IBM – Coursera

Data Science Specialization

Capstone project - Final report

Analysing and comparing student friendly neighbourhoods in USA

Kripa Shankar Muthu Kumar– 2020

I. Introduction:

The students (all graduate levels) who decides to go to a university strive to know the various characteristics and benefits of the city/location to determine the ease of living.

Both international and citizens search for common parameters in the city/state:

1. Housing
2. Transport facilities (public transport), traffic incidents (crash, weather)
3. Corporate companies and tech startups (internship/full time)
4. Sports and recreation
5. Restaurants, malls, music concerts and game night pubs/bars
6. Libraries, medical centres and more

I have added traffic incidents as the students/parents might like to know the safety of driving through the neighbourhoods.

Problem

We must enable the user to determine the best neighbourhood cluster that satisfies most of their parameters

In this project, I try to solve this problem using foursquare location data, New York crash data and machine learning algorithms to compare and find the best places for students to find universities and settle in.

Stakeholders

This project will be useful to students, parents and graduate students (job seekers). This will be useful to other third parties who want to see the behaviour and psychology behind students' requirements that must be around the universities after learning what clusters usually those people fall into.

II. Data description:

We pull location/venue data that belong to above categories from foursquare and traffic incidents data from the selected state website. In this project, we consider only one state, New York city and its neighborhoods.

I obtain Incident/crash data from New York Motor Vehicle Collisions – Crashes dataset. The dataset contains a vast number of features like latitude, longitude, number of persons injured/killed, number of pedestrians/cyclists injured, contributing factor vehicle 1 and more. But we just count the accidents per location or borough and append it to the dataset.

Regarding the venue data, it is same as what we saw in our previous exercise (Week 3). I will extract locations like universities, tech startups/corporate companies, restaurants/pubs, medical centers, events, etc, around Queens, NYC. And, combine with the crash data.

For example, we extract venues using foursquare API: **https://api.foursquare.com/v2/venues/explore.** We get the data, normalize, find the frequency and append to the dataset.

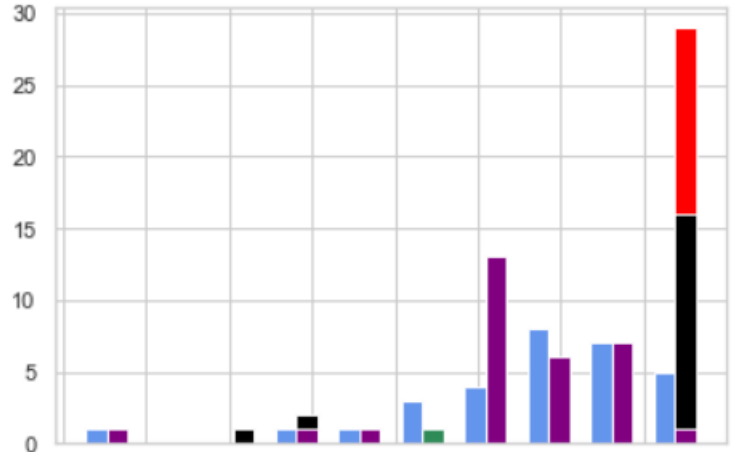
Below is the example of the resultant data.

	neighborhood	latitude	longitude	Medical_centers	Corporate_companies	Sports_venues	Restaurants	Libraries	Accidents
0	Marble Hill	40.876551	-73.91066	15	25	15	150	5	25
1	Carnegie	40.7845	-73.9551	12	15	10	155	6	21

