Los Angeles Neighborhood Analysis Report

Aleksandr Malakhov

March 23, 2021

1. Introduction

Let's imagine that some big travel company decided to expand their business. They plan open new hotel, restaurant and travel agency in Los Angeles City, CA. The stakeholders of the travel company want to know which the most popular places for hotels, restaurants and company offices in the L.A. To get this information we'll should explore and clustering L.A. neighborhoods using neighborhoods geo information and venue information. Finally, we'll make visualization clusters on the L.A. map and give a recommendation to recognize that places to make final decision.

2. Data description.

1. Los Angeles County Neighborhoods - the dataset from the UCLA Geoportal. This dataset exists in the GeoJSON format and contains the data about all segments of the L.A. City, standalone cities and unincorporated areas of L.A. County. we'll be used the geo data and meta data of the segments of the L.A. City only. The geo data contains geographic coordinates of polygons each of neighborhood. We need get the mean each of polygon coordinate set.
2. Foursquare - the venues data for L.A. neighborhoods. To get information about L.A. venues I'm using Foursquare Places API. First, we'll get the full venue category list to select ids of needed venue categories: Hotels, Restaurants, Business Centers. Next, we'll get venues of this three categories for each L.A. neighborhood for clustering and visualization.

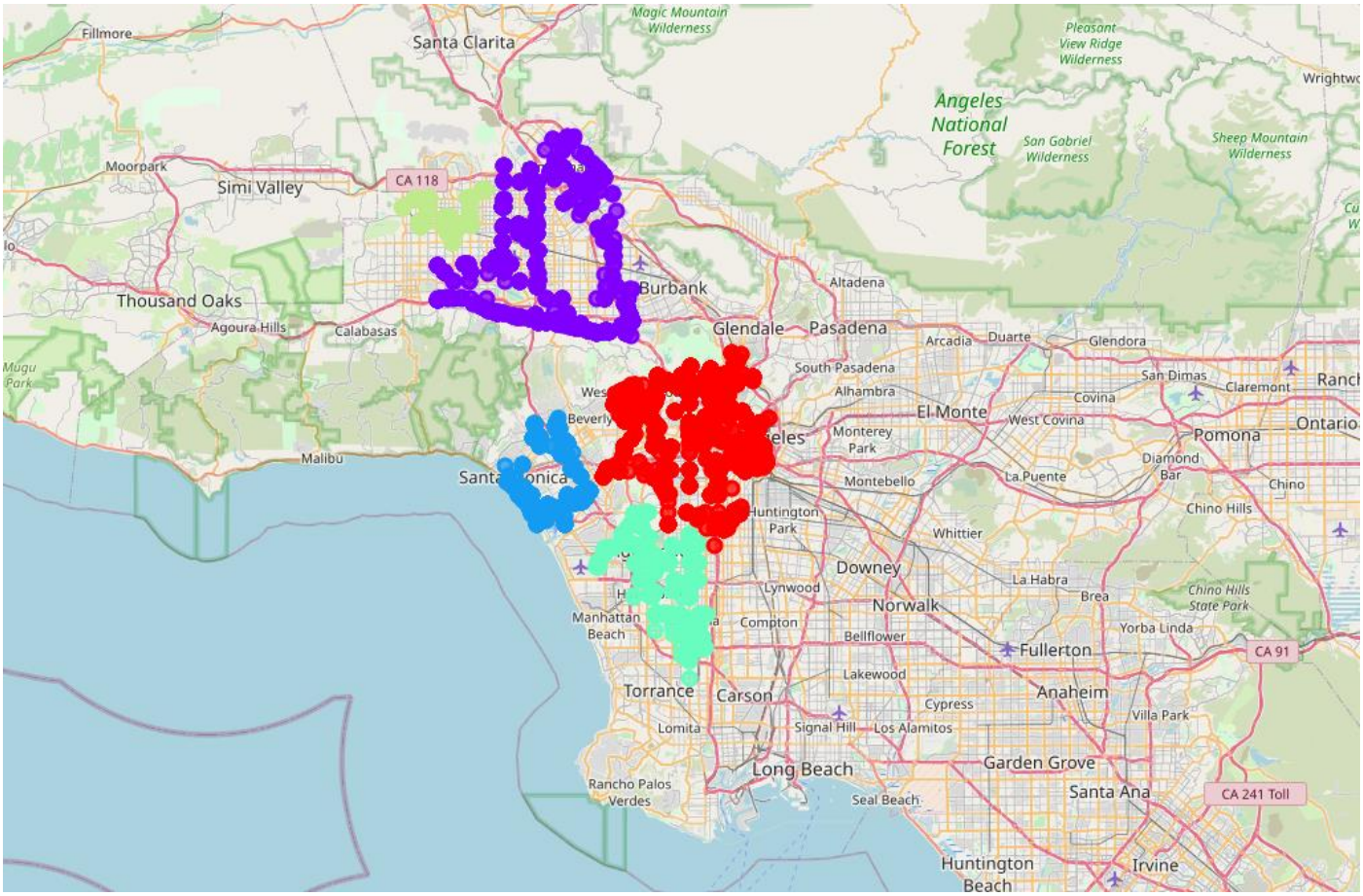
3. Methodology

To get the most popular places of venues we should perform the spatial clustering. This goal is met by the DBSCAN algorithm i.e. Density-Based Spatial Clustering of Applications with Noise. DBSCAN is a density-based clustering algorithm which is appropriate to use when examining spatial data. Density-based clustering locates regions of high density that are separated from one another by regions of low density. Density in this context is defined as the number of points within a specified radius. DBSCAN is particularly effective for tasks like class identification on a spatial context. The wonderful attributes of the DBSCAN algorithm is that it can find out any arbitrary shaped cluster without getting effected by noise. For spatial clustering of venues, we use their geographical coordinates. We will perform clustering separately for each venues category. Next, we will select the 5 largest clusters to show them on the map. In each of the clusters, we will find the neighborhoods with the largest and smallest number of venues. This will allow us to give more precise recommendations regarding the business goal.

