Code ▾

Airbnb & Zillow Data Challenge

Lea Zhang

4/22/2019

Executive Summary

Through the use of data analysis techniques, I was able to form a picture of how profitability on short term rentals changes across New York City, in particular for zip codes. Using the result, the clients will be able to better understand the market and assess risks. These results can also be used in making decisions under different scenarios and time lines. An interactive Shiny app has also been provided to give more intuitive insights.

Key Insights and Conclusions

- From ROI, it's conclude that properties in Staten Island and Queens have higher profitability while those in Manhattan and Brooklyn have annualized ROI (~ 2%) that is less than half of that (>4%) for the former two boroughs.
- For breakeven period, properties in Staten Island and Queens typically will pay back the investment in around 10 years while in Manhattan and Brooklyn the breakeven period ranges from 20 to 40 years.
- The properties in Manhattan that have the longest breakeven periods (such as 10013, 10014) are located in the south part of Manhattan, which also have the highest property price (which can be seen from RShiny App in New York Zillow Map).
- The properties in Manhattan that have relatively shorter breakeven periods (such as 10036, 10025) are located northern side, which is away from the coveted prime downtown Manhattan areas.
- For total number of reviews, both Manhattan and Brooklyn are more popular locations for short term rental. And there are more available listings in these areas, which shows that the short-term rental market is more mature in these areas. It also aligns with the observations from the first two insights that the profit marginal is small in these two areas.

Recommendations

- Based on the three individual metrics, I recommend zip code 10036, 10025 and 11231. Although they are not ranked high in terms of ROI and breakeven period, the demand in these areas are high so it's less risk to invest and they can provide downside protection.
- If the client knows about short-term rental industry or is willing to do market research about Staten Island and Queens, zip code 11434 and 10306 are recommended.

Package Loading

Required Packages

- Tidyverse (dplyr, ggplot2..) - Data Read, Manipulation and visualisation
- Caret - Pre Processing, Feature Selection
- Plotly - Interactive Visualization
- ggmap - For geographic visualization
- KableExtra - Styling for Kable (Interactive Data Tables within Markdown)
- Shiny - for building shiny app later

Data Loading

Hide

Code

Data Preparation & Data Quality Check

Cost and Revenue datasets are handled separately in an attempt to enrich the data quality for exploratory data analysis.

- Account for common data quality issues: Missing Values, Duplicated Rows etc and make relevant changes.
- Filter Zero Variance/imbalanced columns, high missing value columns, columns not associated to the problem statement etc.
- Analyze any redudant columns and aggregate based on reasoning/assumptions.
- Produce clean cost and revenue data for joining and further exploration.

Revenue Data

Revenue data contains a mix of information including details about the properties like address, zipcode, bedrooms, bathrooms to information about host, daily/weekly and monthly price details for stay.

Dimension of Revenue Data, Summary and Check for NAs and unique values.

Remove Columns

- First non-relevant columns are removed
- Remove character columns that contain more than 20% of unique observations given considerations of missing values in the character columns. 11 columns are removed.
- Remove Columns that have near zero variance. 9 columns are removed. Therefore, 33 columns are left in the dataset.

Based on Relevance

Remove the columns that start with “require” or “host” and columns that end with “url” and “nights” as it’s found that these columns are not relevant with the problem in this case. 28 columns are removed.

Hide

Code

```
##  [1] "requires_license"                  "require_guest_profile_picture"
##  [3] "require_guest_phone_verification" "host_id"
##  [5] "host_url"                          "host_name"
##  [7] "host_since"                        "host_location"
##  [9] "host_about"                        "host_response_time"
## [11] "host_response_rate"                "host_acceptance_rate"
## [13] "host_is_superhost"                 "host_thumbnail_url"
## [15] "host_picture_url"                  "host_neighbourhood"
## [17] "host_listings_count"               "host_total_listings_count"
## [19] "host_verifications"                "host_has_profile_pic"
## [21] "host_identity_verified"            "listing_url"
## [23] "thumbnail_url"                     "medium_url"
## [25] "picture_url"                       "xl_picture_url"
## [27] "minimum_nights"                    "maximum_nights"
```

Based on Unique Values - Character Columns

Character columns with near 100% variance (Every Row is different) are removed as they provide no group level information that can be used on a larger population/scale. These columns include textual columns describing the home, host, amenties etc.

Hide

Code

```
##  [1] "name"                  "summary"
##  [3] "space"                 "description"
##  [5] "neighborhood_overview" "notes"
##  [7] "transit"               "access"
##  [9] "interaction"           "house_rules"
## [11] "amenities"
```

Based on Vairance

Get the matrics for vairables that have near zero variance using caret package and remove them.

Hide

Code

```
## [1] "scrape_id"             "experiences_offered" "state"
## [4] "market"                "country_code"        "country"
## [7] "bed_type"              "has_availability"    "license"
```

Remove 14 Columns Manually

At last, 14 columns are removed manually by going through the dictionary of the dataset based on the relevance with business problem.

Given six aspects for data quality check, several data cleaning approaches are taken for important columns.

