A data exploration and visualization of two-bedroom Airbnb properties located within New York City

Table of Contents

Executive Summary

In this document, I will walk through the data analysis of two bedrooms rental properties listed on Airbnb within the New York City with the goal of understanding which zip codes would generate the most profit on short term rentals. The data has been sourced from Airbnb which contains information about the listed rental properties and Zillow which contains seasonally adjusted measure of the median estimated for 2 bedrooms property value across a given zip code.

The Airbnb dataset originally contains 95 columns, but majority of these information are not relevant while the initial investment by the real estate company. The total number of columns has been reduced to 21 which contains information like, property type, price per night, zip code, neighborhood group, availability_365, etc. This dataset has been used to calculate the revenue generated by the properties for a given zip code in a year.

The Zillow dataset contains estimated time series median value of the properties for a given zip code from 1996 to 2017. In order to understand the price change trend over the years, we have focused only on the last 5 years median price. A median of median has been calculated to represent the price of the properties for a zip code throughout the year.

Through the data exploration and visualization, below points are observed:

1. Manhattan has the highest number of properties listed followed by Brooklyn
2. The top three zip code having the highest number of properties are: 11211, 11238, 10002
3. Although Manhattan market has the highest property price, it still has a significant return ratio followed by Brooklyn
4. The property cost in Brooklyn is not as competitive as Manhattan, yet it gives similar return ratio as of Manhattan
5. The zip codes generating maximum profit are located in Manhattan and Brooklyn
6. The zip code 10036 in Manhattan is the best location for investment in properties for short term rentals
7. The top 5 zip codes for investment is 10036, 10025, 1003, 11215 and 10011

Problem Statement

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 has already concluded that two-bedroom properties are the most profitable; however, they want to know which zip codes would generate the most profit on short term rentals within New York City.

Assumptions

Following assumptions are made for the scope of this analysis.

- 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.)

Data Analytics Tools and Packages Used

Jupyter notebook has been used for this project. Following python libraries has been used for data analysis and visualization:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Re for regular expression pattern matching
- uszipcode for finding missing zipcode
- Tabulate

Data Preparation

Two different datasets have been used for this analysis which are sourced from Airbnb and Zillow:

- Revenue Data – The revenue will be generated from the Airbnb dataset which has information on the listing including location, number of bedrooms, room types (entire home/private home/shared home), etc. for each property.
- Cost Data – The cost data has been sourced from Zillow which provides the average property price for 2 bedrooms by zip code.

Dataset Name	**Source**	**No of Rows**	**No of Columns**
Cost Data	Zillow	8946	262
Revenue Data	Airbnb	4893	95

Cost Data

Data Filtering

For the scope of this project, we are only analyzing the properties located in New York city. Hence, we will discard rows which belongs to other State.

Data Cleaning

The median price column has some NA values during the early years. We will replace the NA values with zero.

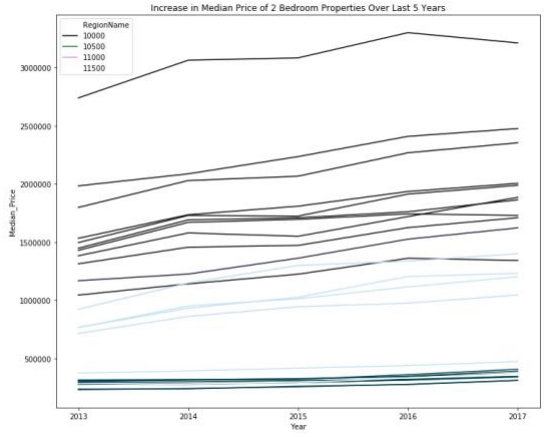
Trend Analysis

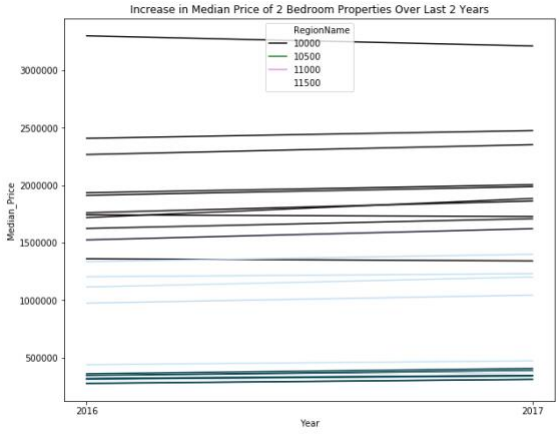
From the 5-year trend analysis, it is observed that few regions show an increase in the median cost price such as zip code 10000, 10500 & 11000 but other regions see little change.

