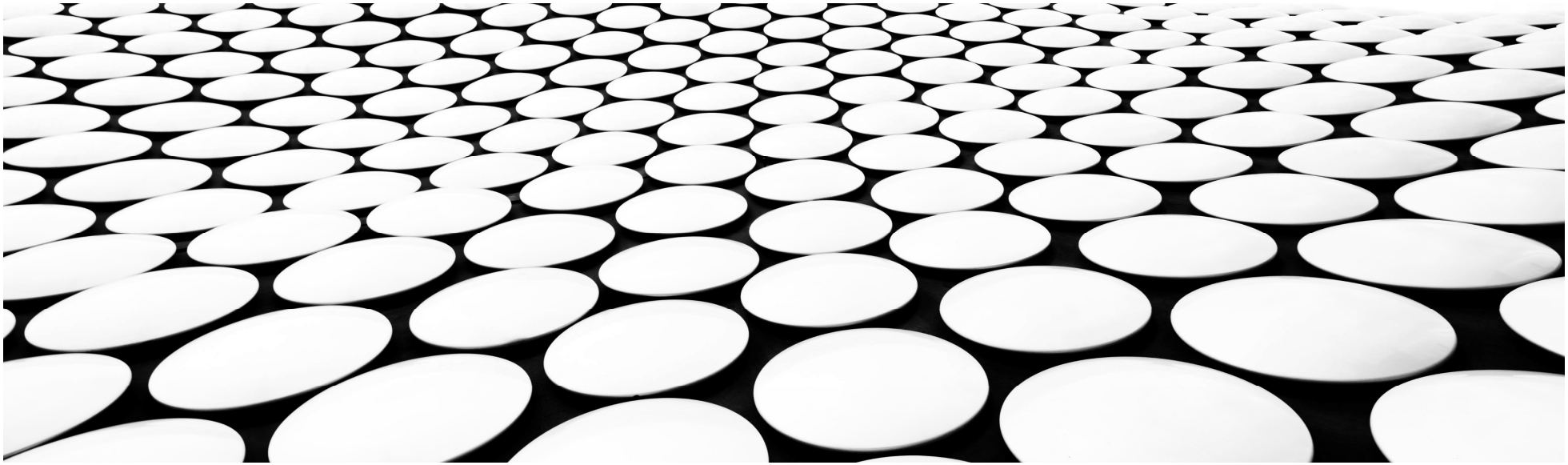
CAPSTONE COURSERA - APPLIED DATA SCIENCE BATTLE OF THE NEIGHBOURHOODS

FINAL PROJECT PRESENTATION



Colin W. Wilsnagh – 06 May 2020

INTRODUCTION

- A large multinational company with HQ in Houston Texas has personnel moving globally between offices and geographical locations
- The Human Resources department constantly has enquiries from personnel who are relocating as to which areas are best to relocate to. These include existing as well a new people
- The people are looking for on direction as to which neighbourhoods have low crime levels but also have a variety of facilities and venues in close proximity

DATA SOURCES

There are 3 major data types used in the model namely Neighbourhood data, Crime Data and Venue / Amenity data

- The neighbourhood coordinate data mapped to area Zip codes was sourced from https://www.zip-codes.com/zip-code-database.asp
- The zip code names for the neighbourhoods comes from the website https://data.mongabay.com/igapo/zip_codes/metropolitan-areas/metro-alpha/Houston%20(TX)1.html and extracted using Beautiful Soup with an HTML parser to render the dataframe
- Crime data was sourced from the Houston Police Department's NIRBS system https://www.houstontx.gov/police/cs/xls/NIBRSPublicView.Jan1-Mar31-2020-FINAL.xlsx
- Crime category groupings came from the FBI's Unified Crime Reporting system https://www.fbi.gov/file-repository/ucr/ucr-srs-user-manual-v1.pdf/view) as this is used by the NIBRS system
- Venue and facility data is sourced from the Foursquare application https://foursquare.com/

