Capstone Coursera Applied Data Science Battle of the neighbourhoods

Final Project- Report Submission

MAY 6, 2020

COLIN W.WILSNAGH

Contents

Introduction

Given that we now live in a "global village", people are highly mobile and relocate extremely easily given the potential for career advancement, better prospects for themselves or family, crises in their country, war, genocide etc.

Large multi-national companies also need talented individuals to be mobile in order to satisfy resource requirements for various projects or work situations.

The overwhelming question that comes to mind of the person that is relocating is, in which area of the major location they are moving to, do they settle.

Business Problem Statement

This question was being posed on numerous occasions to the Human Resources division of a major global consulting company which prompted them to embark on trying to solve it, by examining the feasibility of building an application to model the following:

- What are the crime profiles of the various neighbourhoods surrounding their company head offices in the different locations globally within a given acceptable travel radius?
- Where acceptable levels of crime are found in the said neighbourhoods, what facilities and amenities are available to their prospective employees in a given radius?

The scenario posed was that a person, either an existing employee or a new recruit, would need to relocate to one of the company's offices in another city. The system should profile the crime levels in the local neighbourhoods of the target city within the given travel radius of the office and present the neighbourhoods with the lowest crime levels.

Once the potential neighbourhood(s) for relocation is selected, the system should then model the areas facilities, venues and amenities and present the most acceptable areas within the profile.

This could be an iterative process based on user choice, until the acceptable neighbourhood is shortlisted. The user could then examine the suitable housing choices available for relocation.

Development Approach

After deliberation it was decided that the pilot project would focus on the Houston Texas (USA) office. This is the company's Head Office and the office that initiated the project request.

Once the application has been proved and requirements met, the system would then incorporate further offices in the rollout plan. These would be determined given priority, volume of user's information requests and available data.

The program would also be modular so that even if Crime data was not available, the application would still be capable of providing the available venues and amenities in the given areas.

Data to be used

An initial overview of Houston was sought to get a better idea of the city in question.

Wikipedia states that "Houston is the most populous city in the U.S. state of Texas, fourth most populous city in the United States, most populous city in the Southern United States, as well as the sixth most populous in North America, with an estimated 2018 population of 2,325,502. Located in Southeast Texas near Galveston Bay and the Gulf of Mexico, it is the seat of Harris County and the principal city of the Greater Houston metropolitan area".

The Neighbourhood Scout (https://www.neighborhoodscout.com/tx/houston/crime) states that "With a **crime rate** of 54 per one thousand residents, **Houston** has one of the highest **crime rates** in America compared to all communities of all sizes - from the smallest towns to the very largest cities. One's chance of becoming a victim of either violent or property **crime** here is one in 19".

Given that Crime is a major factor in neighbourhood choice, data that indicates crime levels per neighbourhood areas chosen, will be of paramount importance. The crime statistics need to be representative of the areas and dated as current as possible so as to represent the situation as at the time reported.

The crime incidents will need to have location data associated with it and be reported in such a way as to be at a granular level yet able to be easily rolled up into major crime categories.

The facilities and amenities that are in the various locations need to be representative of the venue categories such as Schools, College & University, Shops, Restaurants, Malls, Health and Fitness facilities, Sports Stadiums etc.

Postal or Zip code data will be needed to map the neighbourhoods to their location coordinates as this provides the correct level of granularity for the analysis.

Data Sources

A major requirement for the system was that data be readily available electronically so that system updates would be seamless and maintainability simplified.

Data sources via the internet was therefore deemed to be a priority.

Co-ordinate and Zip code data

Various sources for Zip code and coordinate data were found to be available and the one chosen was the Zip code database as this provided data for the entire USA area.

https://www.zip-codes.com/zip-code-database.asp

They offer data for sale with various data sets but also a free version which was used for this exercise.

File Data

Field Name	Description
zip	Area Zip Code
type	Type of code
decommissioned	Decommissioned code or in use
primary_city	Primary City
acceptable_cities	Acceptable area name
unacceptable_cities	Unacceptable area name
state	State
county	County in the state
timezone	Time zone

