Predictive Analysis of AirBnB Listings Pricing

Shlok A. Deshpande (TY2 WST2)   MIS number: 112003133

The Problem Statement

To use listings information to analyze and determine what factors decide the listing prices on AirBnB while simultaneously exploring the relationship between socio-economic factors of the location and the Listings data. Building on this information, developing a predictive model for the same.

Requirement Analysis

1.1 Functional Requirements

The Predictive Model developed should be able to identify the amenities that increase the pricing and predict the prices of the new set of listings.

1.2 Technical Requirements

The data and relationships should be realized in a visual form to aid the understanding of the underlying hidden relationships. The visual forms include but are not limited to candle plots, heatmaps, scatter plots and histograms.

Project Objectives

1. Identify amenities and how they affect the pricing.
2. Identify the listings based on the number of rooms and type of the listing.
3. Identifying all airbnb listing factors and basic things that affect pricing so that host can maximize profits
4. Bring to the light, the relationship between Socio-Economic factors of the location and the Pricing of the Listing.
5. Build a predictive model to predict the price of a listing based on the amenities, location etc.

Data Generation

In this project, data generation was accomplished by utilizing a pre-existing dataset from Inside AirBnB website. This approach was necessary to collect the critical attributes required for the project. We also generated data by web scraping to see the recent trends
By leveraging the Inside AirBnB dataset, and the web scraping dataset we were able to obtain a high-quality data source that met the project's requirements and allowed us to perform accurate analysis and modeling.

Data Cleaning

- NaN values
  - Remove the tuples completely
- HTML Tags in Data
  - Selectively remove using regex matching
- Irrelevant Data
  - Do we really need "host_profile_pic"?
  - No important information can be extracted from images and urls

Examples

```
[ ]    1 data['price'] = data['price'].str.replace("$","")
       2 data['price'] = data['price'].str.replace(",","")
       3 data['price'] = pd.to_numeric(data['price'])

    <ipython-input-15-ad1142522408>:1: FutureWarning: The default value of regex will change from True t
      data['price'] = data['price'].str.replace("$","")
```
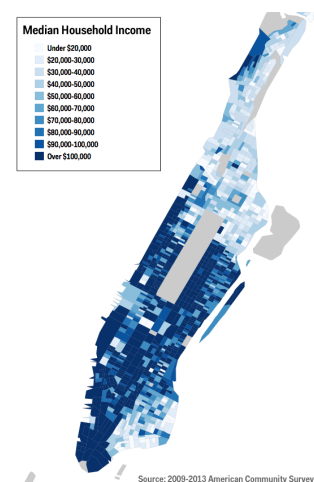
Exploratory Data Analysis

Location Analysis

- How does the Location/Demography affect the Pricings of the Listing?
- Do all the Properties cost the same at different places?
- How do they differ?
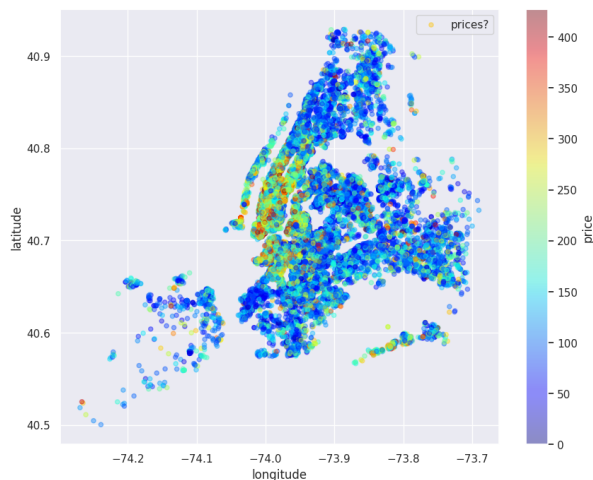
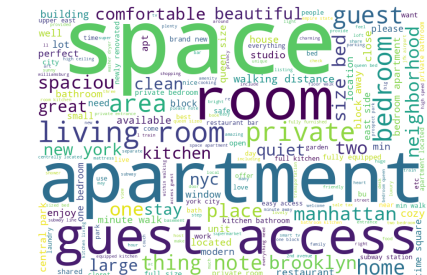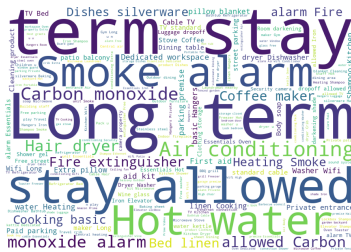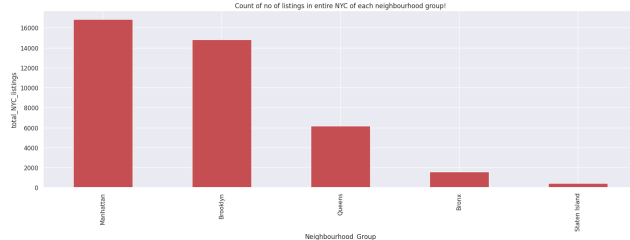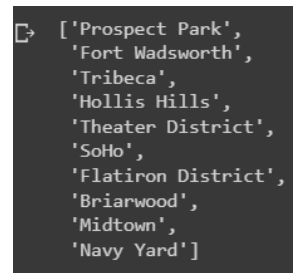Heatmap of Listings Price in New York.

