Locating Coffee Shop in Los Angeles through Data Analysis

Aynaz Mirzazada

Introduction

1.1. Background

It is crucial factor to pick a good location of a coffee shop as it can be considered a vital factor for its survival within the first years of its opening. It would be safe to say that within the first years when there is no usual customer base and loyalty, location of the coffee shop makes it visible to the passing by people and to the customer audience. However, we should not forget about the fact that it is most necessary to make it available to the target customers and make it in the easily accessible location.

1.2. Business Problem

The aim of the client is to open a small coffee shop for the youth, especially the university students and young adults as a quiet place to visit, to study, to have productive time or just to hang out with the groupmates. In this project, the California state is considered and Los Angeles city is the aimed location. With the usage of data we will also be able to see why it is a good decision to open this shop exactly in Los Angeles.

The main objective of this Capstone Project is to locate this coffee shop in optimal location where mostly university students are populated. Using the California Census Data we will consider the densest population for the 20-30 years old California Citizens across neighborhoods, furthermore discussing the borough location where the most universities in LA is located.

For the data analysis thorough California Census Data and Los Angeles Education Facilities data is used from the Los Angeles Open Data Portal. For geolocating our objects Foursquare API will be utilized.

1.3. Interests

Other businesses and non-governmental organizations with similar target audience (youth, university students) can utilize this analysis to understand the densest youth population across California.

2. Data acquisition and Cleaning

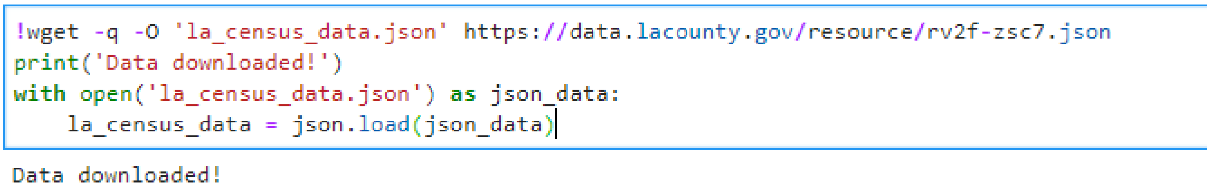
2.1. Data Sources

Estimated population by census tract, city name, ethnicity, gender and age group data for Los Angeles County citizens is obtained from this link; all officially registered education facilities list in Los Angeles city is taken from this link, both from the Los Angeles Open Data portal. Geographical coordinates, latitude and longitude data for each neighborhood in Los Angeles is obtained using this link from the LA City Geohub website.

Foursquare API explore function is used to receive the facilities across the neighborhoods of Los Angeles according to the coordinates. For more information on developer applications of Foursquare API this link can be useful.

2.2. Data cleaning and Feature Selection

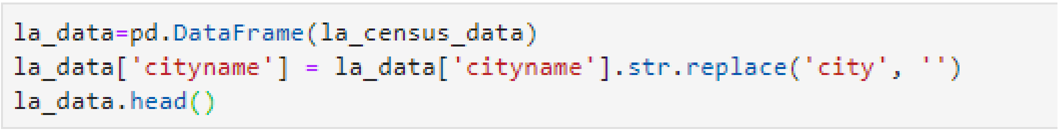
As mentioned previously three sets of data are used for this analysis. Both the LA Census data and University resource data is obtained as a JSON file and opened using the open as JSON Data function.



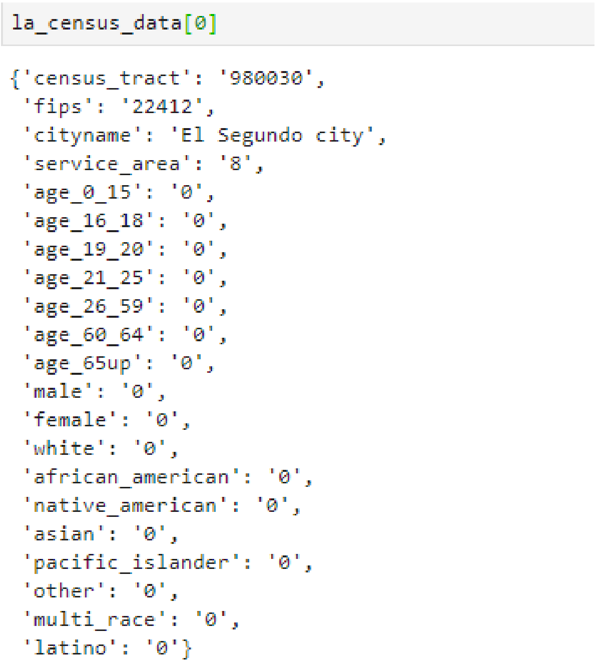
```
!wget -q -O 'la_census_data.json' https://data.lacounty.gov/resource/rv2f-zsc7.json
print('Data downloaded!')
with open('la_census_data.json') as json_data:
    la_census_data = json.load(json_data)
```

```
Data downloaded!
```

We should look at the data to see which columns are necessary for our analysis. It can be seen that each city name is ending with the word ‘city’ which can be removed to achieve more accurate table view.