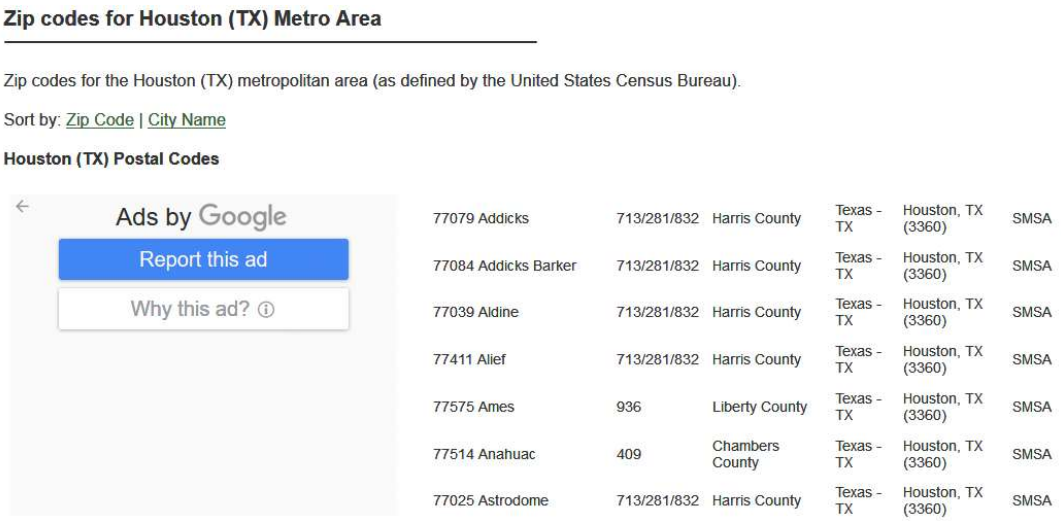
area_codes	Telephone area codes for the zip code
world_region	World region
country	County (USA)
latitude	Latitude
longitude	Longitude
irs_estimated_population_2015	Estimated population for 2015 as estimated by the Revenue department

Area data

It was deemed that for the data to be meaningful area data need to be at a granular level. The data in the zip code file had the mappings for Zip Codes and Coordinates but the individual area names for the neighbourhoods were not available or missing.

A website was found that had these mappings and was used for the additional layer of detail.

https://data.mongabay.com/igapo/zip_codes/metropolitan-areas/metro-alpha/Houston%20(TX)1.html



Zip codes for Houston (TX) Metro Area

Zip codes for the Houston (TX) metropolitan area (as defined by the United States Census Bureau).

Sort by: Zip Code | City Name

Houston (TX) Postal Codes

77079 Addicks	713/281/832	Harris County	Texas - TX	Houston, TX (3360)	SMSA
77084 Addicks Barker	713/281/832	Harris County	Texas - TX	Houston, TX (3360)	SMSA
77039 Aldine	713/281/832	Harris County	Texas - TX	Houston, TX (3360)	SMSA
77411 Alief	713/281/832	Harris County	Texas - TX	Houston, TX (3360)	SMSA
77575 Ames	936	Liberty County	Texas - TX	Houston, TX (3360)	SMSA
77514 Anahuac	409	Chambers County	Texas - TX	Houston, TX (3360)	SMSA
77025 Astrodome	713/281/832	Harris County	Texas - TX	Houston, TX (3360)	SMSA

Crime data

Crime data needs to be up to date in order to be relevant. The USA has a National Incident-Based Reporting System system (NIBRS) for Crime reporting and fortunately the data sets are both easily available and very relevant with the latest dataset which we used being for the period January to March 2020. (refer https://www.houstontx.gov/police/cs/NIBRS_Public_Data_Dictionary_050919.pdf).

Crime Statistics

Crime Statistic data is available in various formats (csv, xls etc) from:

https://www.houstontx.gov/police/cs/xls/NIBRSPublicView.Jan1-Mar31-2020-FINAL.xlsx

File Data

This data comprises the following data fields:

Field Name	Description
Incident	Incident number
Occurrence Date	Data of occurrence
Occurrence Hour	Time of occurrence
NIBRS Class	National Incident Reporting System crime class
NIBRSDescription	National Incident Reporting System crime description
Offense Count	Number of offenses
Beat	Police beat or route
Premise	Type of premises of the crime
Block Range	Block range
StreetName	Street Name
Street Type	Type of street
Suffix	Suffix
ZIP Code	Zip code of crime incident

Crime Groupings

In order to roll the data up into major crime segments, the group coding as used by NIBRS is according to the FBI's Uniform Crime Reporting system (https://www.fbi.gov/file-repository/ucr/a-word-about-ucr-data.pdf/view) and was used.

