AirBnB & Zillow Data Challenge - Metadata

Dataset Level Metadata

Data Set	Description
Airbnb_data	The original data loaded from the provided URL
Zillow_data	The original data extracted from the Zip_Zhvi_2bedroom.csv file
airbnb_req	The filtered dataset with only 2 bedrooms data and the useful columns
zillow_req	The filtered Zillow data with only the estimated costs of the recent year
joined_airbnb_zillow	The single merged dataset with Airbnb data and Zillow data (only required columns)
final_data_req	This dataset only has columns regarding the breakeven data, mean price and revenue
final_data	final_data_req group by the zipcodes
most_profitable_zipcodes_and_revenue	Result dataset with the most profitable zipcodes depending on the number of years which is passed as an argument

Packages Used;

1. **dplyr**
2. **ggplot2**

Metadata – Columns and Variables

Field	Description
Mean_zillow_cost	Mean cost value of the last 12 months estimated Zillow costs of the properties
revenue_per_year	The estimated revenue generated for the property according Airbnb price per night and the assumed occupancy 75%.
breakeven	Number of years it will take to break even(No profit and No loss).
zipcodes_earliest_profit	Top 10 zipcodes which would start generating the profit at the earliest

Metadata – Functions

Function Name	Description
extract_airbnbr_url	Takes the filename containing the Airbnb data URL and extracts the URL after splitting
get_most_profitable	Takes the number of years to be considered and Number of zipcodes to be displayed. Returns the Profit or Loss amount in USD after the particular number of years.

Answer for the main question:

The answer to the main question to find the zipcodes that would generate most profit depends on the time period considered.
For instance, if the time period is **5 years** (displaying the top 5):
get_most_profitable(final_data, num_years = 5, num_zipcodes = 5)

	Profit/Loss in USD	Zipcode
1	-41585.4166666667	10312
2	-166439.583333333	10304
3	-201956.25	10306
4	-222852.083333333	11434
5	-238872.916666667	10305

Most Profitable zipcodes in order: 10312, 10304, 10306, 11434, 10305

#If the time period is **10 years** (displaying the top 5):
get_most_profitable(final_data, num_years = 10, num_zipcodes = 5)

	Profit/Loss in USD	Zipcode
1	252695.833333333	10312

	Profit/Loss in USD	Zipcode
2	-33670.8333333333	10304
3	-74662.5	10306
4	-86637.5	10305
5	-86737.5	11434

Most Profitable zipcodes in order: 10312, 10304, 10306, 10305, 11434

#If the time period is **30 years** (displaying the top 5):
get_most_profitable(final_data, num_years = 30, num_zipcodes = 5)

	Profit/Loss in USD	Zipcode
1	1429820.83333333	10312
2	1245635.41666667	10036
3	1189723.13596491	10022
4	984850	10025
5	642585.171568628	10011

Most Profitable zipcodes in order: 10312, 100346, 10022, 10025, 10011

Other Conclusions:

1. The real estate company should invest in those zipcodes according to their future planning i.e. how many years they want to invest.

2. Overall, the breakeven period of the properties in Manhattan are so high compared to the other regions. The real estate company should avoid investing in Manhattan neighbourhood.

3. The breakeven period of the properties in Staten Island is much lower than other regions. So, the company should prefer to invest in Staten Island.

4. The zipcode 10312 generates the profit at the earliest compared to all the other zipcodes. The client should try to invest in this zip code.

What's Next - Future Steps :

1. Some zip codes in the airbnb_data are missing. Though we have their neighbourhood group, it is very inappropriate to fill the zip codes with the most occurred zip code in that neighbourhood. Doing so would decrease the data quality and induce the bad data. Therefore, it is avoided.
Another possibility of filling the missing zipcodes is through the latitude and longitude information that we have. If the lat and long information given is identified as correct, then we can use free reverse geocoding R packages like ggmap and find the zipcodes. This can be pursued in future work. Having more data refines the results and provides more insights.

2. We observe that though the original data is large , we had to do analysis on much smaller data due to the unavailability of Zillow data for those zip codes.
We can try to incorporate multiple data sources like Zillow, property registration records in the future.