Clustering Los Angeles Neighborhoods by Median Rent Prices and data on Venues using the Foursquare API to assist immigrants to find suitable apartments

May 10, 2020

Introduction

Background

According to the Migration Policy Institute, more than 44,7 million immigrants lived in the United States in 2018. Then, according to Census.gov, 1.21 million people migrated to the US in 2018. Immigration has been a major source of population growth and cultural change throughout much of U.S. history. And one of the questions new arrivals are going to ask is where to settle down to start a new chapter in their life. We will try to answer this question on the example of Los Angeles.

Problem

In this case, we are adopting machine learning tools to assist immigrants to find apartments with optimum conditions. The question we are currently solving is: how could we provide support to immigrants in finding suitable apartments in Los Angeles(further - LA)? To solve this problem, we are going to cluster LA neighborhoods in order to recommend venues with the median renting price amount where immigrants can find suitable apartments. We will recommend neighborhoods according to essential facilities and services surrounding such venues i.e. schools, stores, restaurants, hospitals, and entertainments.

Data

Data about Los Angeles(further - LA) neighborhoods and rent price paid in these neighborhoods was taken from USC Price Center for Social Innovation (https://usc.data.socrata.com/Los-Angeles/Rent-Price-LA-/4a97-v5tx). The dataset comprises LA Neighborhoods, Median Rent Price Amount in LA Neighborhoods.

To discover and choose affordable and convenient locations across different neighborhoods according to the presence of essential facilities and services, we will access data through FourSquare API interface. By merging data frames on LA neighborhoods and the median rent price amount from USC price center and data on essential facilities and services surrounding these neighborhoods from FourSquare API interface, we will be able to recommend neighborhoods with affordable renting apartments.

Methodology

The Methodology section is consist of four following stages:

1. Data Collection
2. Data Exploration and Understanding
3. Data preparation and preprocessing
4. Modeling

Data Collection and Cleaning

Data Sources

Data about Los Angeles(further - LA) neighborhoods and rent price paid in these neighborhoods was taken from USC Price Center for Social Innovation (https://usc.data.socrata.com/Los-Angeles/Rent-Price-LA-/4a97-v5tx).

Data cleaning and Preparation

Dataset was prepared for the modeling process following next steps:

1. **Data downloading and exploration**

Data downloaded from USC Price Center Website. Dataset was in a good condition and did not need a lot of cleaning. Then dataset was explored using several procedures to find out its shape and to know datasets data types. These procedures let us know that "Location" procedures had to be transformed from "object" to "float64".

2. **Columns extraction**

We have extracted from the dataset three following columns:

- "Amount" - information about Median Rent Price Amount
- "Neighborhood" – information about Los Angeles Neighborhoods
- "Location" – information about

3. **Renaming the columns and preparing coordinates**

We renamed column "Amount" to "Median Rent Price" to avoid controversy, and split column "Locations" in two columns "Latitude" and "Longitude" for more convenience. Then we transformed columns "Latitude" and "Longitude" to format "float64" to be able to use in our analysis.

4. **Finding Median Rent Price**

We described our dataset and found Average Median Rent Price among all neighborhoods to find the most frequent price which newcomers could meet and to narrow dataset because of limited abilities of the Foursquare API. After, we narrowed our dataset according to Median Rent Price among all neighborhoods

5. **Map preparation**

We found the latitude and longitude values of Los Angeles to create a map of Los Angeles with neighborhoods superimposed on top. After which defined and version.

Modeling

We used Foursquare credentials to extract information about venues located around explored neighborhoods to use it in our analysis of neighborhoods.

Then, we transformed the data and created Los Angeles Neighborhoods Venues dataframe to improve and ease the utilization of extracted data.

After which explored our dataset by finding amount of venues returned in each neighborhood, discovering frequency of occurrence of each category of venues in neighborhoods and finding the Top 10 venues in each neighborhood. Doing these we got the better understanding of our dataset and prepared data and our self for cluster analysis.