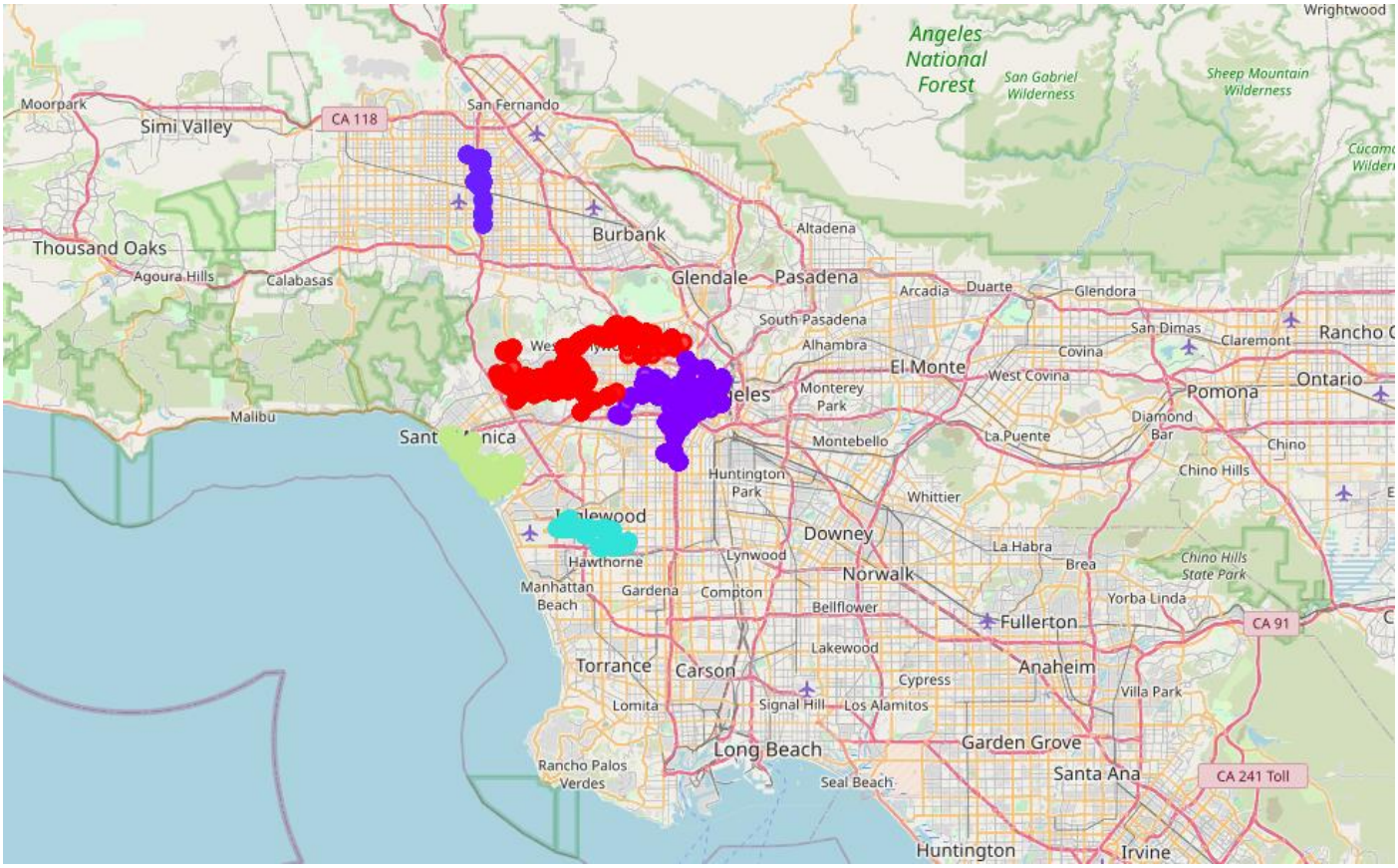
4. Results

In this section, firstly the results of venues clustering are shown on the L.A. map. Next, we are calculating count of clustered venues of each neighborhood