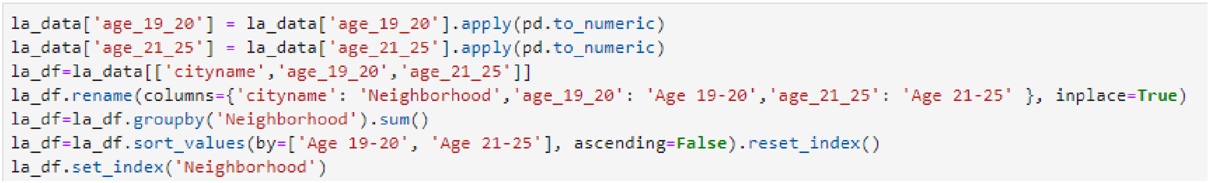
```
la_data=pd.DataFrame(la_census_data)
la_data['cityname'] = la_data['cityname'].str.replace('city', '')
la_data.head()
```



```
la_census_data[0]
```

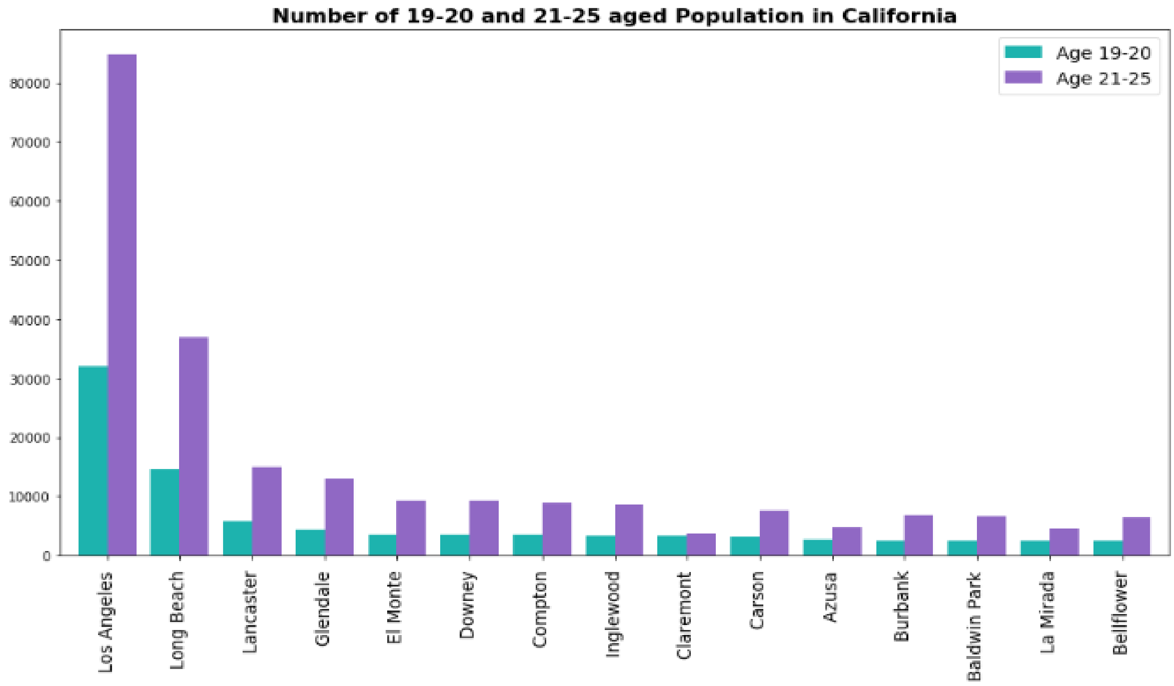
```
{'census_tract': '980030',
 'fips': '22412',
 'cityname': 'El Segundo city',
 'service_area': '8',
 'age_0_15': '0',
 'age_16_18': '0',
 'age_19_20': '0',
 'age_21_25': '0',
 'age_26_59': '0',
 'age_60_64': '0',
 'age_65up': '0',
 'male': '0',
 'female': '0',
 'white': '0',
 'african_american': '0',
 'native_american': '0',
 'asian': '0',
 'pacific_islander': '0',
 'other': '0',
 'multi_race': '0',
 'latino': '0'}
```

We should look at the data to see which columns are necessary for our analysis. It can be seen that each city name is ending with the word 'city' that can be removed to achieve more accurate table view.