However, increase in median cost price measured for recent 2 consecutive years (2016,2017) shares a different story. The market has little or no fluctuation. Median Cost for regionName - 10000 which had an upward trend for over 4-5 years (2013-2016) has dropped slightly. Rest of the region have little or no increase. From this analysis, we are assuming that Median Cost Price for Homes have rather been stagnant for the past 2 years. Hence, we will choose median of medians price as the actual price for year 2017 for the two-bedroom properties across every regionName (zipcode). **All further decisions will be made on this assumption.**

Assumption

Using above assumption, I have calculated the median of given median home values for each month of a given year and this median of median will be taken as the estimated cost of the properties over a particular year.





Final Cost Data

Our final cost data only includes 3 columns, zipcode, PopulationRank and Cost_2017. It has 25 rows and three columns.

	zipcode	PopulationRank	Cost_2017
0	10025	1	1342900.000
1	10023	3	1988700.000
2	10128	14	1622500.000
3	10011	15	2354000.000
4	10003	21	2005500.000
5	11201	32	1400200.000
6	11234	52	473300.000
7	10314	68	345950.000
8	11215	71	1045400.000
9	10028	109	1885350.000

Revenue Data

Data Filtering

For the scope of this project, we are focusing on understanding zip codes which would generate the most profit on short term rentals within New York City for two-bedroom properties. Hence, we will only retain rows which are in NY state, and has two bedrooms.

Also, the dataset has 95 columns, many of these columns contain information which is not relevant during the initial investment in the property by a real estate company. Such columns are removed. Also, columns high number of missing values will be removed. After filtering out, our dataset contains only 21 columns.

Data Cleaning

Missing Values

We have 3 columns with some missing and NA values. For handling these, following strategies have been involved:

1. Use longitude and latitude to find out the missing zip codes.
2. Replace NA values with 0 for remaining columns.

Invalid Values

1. We have ensured that the zipcode is only 5 digits long. Invalid values like 10003-8623 has been treated using regular expression pattern matching where values after '-' were removed.
2. The column price has '$' which is removed using string replace method.

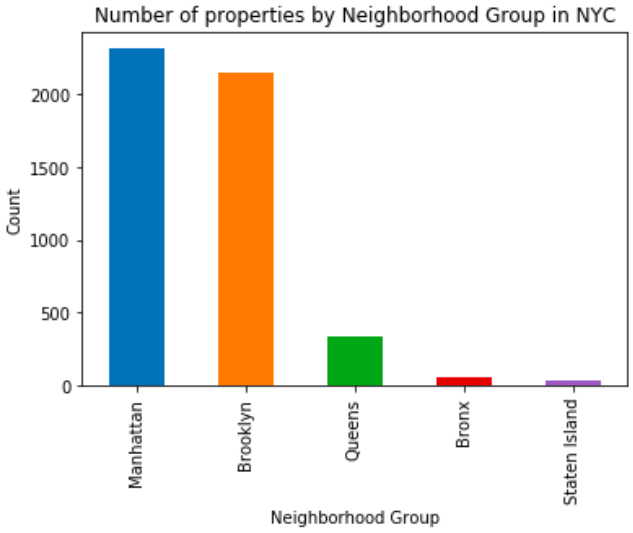
The final revenue data has 4893 rows and 21 columns.

