**FINAL PROJECT**

**GROUP M**

**MEETKUMAR PATEL**

**SHIV PATEL**

**MAAHI PATEL**

**PARTH BAROT**

**ANALYSIS OF AIRBNB LISTINGS IN US CITIES**

**BEST ZIPCODES FOR REAL ESTATE INVESTMENT IN TOP 3 US CITIES USING AIRBNB LISTING**

### 

### 

### 

### 

### **BUSINESS PROBLEM**

### ***An analysis of Airbnb listings of popular US travel destinations in conjunction with Zillow's residential real estate data to suggest the best zip codes for real estate investment in the top 3 cities using Airbnb listings.***

Where to buy the desired property based on various factors( price, trends, neighborhood, etc.) along with their availability in different cities (Short Term Rentals)???

* property - whether apartment or house? will be selected as the best based on the location.

**Analytical Context:**  We mainly focused on the listing, demographic, and venues dataset as it has crucial information for us to extract to answer this question, and using the real estate dataset to study the trend of property prices over the years in the US cities.

**Business Context:** The goal of real estate investment is to put money to work today and allow it to increase so that you have more money in the future. In this data assignment, we will assist real estate companies to understand which zipcodes are more profitable for short-term rentals.

**Business Impact**

After the target analysis is conducted on the dataset, buyers can easily make their decision to buy the right property according to their personal needs. Allowing buyers to see a better area to buy properties in terms of their requirements which is more economically feasible and at the same time provides multiple options from the given dataset.

This allows them not only to have a better property in terms of their own living (i.e., any unit with multiple bedrooms and bathrooms) from their areas of preference multiple options to check the availability of the unit.

It also provides an opportunity for them to profit if they are looking for it in terms of investments, as they would not be able to get price rates, but also data on where they will be able to make long term margins on investments.

The trend will give the buyer an opportunity to look, compare, and choose options among datasets from Airbnb.

The ability to make decisions will drastically improve as the decision will be backed by numbers, and at the same time, the decision will also provide key results which would save them time to carefully consider each option before making any decisions.

**INSIGHTFULLNESS OF CONCLUSION**

Through this exploratory data analysis and visualization, we gained several interesting insights into the Airbnb rental market. This Airbnb dataset appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented. After that, we proceeded with analyzing boroughs and neighborhood listing densities and what areas were more popular than others, their price variations, and their availability as per room types.

After attempting to interpret the provided dataset in a variety of ways, we have selected to investigate this inquiry because we discovered that investment in property is a growing trend and business is related to it. Investors frequently look to these factors before heading into the lodging industry. Our key concern is to help investment businesses to find the best location which is more profitable for short-term rentals.

Key Findings:

1. The majority of zip codes in New York Manhattan and Brooklyn have higher property costs, and some of them don’t. These zip codes are also popular in terms of rental rates, the number of available choices, and prebookings(occupancy rate) - they are excellent choices for investment.
2. Staten Island, Queens, and Sunnyside Neighbourhood have limited choices but they cost lower, are highly popular, and return higher rental prices. These zip codes were explicitly identified.
3. Zipcodes in Manhattan and Brooklyn have also seen high property price rises in the last 4-8 years. Staten Island and Queens have been constant for a long time. If the trends were to continue, both the areas of Manhattan and Brooklyn will make a great choice while considering a re-sale option if renting business does not go well as predicted.
4. Price and Availability are not inversely proportional. Consumers are ready to pay prices for better locations and other facilities provided to them.

**TECHNICAL EXPOSITION**

**WRANGLING AND CLEANING PROCESS**

Several Data mining tools and Techniques are available to analyze the huge volume of data and we used **weka** for data cleaning and preprocessing. We took a good look at our data and get an idea of its overall quality, relevance to your project, and consistency. There are a number of data anomalies and inherent problems to look out for in our data set.

**Data cleaning** will correct all of the inconsistent data you uncovered in your data quality assessment. Depending on the kind of data you’re working with it. Weka got it all, it will filter your data from **missing data and noisy data**, which is data that includes unnecessary data points, irrelevant data, and data that are more difficult to group together.

