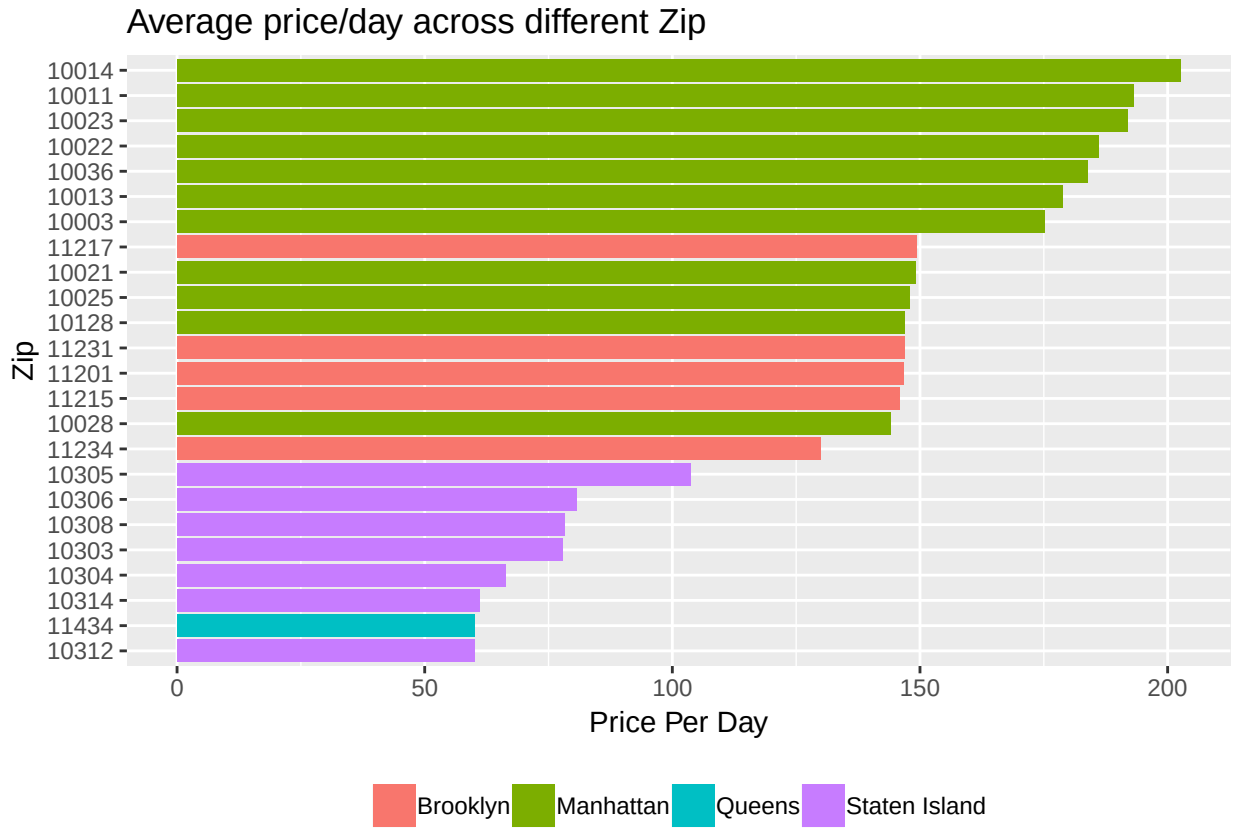
Count of properties across Zips

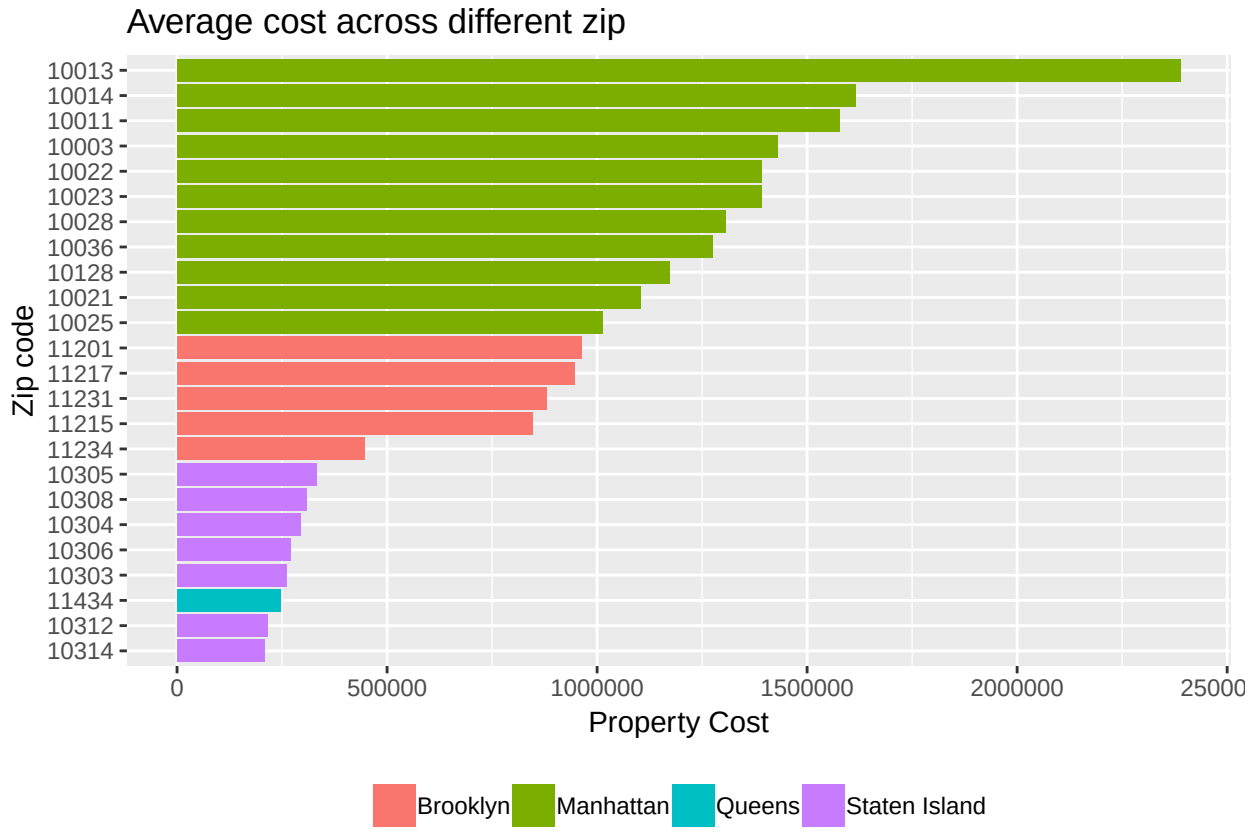
In case same zip was associated with multiple neighbourhood, zip was mapped to neighbourhood with maximum mappings



Insights from the charts creted at neighbourhood level

- **Price per day** - Average price per day is highest for manhattan zips(10014,10011,10023), followed by brooklyn zips(11217)
- **Property cost per day** - Property cost also follows the same order-Manhattan, brooklyn,Staten Island and Queens


Average price/day across different Zip
Zip
10014
10011
10023
10022
10036
10013
10003
11217
10021
10025
10128
11231
11201
11215
10028
11234
10305
10306
10308
10303
10304
10314
11434
10312
0
50
100
150
200
Price Per Day
Brooklyn
Manhattan
Queens
Staten Island