Data Joining

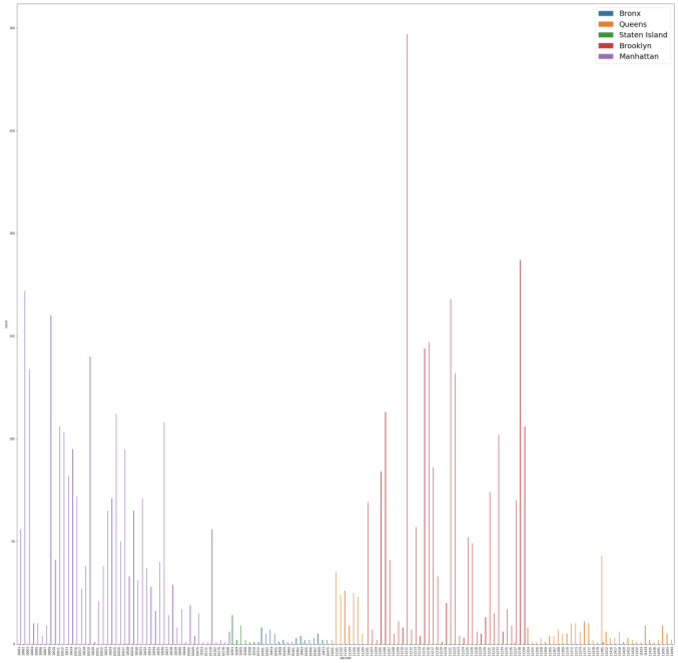
A new data frame has been created by joining the cost and revenue data on zip code. This dataset contains 4893 rows and 23 columns.

Data Exploration

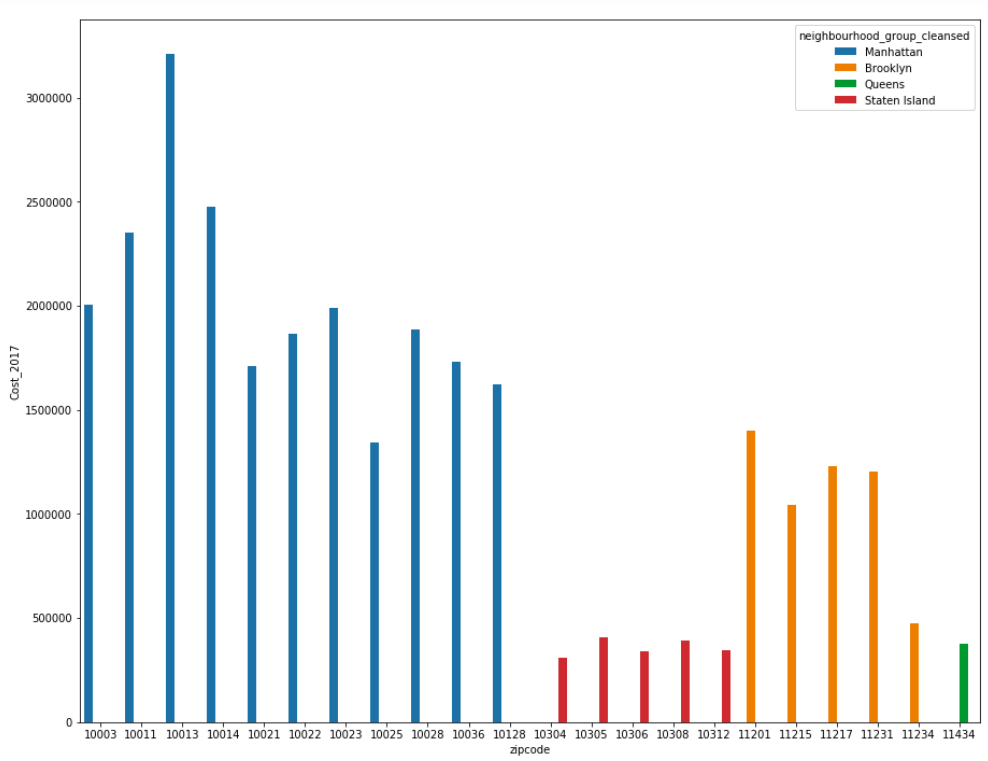
1. Number of properties by neighborhood: Manhattan has the highest number of properties followed by Brooklyn