The final file had to be manually prepared as the available file was in pdf format and also found to be incomplete. (https://www.fbi.gov/file-repository/ucr/ucr-srs-user-manual-v1.pdf/view)

Group "A" Offenses

NIBRS OFFENSES	NIBRS CODES	NIBRS OFFENSES	NIBRS CODES
Arson	200	**Human Trafficking**	
Assault Offenses		-Commercial Sex Acts	64A
-Aggravated Assault	13A	-Involuntary Servitude	64B
-Simple Assault	13B	**Kidnapping/Abduction**	100
-Intimidation	13C	**Larceny/Theft Offenses**	
Bribery	510	-Pocket Picking	23A
Burglary/B&E	220	-Purse Snatching	23B
Counterfeiting/Forgery	250	-Shoplifting	23C
Destruction/Damage/Vandalism of Property	290	-Theft from Building	23D
Drug/Narcotic Offenses		-Theft from Coin-Operated Machine or Device	23E
-Drug/Narcotic Violations	35A	-Theft from Motor Vehicle	23F
-Drug/Narcotic Equip. Violations	35B	-Theft of Motor Vehicle Parts or Accessories	23G
Embezzlement	270	-All Other Larceny	23H
Extortion/Blackmail	210	**Motor Vehicle Theft**	240
Fraud Offenses		**Pornography/Obscene Material**	370
-False Pretenses/Swindle/ Confidence Games	26A	**Prostitution Offenses**	
-Credit Card/Automatic Teller Machine Fraud	26B	-Prostitution	40A
-Impersonation	26C	-Assisting or Promoting Prostitution	40B
-Welfare Fraud	26D	-Purchasing Prostitution	40C
-Wire Fraud	26E	**Robbery**	120
Gambling Offenses		**Sex Offenses (Forcible)**	
-Betting/Wagering	39A	-Forcible Rape	11A
-Operating/Promoting/ Assisting Gambling	39B	-Forcible Sodomy	11B
-Gambling Equip. Violations	39C	-Sexual Assault with An Object	11C
-Sports Tampering	39D	-Forcible Fondling	11D
Homicide Offenses		**Sex Offenses (Non-Forcible)**	
-Murder/Non-Negligent Manslaughter	09A	-Incest	36A
-Negligent Manslaughter	09B	-Statutory Rape	36B
-Justifiable Homicide	09C	**Stolen Property Offenses**	280
		Weapon Law Violations	520

Group "B" Offenses

Group B's MUST have an arrest to be NIBRS Reportable

NIBRS OFFENSES	NIBRS CODES	NIBRS OFFENSES	NIBRS CODES
Bad Checks	90A	**Family Offenses, Non-Violent**	90F
Curfew/Loitering/Vagrancy Violations	90B	**Liquor Law Violations**	90G
Disorderly Conduct	90C	**Peeping Tom**	90H
Driving Under the Influence	90D	**Trespassing**	90J
Drunkenness	90E	**All Other Offenses**	90Z

Source: Association of State Uniform Crime Reporting Programs (ASUCRP). Accessed on June 6, 2014.

On examination of the NIBRS Crime data it was found that some crime types were missing e.g. Animal Cruelty (Code 720), Hacking/Computer Invasion (Code 26G) etc. These were then added to the Crime Categories data file for completeness.

The Crime types are grouped into 2 major groups, i.e. Group A are all serious crimes and Group B are more petty crimes. We only worked with Group A crimes for this model and dropped the Group B types during data preparation.

Venue and Amenity Data

In order to get the various listings of venues, facility and amenities available in each area or neighbourhood, data will be sourced using Foursquare via API requests. Data could also be obtained from other sources such as Google etc. but for the purposes of this exercise we will limit the data requests to the Foursquare app given convenience and the fact that we already have an account setup.

https://foursquare.com/

Data Preparation

Before any analysis could be carried out the data had to be prepared so as to be in the desired format and have the correct subsets of columns and rows available in order to deliver the requisite results.

Unnecessary data would need to be dropped from the dataset and the columns re-arranged and renamed to have meaning and aid in the analysis.

Co-ordinate data

The preparation of the Coordinate data was achieved by reading in the raw Excel file ('Zip_code_database.xlsx') from the source website.

	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	

The dataset was then manipulated and unwanted columns were dropped as well as rows that did not pertain to the State of Texas until the final data set required was derived containing only Texas data.

	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

Zip Codes mapped to their Areas / Neighbourhoods

The website page ('https://data.mongabay.com/igapo/zip_codes/metropolitan-areas/metro-alpha/Houston%20(TX)1.html') was scraped using Beautiful Soup and rendered into a data frame using an HTML parser.

	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

The data was then manipulated, columns split and renamed to deliver the data in the required format.

Some rows had different area names for the same zip code and these were then combined to have a single zip code with the names concatenated together *(see zip code 77005 below*).

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

Final Zip Code, Coordinate and Area Name Dataset

The 2 datasets were then merged to get a final dataset containing Zip code, coordinate and area / neighbourhood name mappings for the City of Houston. *(Note that when merging data, the key fields need to be checked to ensure that they are of the same data type or the merge will not be successful)*.

