**The Challenge:**

The challenge is to determine zip codes in New York area that has potential for high ROI (Return of Investment) in Short term rental business. The client real estate company ‘xxxx Real estate’ has already shortlisted to invest in New York area and on 2-bedroom properties that are rented on short term.

**Source Data:**

We have 2 source data available. A Zillow dataset with median property values for 2-bedroom properties across various Zip codes. The property values are available at monthly frequency for 2 decades (since Jun-1996). The 2nd dataset is the extract from Airbnb listings in New York area during May-1st -May-3rd – 2016. It has around 145K listings spanning across various property types (tent, boat to bungalow) and rental types (shared room to entire building).

**Source Data quality:**

At a first glance the source data is very extensive with thousands of rows and hundreds of columns, however the data had few discrepancies that are naturally expected in publicly available raw data.

Zillow data was confined to 2-bedroom properties, however it had data for many zip codes outside our area of interest. Also, the median property values go back for 2 decades. Overall the data quality was good with some minor datatype issues. The zip code was called as RegionName. We renamed it to suit our analysis. Also, property values for many zip codes were null for the months older than 5 years.

Airbnb Data on the other hand had listings confined to New York area however there were many types of properties like tent, boat, loft to Villas & bungalows also many listings were for shared room, portion of the property. As with any CSV source data many numeric and date type columns had blank space, currency symbols etc. causing them to be inferred as character datatype in the initial read.

**Data Munging process:**

Zillow Data:

* Zillow data was clean in terms of datatypes, we only renamed ‘RegionName’ column to ‘zip code’.
* Filtered Zillow data for only zip codes within ‘NY’ state. We pulled in all zip codes by state code even though we only needed NY city data, so that we don’t miss out any zip codes that may have been incorrectly classified as a different city.
* And there was fairly long history (about 2 decades) of monthly snapshot prices. For our analysis we do not need property prices for that long. Our assumption is that the real estate company is looking to invest in near future, hence a more recent price index would be sufficient. However, in order to study how the pricing is fluctuating historically, we will set a 5-year historical limit and calculate various median price indices.
* SizeRank is a numerical number indicating the rank of the Zip code by population. it does not provide the net difference in population size between 2 zip codes or the population percentage. So, we cannot use that for estimating population density for Zip code area.

Airbnb Data:

* Airbnb Data had lot of columns that are not required for analysis. for ex: listing\_url, host\_url, house\_rules, thumbnail\_url etc. we got rid of unwanted columns.
* The Daily, weekly and Monthly price in the listing had currency symbols and decimal period (.). That would cause issues to run aggregate functions on those columns, so we converted those columns to Numeric.
* The Data had listings on many types of properties and rental types. Also, there were properties with up to 8 bedrooms in the listing. Since our analysis is focused on New York 2-bedroom properties. Hence filtering out only columns related to Revenue, Availability, Customer Score, Property type, Neighborhood etc. by Zip code, State and City. That resulted into around 5000 rows of data.

**Assumptions:**

The analysis is conducted assuming current period as May-Jun-2017 since the datasets are time-stamped for that window and we do not have time-series data for Airbnb for later periods.

We assume the estimation study conducted per May-Jun-2017 would be relevant for future period as well, since the time value of money discount rate is 0%. Also, the ratio of income vs expenses should remain fairly same for near future.

We assume the occupancy rate of the listings for 90-day period has no outside influences such as promotions, seasonal attractions etc.

Due to non-availability of time-series data. we assumed a study of listings for their performance for one quarter would provide insight on profitability.

Since the listing price per night provides the highest revenue, for this study we assume all the listings are rented for daily rental so that the estimated revenue is highest possible for a given listing. Moreover, the monthly and weekly prices are closely related to Daily price. So, comparison of Zip code for profitability should not be impacted.

We assumed that variable costs and minor fixed costs associated with running a short-term rental to be similar across all zip codes. meaning in RoI calculations the only cost considered is the property purchase price and all other cost offset is considered to be uniform across all properties. More over other costs unaccounted are very nominal when compared to property cost.

**Analysis:**

Zillow data has time series data about median price index for over 2 decades. For our study historical prices of the property would not have much influence on budget. We are assuming that the company wants to invest in short term rental in near future (<1year). so, a more stable price estimate would be 1year\_median\_price. Moreover, Current price and historical median prices are in the same relative trend. Also, long historical median price tends to be slightly lower than current price, hence we will use 1year\_median\_price for our analysis

Quick check on match quality revealed that nearly 60% of the listings from Airbnb data is not matched with Zillow data since Zillow dataset does not have all Zip codes in NYC. Some of the Zip codes in 'Williamsburg' and 'East village' neighborhood have more than 100 listings. Cost data is critical to determine profitability, so unfortunately all unmatched Zip codes must be left out from Return of Investment analysis. This is a significant population to be left out. At this point we will split our analysis into 2 phases. In Phase 1 we will consider Zip codes with Cost data (matched with Zillow data) and in Phase 2 we will conduct only revenue analysis on Zip codes without cost data and see how other factors are influence pricing. That could reveal interesting trends within Airbnb Listings.

