UCSD Data Science Bootcamp Project 2 Report, 2/24/20

AirBnB Rental Rates vs Local Income

Team Members:

* Grant Thompson
* Arundhati Chakraborty
* Alexis Perumal

Project Objective

* Groom a dataset that a data analyst can use to evaluate AirBnB rental rates vs. local income and other census characteristics.

**Extraction -**

We pulled data from two data sets:

* [AirBnB rental rates for New York City, includes Lat/Long](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data%20kaggle.com) (Kaggle)
* [Census Income Data by census tract](https://www.kaggle.com/muonneutrino/new-york-city-census-data) (converted to census tract)

The AirBnb dataset describes the listing activity and metrics in NYC, NY. It includes county and township of each listing, the listing price, and latitude and longitude data for the listing.

The Census Data for New York includes information down to the Census Tract level, which includes approximately 8,000 people. It has information on income, income per capita, race, and employment statistics.

Both datasets are formatted in CSV. We used Pandas to extract and explore the data.

**Transformation -**

We wanted to join the two tables together with this schema.

**Intended Schema**

A screenshot of a cell phone

Description automatically generated

However, since the two tables didn’t have matching data for a join, we had to convert the latitude/longitudes of the table 1 AirBnb listings to the appropriate Census Tract codes to match table 2. To do so, we used the geo.fcc.gov API to convert Lat/Long to Census Tract. The API returns Census Block data, which is more granular than Census Tracts (~8,000 people compared to ~1,500 people). To fix this, we simply dropped the last four digits of the Census Block to get the Census Tract.

The API queries took about 5 hours to run to do 48,895 calls.

**API Code and first few responses:**

A screenshot of a social media post

Description automatically generated

This result included 4 low order decimal digits we didn’t need, so we dropped those. That then allowed us to create a modified version of table 1 with the census track data added to each row.

This can be seen in the [schema](https://github.com/alexisperumal/ga2-etl-project/blob/master/schema/QuickDBD-NYC_ABNB_Census.pdf) doc.

**Actual Schema: DB, Table 1 airbnb\_data, Table 2 census\_data**

**A close up of text on a white background

Description automatically generated**

**Resulting Joined Table:**

**A close up of text on a white background

Description automatically generated**

**Load**

We created a SQL database called airbnb\_db with two tables: airbnb\_data and census\_data. The schema for the database and tables is above. We used SQL because both data tables were very structured, and the result was conducive to structured queries of the joined data. To run an analysis, we simply perform a join on the Census Tract column. This allows us to run queries to find, for example, if correlations exist between the average cost of an Airbnb and the average income level for the same area, or an aggregation of average listing prices by census track.

**Example Queries**

A screenshot of text

Description automatically generated