to show on the bar chart most popular and least popular neighborhoods by number of venues.

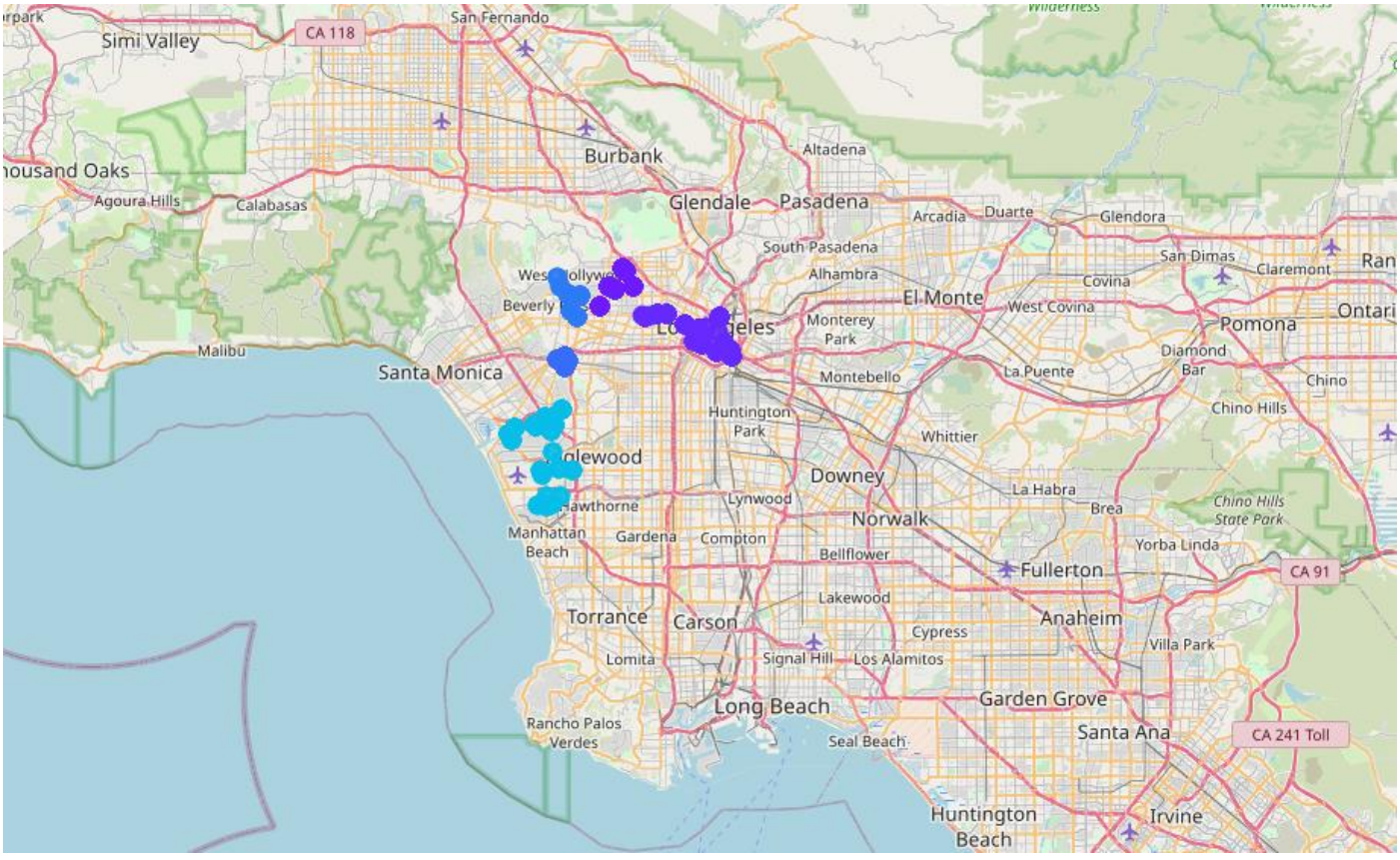
Visualization Top 5 biggest clusters in the Food category on the L.A. map