DATA PREPARATION

Neighbourhood Data

- Data wasread into a dataframe
- Unwanted rows and columns dropped
- Columns renamed as required

	zip	type	decommissioned	primary_city	acceptable_cities	unacceptable_cities	state	county	timezone	area_codes	world_re
0	501	UNIQUE	0	Holtsville	NaN	I R S Service Center	NY	Suffolk County	America/New_York	631.0	
1	544	UNIQUE	0	Holtsville	NaN	Irs Service Center	NY	Suffolk County	America/New_York	631.0	
2	601	STANDARD	0	Adjuntas	NaN	Colinas Del Gigante, Jard De Adjuntas, Urb San...	PR	Adjuntas Municipio	America/Puerto_Rico	787939.0	
3	602	STANDARD	0	Aguada	NaN	Alts De Aguada, Bo Guaniquilla, Comunidad Las ...	PR	Aguada Municipio	America/Puerto_Rico	787939.0	

- Data manipulated into a form that was required

	Zip	State	Latitude	Longitude	Population2015
0	73301	TX	30.26	-97.74	433
1	73344	TX	30.26	-97.74	0
2	73960	TX	36.49	-101.78	0
3	75001	TX	32.96	-96.83	14180
4	75002	TX	33.10	-96.66	65530

DATA PREPARATION

Zip Codes and Area Names data

- Data read from website using Beautiful Soup
- Parsed into dataframe with HTML parser
- Unwanted rows and columns dropped
- Column (0) split to get data correctly
- Columns renamed as required

	0	1	2	3	4	5
0	77079 Addicks	713/281/832	Harris County	Texas - TX	Houston, TX (3360)	SMSA
1	77084 Addicks Barker	713/281/832	Harris County	Texas - TX	Houston, TX (3360)	SMSA
2	77039 Aldine	713/281/832	Harris County	Texas - TX	Houston, TX (3360)	SMSA
3	77411 Alief	713/281/832	Harris County	Texas - TX	Houston, TX (3360)	SMSA
4	77575 Ames	936	Liberty County	Texas - TX	Houston, TX (3360)	SMSA

- Data manipulated into a form that was required namely Zip code and Neighbourhood names

	Zip	Neighborhood
0	77001	Houston
1	77002	Clutch City,Houston
2	77003	Houston
3	77004	Houston
4	77005	Houston,Southside Place,West University Place

DATA PREPARATION

Zip Codes, Area Names and Coordinate data

- The dataframes were merged to give a single dataframe with the desired data

	Zip	Neighborhood	Latitude	Longitude	Population2015
0	77001	Houston	29.76	-95.38	575
1	77002	Clutch City,Houston	29.76	-95.37	5850
2	77003	Houston	29.75	-95.35	8760
3	77004	Houston	29.72	-95.36	21460
4	77005	Houston,Southside Place,West University Place	29.72	-95.42	23920

DATA PREPARATION

Crime Data

- The data was read into a dataframe
- Data was manipulated dropping unnecessary columns and rows
- Columns were renamed and order adjusted

	Incident	Occurrence\nDate	Occurrence\nHour	NIBRS\nClass	NIBRSDescription	Offense\nCount	Beat	Premise	Block Range	StreetName	Street\n
0	8220	2020-01-01	0	23G	Theft of motor vehicle parts or accessory	1	8C50	Residence, Home (Includes Apartment)	9311	BELLA PINE	
1	18920	2020-01-01	0	13A	Aggravated Assault	1	17E30	Residence, Home (Includes Apartment)	8701	GUSTINE	
2	23020	2020-01-01	0	290	Destruction, damage, vandalism	1	11H30	Residence, Home (Includes Apartment)	8064	LENORE	
3	24120	2020-01-01	0	13B	Simple assault	1	17E10	Residence, Home (Includes Apartment)	5930	DASHWOOD	
4	27120	2020-01-01	0	290	Destruction, damage, vandalism	1	14D30	Residence, Home (Includes Apartment)	5218	KENILWOOD	

- Crime category data read into a dataframe
- This will allow the crime data to be consolidated according to major crime categories per zip code