Clustering Neighborhoods

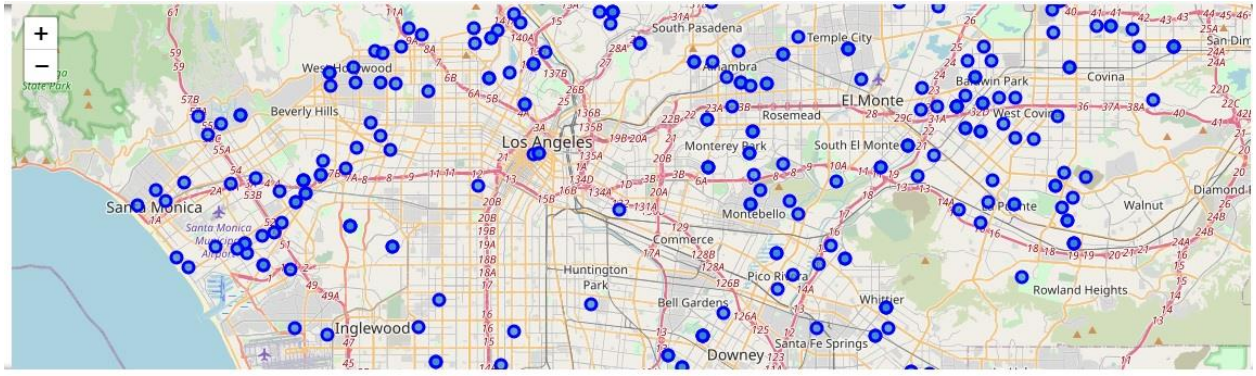
We used the k-means clustering technique because of it is high flexibility to account for mutations in Apartment renting market. And followed severel steps to conduct analysis

1. **Running the k-means to cluster neighborhoods**

We chose to divide dataset to 5 clusters because of widespread usage of 5 clusters among previous projects. Then we ran the k- means to cluster neighborhoods.

2. **Creating new dataset that includes clusters and venues and Map visualization**

To get the better view and understanding of clusters and venues around them, we created clusters dataset and merged it with the dataset of Top 10 venues in each neighborhoods and visualized on the Map.



3. **Clusters Examination**

We examined our 5 clusters to find neighborhoods that could be considered suitable for renting apartment.

Result and Discussion

We examined this dataset using two main perspectives. The first is finding neighborhoods that average rent price fit Median Rent Price among all Neighborhoods. The second is clustering found neighborhoods using data about venues located around them.

As a result, we can see that the Neighborhoods from Cluster 4 can be considered more suitable for newcomers to rent apartments, because they contain a big variety of different venues in close range which can satisfy their needs. Other Clusters are not so diversified and can be suitable for immigrants who have enough experience of living in Los Angeles and familiar with the city.

Conclusion

According to the Migration Policy Institute, more than 14% of the population of the United States were considered immigrants in 2018. From which we can assume that immigrants have become a major part of the US society and one of the drivers of cultural changes. In this context of a continuous influx of migrants it is urgent to adopt machine learning tools in order to assist newcomers to find a suitable place to settle down. As a result, the question we were trying to solve was: how could we provide support to immigrants in finding suitable apartments in example of Los Angeles?

To solve this problem, we clustered LA neighborhoods in order to recommend neighborhoods with the median renting price amount where immigrants can find suitable apartments. We recommended neighborhoods according to essential facilities and services surrounding such venues i.e. schools, stores, restaurants, hospitals, and entertainments.

We gathered LA Neighborhoods and the Median Rent Price data from USC Price Center for Social Innovation (https://usc.data.socrata.com/Los-Angeles/Rent-Price-LA-/4a97-v5tx). Moreover, to explore and target recommended locations across different venues according to the presence of essential facilities and services we accessed data through the FourSquare API interface and arranged them as a data frame for visualization. By merging LA Neighborhoods and the Median Rent Price data and data on essential facilities and services surrounding such properties from the FourSquare API interface, we were able to recommend suitable neighborhoods to rent apartments.

The Methodology section comprised four stages: 1. Data Collection; 2. Data Exploration and Understanding; 3. Data preparation and preprocessing; 4. Modeling. In the modeling section, we used the k-means clustering technique as it is highly flexible to account for changes in the median rent price amount in Los Angeles Neighborhoods.

Finally, we drew the conclusion suggesting suitable neighborhoods to rent apartments. We used two main perspectives to reach our results. First, we examined them according to median price amount paid around all neighborhoods. Second, we analyzed our results according to the five clusters we produced and found that Cluster 4 might be considered as the most suitable.