**DOCUMENTATION OF THS STEPS FOLLOWED**

**Step 1: Quality Check**

Zillow Dataset:

* This dataset has selling prices of 2-BHK flats in different zip codes of USA from April 1996 to July 2017.
* The dataset has 8946 rows and 262 columns.
* The dataset has data for 48 US states.
* Zip 10013 has the highest current selling price (July 2017) and the highest increase in price (rate calculated over 10 years- difference between prices in 06-2007 and 06-2017)
* 59% of the zips have a decline in selling price over the years (rate calculated over 10 years, July 2006 to July 2017).

Airbnb Dataset:

* This dataset has data for properties listed on Airbnb in USA.
* The dataset has 40753 rows and 95 columns.
* The dataset has data for 246 US zipcodes.

**Step 2: Data Munging**

* We started out by ***combining the 4 Airbnb datasets given***. For the Zillow dataset, we ***filtered out for NY, NJ and VT states*** as the Airbnb dataset has data only for these US states.
* We ***filtered out the Airbnb dataset based for number of bedrooms equal to 2*** as it was mentioned in the problem statement that 2BHKs are most profitable
* On combining the zip codes present in Zillow and Airbnb dataset (filtered for 2 BHK flats), we could get only **22 zip codes** in common.

**Step 3: Visualization**

**Metrics we created for calculating Profitablity:**

***Annual\_Rent:***

This is annual rent collected from a property in Airbnb assuming 75% occupancy using rent price per night. This value is assumed to be fixed.

***Payback\_Period:***

This is the number of years required to earn back the initial cost of investment from annual rental income from property assuming 75% occupancy throughout the year. This tells us how fast we can recover investments from our property.

***Rate:***

We know that value of properties appreciate/depreciate depending upon the neighbourhood and locality. The difference between prices in 06-2007 and 06-2017 (in the Zillow data) divided by 10 has been taken as an average rate by which a property rises in that zip code.

***ROI:***

This metric is return on investment in 10 years assuming rent appreciates at a constant rate without including the equity value of the property. This metric is a good metric to assess profitability from a property. Because of variable rates of appreciation for each zip code this metric gives a good estimate how location of zip code determines ROI of a property.

It is calculated as follows:

(Annual\_Rent(1+Rate/100) ^10)/Current Price of the property

***ROI\_Incc:***

This metric takes into account the return on investment through rental as well as equity appreciation i.e. the increase in the price of the house is also taken into account to calculate roi. This has been calculated for 10 years as well. It is is calculated as:

A= (Annual Rent(1+Rate/100) ^10)/Price of the property {Rent amount in 10 years}

B= (Price of Property(1+Rate/100) ^10)-Price Of property {Increase in cost of property in 10 Years}

Roi\_10= (A+B)/Price of Property

**Step 4: Insights**

* Zip 10013 has the highest current selling price ($3316500) with the highest payback period (46.2 years) and ROI (2.2\*10^29). This zip has also seen the highest increase in price over 10 years. Zip code 10013 is primarily located in New York County – hence it is justified in a way.
* ZIP 10312 (Richmond City) has the lowest payback period (~6 years) with an ROI of 1.52\*10^14
* ZIP 11434 (Queens County) has a decline in selling price over the years as seen from its rate (~ -800). This zip has the lowest ROI as well (2.011457e+07) amongst all other zips
* Finding an optimum between payback period and ROI is critical for selecting the zip codes for buying the properties. Zip 11215 (Kings County) has a comparatively low payback period ( ~ 25 years) with good ROI ( ~ 1.639460e+26).

**Next Steps:**

* Since Airbnb property data is available only for 22 zip codes (amongst all the zip codes present in our Zillow Dataset), more data can be procured for zip codes where data is unavailable
* For Airbnb listings without zip codes, zip codes can be estimated by FROM longitude and latitude data, and they too can be included in our analysis.