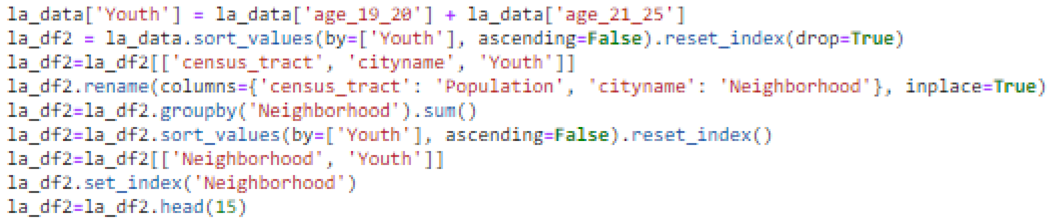


```
la_data['age_19_20'] = la_data['age_19_20'].apply(pd.to_numeric)
la_data['age_21_25'] = la_data['age_21_25'].apply(pd.to_numeric)
la_df=la_data[['cityname','age_19_20','age_21_25']]
la_df.rename(columns={'cityname': 'Neighborhood','age_19_20': 'Age 19-20','age_21_25': 'Age 21-25' }, inplace=True)
la_df=la_df.groupby('Neighborhood').sum()
la_df=la_df.sort_values(by=['Age 19-20', 'Age 21-25'], ascending=False).reset_index()
la_df.set_index('Neighborhood')
```

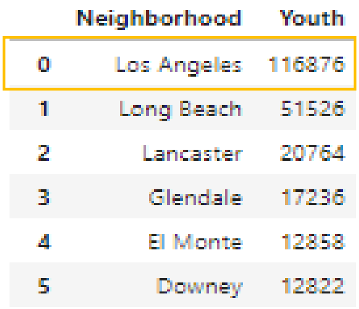
From this table we will need to take the number of young people living in California. Our target age group is 'age_19_20' and 'age_21_25' as most university students are around these ages. All data in this table is in string format so to work with the age groups we will convert them to integer format. Moreover, column names are adjusted respectively to decrease the clutter. Our target age population numbers are summed up according to the neighborhoods as seen above. From where we can achieve our first statistical result as such.



We can go forward and calculate the total number of youth according to the California cities. To do this we can sum up 'Age 19-20' and 'Age 21-25' columns to obtain the sum of all the young population. Consequently, this date should be sorted in decreasing order to have the final table.



```
la_data['Youth'] = la_data['age_19_20'] + la_data['age_21_25']
la_df2 = la_data.sort_values(by=['Youth'], ascending=False).reset_index(drop=True)
la_df2=la_df2[['census_tract', 'cityname', 'Youth']]
la_df2.rename(columns={'census_tract': 'Population', 'cityname': 'Neighborhood'}, inplace=True)
la_df2=la_df2.groupby('Neighborhood').sum()
la_df2=la_df2.sort_values(by=['Youth'], ascending=False).reset_index()
la_df2=la_df2[['Neighborhood', 'Youth']]
la_df2.set_index('Neighborhood')
la_df2=la_df2.head(15)
```

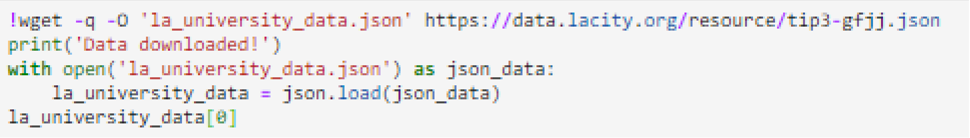


	Neighborhood	Youth
0	Los Angeles	116876
1	Long Beach	51526
2	Lancaster	20764
3	Glendale	17236
4	El Monte	12858
5	Downey	12822

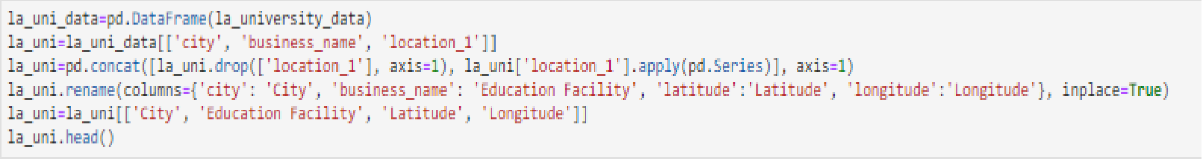
From this table it is obvious that Los Angeles is the top city across California for youth population hence the best location for our coffee shop. As a next step we should locate the highest number of education facilities across Los Angeles city neighborhoods to choose exact location. For this purpose, we will use the university resource data taken from Los Angeles Open Data Portal.

From Los Angeles Education facilities data we are going to need to explore the neighborhood the most facilities are located, name of the facility and geographical data. It is obtained using the data below.