- Zip code
  - Replace 567 missing zip codes with newly generated zip codes using latitude and longitude to have a complete dataset.
  - Unify zip codes to 5 digits to maintain consistency and conformity across the dataset.
- Number of bedrooms
  - Filter out the bedrooms that don’t equal to two to keep integrity for future merger with cost dataset, which only has cost for two-bedroom properties.
- Price
  - Remove dollar and comma sign and convert it to numeric format to maintain conformity.
  - Convert price for listings that have room type as “private room” by multiplying by two and get a new column “price_new”.
  - For visualization, Cap for outliers that lie outside the 1.5 * *IQR* limits, replace those observations outside the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile.
- first review/last review
  - Convert to date format.
  - Change the date to year , I found that the time these listings in all four boroughs started rental business is from 2001 to 2017. Most of them started during 2014 to 2016.
  - There are more than 25% of missing values in this variable.
  - Given these two observations, I used total number of reviews as a proxy for demand later.

After data cleaning, the dataset has 149 unique zip codes.

Zip Code

- Replace Missing Values After replace the missing zip codes, check if there’s missing value in the zip code column.

Hide

Code

```
##         id zipcode latitude longitude
## 1: 4896855    <NA> 40.83314 -73.91888
##                                                result zipcode_new
## 1: Grand Concourse/MC Clellan St, The Bronx, NY, USA        <NA>
```

Hide

Code

```
## Empty data.table (0 rows) of 33 cols: id,last_scraped,street,neighbourhood,neighbourhood_cleansed,neigh
bourhood_group_cleansed...
```

There's no NA in the zip code column

- Unify Zip Code to 5 Digits

Hide

Code

```
## Empty data.table (0 rows) of 33 cols: id,last_scraped,street,neighbourhood,neighbourhood_cleansed,neigh
bourhood_group_cleansed...
```

Number of Rooms

Most of properties in the data is one bedroom home/apt or one bedroom private room. Based on the assumption give in the problme statement, I chose two bed room properties here. In the next step I converted the price by times 2. But we should keep in mind the fact that it includes some price for two bedrooms private room, which should be lower than that for entire home/apt. So I underestimated the price here.

Hide

Code

```
## # A tibble: 21 x 3
## # Groups:   room_type [?]
##    room_type       bedrooms no_properties
##    <chr>              <int>         <int>
##  1 Entire home/apt        0          3369
##  2 Entire home/apt        1         10104
##  3 Entire home/apt        2          4593
##  4 Entire home/apt        3          1349
##  5 Entire home/apt        4           340
##  6 Entire home/apt        5            81
##  7 Entire home/apt        6            27
##  8 Entire home/apt        7             8
##  9 Entire home/apt        8             7
## 10 Entire home/apt        9             2
## # ... with 11 more rows
```

Hide

Code

Price

Hide

Code

Cost Data

First subset the dataset by filtering the city name that is "New York". There are 25 unique zip codes from four boroughs: Brooklyn, Manhattan, Queens, Staten Island in the subset dataset. * County name + Convert it to borough names for further analysis * Quality check: + There're missing values from 1996-04 to 2007-05 for median price. + There's no duplication in the cost dataset.

After data cleaning, there are 25 unique observations, each representing a zip code in the dataset.

Zip Code

Based on the assumption that all properties and all square feet within each locale can be assumed to be homogeneous, created the column of the boroughs each zip code belongs to.

Hide

Code

Quality Check

Historical Prices Change

Hide

Code

Merge Two Datasets

Hide

Code

After capping, the distribution seems still to be influenced by outliers.

- Staten Island and Queens have lower price and narrower distribution and this is because of limited sample size.
- Manhattan has a wider range of distribution of price.

Quality Check for All Data

- Listings in AirBnb data whose last review date is before May 1st, 2015 (two years before now) and availability in 30, 60, 90 and 365 days are 0 are regarded as listings that don't exist anymore or in other words, "fake listings".

Remove "fake listings". there are 17 rows removed.

Hide

Code

Metadata Created for Further Evaluation

Cost

Given the observations above and the assumption I have for the time, I used centered moving average to represent the property price in April 2017 in order to reduce the noise and uncover patterns in the data. In particular, I calculated a 5-point moving average. The code can be found right before the merge code. $y_t = \frac{y_{t-2}+y_{t-1}+y_t+y_{t+1}+y_{t+2}}{5}$

Occupancy Rate

Occupancy rate for a short-term vacation rental is the number of booked nights divided by the sum of the available nights and booked nights.

- Availability is not representative, the time that the host updated the calender range from several days ago to several years ago.
- Review scores based on locations of property.
- Use 0.75 as the occupancy rate for all the properties. Keep in mind that here the occupancy rate is over-estimated for certain properties that have lower demand. Same are Return on Investment and Cap Rate.

Hide

Code

Profitability Metrics

Here I used the unit of time as one year. I calculated the annual revenue based on the occupancy rate and median price for each zip code.

I chose three metrics to evaluate the profitability for each zip code as they focus on different aspects for investment: firstly, it's Return on Investment (ROI), which measures the efficiency of the investment; second is Break-Even Period, which measures how long does it take to pay back the initial cost; third is total number of reviews, which measures the demand for the zip code. I included these three metrics as "there is no single metric that will give all the information you need to make the best choices possible about real estate investment." In addition, since I assume the occupancy to be the same across zip codes, which is not realistic, it's important to incorporate the metric that could reflect actual demand for each zip code in the analysis.