	Category	Offence Group	NIBRS Description	NIBRS Class
0	A	Arson	Arson	200
1	A	Assault Offenses	Assault Offenses	13
2	A	Assault Offenses	Aggravated Assault	13A
3	A	Assault Offenses	Simple Assault	13B
4	A	Assault Offenses	Intimidation	13C

DATA PREPARATION

Crime Data

- The data dataframes were merged
- Data was manipulated dropping unnecessary columns and rows
- Columns were renamed and order adjusted

	Incident	NIBRS Class	NIBRS Description_x	Offense Count	Zip	Category	Offence Group	NIBRS Description_y	Truecc
0	8220	23G	Theft of motor vehicle parts or accessory	1	77078	A	Larceny/Theft Offenses	Theft of Motor Vehicle Parts or Accessories	both
1	18920	13A	Aggravated Assault	1	77031	A	Assault Offenses	Aggravated Assault	both
2	23020	290	Destruction, damage, vandalism	1	77017	A	Destruction/Damage /Vandalism of Property	Destruction/Damage /Vandalism of Property	both
3	24120	13B	Simple assault	1	77081	A	Assault Offenses	Simple Assault	both
4	27120	290	Destruction, damage, vandalism	1	77033	A	Destruction/Damage /Vandalism of Property	Destruction/Damage /Vandalism of Property	both

- Crime incidents pivoted so that all crimes per zip code represented by a single row

	Zip	Animal Cruelty	Arson	Assault Offenses	Bribery	Burglary/B&E	Counterfeiting/Forgery	Destruction of Property	Drug/Narcotics	Embezzlement	Extortion/Blackmail
0	77002	1	2	255	1	35	7	100	73	1	0
1	77003	3	1	99	0	26	3	37	20	1	1
2	77004	0	2	314	0	91	12	109	97	2	0
3	77005	0	0	32	0	7	1	20	1	0	1
4	77006	0	3	127	0	61	9	63	26	1	0

- Data then merged to get the coordinate and neighbourhood information

	Zip	Neighborhood	Latitude	Longitude	Population2015	Animal Cruelty	Arson	Assault Offenses	Bribery	Burglary/B&E	Counterfeiting/Forgery	Destructio of Propert
0	77002	Clutch City,Houston	29.76	-95.37	5850.0	1	2	255	1	35	7	10
1	77003	Houston	29.75	-95.35	8760.0	3	1	99	0	26	3	3
2	77004	Houston	29.72	-95.36	21460.0	0	2	314	0	91	12	10
3	77005	Houston,Southside Place,West University Place	29.72	-95.42	23920.0	0	0	32	0	7	1	2
4	77006	Houston	29.74	-95.39	17710.0	0	3	127	0	61	9	6

METHODOLOGY

Crime Data summarised Statistically

- This returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crime.