Zip code in listings.cvs had around many null values. We used longitude and latitude values to fill those cells.

Geopy library was used to do this.

Missing values of bedrooms and bathrooms columns were replaced using median values.

We also tried to do this using linear regression.

There were many data sets which we never used/ planned to use so we dropped them using the drop function.

Venues and the calendar have the same cities however the names were different. So we changed the names so that we can fully correlate all the data in both datasets

There were also many columns that were empty/ had too many missing values. So the information gained was very low. So we also decided to get rid of them.

After data cleaning, you may realize you need more data for the task at hand. At this point, we also perform **data wrangling** or data enrichment to add new data sets and run them through quality assessment and cleaning again.

before adding them to your original data. We performed data transformation with **feature selection and normalization** methods on weka and performed in-depth analysis and visualization on the modified dataset from weka.

Zipcode was common between real estate data sets and listings.

So that served as a primary key to join to data sets and extract information.

**INVESTIGATIVE DEPTH**

As we know, usually EDA is applied to investigate the data and summarize the key insights. It provides a brief understanding of the data and its distribution. We started by looking at the distributions of the available variables to look for interesting patterns and/or outliers.

The info() function will give us basic information about the dataset. For any data, it is good to start by knowing its information. The describe() function will give us the number of null values, data types, and memory usage as shown in the above outputs along with descriptive statistics.

First we looked for cities which were in high demand. That was discovered using the Zillow data set.

Then we looked for cities which had high supply. This was found from the airbnb data set.

We selected the city with the highest supply and highest demand.

It turned out that it was New york. So we empathize with properties in New york.

We also analyzed customer reviews of location with the price. Interestingly it was found that the price was high when customers loved the location. Which definitely makes sense.

Since our main goal is also to earn more profit from the investment. We decided to look for properties in the prime location. As customers will be willing to pay more over there.

**Correlation Plot** -to find the correlation among the variables, we can make use of the correlation function. This will give you a fair idea of the correlation strength between different variables. And then we can visualize it by using a heatmap.

Also, when we do an in-depth EDA, we might be able to find **frequent patterns**

associated between the ratings and the frequent amenities. This will help us to find out which types of p**roperties are in high demand** and will give us better insights on the type of property to invest in. Frequent pattern mining can also be done using the column’s bedrooms and bathrooms in the listings dataset. This will give us better insights.

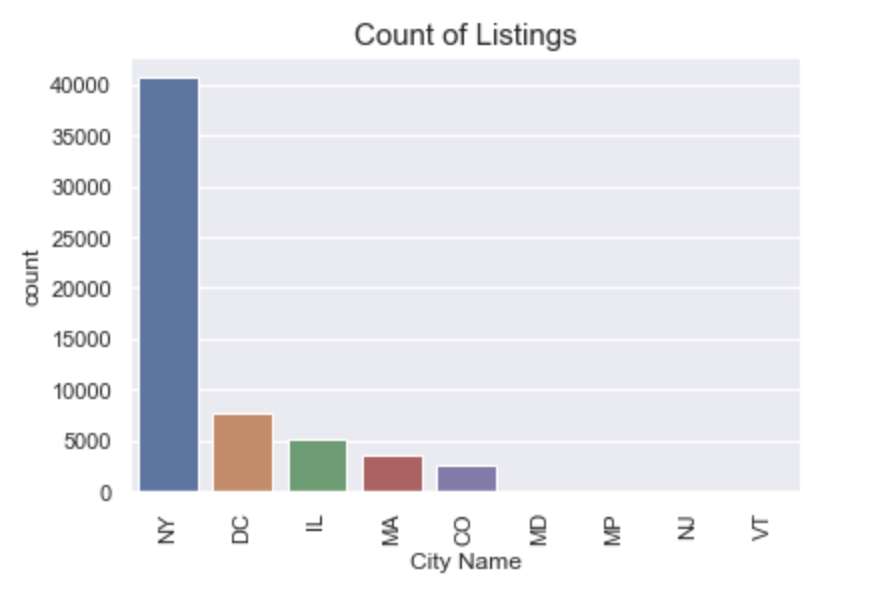
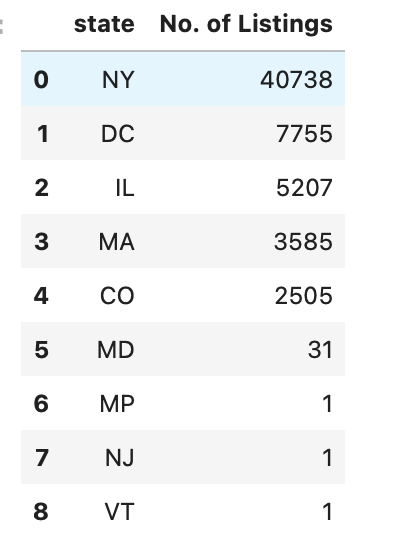
**ANALYTICAL AND MODELING RIGOR**