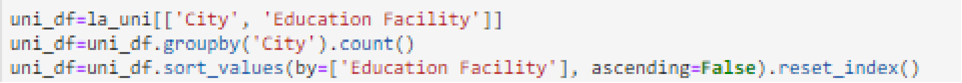
We can group the facilities located in the same neighborhood and sort them in descending order to get the top locations.



```
!wget -q -O 'la_university_data.json' https://data.lacity.org/resource/tip3-gfjj.json
print('Data downloaded!')
with open('la_university_data.json') as json_data:
    la_university_data = json.load(json_data)
la_university_data[0]
```

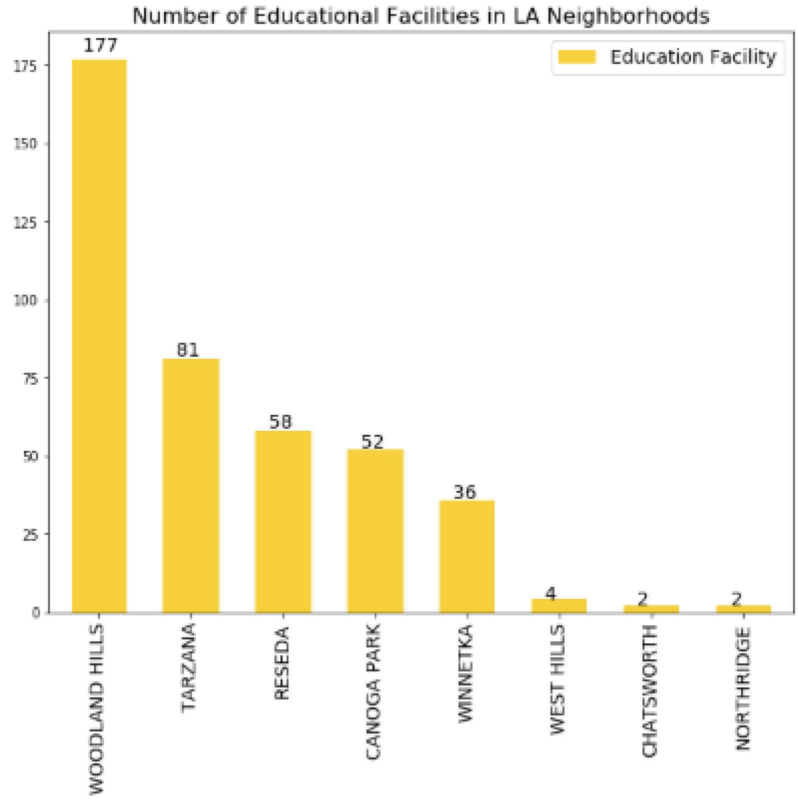


```
la_uni_data=pd.DataFrame(la_university_data)
la_uni=la_uni_data[['city', 'business_name', 'location_1']]
la_uni=pd.concat([la_uni.drop(['location_1'], axis=1), la_uni['location_1'].apply(pd.Series)], axis=1)
la_uni.rename(columns={'city': 'City', 'business_name': 'Education Facility', 'latitude':'Latitude', 'longitude':'Longitude'}, inplace=True)
la_uni=la_uni[['City', 'Education Facility', 'Latitude', 'Longitude']]
la_uni.head()
```



```
uni_df=la_uni[['City', 'Education Facility']]
uni_df=uni_df.groupby('City').count()
uni_df=uni_df.sort_values(by=['Education Facility'], ascending=False).reset_index()
```

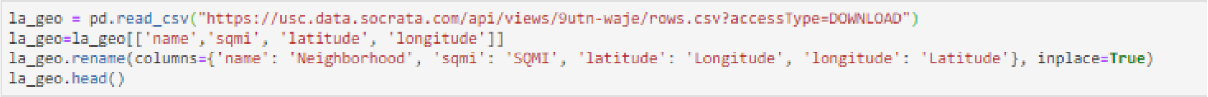
We can visualize our table to see the top location for educational facilities in Los Angeles. From the bar chart below, we can conclude that with 177 facilites top location is Woodland Hills to open our coffee shop targeting the university students.



3. Exploratory Data Analysis

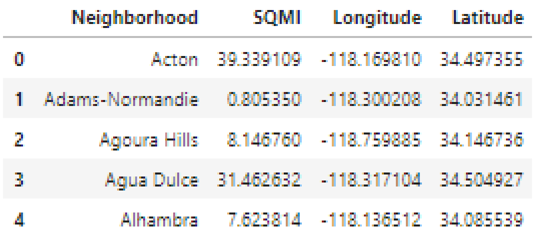
3.1. Folium mapping

Now we can start mapping our neighorhoods on map. Firstly CSV data is obtained from LA City Geohub and opened by pandas open csv function.