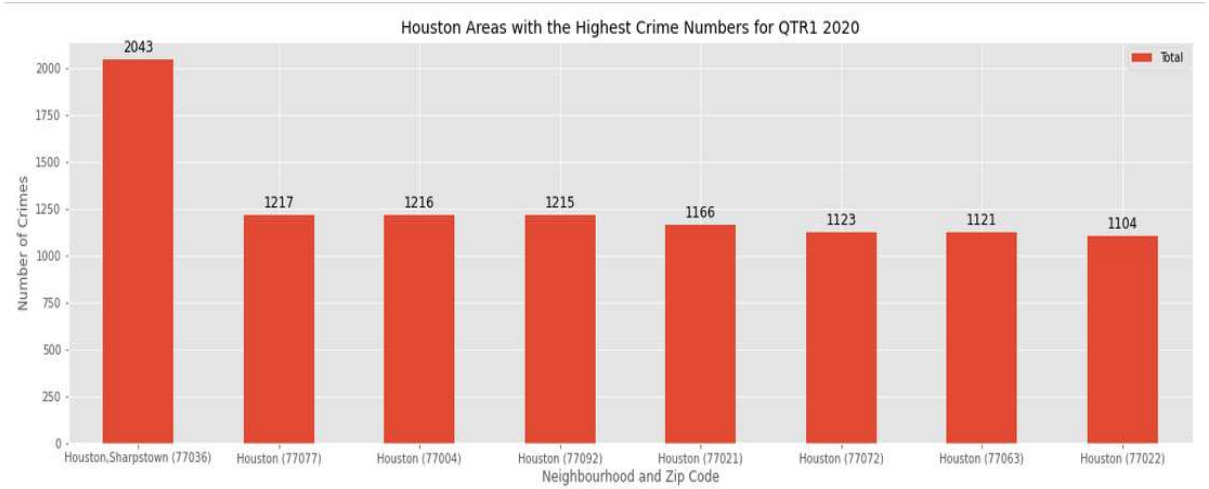
	Latitude	Longitude	Population2015	Animal Cruelty	Arson	Assault Offenses	Bribery	Burglary/B&E	Counterfeiting/Forgery	Destruction of Property
count	133.000000	133.000000	133.000000	133.000000	133.000000	133.000000	133.000000	133.000000	133.000000	133.000000
mean	29.778571	-95.409549	32118.496241	0.578947	0.804511	113.210526	0.052632	27.578947	4.030075	38.684211
std	0.147711	0.189392	19163.591762	0.897898	1.189975	123.437449	0.224141	30.620901	4.643242	40.271685
min	29.390000	-95.920000	730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	29.680000	-95.520000	18820.000000	0.000000	0.000000	4.000000	0.000000	1.000000	0.000000	1.000000
50%	29.750000	-95.420000	29830.000000	0.000000	0.000000	71.000000	0.000000	19.000000	2.000000	26.000000
75%	29.870000	-95.280000	38070.000000	1.000000	1.000000	199.000000	0.000000	45.000000	7.000000	66.000000
max	30.150000	-94.960000	109280.000000	4.000000	5.000000	561.000000	1.000000	153.000000	21.000000	219.000000

METHODOLOGY

Neighbourhoods with Highest crime charted

- Data sorted according to Crime totals and the Top 8 neighbourhoods represented graphically

Zip	Neighborhood	Latitude	Longitude	Population 2015	Total Crime	ZipNeighbourhood
77036	Houston,Sharpstown	29.7	-95.53	59100	2043	Houston,Sharpstown (77036)
77077	Houston	29.75	-95.62	51970	1217	Houston (77077)
77004	Houston	29.72	-95.36	21460	1216	Houston (77004)
77092	Houston	29.83	-95.47	31450	1215	Houston (77092)
77021	Houston	29.7	-95.36	22520	1166	Houston (77021)
77072	Houston	29.7	-95.58	53090	1123	Houston (77072)
77063	Houston	29.74	-95.52	29830	1121	Houston (77063)
77022	Houston	29.83	-95.38	25150	1104	Houston (77022)