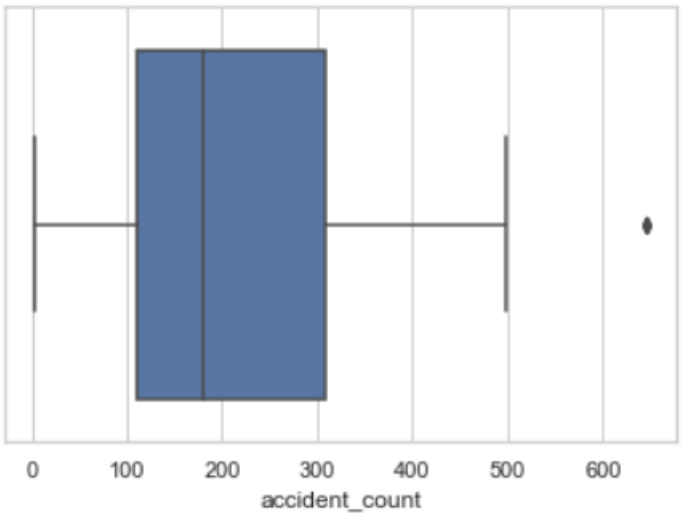
Also, this is a sample dataset for now. Dataset features can get modified as the project gets progressed.

III. Methodology:

After integrating the data (including accidents), that is data preparation, we must understand the data.



Venue distribution



Accident count

Venue distribution plots the following venue facilities data

	transport	housing	shops	night_life	sports	office
0	45	50	50	50	47	49
1	45	50	50	50	46	48
2	42	50	50	50	44	44
3	44	50	50	50	44	42
4	47	50	50	49	44	42
5	45	50	50	49	45	39
6	50	48	49	50	46	46
7	44	44	50	50	47	44
8	40	50	50	48	45	46
9	49	46	50	47	44	41
10	49	47	50	48	44	36

The final data are formed by combining accident count, neighbourhood data and the facilities data. Zip is obtained using google maps

	level_0	index	Borough	Neighborhood	Latitude	Longitude	zip	accident_count	transport	housing	shops	night_life	sports	office
0	0	0	Queens	Astoria	40.768509	-73.915654	11103	130	45	50	50	50	47	49
1	1	1	Queens	Woodside	40.746349	-73.901842	11377	394	45	50	50	50	46	48
2	2	2	Queens	Jackson Heights	40.751981	-73.882821	11372	370	42	50	50	50	44	44
3	3	3	Queens	Elmhurst	40.744049	-73.881656	11373	387	44	50	50	50	44	42
4	4	7	Queens	Kew Gardens	40.705179	-73.829819	11415	81	47	50	50	49	44	42
5	5	8	Queens	Richmond Hill	40.697947	-73.831833	11418	199	45	50	50	49	45	39
6	6	12	Queens	East Elmhurst	40.764073	-73.867041	11369	219	50	48	49	50	46	46
7	7	13	Queens	Maspeth	40.725427	-73.896217	11378	257	44	44	50	50	47	44

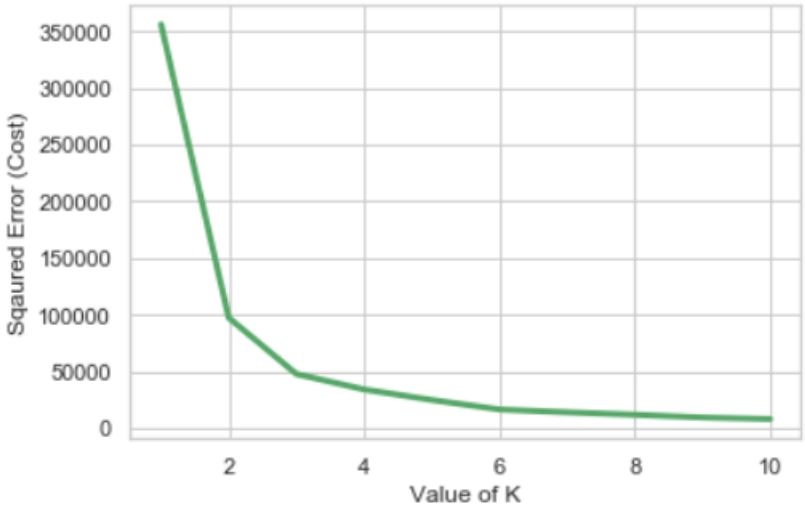
K means clustering

Using K means clustering, we cluster the following queen's data to find areas with both facilities and accident-less prone areas.