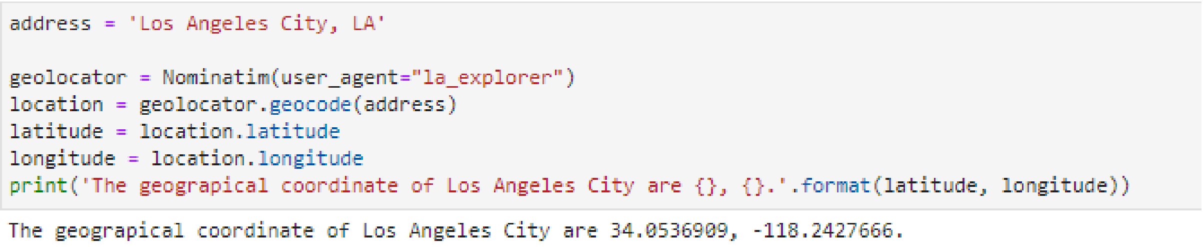
```
la_geo = pd.read_csv("https://usc.data.socrata.com/api/views/9utn-waje/rows.csv?accessType=DOWNLOAD")
la_geo=la_geo[['name','sqmi', 'latitude', 'longitude']]
la_geo.rename(columns={'name': 'Neighborhood', 'sqmi': 'SQMI', 'latitude': 'Longitude', 'longitude': 'Latitude'}, inplace=True)
la_geo.head()
```

We will take the neighborhood names, square meters per inch, latitude and longitude data.



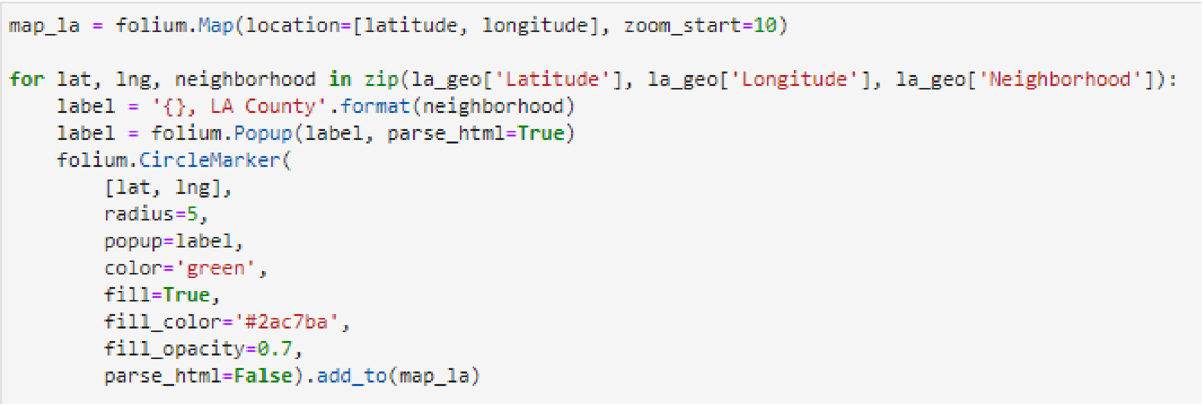
	Neighborhood	SQMI	Longitude	Latitude
0	Acton	39.339109	-118.169810	34.497355
1	Adams-Normandie	0.805350	-118.300208	34.031461
2	Agoura Hills	8.146760	-118.759885	34.146736
3	Agua Dulce	31.462632	-118.317104	34.504927
4	Alhambra	7.623814	-118.136512	34.085539

To locate our neighborhoods firstly empty map is built using the geonominatim function to create instance of a map. Firstly, coordinates of Los Angeles city is obtained.



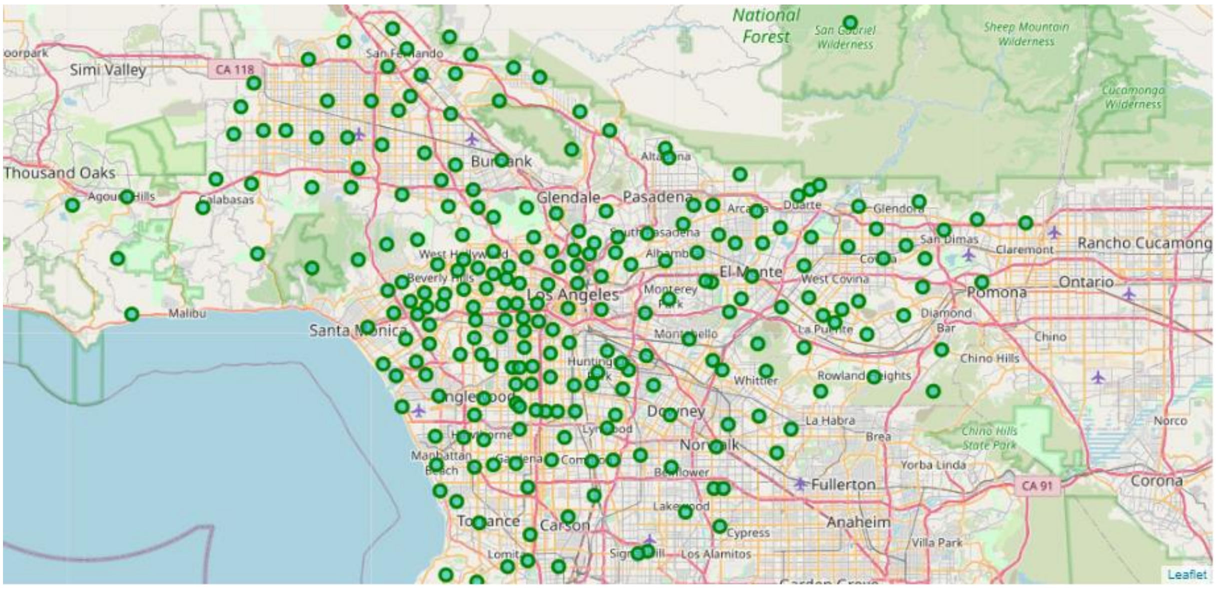
```
address = 'Los Angeles City, LA'

geolocator = Nominatim(user_agent="la_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Los Angeles City are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Los Angeles City are 34.0536909, -118.2427666.
```



```
map_la = folium.Map(location=[latitude, longitude], zoom_start=10)

for lat, lng, neighborhood in zip(la_geo['Latitude'], la_geo['Longitude'], la_geo['Neighborhood']):
    label = '{}, LA County'.format(neighborhood)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='green',
        fill=True,
        fill_color='#2ac7ba',
        fill_opacity=0.7,
        parse_html=False).add_to(map_la)
```