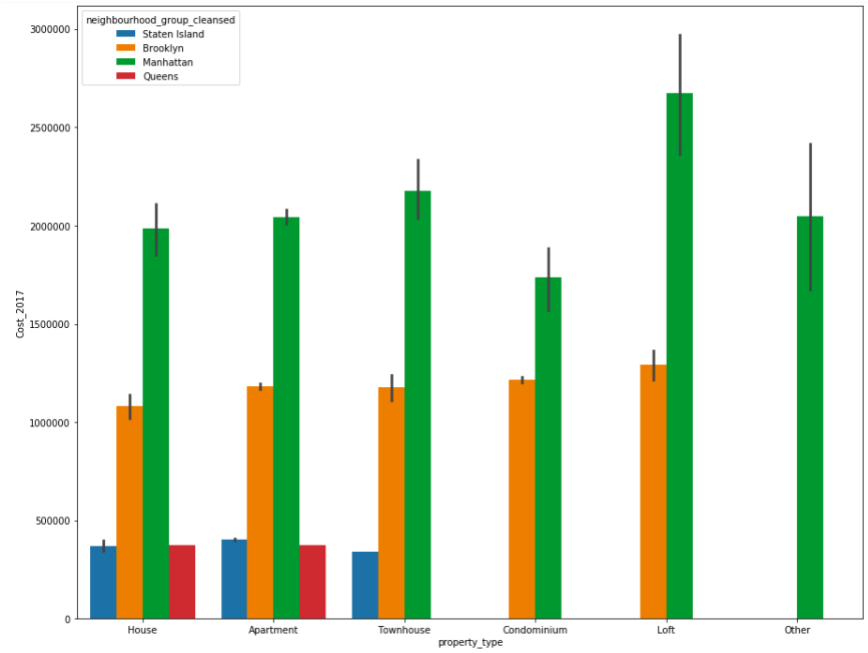
2. Number of properties by zip code. The top three zip code having the highest number of properties are: 11211, 11238, 10002.



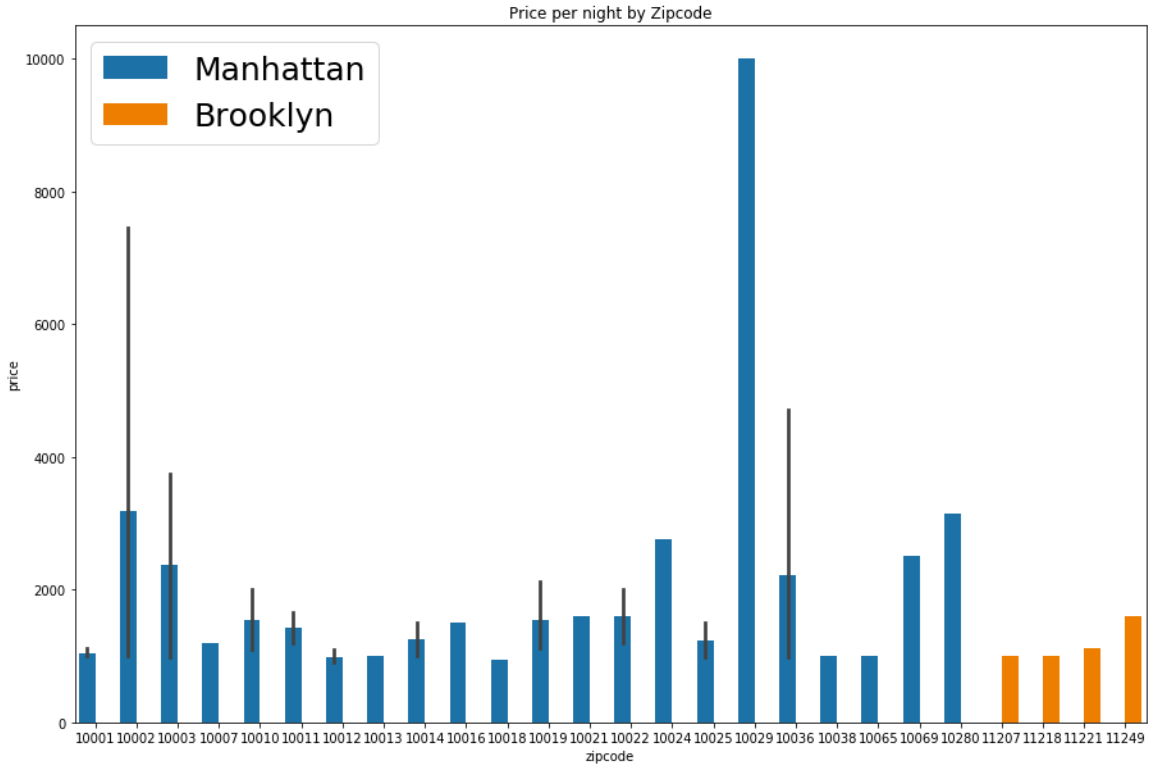
3. Property cost by Zip code: It can be observed that properties in Manhattan are the most expensive.

