new Computer Science book stores in LA Neighborhoods using Clustering

Capstone: **Week 5** Raymond Samuel: rsam411@yahoo.com

.

Introduction

Description of the Problem and Context:

A group of investors started a small Data Science book store in Alabama. One of the decisions made during last business strategy company meeting was to confirm a plan to establish and deploy new stores on the west coast via launching stores in Los Angeles. As a consequence, a first pilot project is launched, and Raymond, project leader was nominated with first objective to launch 10 new stores in Los Angeles, CA. The success of this initiative is crucial for the next phases of the project and for the development of additional stores.

Among Raymond's key tasks, he needs to work with a real estate consultant to identify available venues in Los Angeles and close relevant deals as soon as possible to deploy the stores. After a first discussion with Raymond's preferred real estate broker in LA, he realized that the criteria he defined for finding the stores are not accurate enough and his broker shared additional information. Here are a few problems shared:

A) He may spend too much time finding the right places in such a large city like LA, so he needs to identify some preferred zones within Los Angeles to focus his search.

B) He may not be the best person to define priorities and most relevant areas for the Data Science store criteria, so he needs more views with targeted areas.

As a consequence, Raymond contacted me and asked for some help on where I would recommend they should open the first stores in LA. Raymond and I discussed and we came to a conclusion that the problem could be solved with defining a list of preferred areas in Los Angeles issued from classifying neighborhoods based on exploring existing venues and most frequent categories of venues in each candidate zone. This way, we can identify similar neighborhoods, gather them within several clusters and choose the right cluster of areas within Los Angeles to focus on. Such output will serve as a view for Raymond and a list of target zones for Raymond's real estate broker.

Description of the data that will be used

As Los Angeles is a large city, we have many neighborhoods to explore within the city itself (almost 100 zip codes). During the study, we will exchange on a regular basis together with Raymond on first data extracted and first explorations in order to decide if we need to extend to Los Angeles county or keep the research within Los Angeles city boundaries. Los Angeles county will be a larger data of more than 400 zip codes. Online search for location data, let us find a database gathering all zip codes in Los Angeles, CA and each corresponding latitude, longitude coordinates. This database use is free to download and use; the only request is to clearly mention the link it is coming from https://simplemaps.com/data/us-zips so we will mention on each notebook or page or presentation related to this assignment. At this stage, the main info we plan to extract from this database are zip code, latitude, longitude, city, county, state. We also have access to the population

and density for each zip code. We will extract such information as potential additional data to be used during the study, if needed. Such parameter like population and density are valuable info as they may come up in some discussions with the client during the project. Knowing each zip code latitude and longitude, we can then explore categories of venues thanks to Foursquare API. We use the Foursquare location data to explore neighborhoods of LA, specifically categories of venues, in a similar way we did with Toronto area in the previous lab and assignment. We plan to use unsupervised machine learning method for classification, like k-means algorithm. This topic will be developed further in the next Methodology section of this report.

Some additional kind of data may be relevant to add for a deeper exploration: the foot traffic data. Such data is available if we build a process to get it for each location and we can even explore the foot traffic depending on the time in the day and the day in the week. Adding such data would mean defining the right method of usage of Foursquare API and would mean more time and effort, whereas the method we are using has been proven with previous examples like the one in Toronto during the lab. Knowing this, Raymond asked me to focus for now on the current method without foot traffic data and we can extend our study further if needed. Another reason why we did not add such data is that the consumer habits in Alabama and in California are fairly different. As an example, the stores operate in Alabama are in zones with very high foot traffic. But this does not mean Raymond wants to follow a location strategy in LA necessarily similar to the location strategy used in Alabama. So, we focus for now on unsupervised clustering method based on venues around candidate locations, and within each zip code in Los Angeles, CA. Additional tools for solving the problem: In addition to Foursquare API we already mentioned, we will use Jupyter notebook for all the coding and explanations of our method, process and computations.

Coding will be done in python 3, and leveraging usual libraries: NumPy and SciPy for scientific computing, Pandas for data extracting, cleaning and analysis, Matplotlib and Folium for figures, plots, maps and visualization, Scikit-learn for machine learning, in particular we plan to use clustering k-means algorithm In this report, we mainly present explanations, notes and comments as described above, but we don't insert actual code. Another note book is provided as part of the assignment and gathers all the python code, dataframes, plots and maps involved in the capstone project.

Summary of Intro and Data section

The objective is to build clusters to partitions Los Angeles, CA in similar areas and identify the most suitable areas for launching new stores for. We will leverage unsupervised method like k-means clustering algorithm to classify all Los Angeles, CA zip codes. For that, we will clean and leverage Foursquare location data - specifically to explore categories of venues in the neighborhood - as well as Los Angeles zip code location database. Once clusters are created, we will review the clusters and identify similarities within a given cluster and similarities between two different clusters. Then the target cluster of areas (zip codes) will be defined and validated with the client.

Methodology

Review our sources of data Let's review our data source and see what data preparation process is needed. As said above, we have two main sources of data: a) one database gathering all zip codes in Los Angeles, CA and each corresponding latitude, longitude coordinates. This database is coming from https://simplemaps.com/data/us-zips b) another database we'll build thanks to Foursquare API to get neighborhood information for each (latitude, longitude) point that will be considered, and coming from the first source. Foursquare requests will help gather venues information and categories of venues in each considered area. In particular, computing the n most frequent categories of venues in a predefined radius will help build main features used by our clustering algorithm, and for achieving

our classification objective of all considered areas (zip codes). The first source of data (from simplemaps) is easily downloadable as a .csv file and transformed in a pandas dataframe.