Later on using the folium library, we can label each of the neighborhoods on the map using our la_geo table we created earlier. In the next picture we can see Los Angeles map with all of the neighborhoods from our table labeled with green-blue marker.

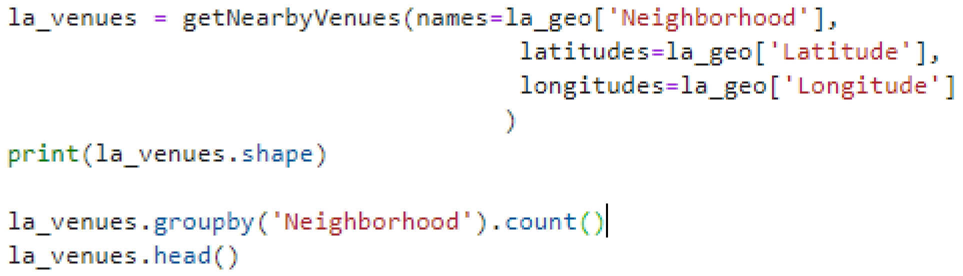


4. Predictive Modeling

4.1. Data Pre-processing

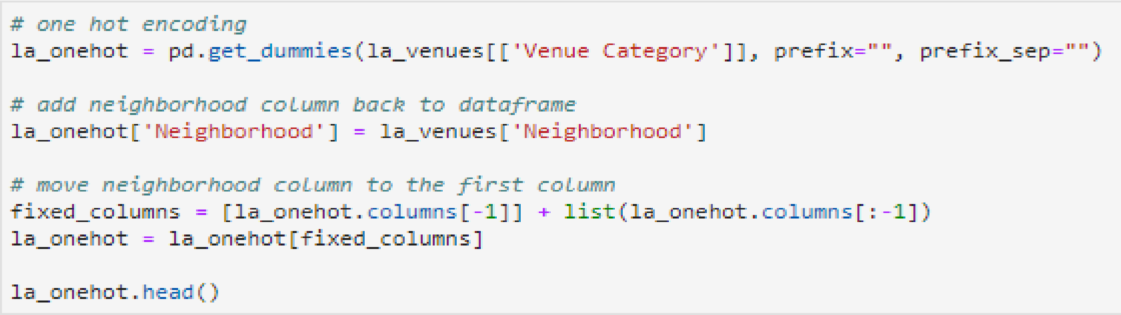
As a next stage of our analysis, we should use Foursquare API by customly created getNearbyVenues function to get the list of the venues located in each Los Angeles neighborhood. K-means clustering method is going to be used to achieve the most common venues.

After getting all the venues located in 500 m radius to our neighborhood coordinates we can group them by neighborhoods.



```
la_venues = getNearbyVenues(names=la_geo['Neighborhood'],
                                   latitudes=la_geo['Latitude'],
                                   longitudes=la_geo['Longitude']
                                  )
print(la_venues.shape)

la_venues.groupby('Neighborhood').count()
la_venues.head()
```