- ROI

$$ROI_i = \frac{CurrentInvestmentValue - CostofInvestment}{CostofInvestment} = \frac{Revenue_i + AppreciationValue_i - PurchasePrice}{CostofInvestment}$$

in which $Revenue_i$ represents the total revenue generated until $year_i$ $Appreciaionvalue_i$ represents the total appreciation value of the property until $year_i$ For now, I only consider $Revenue_i$ given time limitation. I set i as 1, 5, 10, 15, 20, 25, 30, 35, 40 respectively so that I could analyze the profitability for each zip code in different time frames. And I chose and divided it by 40 to get the annuzlized ROI rate for convenience of comparison between different zip code.

- Breakeven Period measure how long does it take for the investment to pay back for each zip code.

$$AnnualNetOperatingIncome(NOI) = 365 \times OccupancyRate \times DailyRevenue$$

$$BreakevenPeriod = \frac{InnitalCost}{AnnualNOI}$$

- `rev_peryear` : Calculate revenue per year
- `beperiod` : Calculate break even period

Hide

Code

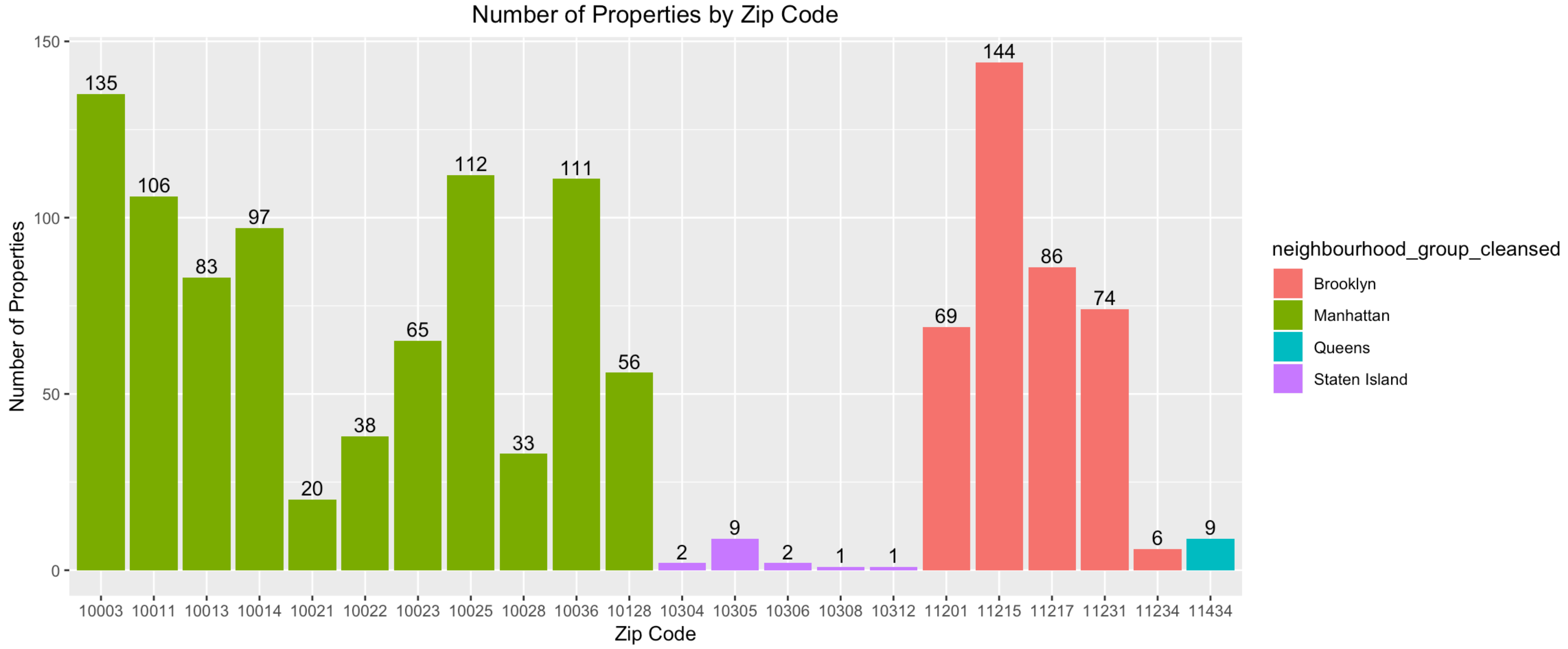
Hide

Code

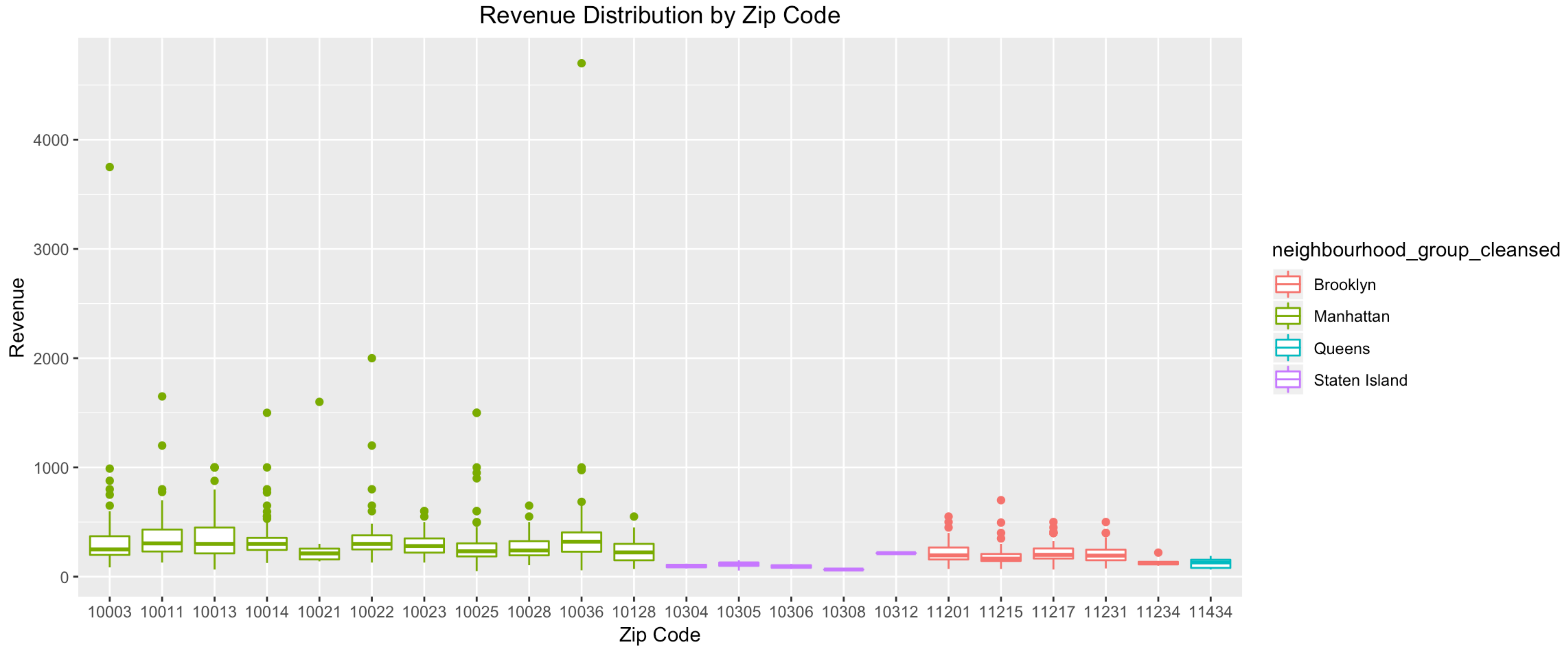
Explotary Data Analysis

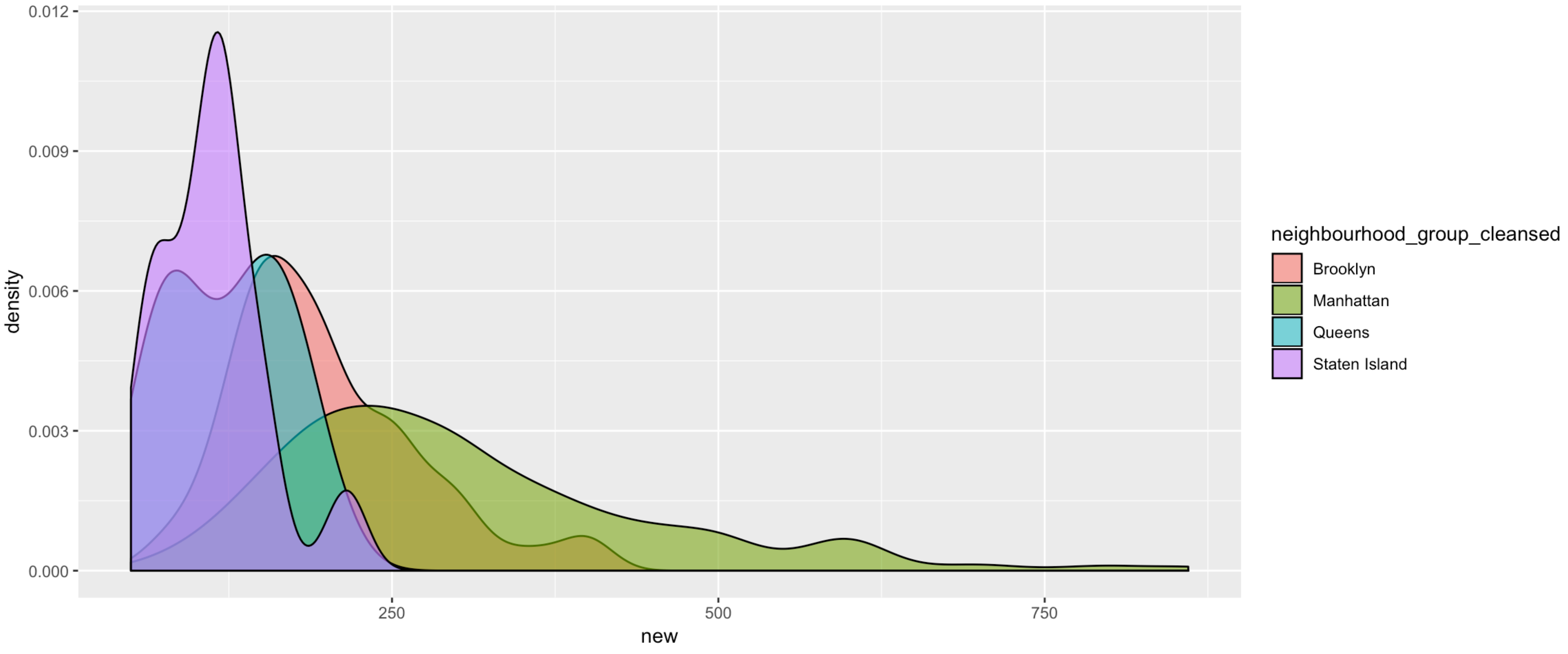
Revenue Analysis

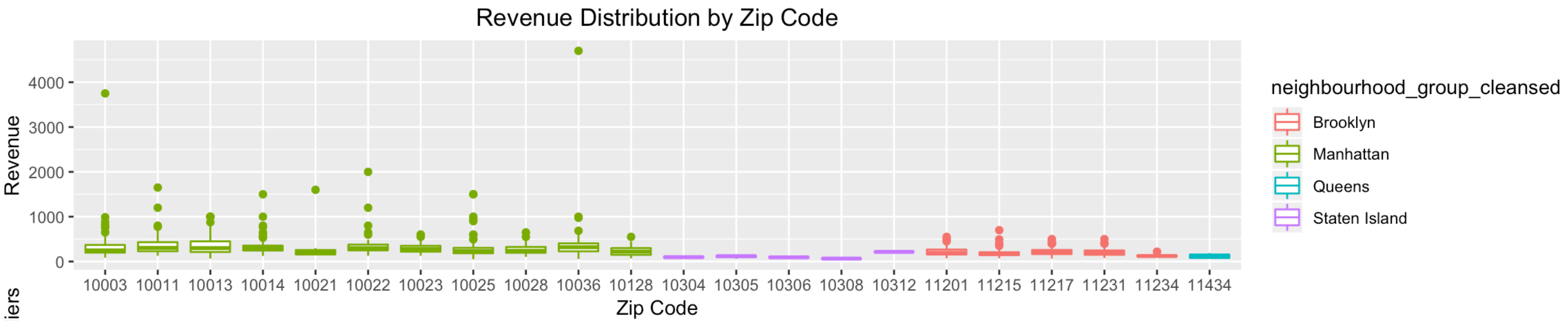
Quantity Analysis

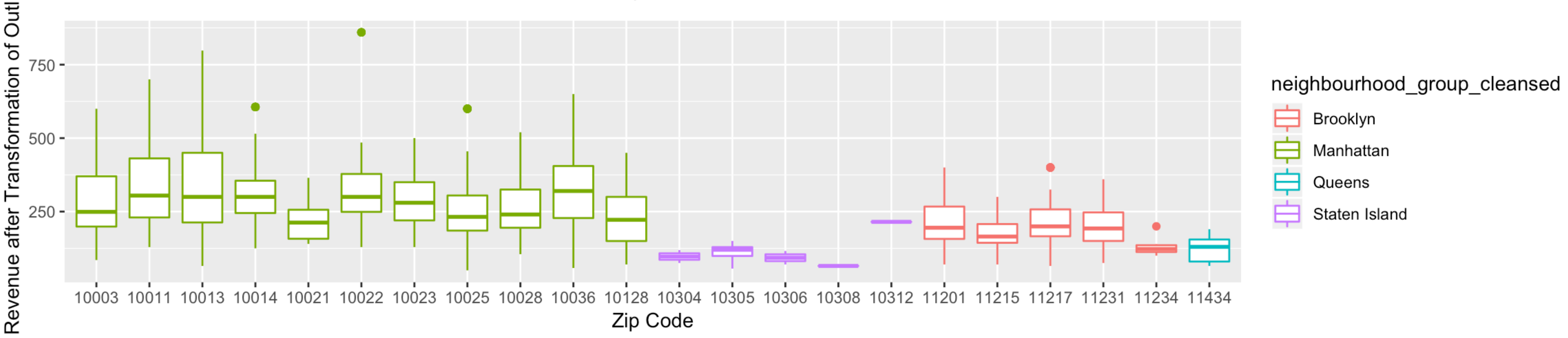


Renting Price Analysis







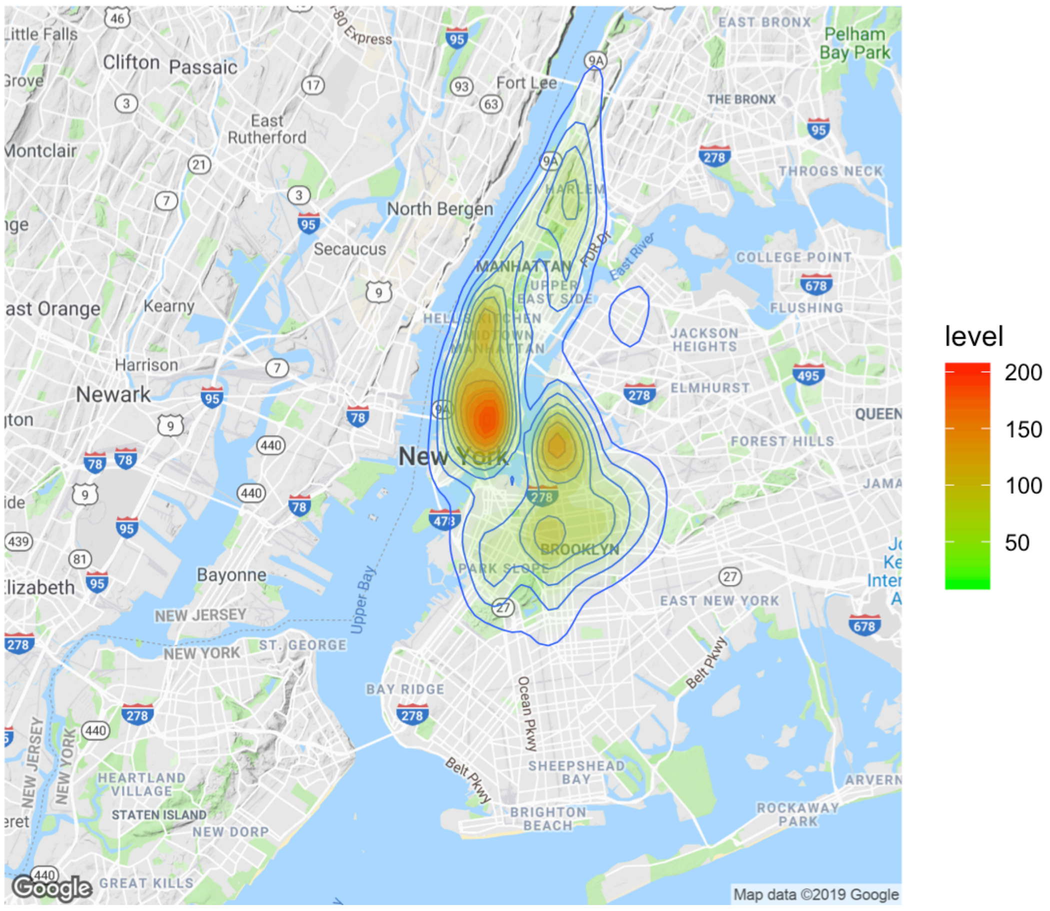