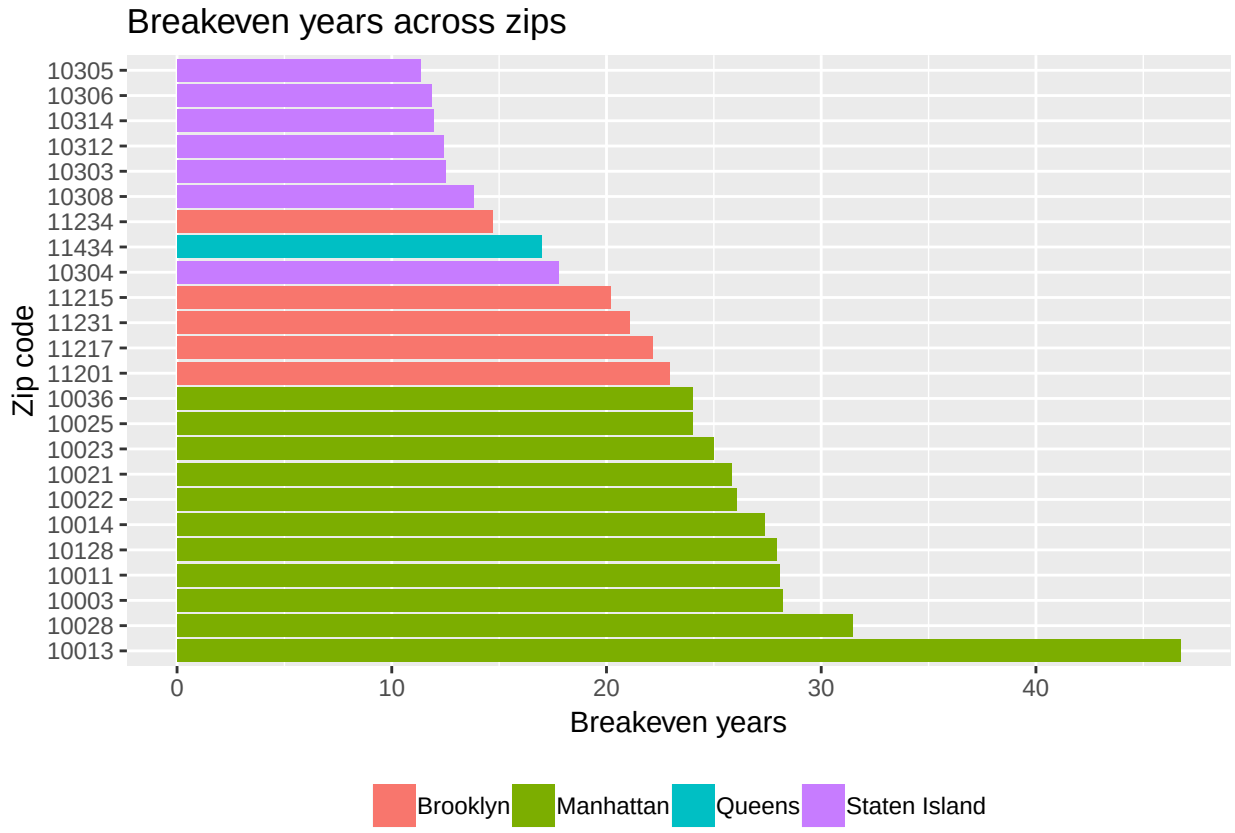


- **Years to breakeven** - Breakeven Years is minimum for staten Island followed by Queens, Brooklyn and manhatten. Graph suggests that it would be best to invest in following top 10 zips-

1. 10305(SI)
2. 10306(SI)
3. 10314(SI)
4. 10312(SI)
5. 10303(SI)
6. 10308(SI)
7. 11234(B)
8. 11434(Q)
9. 10304(SI)
10. 11215(B)

Graph showing Breakeven time period across Zips in order from lowest to Highest

Zips present in Staten Island gives the best ROI



4.Recommendations And Next Steps

4.1.Observations

- Although, most of the properties are in Manhattan as it receives the highest number of guests . Our Analysis suggests that the breakeven is also longest for Manhattan due to high property cost
- Manhattan presents tough competition as there are many competitors and it has high property cost. Overall, it has long breakeven point and doesn't present a good opportunity in terms of ROI
- Staten Island has considerable lower property cost but it also gets fewer guests. ***Also, travellers and tourist who are visiting NY for business and leisure, it is much better for them to stay in manhattan which is closer to offices and other tourist attractions***
- ***But going just by numbers it seems it is much safer to invest in a property at Staten Island as it offers shorter Breakeven point and is less risky compared to other locations***
- In case the company wants to diversify, they should pick the top zips from different neighbourhoods to minimize risk

4.2.Recommendations

- This is the list of top 10 performing zips that came in the analysis-

1. 10305(SI)

2. 10306(SI)
3. 10314(SI)
4. 10312(SI)
5. 10303(SI)
6. 10308(SI)
7. 11234(B)
8. 11434(Q)
9. 10304(SI)
10. 11215(B)

- ***Our recommendation would be to diversify and buy properties in top performing zips of different neighbourhoods with prime focus on staten Island***
- For instance, If we need to recommend 8 zips then these are the zips we would recommend-

1. 10305(SI)
2. 10306(SI)
3. 10314(SI)
4. 10312(SI)
5. 11234(B)
6. 11434(Q)
7. 11215(B)
8. 10036(M)

- Profitability should be evaluated in a year to understand whether these properties are providing expected returns

4.3.Next Steps

- ***Property cost*** - Property cost used for this analysis was from Jun'17. Latest property cost should be used if it is available. We can also use forecasting techniques like ***ARIMA*** to predict latest property cost from past trend. We can also take the actual cost of properties which would give more accurate results
- ***Factors impacting Occupancy***- In the present analysis, although we have used occupancy rate based on review score location, there are other factors which impact occupancy rate. We should also normalize our results for other factors like cleanliness, staff behavior etc which definitely impacts Occupancy rate using regression model.We can also use Availability_30 at property level to get more accurate results
- ***Discount/Interest Rate***- In this case, we have taken 0% discount rate as our assumption but that assumption is not practical. ***Some reasonable percentage rate can be taken to calculate NPV value and make a more accurate prediction***