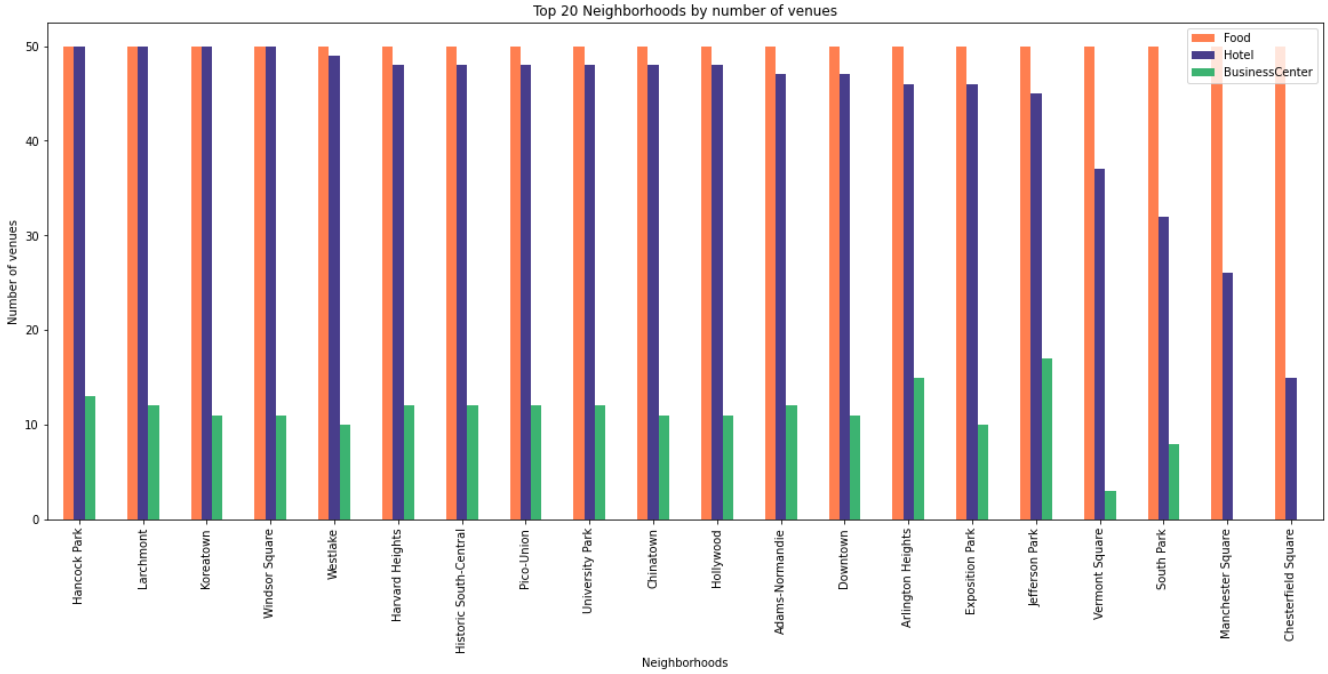
Visualization Top 5 biggest in the Hotel category clusters on the L.A. map