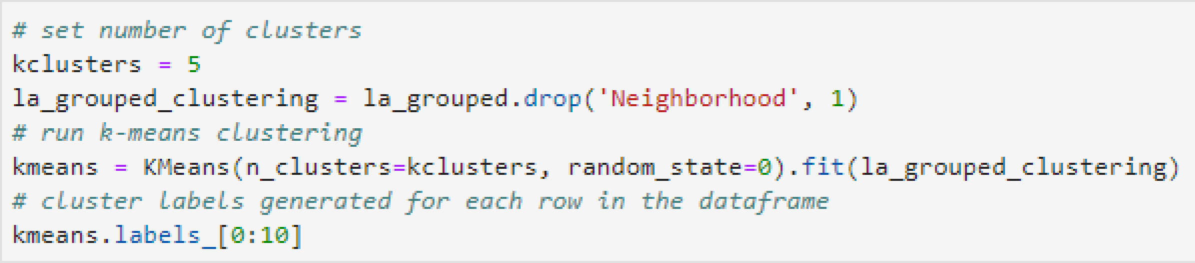
Consequently, for each venue category dummies were created and grouped under each neighborhood. Furthermore, on for each neighborhood we could obtain the top 10 venues.



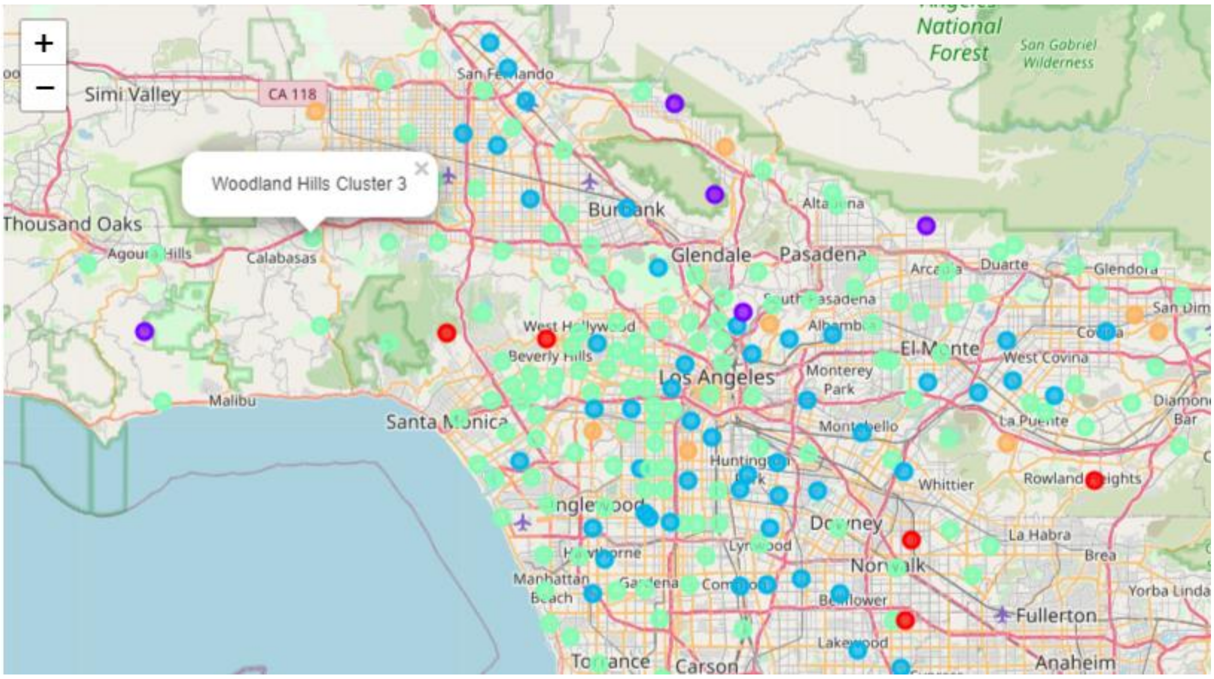
```
# one hot encoding
la_onehot = pd.get_dummies(la_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
la_onehot['Neighborhood'] = la_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [la_onehot.columns[-1]] + list(la_onehot.columns[:-1])
la_onehot = la_onehot[fixed_columns]

la_onehot.head()
```

4.2. k-Means Clustering

After the trial and error method the optimal k-value is discovered to be 5 for Los Angeles Neighborhoods. The resulting cluster table is joined with our previous neighborhood table to get the final table. This final table is mapped over a folium map and each cluster is colored with different colors respectively.

