LOCATION SELECTION FOR A CAR DEALERSHIP BUSINESS IN ATLANTA, GEORGIA, USA

Donah Simiyu,

25th June 2020

1. Data acquisition and cleaning

1.1. Data sources and contents

The following datasets were obtained from the given sources:

i. Names of neighborhoods in Atlanta: The source of data for neighborhoods in Atlanta was the ARC's Open Data and Mapping Hub [2]. The data file was obtained in .xlsx format and contained data on the Neighborhood Planning Unit (NPU), names of neighborhoods and the geometry coordinates of each neighborhood. This source did not have the coordinates of each neighborhood.

ii. Coordinates of neighborhoods in Atlanta: To find coordinates of each neighborhood, Nominatim API and Googlemaps API was used. The latitude and longitude of neighborhood was returned.

iii. Places in the neighborhoods of Atlanta: After getting the neighborhoods in Atlanta, **FOURSQUARE LOCATION API** was used to explore the neighborhoods to find out the places, businesses and social activities going on. The data was put in JSON (JavaScript Object Notation) format to allow for cleaning.

iv. Crime data of neighborhoods in Atlanta: Crime data was sourced from Data World, a cloud-native data catalog [3]. The data was in .csv (comma separated values) format and showed the type of crime, date it occurred, location, latitude, longitude and the neighborhood.

v. To visualize various aspects in Atlanta, a GeoJSON (geospatial data interchange format based on JSON) file that defines the boundaries of each neighborhood was sourced from the Fulton County government GIS Portal [4].

1.2. Data cleaning

i. The data on neighborhoods had many columns including the population, NPU, statistical area, population, neighborhood, URL, area, races (white, black, Asian, other, Hispanic), Global ID, and last edited date. Data type for each column was checked. In the neighborhood column, some neighborhoods were grouped together in one row since they belong to the same NPU.

These were split so that each row had one neighborhood. The spellings of column names were checked and modified.

ii. Data from from Nominatim API and Googlemaps API did not need any cleaning. However, before fetching the coordinates from the API's, the words 'Atlanta, Georgia' were appended to each neighborhood to give the correct address. This is because a neighborhood like 'Downtown' in Atlanta can also be found in other areas such as Los Angeles.

iii. Data from Foursquare API was obtained in JSON format. The data was converted to a Python pandas dataframe. Category names were extracted from the categories column and a separate column created to hold this data.

iv. Crime data in the csv file was loaded as a pandas dataframe. It has columns including crime, number, date, location, beat, neighborhood, NPU, latitude and longitude. Columns 'number' and 'beat' were not necessary in this project and were therefore deleted. Spellings of column names were corrected.

v. The GeoJSON file containing boundary coordinates had characters such as braces and commas that were preventing it from loading. Using Visual Studio Code, all the unnecessary characters were removed.

1.3. Feature selection

After cleaning, examination of all the datasets revealed some redundancies. Feature selection was done as follows.

i. Examination of dataframe on neighborhoods showed that the population, NPU, statistical area, population, neighborhood, URL, area, races (white, black, Asian, other, Hispanic, Global ID, and last edited date columns contained data that is important but not necessary in this project. The necessary features selected were the NPU and neighborhoods. The rest of the columns were therefore dropped.

ii. The NPU, Neighborhood from part i. above were concatenated with Latitude and Longitude data; these 4 features were then selected for use this project. This dataset was used to display each neighborhood on a map.

iii. From Foursquare output, the name, category, latitude, longitude and address were selected for use in this project to analyze the locations and provide information that will aid potential investors in selection of a location.

iv. From the crime data, crime number and beat were found to be redundant for this project and therefore dropped. The crime data was listed from the year 2010 to 2017. Data from 2010 to 2014 was found to be a little outdated and was therefore dropped. For the remaining dates, the features selected were crime, location (street address), neighborhood, NPU, latitude and longitude.

v. From the cleaned GeoJSON file, the necessary variable was the name of the neighborhood. This was selected together with the corresponding boundary coordinates and used in data visualization.

2. References

[1] Consumer Expenditures for the Atlanta Metropolitan Area: 2017–18, U.S. Bureau of Labor Statistics

[2] City of Atlanta Neighborhood Statistical Areas, retrieved on 24th June 2020, from https://opendata.atlantaregional.com/datasets/d6298dee8938464294d3f49d473bcf15_196

[3] Crime in Atlanta 2009-2017, dataset by Alexander Bryant, retrieved on 24th June 2020, from https://data.world/bryantahb/crime-in-atlanta-2009-2017

[4] Atlanta City Limits, retrieved on 24th June 2020, from https://gisdata.fultoncountyga.gov/datasets/b711dbcf5a1d4f479e275b1781c4bda0_1?geometry=-91.846%2C32.119%2C-77.025%2C35.316