```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=New%20York&zoom=11&size=640x640&scale=2&
maptype=terrain&language=en-EN&key=xxx
```

```
## Source : https://maps.googleapis.com/maps/api/geocode/json?address=New+York&key=xxx
```

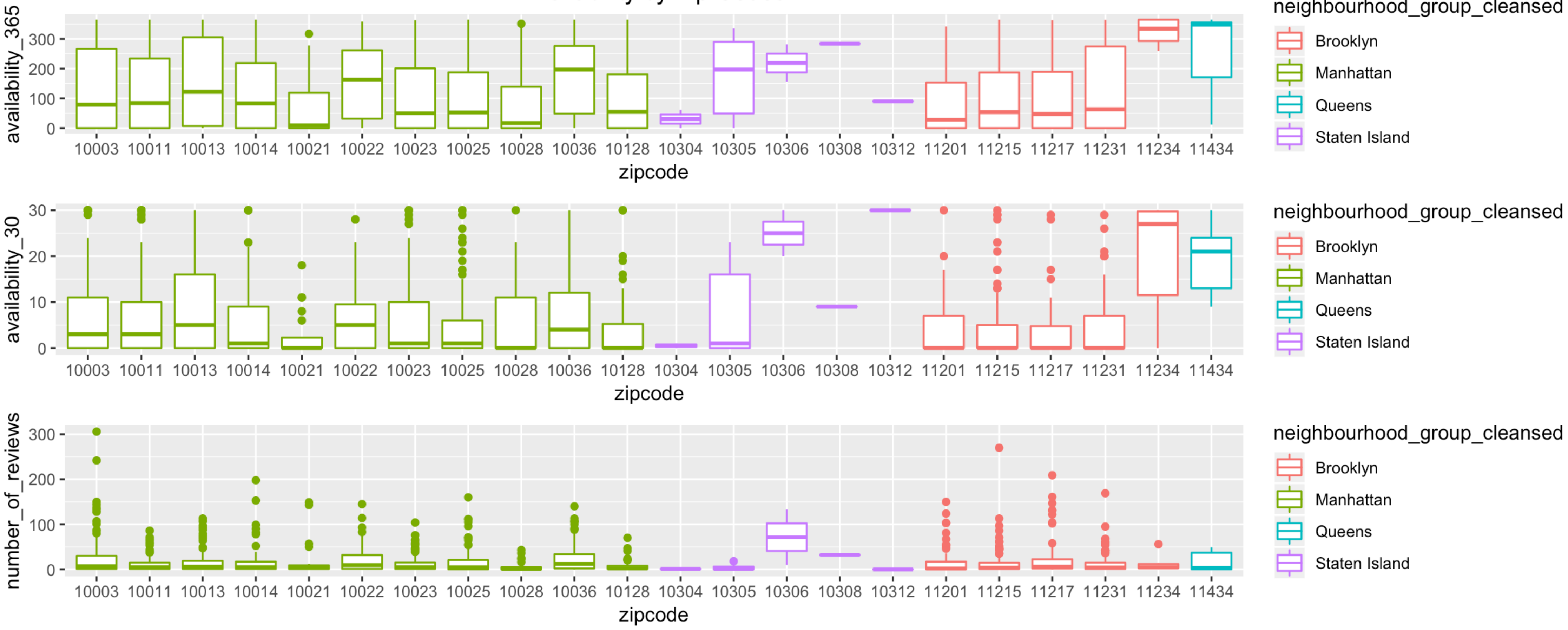
```
## Warning: Removed 37 rows containing non-finite values (stat_density2d).

## Warning: Removed 37 rows containing non-finite values (stat_density2d).
```