	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

Crime Data

Crime incident data was picked up from the Houston Tx police website in Excel format and read into a dataframe (https://www.houstontx.gov/police/cs/xls/NIBRSPublicView.Jan1-Mar31-2020-FINAL.xlsx).

	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	

The data was then manipulated with columns renamed and unwanted columns dropped.

	Incident	NIBRS Class	NIBRS Description	Offense Count	Zip
0	8220	23G	Theft of motor vehicle parts or accessory	1	77078
1	18920	13A	Aggravated Assault	1	77031
2	23020	290	Destruction, damage, vandalism	1	77017
3	24120	13B	Simple assault	1	77081
4	27120	290	Destruction, damage, vandalism	1	77033

Where the crime data was not associated with a Zip code, those rows were also dropped as we would not be able to map it to a neighbourhood.

The Crime Categories data file ('Crime_cats.xlsx') was then read into another dataframe. This would allow the data to be rolled up from a granular level into major crime categories per zip code and allow the data to be totalled.

	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

Once the key fields (NIBRS Class) of the Crime and the Crime Categories data were checked for correct data types, the data was merged.

	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

Category B crimes were dropped as they represent petty crime. The data was then grouped by Zip code and the crimes per category totalled for each Zip Code.

	Zip	Offence Group	Offense Count
0	77002	Animal Cruelty	1
1	77002	Arson	2
2	77002	Assault Offenses	255
3	77002	Bribery	1
4	77002	Burglary/B&E	35

The data was then pivoted to represent all crime per category for each zip code in 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

The dataframe with the coordinates and neighbourhood names was then merged with the Crime data to give a final Crime dataset for analysis.

	Zip	Neighborhood	Latitude	Longitude	Population2015	Animal Cruelty	Arson	Assault Offenses	Bribery	Burglary/B&E	Counterfeiting/Forgery	Destructio of Properi
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 Followed

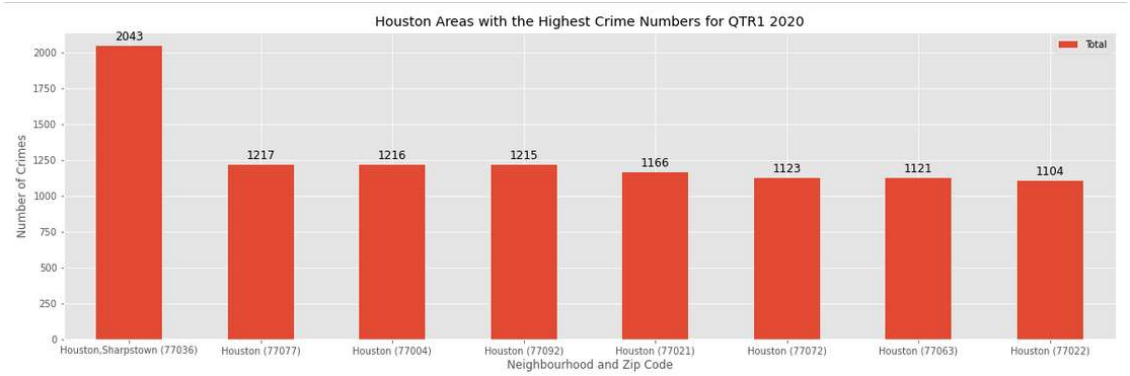
Crime data Analysis

The describe function in python is used to get statistics of the Houston crime data. 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

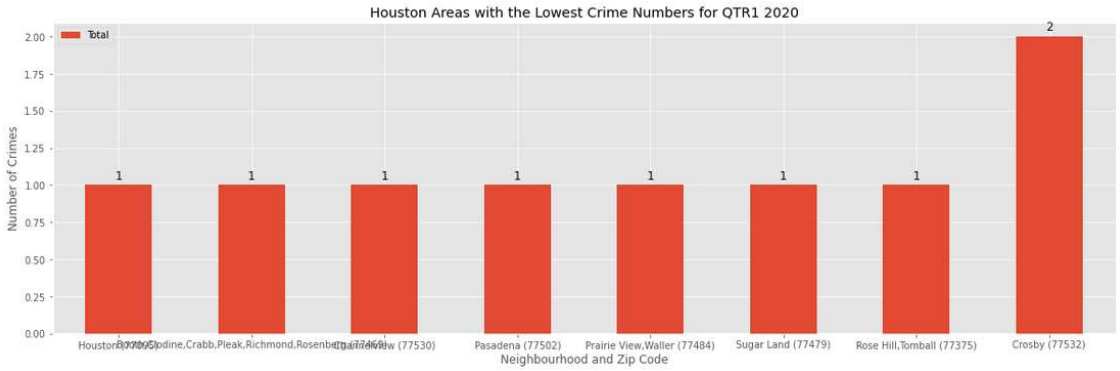
In order to get a visual representation of the top and bottom crime areas, these were then charted with the 8 highest crime ridden areas represented below (as Houston has multiple zip codes associated with it, we included them in the labels for clarity).