We perform clustering because we don't have labels for the process that we do. And, this is a segmentation procedure. K means calculates all the possibilities and positions the centroids accordingly.

For clustering, we index data frame from accident_count to office column and perform.

To find an optimal value of k, we use elbow method

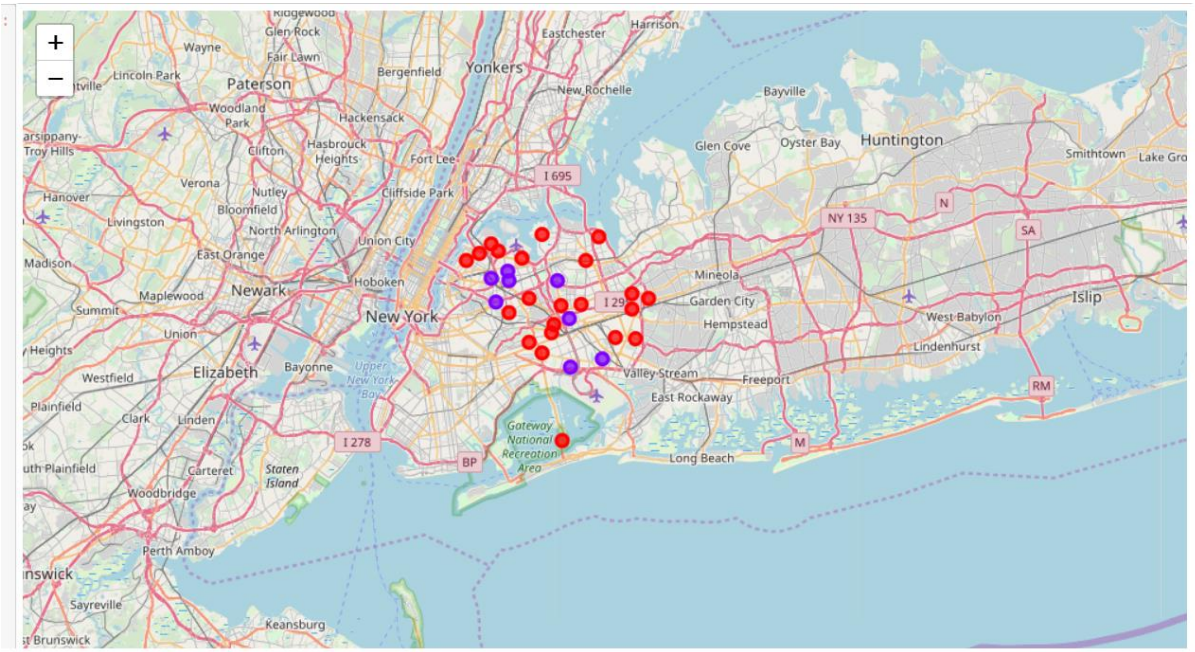


From this graph, we see that optimal number of clusters could be 2.

IV. Results:

K means assigns 0 cluster_label to less accident-prone areas and 1 to more accident prone areas.

Here is the accident map plotted for queens, New York city, New York.



Purple belongs to cluster label 1. Read belongs to cluster label 0.