Listings Offered by Airbnb at Borough Level



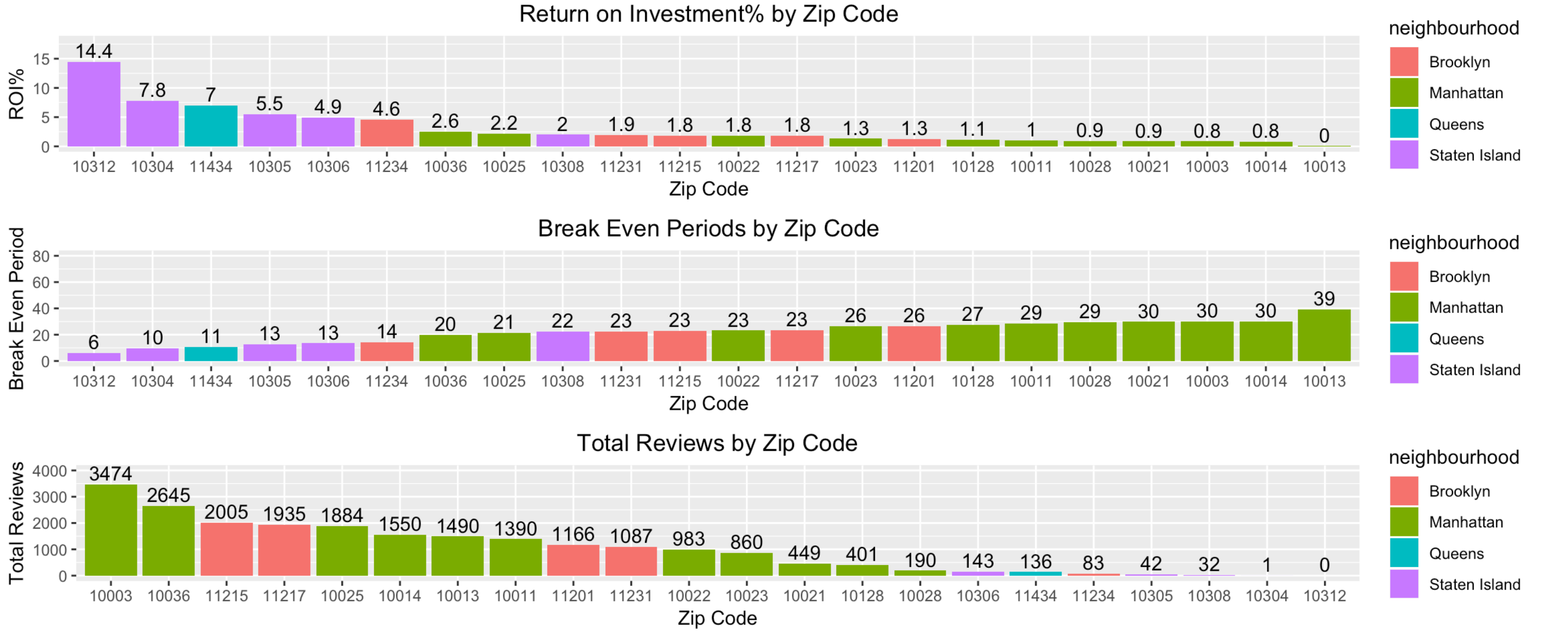
I didn't use availability to forcast occupancy rate because hosts updated their properties' availability different time.

Availability by Zip Codes



Visualization of Profitability Metrics

Combine each metric together and plot them as below.

