Author: Milind Ikke

- ## Objective

  The objective of this report is to find hidden insights from publicly available data of Airbnb and Zillow and further provide recommendations on which zip code have the highest return ratios within New York City and what are the most significant factors contributing to the revenue. Our goal is to help develop a data product and tailor client's data strategies, whose part of its business model relies on short term rentals, to the market demand. Data visualization and predictive statistical models would be employed in the report to better answer our business questions.

- ## Problem Statement

  A real estate company has a niche in purchasing properties with two-bedroom to rent out short term as part of their business model specifically in New York City. They want to learn more about which zip codes are the best to invest in and build a data product with a rational conclusion to find out where their targets are and what factors they should consider when purchasing properties within New York City.

- ## Dataset Description

  The publicity available data from Zillow and Airbnb would be used for our analysis.

  - **Zillow Data:** Cost data to provide us an estimate of value for two-bedroom properties
  - **Airbnb Data:** We can treat the data from Airbnb as our revenue data that contains information about properties located in Greater New York area, such as zip code, price per night for each listing, and geographic information.

- ## Data Cleaning and Preprocessing

  The importance of data integrity would never be overemphasized. Data preparation accounts for 80% of the works of analyst or data scientist. The tasks involve cleaning and organizing data, and we rely heavily on quality data to conduct data analysis. The better the data quality, the more confidence users will have in the output they produce. We will discuss how we preprocessed the cost and revenue data respectively in six data quality dimensions: Accuracy, Integrity, Cleanliness, Correctness, Completeness, and Consistency.

In this section, we would discuss how we deal with messy data of the Zillow and Airbnb data respectively. The scope of data preprocessing involves feature selection, missing value imputation, and data cleaning.

1. Zillow Dataset

Features	Data Preprocessing
RegionName	Rename it to zipcode, Unify zipcode to 5 digits, check uniqueness
City	Consider only New York City
Price Trend	Predict price over the period of 5 years
Estimated Price	Median for last 6 months and drop rest of the months

2. Airbnb Dataset

Zipcode	**Unify zipcode to 5 digits, convert latitude and longitude to identify missing zipocdes, check uniqueness**
Price	Perform outlier analysis
Bedroom	Select only 2 bedrooms listings

- **Metadata for the created data**

estimated_price	**Median for last 6 months of property cost data**
occupancy_rate	Occupancy rate for each zipcode
expected_return	Net return on the investment
return_ratio	Profitability ratio that measures the net income generated by the total assets in a certain period
breakeven	Reciprocal of return ratio.
Total Rank Score	Cumulative rank for all analysis

- **Tools**

  Python is the main tool in our analysis. Here are the packages used for data cleaning, data manipulation and data visualization.

Platform	Packages
Anaconda Python Distribution	NumPy, Pandas, Matplotlib, Seaborn, uszipcode, warnings

- **References**

1. https://pypi.org/project/uszipcode/
2. https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm
3. https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
4. https://www.mashvisor.com/blog/breaking-even-real-estate-investment-property/