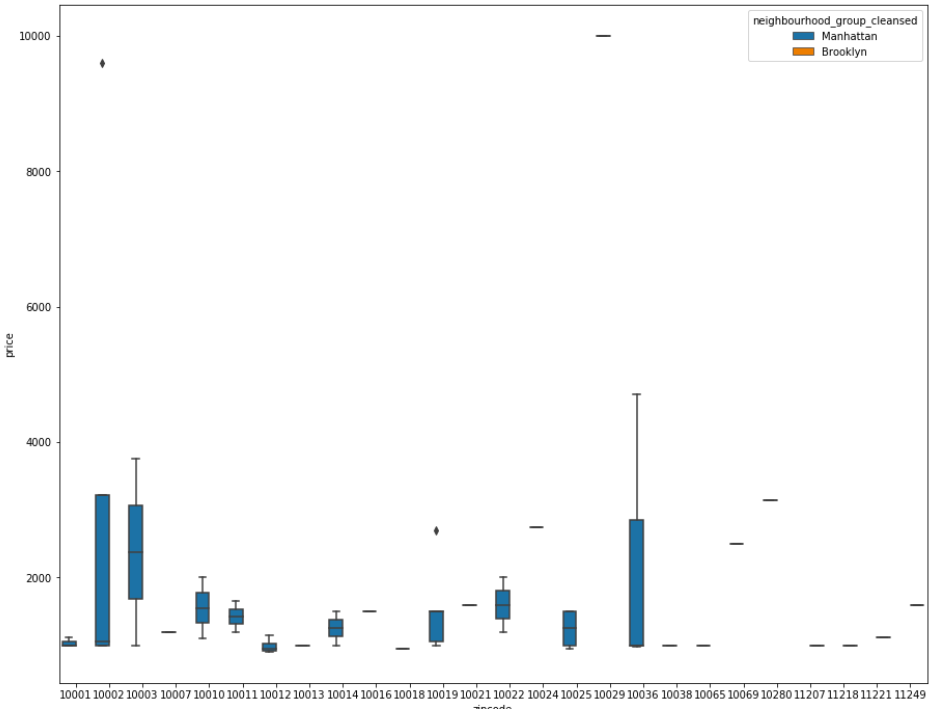


4. Cost by Property Type: The below graph indicates that loft in Manhattan are the most expensive property type. However, there is no pattern between property type and cost.