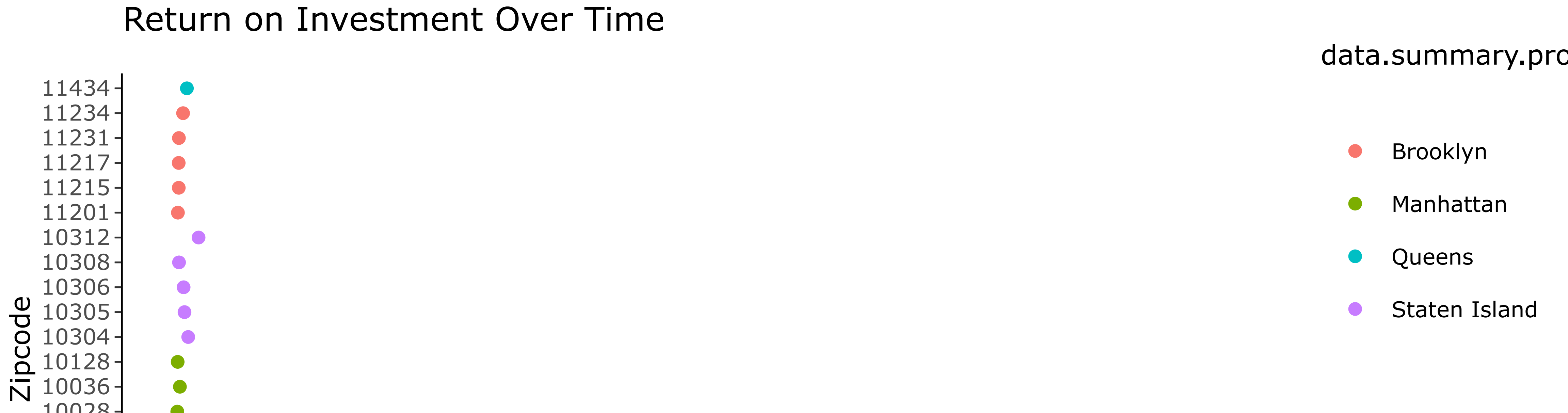


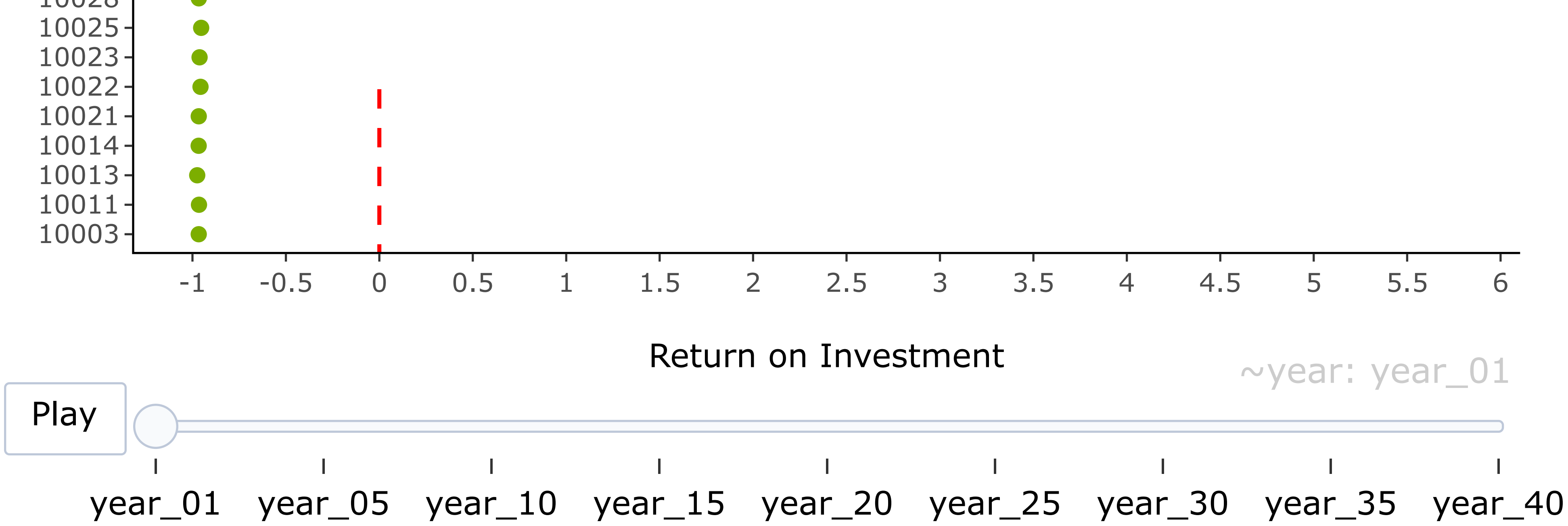
I listed the top 10 zip codes for each metrics, and the zip codes that are listed in each metrics will be recommended.
By viewing the aggregated table, I find that 10036, 10025, 11231 meet the requirments: a relatively high annual ROI, a relatively short break-even period.

Top 10 Zip Codes for Individual Metric

return_on_investment	break_even_period	total_reviews
10312	10312	10003
10304	10304	10036
11434	11434	11215
10305	10305	11217
10306	10306	10025
11234	11234	10014
10036	10036	10013
10025	10025	10011
10308	10308	11201
11231	11231	11231

Tidying up profit summary data for visulization. Using the plot below we can see how ROI changes over time for each zip code. By year 40, ROIs for all the zip codes are above zero and 10312 has the highest ROI, which aligns with the figure above.





Next Steps

- Data
  - Collect more up to date and more balanced data.
  - Include the time series for revenue data so that revenue can also be predicted.
  - Collect data from other creditable sources, such as HomeAway, Redfine, which are competitors for AirBnb and Zillow respectively.
  - Collect historical availability and get prediction of occupancy rate.
- Metrics
  - Incorporate seasonality analysis in revenue data if time series data is given. Then revenue per year is calculated based on different prices for workdays and vocations and also conditional occupancy rate.
  - Perform time series modeling (ARIMA) to predict future property cost and add this part to construct return on investment.
  - Consider fixed cost such as property management, maintenance, taxes in return calculation.
  - Get the annualized total number of reviews as the evaluation metric by calculating the duration for each listing using the time difference between today and the date for first review.