**Airbnb Popularity Analysis**

*User Instruction*

By **0% Risk Team**

**Zhexin Hu** (zhexinh), **Yijing Lin** (yijingli), **Lingqin Yang** (ylingqin), **Chengyu Wu** (chengyu2)

**Content**

[Abstract 1](#_Toc58582106)

[Data Collection 1](#_Toc58582107)

[Nearby Analysis 2](#_Toc58582108)

[Listing cleaning 2](#_Toc58582109)

[Bayesian average rating 3](#_Toc58582110)

[Analysis on price, superhost, amenities, and response rate 3](#_Toc58582111)

[Crime Cleaning and Merging 3](#_Toc58582112)

[Analysis on crime, roomtype, and roomno 3](#_Toc58582113)

# Abstract

The objective of this project is to identify characteristics that contribute to the popularity of a listing on Airbnb in New York City, eventually to help Airbnb listing owners efficiently improve their listing popularity. Specifically, we will explore what aspects of listings influence popularity most and construct a predictive model to capture potential popularity of an input listing. The listing popularity will be measured by ratings and number of reviews.

# Data Collection

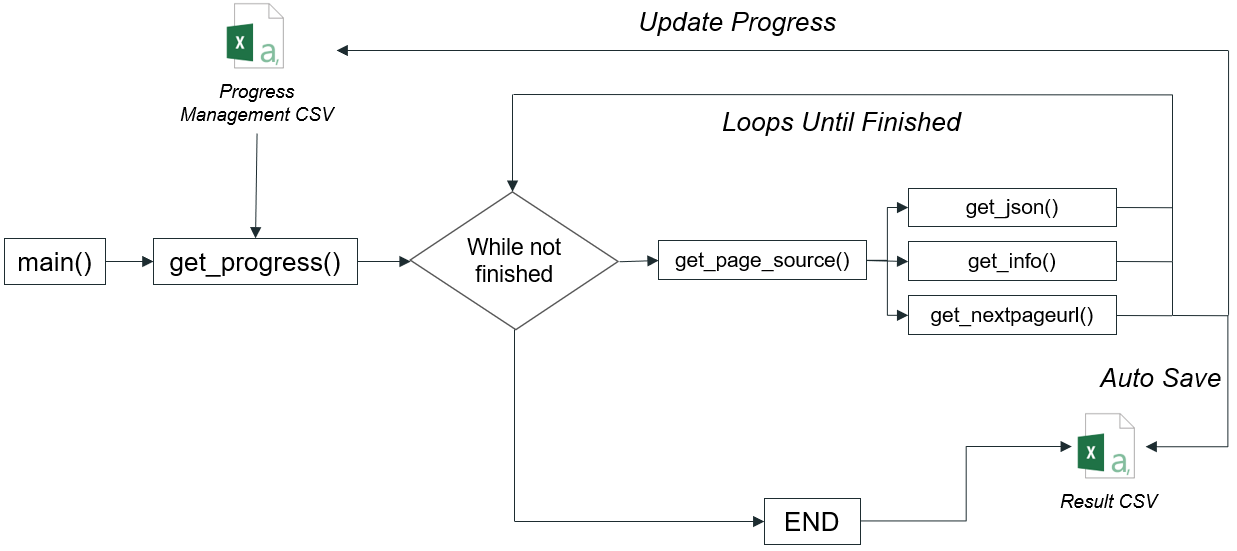
* **Data Sources:** Our data sources include the following three main types: web scraping from Airbnb.com (Search Page), Google Maps API (Nearby Search), and online CSV files (NYC Open Data for Crime Rate data and Insideairbnb.com for response rate data). Variable-level list is shown below.

|  |  |  |
| --- | --- | --- |
| **Name of the Source** | **Type** | **Variable(s)** |
| Airbnb.com (Search Page) | Web Scraping | id, name, city, lat, lng, avgRating, reviewsCount, bathrooms, bathroomLabel, bedrooms, bedroomLabel, beds, guestLabel, personCapacity, roomType, isNewListing, isSuperhost, previewAmenityNames, priceString, rateType. |
| Insideairbnb.com | CSV | host\_response\_time, host\_response\_rate, host\_acceptance\_rate. |
| NYC OpenData | CSV | Crime rate of each borough. |
| Google Maps API (Nearby Search) | API | Nearby subway and nearby tourist attractions. |

* **Running Dependency:**
  + The module *searchpagefetch.py* relies on the progress management CSV file to be placed with in the same directory. Default file name is “neighborhood\_id.csv”.
  + Each search page normally has information of 20 listings, information in detail page is NOT included here due to limited time and resources.
  + If users want to customize a start point, “Done” marks should be labelled in the second column of all the previous rows (like the example below, the function would start from neighbourhood “Baychester”).



* Structure of the web scraper:
  + A brief graph can be used to understand the logic of the web scraper in searchpagefetch.py.



# Nearby Analysis

* Nearby Search: Register with Google Cloud Platform and get an API for requests. User could change the location, category, search radius and page token when calling functions. JSON response is received. A list will record whether nearby has such kind of places or not.
* Nearby Analysis: Set the variables that user wants to test the relationship. A boxplot will be plotted with corresponding y-variable and x-variable chosen.
* Hotmap plot: Set the variable that user wants to see the distribution on NYC map. A map with distribution of variables (may have categories/quantiles) in NYC will be shown.

# Listing cleaning

* To calculate Bayesian average rating, merge host response information and convert amenity into four separate columns

# Bayesian average rating

* Create the histogram plot of avgRating and bysAvgRating (for presentation purpose)

# Analysis on price, superhost, amenities, and response rate

* Output includes:
* boxplot of number of amenities to Bayesian average rating
* boxplot of number of amenities to Bayesian average rating under superhost or not
* boxplot of superhost
* scatter plot of response rate to Bayesian average rating
* correlation heatmap of attributes

# Crime Cleaning and Merging

* Clean the ‘borough’ column in the cleaned listing data frame to merge with the crime data frame
* Create a new column for population and compute the crime rate for each borough
* Select desired columns and return a final dataset

# Analysis on crime, roomtype, and roomno

* Output includes:
* boxplot of crime rates to Bayesian average rating by borough
* bar plot of the distribution of number of listings by borough
* bar plot of the distribution of the Bayesian average rating by borough
* bar plot of price by borough, boxplot of roomtype to Bayesian average rating
* boxplot of roomtype to Bayesian average rating by borough
* boxplot of number of bedroom to Bayesian average rating
* scatter plot of number of bedroom to Bayesian average rating
* boxplot of number of bathroom to Bayesian average rating
* scatter plot of number of bathroom to Bayesian average rating