## The Battle of Neighborhoods in LA

**A. Introduction:**

**A.1. Background**

Los Angeles, officially the City of Los Angeles and often known by its initials L.A., is the most populous city in California; the second most populous city in the United States, after New York City. With an estimated population of nearly four million people, Los Angeles is the cultural, financial, and commercial center of Southern California. The city is known for its Mediterranean climate, ethnic diversity, Hollywood, the entertainment industry, and its sprawling metropolis. It’s a city full of restaurants, bar and many more, one of the most demanded city to live in. Huge convenience to shops and restaurants does come with a high staying cost, especially when it comes to house price.

Being in such a great city with huge business potentials, as a property developers, you will expect to look for a more preferred districts where current real estate cost are lower but with huge potential. As a business owner, you will like to know any potential business opportunity such as cafes, are huge demand for residents in certain borough. If we think about city residents, they will like to choose to live in a region where real estate is of good values i.e. closed to shops and within their purchase capabilities. However, there is no direct route to obtain these insights.

**A.2. Problem**

To consider these problems, since there are lots of neighborhoods in LA, we will try to detect locations have venues that meet one individual’s specific needs, e.g. someone may prefer to staying in place with bar, cafe or shop center, and it really depends on the locations and one’s affordability.

We will use data analysis capabilities and machine learning methods to examine the situation and potentially generate some suggestion for you to decide on which location could be best for your needs and budgets. As for business developers, this report can provide some insights to decide which neighborhood could have potentials for business.

**B. Data Description**

**B.1. Data Sources**

To consider the problem, we have listed the data as below:

* We will need to collect data on Rent Price that have been recorded in the past, open data website Neighborhood Data for Social Change[1] should be the best place to answer this question as median rent prices of all the neighborhoods in LA are available there.
* We will need to find out all neighborhood names in LA as well as their geographical coordinates. Luckily, all of this information are also provided by Neighborhood Data for Social Change. We only need to seperate the coordinates into latitudes and longitudes.
* In order to accurately locate each neighborhood, I will need geolocation coordinates for each neighborhood. Google MapsAPI[2] geocoding is the most efficient tools to do the job.
* Lastly, we will need to obtain data for the venues of a given neighborhood or the number of restaurants nearby, Foursquare API[3] is well-known to provide those information even for big companies such as Apple and Google.

**B.2. Data Cleaning**

Data downloaded or scraped from multiple sources were combined into one table. There were lots of unwanted or missing values, I decided to drop some of the columns and replace some of the missing values. Some of the data are not in the correct type, e.g. float to integer or string type to float. Generally, the data obtained are in good shape.

**B.3. Feature Selection**

After data cleaning, there are 3,411 samples and 325 features in the data. Upon examining the meaning of each feature, it is clear that there should be some redundancy in the features as I will only need the top 10 common venues. Grouping and create methods for calculation of frequency are required as clustering later will only takes float values.

**C. Methodology**

**C.1. Exploratory Data Analysis**

I have created a horizontal bar chart to better display median rent price of 30 neighborhoods to get an understanding of the data. And before creating the chart, I managed to sort all the data based on price, so I will have good knowledge of the 30 neighborhoods with the top rent price. This can be useful focus points when it comes to venues data for these neighborhoods.

A picture containing screenshot

Description automatically generated

To get a visual understanding of the boroughs in London, I used python folium library to create a map of LA with geographical details for each neighborhood, superimposed on top on the map and indicated with green circle markers. I used latitude and longitude coordinate values that I obtain using Google Map APIs to get the visual display as below:

A picture containing text, map

Description automatically generated

I utilized the Foursquare API to explore the neighborhoods and extract those data. I set the limit as 100 venues and a radius of 500 meters for each neighborhood from their latitude and longitude values. Here is a snapshot of the list with Venue name, Venue Category, its latitude and longitude coordinates.

A screenshot of a cell phone

Description automatically generated

Some neighborhoods such as the Downtown and the Beverly Grove return the maximum limit of 100 venue data. On the other hand, some neighborhoods like the Del Rey and the Mar Vista each only return less than 30 venue data with the given coordinates values and radius point of 500 meters. Some even has no venues around within this distance. This may be partly due to the fact that venues in these neighborhoods are located further apart and shops are not as situated as those closed to LA downtown area.