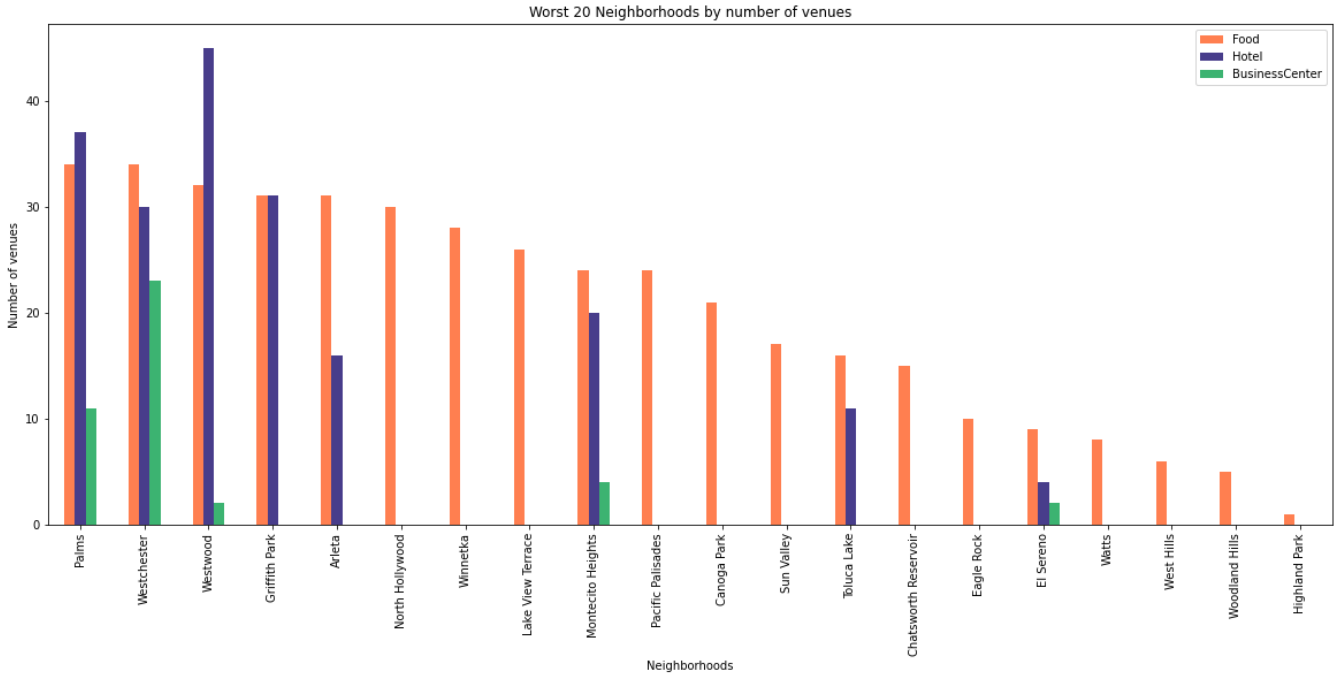
Visualization Top 5 biggest clusters in the Business Center category on the L.A. map.



Visualization Top 20 Neighborhoods by number of venues on the bar chart.



Visualization Worst 20 Neighborhoods by number of venues on the bar chart



5. Discussion

In this section, we look at what the data analysis showed, how its results help us to answer to the business question, and what recommendations we can give to stakeholders.

The chosen density-based spatial clustering methodology (DBSCAN) made it possible to identify clusters with the highest location density venues in Los Angeles City. This shows us which places are most popular in each venues category on the Los Angeles map. For business expansion purposes, these clusters are preferred. However, stakeholders should to know that there is a very high level of competition in these places. Additional analysis showed that inside the clusters there are neighborhoods with different numbers of venues. This information presents in the bar charts Top 20 and Worst 20 neighborhoods. This additional information allows us to make the following recommendations:

1. For the neighborhoods with a large number of venues, we recommend buying an existing business (hotel, restaurant or office) instead of opening a new one. This will avoid hard competition press on a new business.
2. For the neighborhoods with a small number of venues, we recommend opening a new business to develop it in a preferential environment.

6. Conclusion

This data analysis showed the practical application of the chosen algorithm and allowed us to draw conclusions and make recommendations. In the future, we can expand this analysis by using additional information about the venues, such as the rating, foot traffic and so on. This will allow us to use regression algorithms to predict certain features based on additional parameters.