5 factors that influence investment strategy: Demand, popularity, customer value, Revenue, Cost. An Ideal/profitable investment would be in a Zip code that has High Demand, Medium - high popularity, High customer value, High Revenue and Low cost (or High ROI)

The Occupancy rate is used for estimating demand. Unlike noted in the exercise we will not assume 70% occupancy, because by doing so will rate all the zip codes at same performance, then high priced Zip codes will have advantage over other low-price segment Zip codes. Hence, we will use the availability data in Airbnb dataset as occupancy rate.

Airbnb Dataset has at-least 4 availability metrics at different time period 30 days, 60 days, 90 days and 365 days. A quick data check reveals that availability data up to 90 days are realistic and in most of the cases 365-day availability has same value as 90 days availability. Since we are setting our analysis period for a quarter, 90-day availability metric would be an appropriate estimation.

We derived metrics like Demand, Popularity, Aggregate Score, Neighborhood rating, RoI. the definitions of these metrics are provided in the addendum at the end of this document.

Zip codes 10036, 10013,10022 are in the top of Revenue stack earning $400 per night, however if they have very low demand, the high daily rent would not realize into a high revenue for the quarter. so, we proceeded to analyze other factors that influence revenue. From 'Cost to Revenue' Chart we can see as the revenue increases the cost increases too, except for Zip codes: 10306, 10305 and 10312. Those could be anomaly or hidden gems with relative low cost and high revenue. The true measure of profitability is when Revenue is normalized for price that is Revenue/Cost = Return of Investment.

Another trend to notice here, as pricing increases from $200 to $400, the cost increases **exponentially at the rate of 2x**.

In order to calculate Revenue, we have to determine the rate of occupancy. A good way to estimate revenue is to see the current listings and determine how they perform at set periods. The availability metric in Airbnb data can be used for this. as we have established in data quality check the availability data for future 90 days are reliable. So, we used that to estimate the performance of the listings by quarter

The 'Demand by Revenue’ chart reveals that Demand is high for properties below $100 per night or $225-$385 per night. Lot of Zip codes clustered above 80% demand indicates that $250 - $350 is the best pricing band. Assuming no impacts from external factors (like seasonality, attractions etc.) investing in new properties would increase supply decrease demand. Hence Zip code with demand above 80% is ideal for investment and when the demand comes down it could still be around 70% occupancy rate. Based on this Zip codes 10308,11215,10201,11217,10028 etc. are top contenders.

Investing in new properties would add to supply and hence impacts demand and revenue. Total number of listings in a Zip code is a blended indicator for Supply and Popularity of neighborhood/Zip code. That means supply is relative to Zip code for example around 80 listings in Zip code 11217 generates 100% demand whereas the same number of listings in Zip code 10013 generates 60% demand. This can be attributed to price. Since the pricing is high in 10013 it is not generating that much demand. Hence the number of listings is an indicator of both supply and popularity of the neighborhood.

The 'Popularity vs Demand' chart reveals that high demand Zip codes have varying supply. For example, 10308 has 80% demand, but it has only 1 listing. Adding another property will bring down the demand to 40%. Investing in Zip code 11217 is better since the demand is very high (97%), adding more supply would bring down the demand only a little. On the other hand, Zip code 10013 has similar supply but medium demand. Adding more supply would push the demand low and hence price/revenue. With high cost coupled with low demand the ROI is going to go down.

Hence it is advisable to invest in medium to high demand with high supply/popularity Zip codes: 11215,11217,10025,1003,10014 etc. are top contenders.

Overall Customer Affinity chart follows similar trend as Demand chart with the exception of Zip code 10304. Almost all Zip codes have median review score of above 90. Interestingly Zip code 10308 has a high score even though there is only one property listed there. It could probably be a skewed metric and also since the pricing is very low customers are happy with the services they get in that price range. Similar check on Location score indicates that almost all the zip codes are scored at 100%. so, Customer affinity does not play an active role in investment, however we can use this rating to prioritize Zip code. Aftercall happy customer brings more business.

Three Zip codes that emerged profitable 10305,10306,10312 in Cost to Revenue chart are not in high demand though. If there is no high demand than supply is high, even though they have only few listings each, investing in those Zip codes will increase supply and thereby reducing demand and profitability, unless the cost is so low that even with low demand they could emerge profitable.

We went on to analyze the larger population of Zillow unmatched Zip codes. These Zip codes do not have cost information so we cannot conclusively estimate quarterly performance. However, we can conduct study on factors influencing Revenue to see any interesting trends.

The 'Demand by Revenue chart for Zip code without cost' is following similar trend, but the appealing pricing band is widened a bit from \$150 to \$450. Zip code with revenue outside that range have overall lesser demand. Also, this chart reveals there are more than 50% of Zip codes with revenue band \$100 to \$200. Properties with median cost of \$500,000 in these Zip codes would be highly profitable for this price band. Due to unavailability of property cost, we cannot conclusively determine best performing Zip codes.

