C9_WK4,5_ASSIGNMENT_RD

TITLE: USING FOURSQAURE APP TO CHOOSE A CITY NEIGHBORHOOD TO RENT IN.

REPORT BY: RYAN DESOUSA

1) DATA

The average rental rates for different neighborhoods in the city of Los Angeles are available at:

https://www.rentcafe.com/average-rent-market-trends/us/ca/los-angeles/

This neighborhood and rental data will be extracted to csv and then into a dataframe and cleaned.

I will try to pair this data, with information on the different LA neighborhoods, obtained through the Four Square API app.

I will use this rental data and the FourSquare API app to explore different venues in the LA neighborhoods to try get an idea of what are the high traffic offerings in the various neighborhoods.

I will use one hot encoding to try to find the ten most frequently occurring venues in the different neighborhoods.

I will then apply K-Means Clustering Machine Learning algorithm on the acquired neighborhood-venue data to cluster the data into five clusters.

The Neighborhood cluster will then be plotted on a map of LA city using Folium.

I will then filter the Neighborhood-Venue frequency data set, using venues of my choice and look for neighborhoods where these venues are among the top three most frequent venues.

Using the FourSquare API, I will then search any two of these neighborhoods further, with five venues of my choice and try to determine the frequency of occurrance of these venues, within a distance of 2000m of their neighborhood coordinates.

Once I have all this information, I will try to put all this information together to look for similarities/differences in the two neighborhoods and see if any one neighborhood is more preferred to rent in than the other.

The final goal is to see whether this data can be used to differentiate between the neighborhoods and be used by people to make a better decisions on choice of city neighborhood to rent in.

2 . DATA HANDLING – Location, Acquisition and Treatment.

The average rental rates for different neighborhoods in the city of Los Angeles are available at:

https://www.rentcafe.com/average-rent-market-trends/us/ca/los-angeles/

This neighborhood and rental data was extracted to csv and then into a dataframe and cleaned.

As I found the data was not scrapable I manually had to copy the table to excel and save it as a csv.

The csv was converted to a dataframe.

The data was cleaned to remove any missing values and the names LA_Neighborhood column were formatted as Neighborhood Name, CA, USA.

This format allowed for more accurate coordinate extraction using the FourSquare API

The data in the AverageMonthlyRent Column had a '$' sign that had to be deleted and the data converted to float from object.

The cleaned dataframe is shown in Figure 2.1

:

	LA_Neighborhood	AverageMonthlyRent_USD
0	Jefferson Park, CA, USA	1355
1	El Sereno, CA, USA	1396
2	Vermont Vista, CA, USA	1445
3	Vermont Knolls, CA, USA	1445
4	Hyde Park, CA, USA	1484

Figure 2.1: LA Neighborhood rent dataframe la_rent, after treatment.

I then paired neighborhood information, with coordinate information on the different LA neighborhoods, obtained through the Four Square API app to get dataframe LA_lat_long, as shown in Figure 2.2

	LA_Neighborhood	Latitude	Longitude
0	Jefferson Park, CA, USA	34.027234	-118.317576
1	El Sereno, CA, USA	34.081121	-118.177849
2	Vermont Vista, CA, USA	33.941947	-118.285814
3	Vermont Knolls, CA, USA	33.966819	-118.291670
4	Hyde Park, CA, USA	33.980569	-118.330631
...	...	...	...
93	Beverly Grove, CA, USA	34.076034	-118.369972
94	Historic South-Central, CA, USA	34.016230	-118.267308
95	University Park, CA, USA	34.027449	-118.283949
96	Pico, CA, USA	34.040672	-118.266192
97	Santa Monica, CA, USA	34.025072	-118.496513

Figure 2.2: LA Neighborhood and coordinate dataframe La_lat_long.

I then added the rental information to this dataframe to get the dataframe la_data, with all the necessary neighborhood information in one dataframe

	LA_Neighborhood	AverageMonthlyRent_USD	Latitude	Longitude
0	Jefferson Park, CA, USA	1355.0	34.027234	-118.317576
1	El Sereno, CA, USA	1396.0	34.081121	-118.177849
2	Vermont Vista, CA, USA	1445.0	33.941947	-118.285814
3	Vermont Knolls, CA, USA	1445.0	33.966819	-118.291670
4	Hyde Park, CA, USA	1484.0	33.980569	-118.330631

Figure 2.3: LA Neighborhood, rent and coordinate information dataframe, la_data.

I will use this basic dataset and the FourSquare API to to see whether this data can be used to differentiate between the neighborhoods and be used by people to make a better decisions on choice of city neighborhood to rent in.