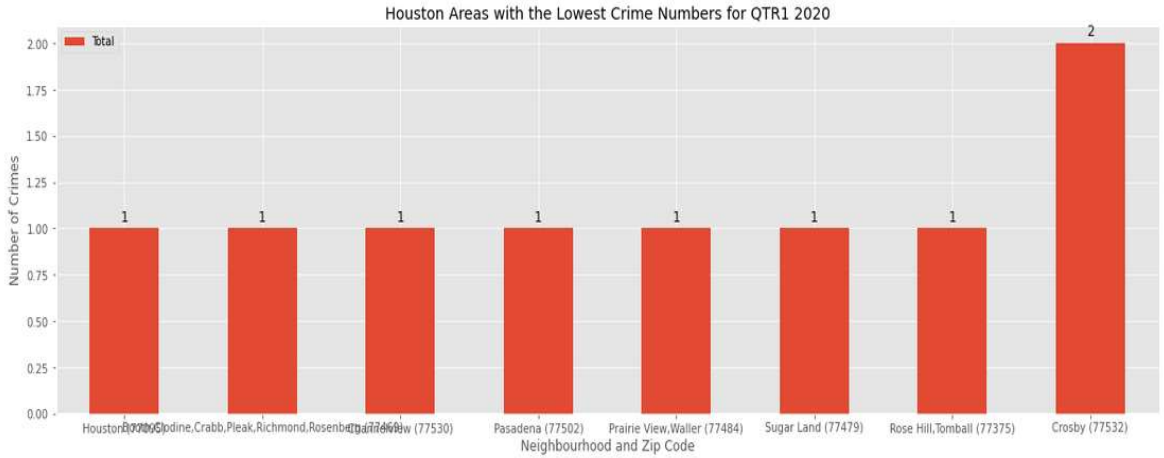


METHODOLOGY

Neighbourhoods with Lowest crime charted

- Data sorted according to Crime totals and the Bottom 8 neighbourhoods represented graphically

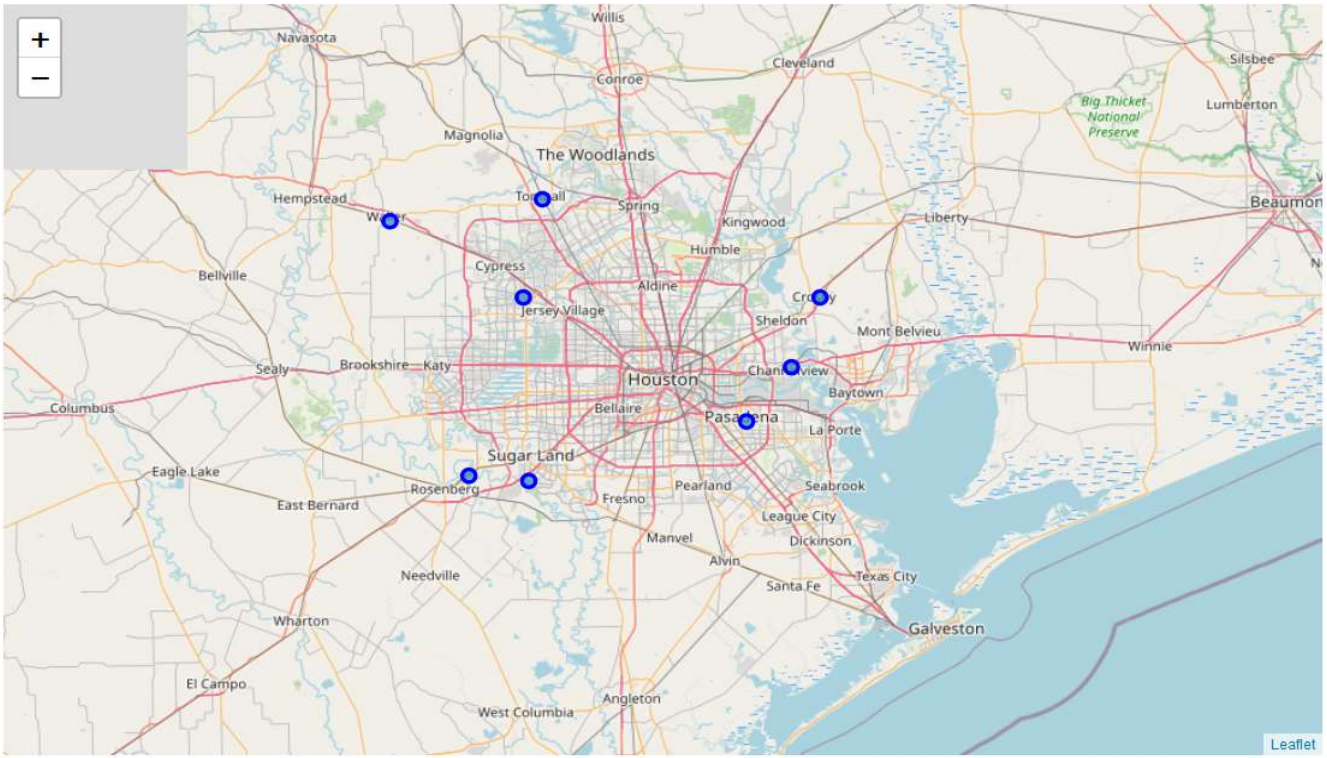
Zip	Neighborhood	Latitude	Longitude	Population 2015	Total Crime	ZipNeighbourhood
77095	Houston	29.91	-95.65	67710	1	Houston (77095)
77469	Booth,Clodine,Crabb,Pleak,Richmond, Rosenberg	29.58	-95.76	43300	1	Booth,Clodine,Crabb,Pleak,Richmond,Rosenberg (77469)
77530	Channelview	29.78	-95.11	31680	1	Channelview (77530)
77502	Pasadena	29.68	-95.2	36960	1	Pasadena (77502)
77484	Prairie View,Waller	30.05	-95.92	11510	1	Prairie View,Waller (77484)
77479	Sugar Land	29.57	-95.64	85720	1	Sugar Land (77479)
77375	Rose Hill,Tomball	30.09	-95.61	48470	1	Rose Hill,Tomball (77375)
77532	Crosby	29.91	-95.05	27850	2	Crosby (77532)