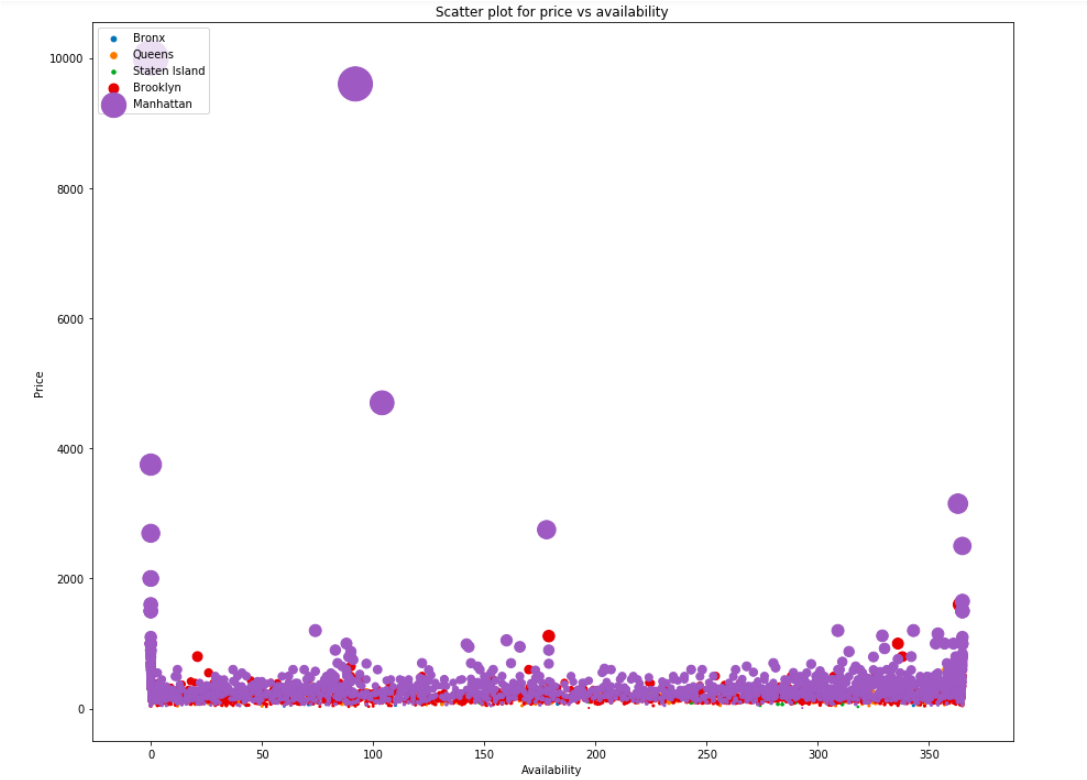


5. Price per night of each properties by Zip code. We observe that zip codes in Manhattan has the highest price per night. Also, looking at the boxplot, it can be observed that in Manhattan there are properties which are quite expensive than the rest. We see a number of outliers.





6. With below scatter plot, we tried to understand availability of a property in terms of price. However, there is no set pattern between availability and price per night.



Overall Revenue Generated for a Region

To calculate the overall revenue generated by properties for a region, following steps are involved:

1. Total price per night of a property: The provided price per night is not the price for the entire property. In order to find the price per night for the entire property, we have to check the field room_type. It indicates whether the rental is entire home/apt or a private room
2. Assumption: Based on above, we have assumed that if the property type is equal to the private room, the total price will be equal to the price per night * number of bedrooms

   Total Price = price per night * no of bedrooms (in our case, no of bedroom = 2)

3. Hence, revenue generated by each property in the first year is:
   Revenue by each property = Total Price * availability_365 * Occupancy Rate
   We are assuming occupancy rate as 75%, hence:

   Revenue by each property = Total Price * availability_365 * 0.75
4. ***Total Revenue = Sum of revenue generated by all the properties for a zip code***

Top 10 Zip Codes by Revenue Generated

	zipcode	Revenue	median_cost	neighbourhood_group_cleansed
0	10036	4959129.75	1729150.00	Manhattan
1	10011	4264544.25	2354000.00	Manhattan
2	10013	4109049.00	3212450.00	Manhattan
3	10003	4035651.00	2005500.00	Manhattan
4	10014	3238188.00	2476250.00	Manhattan
5	10025	2998412.25	1342900.00	Manhattan
6	11215	2094783.00	1045400.00	Brooklyn
7	10022	1635963.00	1863650.00	Manhattan
8	10023	1588314.00	1988700.00	Manhattan
9	11231	1523037.00	1202550.00	Brooklyn

Top 10 Zip Codes by Median Properties' Cost

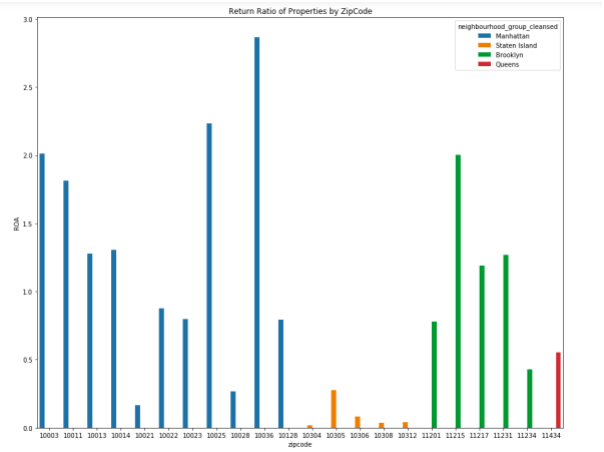
	zipcode	Revenue	median_cost	neighbourhood_group_cleansed
0	10013	4109049.00	3212450.00	Manhattan
1	10014	3238188.00	2476250.00	Manhattan
2	10011	4264544.25	2354000.00	Manhattan
3	10003	4035651.00	2005500.00	Manhattan
4	10023	1588314.00	1988700.00	Manhattan
5	10028	502900.50	1885350.00	Manhattan
6	10022	1635963.00	1863650.00	Manhattan
7	10036	4959129.75	1729150.00	Manhattan
8	10021	279475.50	1709950.00	Manhattan
9	10128	1284585.75	1622500.00	Manhattan

