Sagari Datta

Data Wrangling, MUSA620

Professor Nick Hand

May 13, 2019

**Deliverable 2: Document describing the project**

**Project Name**: NYC Airbnb Data Analysis

**Project Description**: The final project is an attempt to tell the story about how Airbnb is doing in NYC. I downloaded datasets from different sources for this purpose to determine which neighborhoods in the city have the greatest concentration of Airbnb listings and to find out the demographic and real-estate characteristics of these neighborhoods. Furthermore, I performed a sentiment analysis to determine how Airbnb is perceived in NYC given some of the recent controversies that have surrounded Airbnb relating to the issue of affordable housing.

**Datasets used**:

*Structured Data*

1. Kaggle.com – csv files downloaded
2. Open Data NYC – API query using Sodapy

*Unstructured Data*

1. Twitter – Scrapping twitter data by

Note: Census Data not used but code provided.

**Methods/Steps:**

I was supposed to use to the census API data but due to the Census API server being down, I decided to use data available on Open Data NYC website to perform the analysis

1. Download data from different sources (csv, API)
2. Convert relevant data frames into geopandas data frames
3. Spatial join and aggregate data as needed
   1. At the neighborhood level: Neighborhood population, Airbnb listings and property values were aggregated to each neighborhood and the different joins (beside spatial joins) were based on neighborhood name.
   2. At the community district level: The demographic and housing information were used as proxy variables for census variables relating to race, income, rent, median income, etc. The demographic and evictions data had no geographic information other than zipcode. The housing information had latitude and longitude values, but these were in an incorrect format. Therefore, the aggregation and joins had to be all made based on zipcode i.e., not spatial join but a pd.merge based on the zipcode column. These were then further aggregated at the community/council district level since the housing data had community district information. Furthermore, since the community district numbering formats were not consistent for between the housing information dataset and the community/council district shapefile, these were rectified as well for the purpose of performing the joins.
4. Exploratory data analysis and visualizations – maps, plots, charts, etc.
5. Twitter data analysis – Airbnb in NYC and Airbnb in general
   1. Scrap for tweets for relevant search words
   2. Clean twitter data and perform word frequency analysis (plot and word cloud)
   3. Conduct sentiment analysis based on ‘polarity’ and ‘subjectivity’ benchmarks (plots)
6. Dashboard – Dashboard created using Dash. Folium map and plotly charts used for the dashboard. The map display changes according to the borough and variables selected. The scatterplot showing Airbnb list price as a function of property market value changes depending upon which borough is selected. The display of the bar plot showing the mean of a variable selected each of the 5 boroughs changes as and when the variables selected changes.

**Results:**

Structured data: Most of the listings are located in Midtown Manhattan and in the area between Queens and Brooklyn near the East River, with the highest concentration in the Brooklyn neighborhood of North Side-South Side. However, the spatial distribution of Airbnb prices does not seem to have much in common with the spatial distribution of Airbnb listings. Furthermore, the property market values only have some spatial similarity with Airbnb listing prices. Not surprisingly, Airbnb listings in the outer boroughs are not very popular with tourists. The data from NYC Open Data that was used to map white and non-white population across the city of New York seemed a bit unreliable. Staten Island which is predominantly white showed a small count of white population thus raising some concerns about the reliability of the demographics data available obtained through NYC Open Data. The percentage of low-income housing (as a proportion of total housing units) was mainly situated in the outer boroughs particularly Brooklyn and was not close to any of the Airbnb listing areas.

Unstructured data/Twitter: There were only 13 tweets about Airbnb NYC (you can only scrap 7 days data at a time and this was the total number of tweets in the past week). Given the limited data, I decided to perform an additional twitter data analysis on Airbnb (in general not just focused on NYC) and to compare the results of the two analyses. The results show that most people feel positively about Airbnb in NYC more so than Airbnb in general. However, the result should be considered with caution given the limited amount of twitter data on Airbnb NYC I had to work with for this project.

Note: More details about the code and the steps performed can be found in Jupyter Notebook

**Guidelines met**

|  |  |
| --- | --- |
| **Requirement** | **Met (Y/N)** |
| Data is collected through a means more sophisticated than downloading (e.g. scraping, API) | Yes |
| At least one of the datasets contains more than 1,000,000 rows | No |
| It combines data collected from 3 or more different sources | Yes |
| The analysis of the data is reasonably complex, involving multiple steps (geospatial joins/operations, data shaping, data frame operations, etc.) | Yes |
| You use one of the analysis techniques for urban street networks (e.g., osmnx, pandana) or clustering (e.g., scikit-learn) | No |
| The webpage includes a significant interactive component (cross-filtering, interactive widgets, etc.) | Yes |