METHODOLOGY

Neighbourhoods with Lowest crime charted

- Lowest Crime neighbourhoods also mapped graphically



VENUE ANALYSIS

Venues selected according to neighbourhoods with lowest crime

Foursquare API calls made for venue data within 2500 meters each of the neighbourhood centers

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Houston	29.91	-95.65	Chick-fil-A	29.901682	-95.633868	Fast Food Restaurant
1	Houston	29.91	-95.65	Mod Pizza	29.903473	-95.632576	Pizza Place
2	Houston	29.91	-95.65	Catfish Station	29.915821	-95.630031	Cajun / Creole Restaurant
3	Houston	29.91	-95.65	Bonsai Fusion Japanese Steakhouse	29.904144	-95.631484	Japanese Restaurant
4	Houston	29.91	-95.65	Langham Creek Family YMCA	29.898865	-95.667855	Gym

VENUE ANALYSIS

On-hot encoding conducted on the venues data

This will enable K-Means Clustering to be run on the data

	Neighborhood	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	Bakery	Bank	Bar	Big Box Store	Bike Shop	Boat or Ferry	Breakfast Spot	Brewery
0	Houston	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Houston	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Houston	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Houston	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Houston	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Data grouped according to venue categories per neighbourhood

	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	
0	Booth,Clodine,Crabb,Pleak,Richmond,Rosenberg	Fast Food Restaurant	Gym / Fitness Center	Gas Station	Mexican Restaurant	Pizza Place	Convenience Store	Discount Store	Automotive Shop	
1	Channelview	Fast Food Restaurant	Gas Station	Fried Chicken Joint	Boat or Ferry	Mexican Restaurant	Hotel	Burger Joint	Convenience Store	
2	Crosby	Fast Food Restaurant	Pizza Place	Cajun / Creole Restaurant	Discount Store	Gas Station	Pharmacy	Mexican Restaurant	Taco Place	B
3	Houston	Pizza Place	Fast Food Restaurant	Automotive Shop	Pharmacy	Sandwich Place	Burger Joint	Cosmetics Shop	Bar	In
4	Pasadena	Fast Food Restaurant	Mexican Restaurant	Discount Store	Pharmacy	Pizza Place	Wings Joint	Fried Chicken Joint	Grocery Store	

VENUE ANALYSIS

K-Means Clustering run on the data

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
91	Houston	29.91	-95.65	1	Pizza Place	Fast Food Restaurant	Automotive Shop	Pharmacy	Sandwich Place	Burger Joint
113	Booth,Clodine,Crabb,Pleak,Richmond,Rosenberg	29.58	-95.76	1	Fast Food Restaurant	Gym / Fitness Center	Gas Station	Mexican Restaurant	Pizza Place	Convenience Store
126	Channelview	29.78	-95.11	0	Fast Food Restaurant	Gas Station	Fried Chicken Joint	Boat or Ferry	Mexican Restaurant	Hotel
121	Pasadena	29.68	-95.20	1	Fast Food Restaurant	Mexican Restaurant	Discount Store	Pharmacy	Pizza Place	Wings Joint
118	Prairie View,Waller	30.05	-95.92	2	Fast Food Restaurant	Mexican Restaurant	Discount Store	Pizza Place	Pharmacy	Bakery

CLUSTER ANALYSIS

Cluster 0

The first cluster (0) consists of one neighbourhood and has a mix of Fast food and restaurants for eating. It has a Gas Station, Hotel and some Stores.