Income Heat Map

Feature Importance   Amenities Analysis   Description Analysis

feature importance for the first 20 features

```
1 # Some data cleaning
2
3 words=''
4 for index,row in bot1000F.iterrows():
5     words += row['description']
6
7 string_punctuation = string.punctuation
8 ignoreChars=['\r','\n',' ',' ','"s']
9 nums=['0','1','2','3','4','5','6','7','8','9']
10 summary_data=nltk.word_tokenize(words)
11 words_only = [l.lower() for l in summary_data if l not in string_punctuation if l not in ignoreChar if l not in nums]
12 filtered_data=[word for word in words_only if word not in stopwords.words('english')]
13 wnl = nltk.WordNetLemmatizer()
14 final_data=[wnl.lemmatize(data) for data in filtered_data]
15 final_words=' '.join(final_data)
16 print(final_words)
```

Most Popular/Examples Neighbourhood Groups

```
['Prospect Park',
 'Fort Wadsworth',
 'Tribeca',
 'Hollis Hills',
 'Theater District',
 'SoHo',
 'Flatiron District',
 'Briarwood',
 'Midtown',
 'Navy Yard']
```

Neighbourhood & Neighbourhood Group Analysis

count of each room types in entire NYC

Count of no of listings in entire NYC of each neighbourhood group!

Price Distribution

Violin Plot of Listings in Each Neighbourhood Group

Neighbourhood Group Average Price

Neighbourhood Group vs. Room Availability

Reviews Per Month & Property Type

Most Reviewed room_types in each Neighbourhood Groups

Availability 365 vs Location
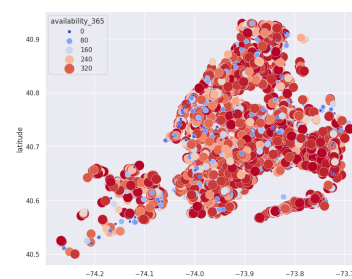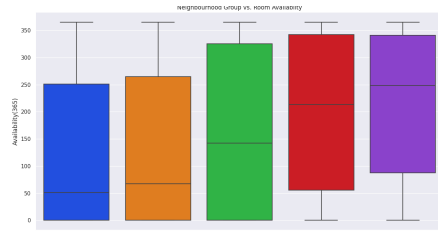
Through the Exploratory Data Analysis and based on our objectives and the relevancy to the normal user, the most relevant features for our Predictive Model are :

- ○ Amenities
- ○ Neighbourhood
- ○ Neighbourhood Group

Necessary Transformations

1. Amenities
   a. Optimisation of the amenities attribute as predicting from a list of strings can get difficult and time consuming.
   b. We first get a list of sorted amenities based on the price.
   c. Then we get all the individual strings and then we use the nltk work_tokenizer and word_lemmatizer to get the proper amenities and disregard outliers.
   d. Then we sort the unique amenities according to the count.
   e. This count is used to calculate the score.
   f. If a listing has a higher number of popular amenities present in expensive listings then its amenity score will increase.
2. Neighborhood -
   a. We assign a unique number to each neighborhood as it is a categorical variable. The neighborhoods are also ordered based on the price to ensure accuracy.
   b. We also separate the numbering per group so that the analysis can use Clustering to obtain a more accurate price.
   c. The neighborhood also follows a Regression model with the price.
3. Neighbourhood Group -
   a. We also assign a no. from 1-5 to each neighborhood group.
4. All these optimisations & transformations allow us to gather a more accurate output and generate a better model.

Data Modeling

Regression

- A Regression model can be used to predict the price of listings in New York based on the amenities, neighborhood, and the neighborhood group.
- The model would take as input a set of amenities like alarms, carbon monoxide etc. It would also include categorical variables like the neighborhood and neighborhood group to capture the location-based factors that influence housing prices.
- The selection of a Regression model is appropriate in this case because it is a simple, interpretable, and widely-used technique for predicting continuous variables.
- The model can be easily extended to include additional features, and it allows for the interpretation of the contribution of each feature to the predicted price.
- The findings of the model can help to identify which features are most important in determining the price of a listing in New York.
- For example, the model might reveal that the neighborhood group is the most significant factor in determining the price of a listing, while amenities like alarms, TV and wifi have a smaller impact.

Web Application of the Data Model

Conclusion

- The project aimed to analyze and determine the factors that influence the listing prices on Airbnb, while also exploring the relationship between socio-economic factors of the location and the listings data.
- The objectives of the project were to identify the amenities that affect pricing, categorize the listings based on the number of rooms and the type of the listing, bring to light the relationship between socio-economic factors and pricing of the listing, and build a predictive model to predict the price of a listing based on amenities and location.
- Through the data analysis process, we have identified the key factors that impact the pricing of Airbnb listings, including amenities such as Wi-Fi, air conditioning, and parking locations (neighborhoods). We also explored the relationship between socio-economic factors such as the neighborhood group, median household income etc.
- We also compared location with many more factors including room type, availability.

We also saw the relationship with prices and the ratings.

- We pinpointed the important amenities and expensive neighborhoods and the relation between the neighborhood groups and factors and much more.
- Based on these findings, we have developed a predictive model that can be used to predict the price of a listing based on the amenities, location, and other factors.
- This model can be used to help Airbnb hosts to set prices for their listings, and also for investors and real estate professionals to make more informed decisions about which properties to invest in or how to price a listing.
- In conclusion, the project has successfully achieved its objectives and provided valuable insights into the factors that influence Airbnb listing prices.
- The developed predictive model can be used to improve pricing strategies for hosts and enhance decision-making for investors and real estate professionals in the Airbnb market.