**Portfolio Project: Determine the most lucrative Zip codes within NYC**

The contents of this document are as follows:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Data Manipulation/Calculation
5. Data Visualization
6. Key Insights & Conclusion
7. Next Steps

**1) Business Understanding:**

The objective of this project is to design a data product to help the real-estate company gain valuable insights as to which zip codes would generate the most profit on short-term rentals within New York City.

**2) Data Understanding:**

There are 4 Airbnb data files (csv) and 1 Zillow data file (csv).

The Airbnb files together contain a total of 40,757 records and has about 95 columns (attributes).

The Zillow data file contains a total of 8,947 records with 262 columns.

**3) Data Preparation:**

* Data Imputation:

Here, we handle missing values for both numerical and categorical variables. For numerical variables, we replace the missing values with the mean () of that column. For categorical variables, missing values are replaced with blank.

* Data Quality & Data Cleaning:

As a part of the process, for the ‘price’ column in our Airbnb dataset, we convert the datatype from object to float. We remove the ‘$’ sign from the ‘price’ column before the conversion.

The ‘zipcode’ column contains 5,7,10 digit zip codes. We retain only 5-digit zip codes and delete the rest.

We also retain the data for ‘country’ => United States and ‘bedrooms’ => 2.

* Exploratory Data Analysis (EDA):

With the help of EDA, we identify and remove the outliers.

For our Airbnb dataset, we considered the ‘price’ column values above $1000 as outliers. Also, for the ‘number\_of\_reviews’ column, values above 150 were considered as outliers.

**4) Data Manipulation/Calculation:**

For the Airbnb dataset, we compute the number of listings, total number of reviews and price\_per\_year (the dollar amount made by each zip code in 1 year) for each zip code.

For computing ‘price\_per\_year’, we assume the occupancy rate to be 75%.

The formula to compute ‘price\_per\_year’ is as follows:

‘price\_per\_year’ = ‘price’ \* 365 \* 0.75

For the Zillow dataset, we compute the Percentage Increase for each zip code over the period of 5 years.

To do this, we retrieve 2 columns;

1. last column of our Zillow dataset, which is the latest year (06/2017) as per the dataset
2. column of exactly 5 years (07/2012) prior to the latest year (06/2017)

The formula to compute ‘price\_per\_year’ is as follows:

‘Percent\_Increase’ (5 yrs.) = [(Latest year - 5 years prior) / 5 years prior] \* 100

For our combined dataset, which is our final dataset, we calculate the Return On Investment (ROI) and ‘years\_to\_repay’ for each zip code.

The formula to compute ROI is as follows:

ROI (%) = [‘price\_per\_year’ / ‘Median Price’] \* 100

The formula to compute ‘years\_to\_repay’ is as follows:

‘years\_to\_repay’ = [‘Median Price’ / ‘price\_per\_year’] \* 100

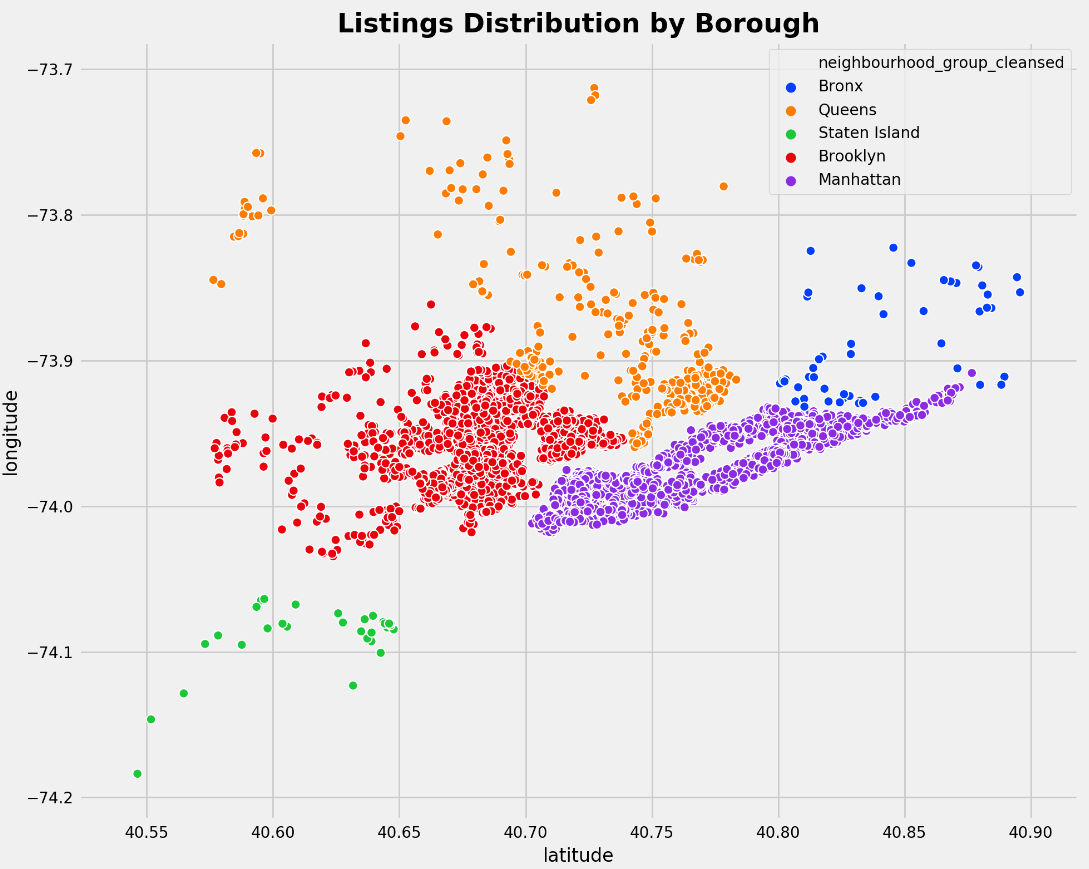
We have defined some data constraints on important columns such as Percent\_Increase, number\_of\_reviews, review\_ratings\_mean, and ROI (%) as follows:

1. **Percent\_Increase => greater than 60%**
2. **number\_of\_reviews => greater than 1500**
3. **review\_ratings\_mean => greater than 94**
4. **ROI (%) => above 6%**

**5) Data Visualization:**

**Observation 1:**

The plot below represents the distribution of Airbnb property listings for 5 boroughs.