Neighborhood	Latitude	Longitude	Cluster Labels	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
Channelview	29.78	-95.11	0	Fast Food Restaurant	Gas Station	Fried Chicken Joint	Boat or Ferry	Mexican Restaurant	Hotel	Burger Joint	Convenience Store	Discount Store	Pizza Place

CLUSTER ANALYSIS

Cluster 1

The second cluster (1) consists of the neighbourhoods Houston(zip code 77095); Booth, Clodine, Crabb, Pleak, Richmond, Rosenberg (Zip Code 77469); Pasedena (Zip 77502);Rose Hill,Tomball (Zip code 77375). The venues provide a good mix of Fast foods, Restaurants, Gas Stations and other Stores.

.

Neighborhood	Latitude	Longitude	Cluster Labels	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 Mo Commo Venu
Houston	29.91	-95.65	1	Pizza Place	Fast Food Restaurant	Automotive Shop	Pharmacy	Sandwich Place	Burger Joint	Cosmeti Sho
Booth,Clodine,Crabb,Pleak,Richmond,Rosenberg	29.58	-95.76	1	Fast Food Restaurant	Gym / Fitness Center	Gas Station	Mexican Restaurant	Pizza Place	Convenience Store	Discou Sto
Pasadena	29.68	-95.20	1	Fast Food Restaurant	Mexican Restaurant	Discount Store	Pharmacy	Pizza Place	Wings Joint	Frie Chicke Joi
Rose Hill,Tomball	30.09	-95.61	1	Fast Food Restaurant	Pizza Place	Mexican Restaurant	American Restaurant	Fried Chicken Joint	Sandwich Place	Mobi Phor Sho

CLUSTER ANALYSIS

Cluster 2

- The third Cluster (2) is linked to one Neighbourhood Prairie View, Waller (Zip 77484) and has groupings of Fast food, Pharmacy, Bakery and Gas Station.

Neighborhood	Latitude	Longitude	Cluster Labels	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
Prairie View,Waller	30.05	-95.92	2	Fast Food Restaurant	Mexican Restaurant	Discount Store	Pizza Place	Pharmacy	Bakery	Breakfast Spot	Fried Chicken Joint	Sandwich Place	Gas Station

Cluster 3

- The fourth Cluster (3) has a mix of Food and Restaurant venues, Salon, Pharmacy and other Stores.

Neighborhood	Latitude	Longitude	Cluster Labels	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
Crosby	29.91	-95.05	3	Fast Food Restaurant	Pizza Place	Cajun / Creole Restaurant	Discount Store	Gas Station	Pharmacy	Mexican Restaurant	Taco Place	Salon / Barbershop	Burger Joint