Let's look at the features we had from the listing Data set and real\_estate Data Set and plot them against price to determine which features, or independent variables should be used to predict price, properties dimensions, and availability. We were looking for a linear relationship that may have to be achieved through transformation.

For that, we had to plot the independent variables versus price to determine which features have the strongest linear relationship with price and which we may have to transform to achieve linearity with a price. The key motivation was to check supply and demand of the units along with their prices in regions, to get better metrics in order to make data-driven, well-researched decisions

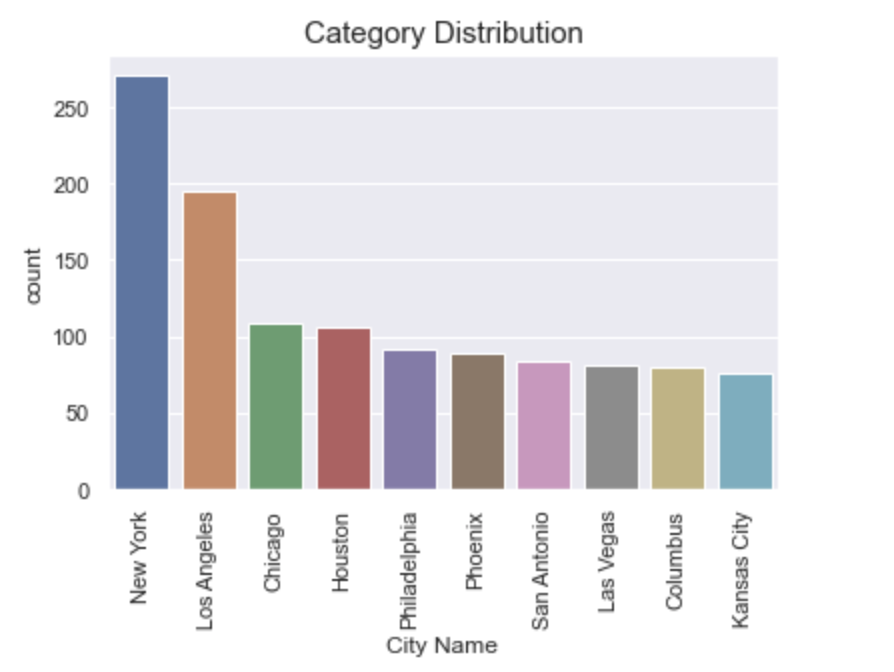