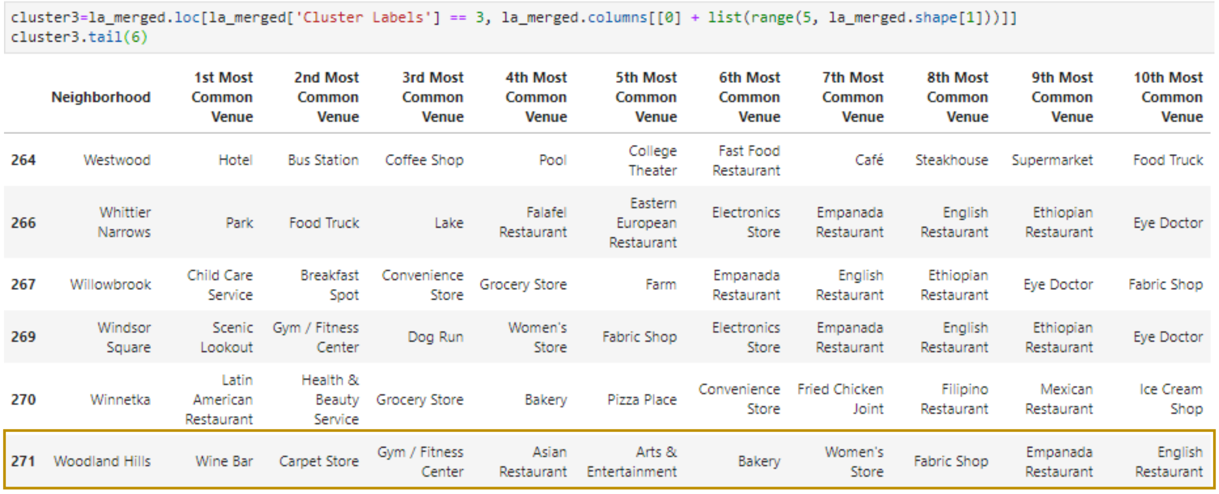


```
# set number of clusters
kclusters = 5
la_grouped_clustering = la_grouped.drop('Neighborhood', 1)
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(la_grouped_clustering)
# cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```



5. Conclusions

As it can be seen from our Folium map Woodland Hills falls into Cluster 3 where the top of the list does not contain any other coffee shops, meaning an highly competitive business environment is not forecasted. It can be considered good news for a brand new coffee shop establishment. Furthermore, it should be noticed that none of the most visited places is repetitive. Hence, it is also safe to say that Cluster 3 neighborhoods are in wide range in their most visited venues and often visited by the Los Angeles citizens for all different kinds of activities. Let's take a look at top Cluster 3 neighborhoods for instance.



```
cluster3=la_merged.loc[la_merged['Cluster Labels'] == 3, la_merged.columns[[0] + list(range(5, la_merged.shape[1]))]]
cluster3.tail(6)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
264	Westwood	Hotel	Bus Station	Coffee Shop	Pool	College Theater	Fast Food Restaurant	Café	Steakhouse	Supermarket	Food Truck
266	Whittier Narrows	Park	Food Truck	Lake	Falafel Restaurant	Eastern European Restaurant	Electronics Store	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Eye Doctor
267	Willowbrook	Child Care Service	Breakfast Spot	Convenience Store	Grocery Store	Farm	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Eye Doctor	Fabric Shop
269	Windsor Square	Scenic Lookout	Gym / Fitness Center	Dog Run	Women's Store	Fabric Shop	Electronics Store	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Eye Doctor
270	Winnetka	Latin American Restaurant	Health & Beauty Service	Grocery Store	Bakery	Pizza Place	Convenience Store	Fried Chicken Joint	Filipino Restaurant	Mexican Restaurant	Ice Cream Shop
271	Woodland Hills	Wine Bar	Carpet Store	Gym / Fitness Center	Asian Restaurant	Arts & Entertainment	Bakery	Women's Store	Fabric Shop	Empanada Restaurant	English Restaurant

As seen, none of the top visited places in Woodland hills is coffee shop. Also, the general trend is not repetitive. According to the census data many young people live in this neighborhood, moreover most education facilities are located here. According to the Foursquare this location seems to be promising location for our purposes. In conclusion, considering both the census data and the location data Woodland Hills is optimal location to establish a brand new coffee shop.

6. Future directions

In future, to expand the coffee shop chain wider census data can be considered to be in close location to number of different age groups. Moreover, in future in location analysis such factors can be considered as having place not close to the highways and more into the city center to make it easier for students to arrive by walking or on bikes.

7. References

1. Applied Data Science Capstone Week 3 Neighborhood Segmenting and Clustering - https://www.coursera.org/learn/applied-data-science-capstone/home/week/3
2. Advanced Visualizations and Geospatial Data - https://www.coursera.org/learn/python-for-data-visualization/home/week/3
3. “What Makes for a Great Restaurant Location?” - https://restaurantengine.com/great-restaurant-location/
4. Los Angeles Open Data Portal - https://data.lacity.org/
5. Los Angeles Census Data - https://data.lacounty.gov/Mental-Health/County-of-Los-Angeles-Estimated-Population-by-Cens/rv2f-zsc7
6. Council District 3 - 611000 Education Facilities (Including Schools, Colleges, & Universities) - https://data.lacity.org/A-Prosperous-City/ACoolDATASET-Council-District-3-611000-Education-F/tip3-gfjj
7. 2010 Census Data By Block - https://geohub.lacity.org/datasets/lacounty::2010-census-data-by-block/data
8. Foursquare Developer - https://developer.foursquare.com/places