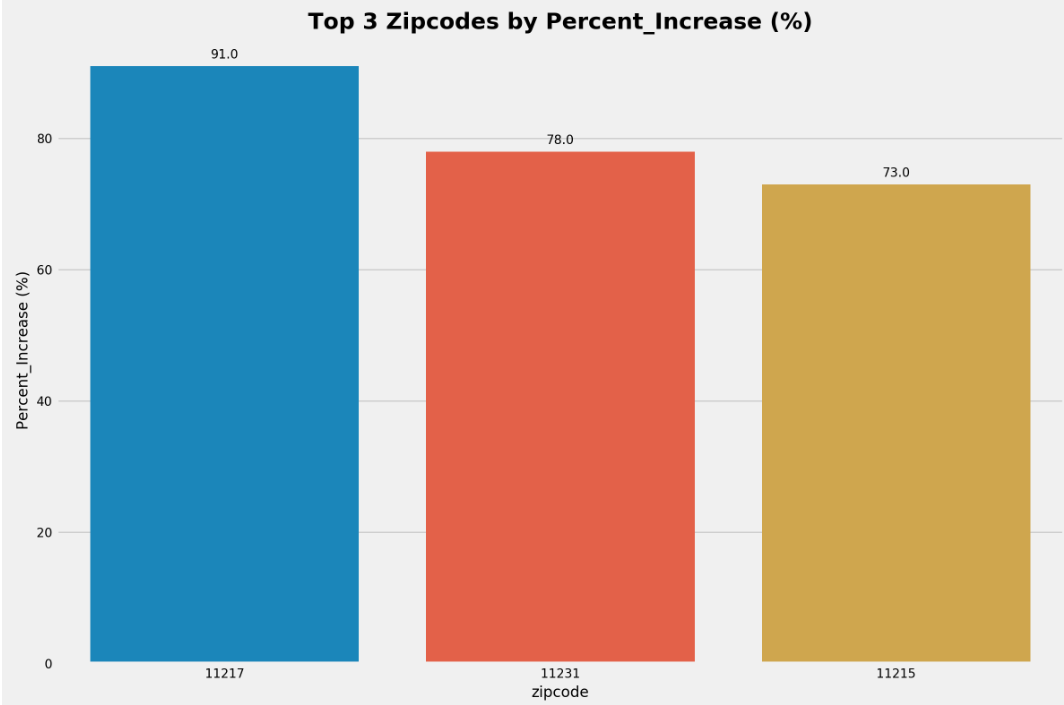


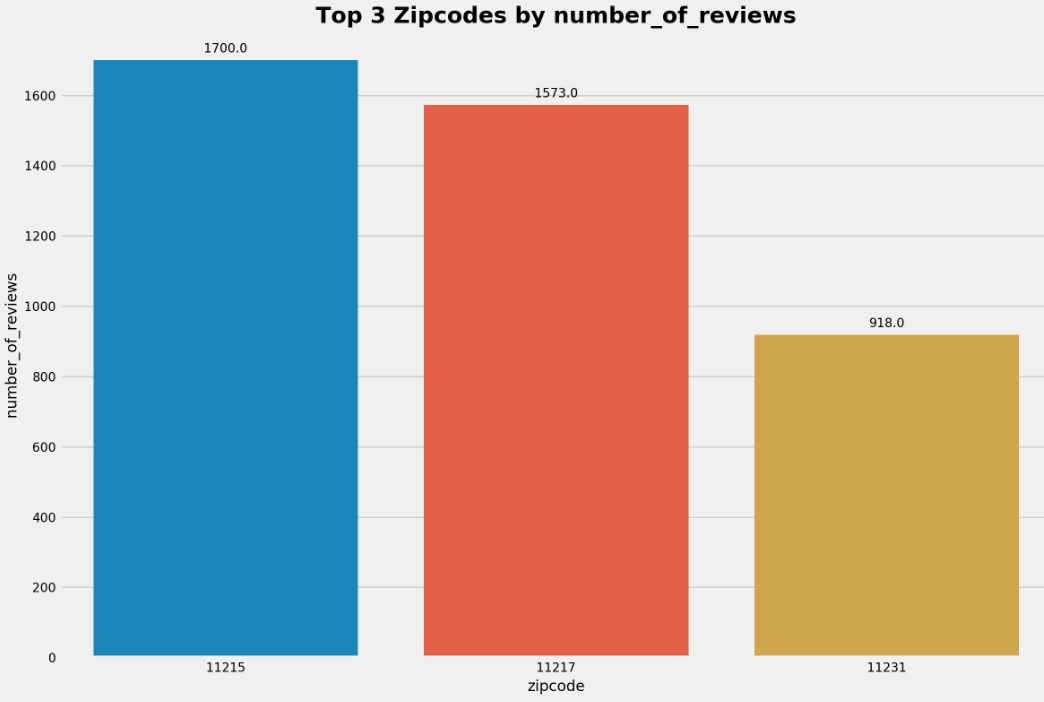
###### From the above plot, we can observe that Airbnb property listings are high in Brooklyn and Manhattan.

###### This is supported by the fact that both Brooklyn and Manhattan are the two most populous of the 5 boroughs.

###### So, it's highly likely that the most lucrative zip code(s) belongs to either Brooklyn or Manhattan.

**Observation 2:**





###### 

###### 

###### From the above 3 plots, Zipcodes 11217 and 11215 satisfies 3 out of the 4 data constraints.

###### **Therefore, Zipcodes 11217 and 11215 are the top 2 zip codes that are recommended.**

**6) Key Insights & Conclusion:**

* For **Zipcode 11217**, the property ratings given by 1,580 users are high [**95**].

Also, the Percentage Increase over the last 5 years is **91%**. This signifies the value of that Zipcode.

Suppose we assume the value [Percentage Increase] to keep growing consistently at the same rate [91%] for the next 5 years, then the ROI will significantly increase and 'Years to Repay' will reduce drastically.

Now, for **Zipcode 11215**, 1,700 users have given a cumulative ratings score of **95**. Also, considering the Percentage Increase of 73%, we can conclude that Zipcodes' property values are great!

###### Both Zipcodes [11217, 11215] belong to Brooklyn, which makes perfect sense as Brooklyn is the most populous of the boroughs with its many tourist attractions.

**Conclusion:**

#### **Clearly, Zipcode 11217 is the most profitable zip code for the real-estate company to invest in!**

**7) Next Steps:**

* We can further enhance our data product by adding more data to it
* We can extend the capabilities of our data product to provide us with zip codes that are the most profitable across the US.
* In our case, we didn't include 4-digit zip codes. For Zip codes that are 4 digits, we can append 0 at the start of it and include the 4-digit zip codes as well
* The ‘name’ and ‘summary’ columns of our Airbnb dataset could be analyzed to see whether these columns have any influence on ‘price’ or ‘review\_ratings’