**CONCLUSION**

**TOP STATES IN THE USA FROM THE LISTING.**



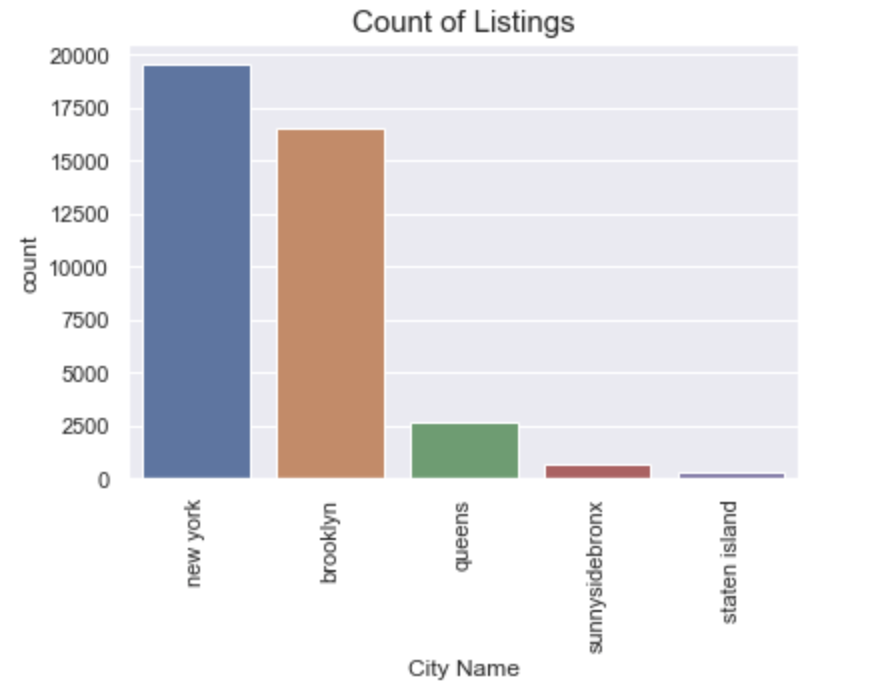
This are the top cities with highest supply

## **Top 10 States for renting properties as per ZHVI Index**

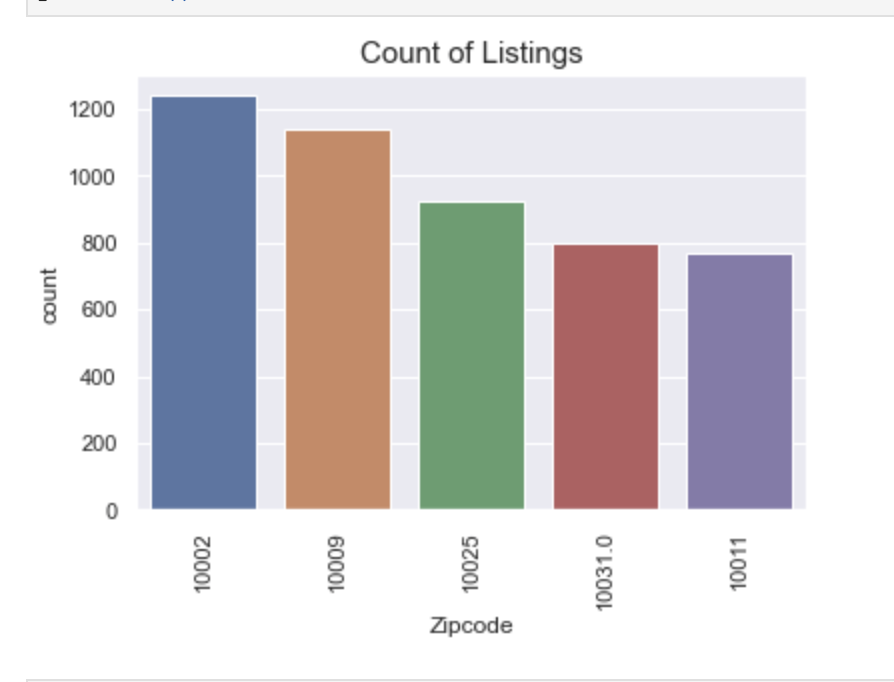


These are the cities with highest demand

After analyzing, the hot market for both markets is New York State. So, our further analysis is on New York state and neighborhood.



**BEST ZIP CODES FOR REAL ESTATE INVESTMENT USING AIRBNB LISTING**





Above are the list of zip codes with counts of property that suggest which type of property to buy under which zipcode.

Interestingly it was found that Best property type was apartments and that was common for all top demanding postal codes.