CLUSTER ANALYSIS

Cluster 4

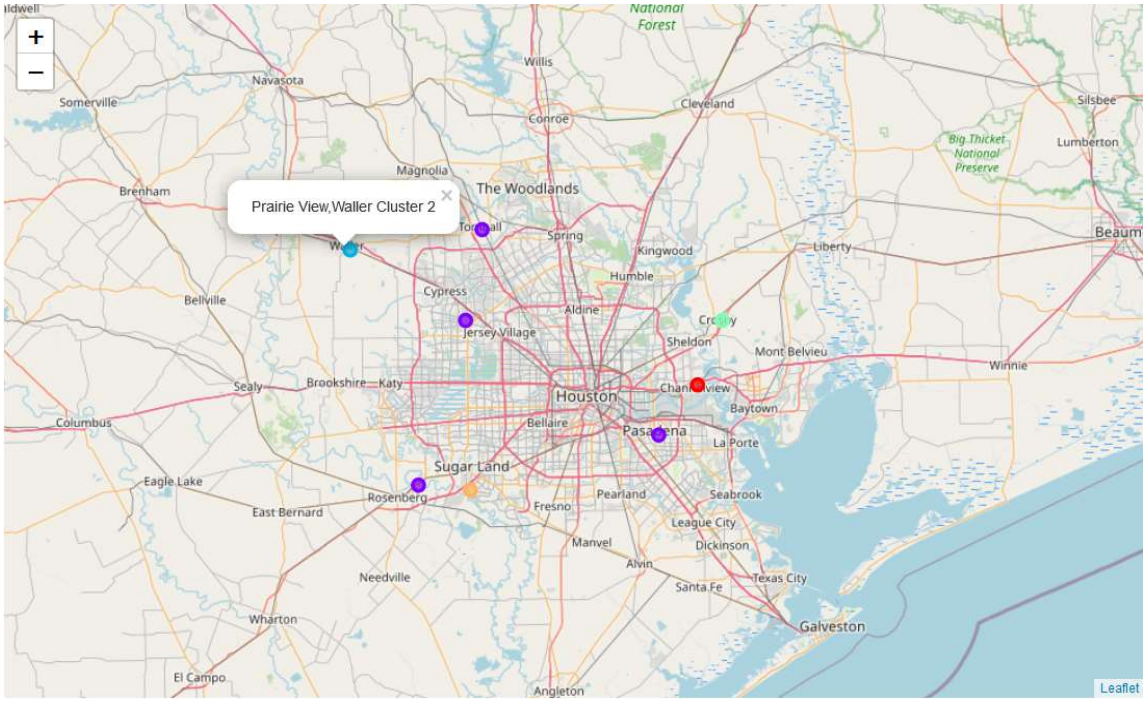
- The final cluster (4) in the analysis has a Tea Shop, Park Home Service and a River.

eighborhood	Latitude	Longitude	Cluster Labels	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
Sugar Land	29.57	-95.64	4	Bubble Tea Shop	Fried Chicken Joint	Ice Cream Shop	Cosmetics Shop	Park	Mediterranean Restaurant	Intersection	Sandwich Place	Home Service	River

CLUSTER ANALYSIS

Clusters mapped

- As can be seen by the map groupings the neighbourhoods and clusters are fairly scattered around the centre of Houston with some having River proximity.



Cluster 0 - Red
Cluster 1 - Purple
Cluster 2 – Light Blue
Cluster 3 – Jade Green
Cluster 4 - Orange

DISCUSSION AND OBSERVATIONS

- Given that the brief for this model was to attempt to analyse crime levels and select neighbourhoods with low acceptable crime levels, the model was able to profile these and map out the various areas according to crime level.
- The model is also able to profile the venues selected in the said neighbourhoods with acceptable crime levels and profile these to indicate what is available at these locations in terms of venues etc.
- The area of Pasedena would probably suit the requirements for a family seeking a neighbourhood to relocate to.
- It has a good mix of facilities and also has a High School. It is about 25 km's from the Houston City centre but has extremely low crime levels with only one crime report (larceny) being reported in the first 3 months of 2020.

CONCLUSION

- The model proves that it is able to satisfy the original question that was posed by the Human Resources department of the company.
- The user will be able to get a view of crime levels around a central Head Office and then make an informed decision on where to relocate to based on the facilities available in the neighbourhoods.
- Further profiling of the Crime data could also have been undertaken using K-Means for clustering purposes as the data had the correct format once prepared. It was decided that this insight would add no further value to the required outcome of this project, and was not undertaken.
- Future developments and enhancements for this application could be to link it to properties available for rent or sale in the selected areas and profile them based on the user requirements and affordability to recommend suitable accommodation to the user.
- The application would be able to be rolled out to the other offices globally once the correct data has been sourced.

THANK YOU