V. Discussion

From examining clusters, we obtain the following data frame

This belongs to cluster 0 – less accident-prone areas

	level_0	Latitude	Longitude	zip	accident_count	transport	housing	shops	night_life	sports	office
0	0	40.768509	-73.915654	11103	130	45	50	50	50	47	49
4	4	40.705179	-73.829819	11415	81	47	50	50	49	44	42
5	5	40.697947	-73.831833	11418	199	45	50	50	49	45	39
6	6	40.764073	-73.867041	11369	219	50	48	49	50	46	46
8	8	40.728974	-73.857827	11374	198	40	50	50	48	45	46
9	9	40.689887	-73.858110	11421	178	49	46	50	47	44	41
10	10	40.680708	-73.843203	11417	204	49	47	50	48	44	36
12	12	40.784903	-73.843045	11356	150	49	16	50	47	47	45
13	13	40.761730	-73.791762	11358	210	45	44	50	50	43	36
14	14	40.728573	-73.720128	11426	82	45	13	50	45	43	43
15	15	40.722578	-73.820878	11367	211	47	50	50	49	45	43
17	17	40.718893	-73.738715	11428	103	44	17	50	45	43	36
18	18	40.694445	-73.758676	11412	185	42	13	50	38	32	37
20	20	40.692775	-73.735269	11411	126	41	7	50	38	38	25
21	21	40.603027	-73.820055	11693	55	31	16	40	19	47	12
22	22	40.775923	-73.902290	11105	121	45	48	50	50	49	47
23	23	40.782843	-73.776802	11360	55	37	42	50	39	45	35
25	25	40.723825	-73.797603	11366	111	47	49	50	48	44	41
26	26	40.761705	-73.931575	11106	164	47	50	50	50	47	50
27	27	40.733014	-73.738892	11427	126	48	23	50	45	44	39
28	28	40.770317	-73.894680	11370	146	46	48	50	50	48	46
29	29	40.716415	-73.881143	11379	169	44	45	50	50	47	43

This belongs to cluster 1 – more accident-prone areas

	level_0	Latitude	Longitude	zip	accident_count	transport	housing	shops	night_life	sports	office
1	1	40.746349	-73.901842	11377	394	45	50	50	50	46	48
2	2	40.751981	-73.882821	11372	370	42	50	50	50	44	44
3	3	40.744049	-73.881656	11373	387	44	50	50	50	44	42
7	7	40.725427	-73.896217	11378	257	44	44	50	50	47	44
11	11	40.668550	-73.809865	11420	307	48	8	50	43	44	39
16	16	40.710935	-73.811748	11435	309	47	49	50	50	44	42
19	19	40.675211	-73.772588	11434	475	49	27	50	36	39	40
24	24	40.744572	-73.825809	11355	352	45	50	50	50	44	36

The K means clustering correctly partitioned the high volume of accidents area to low volume of accidents area. Firstly, I thought of finding the correlations between the venues/facilities to the accident_prone areas. However, just by a glance, it looks like there is not a significant impact of venues, but, still, notably, in hight volume of accidents area, there are a lot of transport options and office areas. On the map, we could see that there are a lot of accidents around Airport area.

The lack of significant impacts of venues could be because of my data collection. Even though, I tried to get the accident count for zip code, I couldn't get the venue spots zip code wise. It is due to some API and data model issues. However, I will try to diagnose this issue more and try to fix it.

Correlations matrix

[75]:

	Cluster_Labels	level_0	index	Latitude	Longitude	zip	accident_count	transport	housing	shops	night_life	spo
Cluster_Labels	1	-0.287388	-0.285268	-0.0746608	-0.172806	0.105282	0.865242	0.0933473	0.104333	0.122986	0.118323	-0.05606
level_0	-0.287388	1	0.98352	-0.0263071	0.302281	0.0221526	-0.336412	-0.125689	-0.236873	-0.121028	-0.278259	-0.03998
index	-0.285268	0.98352	1	0.0145025	0.213873	-0.00383423	-0.322965	-0.094842	-0.155334	-0.0888462	-0.194486	0.02323
Latitude	-0.0746608	-0.0263071	0.0145025	1	-0.348999	-0.715936	-0.0292386	0.256535	0.41826	0.557934	0.663562	0.3625
Longitude	-0.172806	0.302281	0.213873	-0.348999	1	0.51639	-0.230579	-0.115039	-0.684651	-0.0244309	-0.470767	-0.6554
zip	0.105282	0.0221526	-0.00383423	-0.715936	0.51639	1	0.0231896	-0.337831	-0.426574	-0.547119	-0.609062	-0.2967
ıccident_count	0.865242	-0.336412	-0.322965	-0.0292386	-0.230579	0.0231896	1	0.292028	0.225279	0.252039	0.229153	-0.16
transport	0.0933473	-0.125689	-0.094842	0.256535	-0.115039	-0.337831	0.292028	1	0.196964	0.640509	0.6317	0.03971
housing	0.104333	-0.236873	-0.155334	0.41826	-0.684651	-0.426574	0.225279	0.196964	1	0.257029	0.652343	0.4899
shops	0.122986	-0.121028	-0.0888462	0.557934	-0.0244309	-0.547119	0.252039	0.640509	0.257029	1	0.762124	-0.1649
night_life	0.118323	-0.278259	-0.194486	0.663562	-0.470767	-0.609062	0.229153	0.6317	0.652343	0.762124	1	0.3404
sports	-0.0560699	-0.0399861	0.0232352	0.362542	-0.655465	-0.296756	-0.1652	0.0397152	0.489991	-0.164924	0.340403	
office	0.121597	-0.266401	-0.189375	0.678096	-0.536049	-0.703242	0.249648	0.622173	0.520361	0.705657	0.823877	0.3197

Cluster_Labels	level_0	index	Latitude	Longitude	zip	accident_count	transport	housing	shops	night_life	sports	office
1	-0.287388	-0.285268	-0.0746608	-0.172806	0.105282	0.865242	0.0933473	0.104333	0.122986	0.118323	-0.0560699	0.121597
-0.287388	1	0.98352	-0.0263071	0.302281	0.0221526	-0.336412	-0.125689	-0.236873	-0.121028	-0.278259	-0.0399861	-0.266401
-0.285268	0.98352	1	0.0145025	0.213873	-0.00383423	-0.322965	-0.094842	-0.155334	-0.0888462	-0.194486	0.0232352	-0.189375
-0.0746608	-0.0263071	0.0145025	1	-0.348999	-0.715936	-0.0292386	0.256535	0.41826	0.557934	0.663562	0.362542	0.678096
-0.172806	0.302281	0.213873	-0.348999	1	0.51639	-0.230579	-0.115039	-0.684651	-0.0244309	-0.470767	-0.655465	-0.536049
0.105282	0.0221526	-0.00383423	-0.715936	0.51639	1	0.0231896	-0.337831	-0.426574	-0.547119	-0.609062	-0.296756	-0.703242
0.865242	-0.336412	-0.322965	-0.0292386	-0.230579	0.0231896	1	0.292028	0.225279	0.252039	0.229153	-0.1652	0.249648
0.0933473	-0.125689	-0.094842	0.256535	-0.115039	-0.337831	0.292028	1	0.196964	0.640509	0.6317	0.0397152	0.622173
0.104333	-0.236873	-0.155334	0.41826	-0.684651	-0.426574	0.225279	0.196964	1	0.257029	0.652343	0.489991	0.520361
0.122986	-0.121028	-0.0888462	0.557934	-0.0244309	-0.547119	0.252039	0.640509	0.257029	1	0.762124	-0.164924	0.705657
0.118323	-0.278259	-0.194486	0.663562	-0.470767	-0.609062	0.229153	0.6317	0.652343	0.762124	1	0.340403	0.823877
-0.0560699	-0.0399861	0.0232352	0.362542	-0.655465	-0.296756	-0.1652	0.0397152	0.489991	-0.164924	0.340403	1	0.319761
0.121597	-0.266401	-0.189375	0.678096	-0.536049	-0.703242	0.249648	0.622173	0.520361	0.705657	0.823877	0.319761	1

VI. Conclusion

Thus, this project will be useful for students to select the areas that they would like to get housing, jobs or go to places to enjoy. Accidents clustering also helps students and parents discern places that they have to be cautious about. Whole data collection and preparation were done using foursquare API and Google Maps (reverse geo coding)