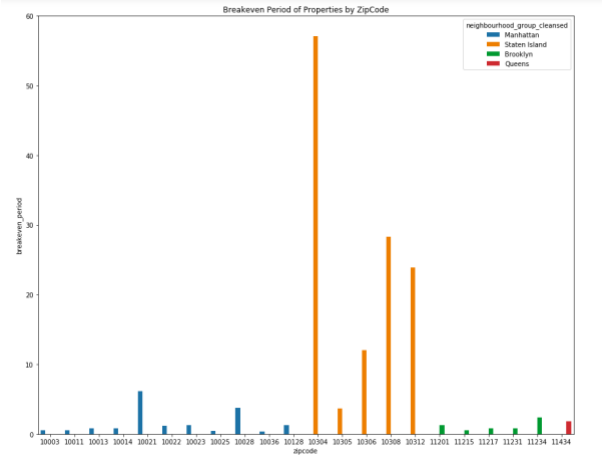
Breakeven Period Analysis

To understand the profitability for a zip code, I am using the return on assets ratio (ROA), it is a profitability ratio that measures the net income produced by total assets during a period by comparing net income to the average total assets. Breakeven Period represents the time that net income can cover total cost.

Return Ration = Total Revenue Generated or Expected Return / Total Assets (Cost Price)
Breakeven Period = 1/Return Ratio

It can be observed from below that properties in Manhattan has the highest return on assets ratio despite the competitive price.





Profit Calculation

1. We have calculated the yearly revenue generated by properties for a specific zip code
2. We have the median cost of the properties for a zip code
3. For the scope of this project, I am assuming that there is no overhead expense apart from the cost of the properties and the revenue generated is only from the rentals occupied at an occupancy rate of 75%
4. Hence, ***Profit = Revenue – Cost***

Zip codes generating profit in the first year of investing in properties

	index	zipcode	Revenue	median_cost	neighbourhood_group_cleansed	ROA	breakeven_period	Profit
0	0	10003	4035651.000	2005500.000	Manhattan	2.012	0.497	2030151.000
1	1	10011	4264544.250	2354000.000	Manhattan	1.812	0.552	1910544.250
2	2	10013	4109049.000	3212450.000	Manhattan	1.279	0.782	896599.000
3	3	10014	3238188.000	2476250.000	Manhattan	1.308	0.765	761938.000
4	7	10025	2998412.250	1342900.000	Manhattan	2.233	0.448	1655512.250
5	9	10036	4959129.750	1729150.000	Manhattan	2.868	0.349	3229979.750
6	17	11215	2094783.000	1045400.000	Brooklyn	2.004	0.499	1049383.000
7	18	11217	1463346.750	1231850.000	Brooklyn	1.188	0.842	231496.750
8	19	11231	1523037.000	1202550.000	Brooklyn	1.267	0.790	320487.000

Conclusions and Recommendations

1. Although Manhattan market has the highest property price, it still has a significant return ratio followed by Brooklyn.
2. The property cost in Brooklyn is not as competitive as Manhattan, yet it gives similar return ratio as of Manhattan.
3. For example, properties' cost in the zip code 10003 in Manhattan is almost twice the properties' cost in 11215 but has the same return ratio.
4. The zip codes generating maximum profit are located in Manhattan and Brooklyn.

5. The zip code 10036 in Manhattan is the best location for investment in properties for short term rentals.
6. The top 5 zip codes for investment is 10036, 10025, 1003, 11215 and 10011.

Future Steps

1. For the simplicity of this analysis, we have assumed as occupancy rate of 75%. However, in real world scenarios, many factors might affect the occupancy rate, including the weather, time of the year, property location, etc.
2. In future, we could calculate a more robust occupancy rate depending on these additional features. We can also see additional historical data which includes the actual occupancy days in a year and build a predictive model for occupancy rate.
3. In our dataset, we only had cost value for 22 zip codes. This can be eliminated by combining data from different publicly available data sources.
4. To determine profitability, we could use more sophisticated approach by combining other data which includes overhead expenses of properties, maintenance, etc.