ETL Project

By: Adesuwa , Kyndall, Krishna

**Pre-Processing**

The use of Airbnb’s has been extremely popular over the past couple of years. With its simple booking features, and most times being cheaper than hotels, most people prefer to use Airbnb for their vacation stay. With a surge in demand for Airbnb’s, people have been buying properties and applying to Airbnb to become hosts in very large numbers. To assist Airbnb to accept new homes into their database, our team has been hired to analyze what makes Airbnb popular in certain markets and not so popular in other markets.

**Extraction**

We used 2 different datasets from the public platform Kaggle which lead us to the the data in the three files included the following information:

* Number of Airbnb’s in Each Borough
* Average Price of Airbnb in Each Borough
* Average Number of Reviews Left Behind in Each Borough
* Population Demographics

The fields of interest include the following:

* Room Type
* Borough
* Minimum Nights
* Reviews Per Month
* Total Population
* Race
* Age
* Population Change

The following sources for our datasets used:

1. <https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data>
2. <https://drive.google.com/file/d/111VC5lD_0_AGJHzLNlsnYY9upyt0JQAh/view?usp=sharing>

**Transformation**

In order to transform the public data and use it in our study we performed the following:

* Used Pandas functions in Jupyter Notebook to load two CSV files.
* Reviewed the files and transformed into two separate data frames
* In the first dataset, we removed excessive variables that we considered not relevant for our data set. We simplified the dataset to only include Neighborhood Group, Room Type, Price, Minimum Nights, and Reviews Per Month.
* The second dataset was for the population census. Since we didn’t need most of the variables, we removed all the extra variables and only included; Age, Race, and Total Population.
* We made sure our datasets did not include any duplicates or other unnecessary variables.
* Created queries to address our hypothesis by grouping the data by each borough in New York. We compared the number of Airbnb to the total population so we could see where the most popular Airbnb’s were located.

**Table

Description automatically generatedAirbnb Data**

**Population Data**

Table

Description automatically generated

**Load**

After we pulled in the CSV files and loaded them into the data frames, we did an initial connection to the Postgres database using PG admin to store our original clean data sets. We used the quick database website to create the initial table schema that got loaded into the Postgres database that generated the first set of tables. After running the queries and created the new tables with only the relevant information we reconnected to the database and generated additional tables for the data frames.

**Postgres Database:**

**Summary**

There were some limitations in our findings due to the available data on Airbnb’s. However, we were able to address our hypothesis question in our initial project proposal below:

H0: There is a relationship between neighborhood group, room type, an area’s age distribution, and AirBnB bookings in that area.

We investigated Airbnb bookings in only one city, New York, and broke down New York by its five boroughs to differentiate Airbnb bookings within the city. We specifically did this to help Airbnb decide which areas they need to accept more hosts compared to other areas where they did not need to accept as many hosts. In conclusion, we realized that Airbnb popularity heavily depends on the popularity of an area. Thus, we recommend the following:

More Airbnb’s in the following areas would increase locations and money for the company.

1. Heavily Populated Areas
2. Areas That Have Younger People
3. More Affluent Areas
4. Areas That Are More Diverse In Population