'Popularity vs Zip code chart for Zip codes without cost' reveals an interesting trend. Even though $100 to $200 is a dense price band, more than 90% of those Zip codes have very few listings (in single digits <10). It could mean that the demand is very high because of less supply and adding new listings in those zip codes would bring down the demand drastically. It also indicates that $100 - $200 is an economically sensitive price band. slight increase in price would drastically bring down the demand. Again, price band $200 - $300 performs better here, consistent with Zip codes that we know their median property cost. One other interesting thing to note here is Zip code 11211 has relatively large number of listings and yet it has more than 80% demand rate. If the median property cost in this Zip code is comparable to other Zip codes in that range, it could be one of the potential Zip codes for investment

Profitability vs Revenue chart reveals that customers like listings with lower price band. An indication towards economically sensitive customers. According to NYC government planning Hotel Market Analysis (Ref: <https://www1.nyc.gov/assets/planning/download/pdf/plans-studies/m1-hotel-text/nyc-hotel-market-analysis.pdf>, Page#38) the average price per night is consistently above $250 in past 6 years, So Airbnb listings turns out to be economical for short term renters and our chart indicates the same with high customer affinity to that price band. The current Airbnb listings in this price band enjoys huge demand and hence could provide high Return of Investment provided the median property cost in those zip codes are in the range of $500,000 - $750,000 (based on our study earlier with Zip codes with cost data).

Overall conclusion is that due to non-availability of cost data for all such Zip codes, it is not possible to conclusively determine the profitability. However, based on the trend revealed through series of charts, the $100 - $200 price band is one of the promising segments to consider for investment.

**# Conclusion:**

Based on detailed analysis looking at various factors influencing profitability and analyzing through 100+ Zip codes with 40000+ listings, we can conclusively shortlist the following

7 top performing Zip codes in their order of preference:

11215

10003

10025

10036

10011

10014

11217

The above-mentioned Zip codes have an average occupancy rate higher than 80% and have high customer affinity. The median cost of property in these Zip codes ranges from $900,000 to $2.3 Million thereby quarterly ROI is estimated to be around 1.25% to 2.5%

In addition to these, based on our findings from analyzing Zip codes that did not have cost information, it is indicating that there is a high demand and high review ratings for listings at $100 - $200 range. This price range is below the average daily rate in NYC hotel market which is $250 and above per <https://www1.nyc.gov/assets/planning/download/pdf/plans-studies/m1-hotel-text/nyc-hotel-market-analysis.pdf>, Page#38. and is appealing for short term renters. In-fact it is one the edges that Airbnb listings have over traditional Hotel businesses. Zip codes 10304, 10305 and 10308 have high demand and high ROI. Median Property cost in those Zip codes are very low (less than $500,000) and hence high ROI. These Zip codes are ideal for low cost investment strategy.

**Next steps and future analysis:**

1.We currently have only a snapshot of revenue data and historical cost data. As next steps we could augment this with more input data related to Zip code population index, seasonality, attractions etc. and further refine the analysis.

2.The cost information from Zillow data matched for less than 40% of Zip codes. There were more than 70 Zip codes without median property cost and hence could not be included into this analysis for conclusive finding. The property cost data can be expanded to include all zip codes

3.The most appealing price band for NY rentals is $250 - $350. This puts Airbnb listings in direct competition with hotel market. Our understanding is that the Airbnb listings are doing well because they do not have overhead costs involved in traditional hotel business. However, an extensive analysis considering those factors could reveal new patterns.

4.Airbnb listings are very popular in $100 - $200 price band. This economical segment could turn out to be most profitable since the properties in such Zip codes tend to cost less. Also, overall trend of rental price vs property cost shows that the property cost increases at a rate of 2x - 3x times that of rental price increase. An extensive analysis into that price segment is advisable.

5.This analysis is performed entirely within Airbnb market, we need to include data from traditional hotel market and other competitive business to realistically gauge the performance.

**Addendum**

**Metadata of derived metrics:**

* **Revenue**: Median price of a property for daily rental at Zip code level. It is Sum of Daily rental price and cleaning fee if any. Even though cleaning fee is technically not a revenue it goes against variable cost and since some properties collect cleaning fee and some do not. A fee collected is additional income that would offset expense. Hence, we are considering the fees as part of Revenue
* **Demand**: Median demand rate by Zip code. It is percentage of number of booked days over 90 days. This gives as the demand rate averaged at Zip code level. 100% demand means the property is booked for all 90 days, in other words availability is 0.
* **Agg**\_**score**: Median score of all customer rating given based on accuracy, cleanliness, check-in, communication, location, and value averaged at Zip code level.
* **Location**\_**rating**: Median score of all customer ratings on Neighborhood.
* **Popularity**: Total number of listings by Zip code.
* **ROI**: Return of Investment: It is the percentage of revenue earned in a quarter against cost of the property. This indicates how many years are required to recover the entire investment made. This rate standardizes revenue and cost at different rate into one scale for easy comparison between Zip codes. Higher the RoI, better the performance of the Zip code.