A screenshot of a cell phone

Description automatically generated

In total, 3,411 rows of data are collected from Foursquare with 325 unique categories. I then created a table which shows list of top 10 venue category for each neighborhood in table below:

A screenshot of a computer

Description automatically generated

We are going to categorize our venue data with house price data into separate clusters. As a result, I am going to use unsupervised machine learning K-means model to do the clustering. There are many models for clustering, K-means is one of the most vastly used clustering models in data science and machine learning fields. Some real-world applications are customer segmentation and pattern recognition.

First I will run K-means to separate the neighborhoods into 4 clusters because when I analyze the K-Means with Elbow Method, it indicates that 4 degree is the optimum value for k in our K-Means function, where the line bents extensively when x = 4 as show on the graph below:

A close up of a map

Description automatically generated

Here is my merged table with cluster labels for each neighborhood:

A screenshot of a cell phone

Description automatically generated

**D. Results**

Our analysis shows that out of 259 neighborhoods in LA, only 228 of them have venues with distance of 500m and these 228 neighborhoods can be divided into 4 distinct clusters. To visualize this result, I have created another folium map with colors to distinct each cluster group for every neighborhood and superimposed them on the map for data visualization.

A picture containing text, map

Description automatically generated

**Cluster 0**

Based on observation, Cluster 0 refers to the most luxurious neighborhoods in LA that have lots of venues and facilities around, where house prices are the most expensive as compare to other clusters. It's a area for the rich, and it is very convenient to live with high cost.

A screenshot of a cell phone

Description automatically generated

**Cluster 1**

Based on observation, Cluster 1 refers to neighborhoods that have lots of venues and facilities around, rent prices are very expensive as compare to other clusters. They are mainly neighborhoods in central, very convenient place to stay with cafe, cinema, supermarket nearby.

A screenshot of a cell phone

Description automatically generated

**Cluster 2**

Based on observation, Cluster 2 refers to boroughs that have quite a few venues and facilities around neighborhoods, rent prices are generally cheaper among other cluster. It is not very close to central LA and the rent prices are more acceptable for living.

A screenshot of a cell phone

Description automatically generated

**Cluster 3**

Based on observation, Cluster 3 refers to neighborhood that have only a few venues and facilities around, rent prices are generally the cheapest among other cluster. It is also the most further to the central, more likely outskirts of LA.

A screenshot of a cell phone

Description automatically generated

**E. Discussion**

As each neighborhood has its own set of data, different classification approaches in segmenting and clustering can yield different results. I used the K-means model as part of this clustering research. The model provide very accurate clustering and extremely close to real-life situation in which how LA are divided into Cluster 0, 1, 2, 3 where those neighborhoods are belonging to. More classification models should be applied to further examine the search to see any difference in results. To further improve the model, not only venue and rent price data are used, but other datasets such as population density or demographic information should be added in to create a more fruitful results.

For the simplicity of this research, we use Google Maps API to get us the geo-coordinates for each neighborhood and what it returned is just one pair of coordinates. These coordinates may not have cover the whole area of the neighborhood and thus multiple coordinates should be used instead. Furthermore, radius limit are set to 500 meters, so only venues data within this areas are being covered.

**F. Conclusion**

Purpose of this project was to target individuals/families interested and looking in house purchase potential within their budget and venue convenience, and also for business developers to discover any potential businesses e.g. opening a restaurant or potential gap in property market in less popular neighborhoods in LA as a general study. By calculating venue density distribution from Foursquare data and rent price data from open data website, we have prepared ourselves with these useful data for clustering model. Clustering using those data though K-Means model enable us to create a zone of interest for the neighborhoods within the cluster. This is important as a foundation point for more detailed analysis which result in narrowing down into more specific neighborhoods for further analysis within the area.

**G. References:**

* [1] [Neighborhood Data for Social Change](https://usc.data.socrata.com/Los-Angeles/Rent-Price-LA-/4a97-v5tx/data)
* [2] [Google Map](https://www.google.com/maps/)
* [3] [Forsquare API](https://developer.foursquare.com/)

Xu Wang