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)

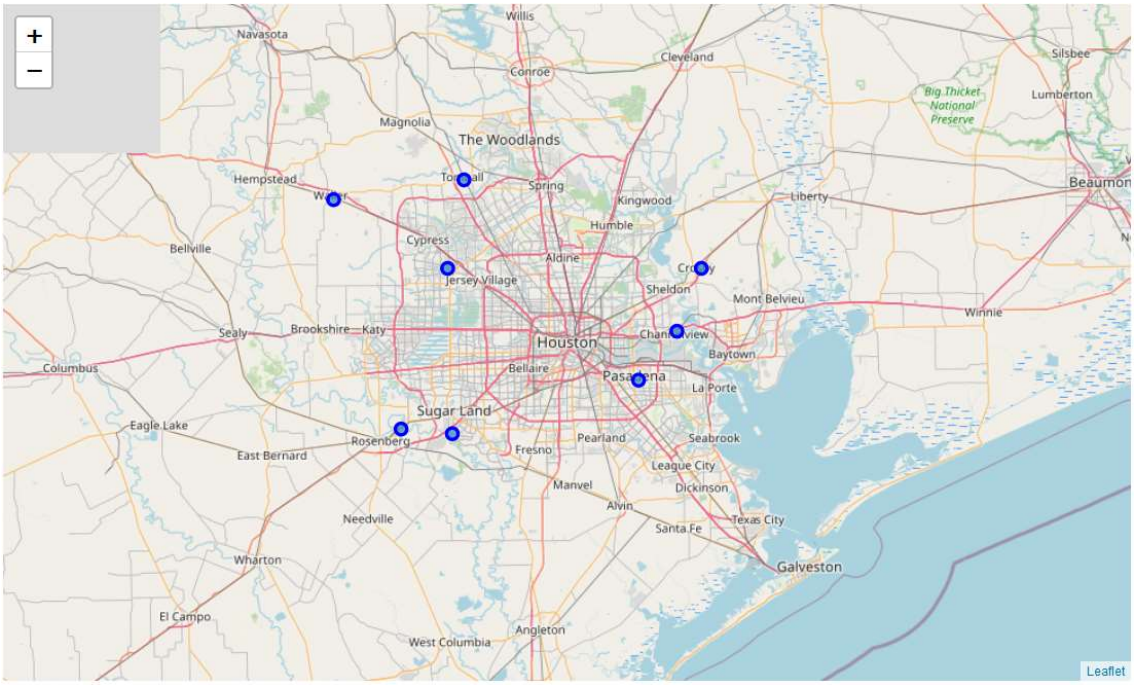


The 8 safest neighbourhoods were then also represented as we will be using them going forward for further analysis of the venues.

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)



We can then visually represent them graphically to get a view of their geographical layout in relation to the centre of Houston



Venue and Amenity Analysis

Now that we have a dataset of the neighbourhoods with the least crime (see above), associated venue / amenity data can be sourced from Foursquare. This will allow the user to see what facilities are available for each of the selected neighbourhoods and have a better understanding of where to look to relocate to.

We then find all the venues within a specified radius of each neighbourhood (in this case we used 2500 metres which is about 1.5 miles) by connecting to the Foursquare API. This returns a json file containing all the venues in each neighbourhood which is converted to a pandas dataframe. This dataframe contains all the venues along with their coordinates and category.

	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

One-hot encoding is then then conducted on the venues data.

One hot encoding is a process by which categorical variables are converted into a form that could be used by Machine Learning (ML) algorithms to analyse and provide clustering according to the data's similarities.

	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

The Venues data is then grouped by the Neighbourhood and the mean of the venues are calculated, and finally the 10 most common venues are calculated for each of the neighbourhoods.

	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	

To help people find similar neighbourhoods in the safest borough we will be clustering similar venue and amenity profiles per neighbourhood using K-Means clustering.

K-Means clustering is an unsupervised machine **learning algorithm** that, as the name hints, finds a fixed number (**k**) of **clusters** in a set of data. A **cluster** is a group of data points that are grouped together due to similarities in their features.

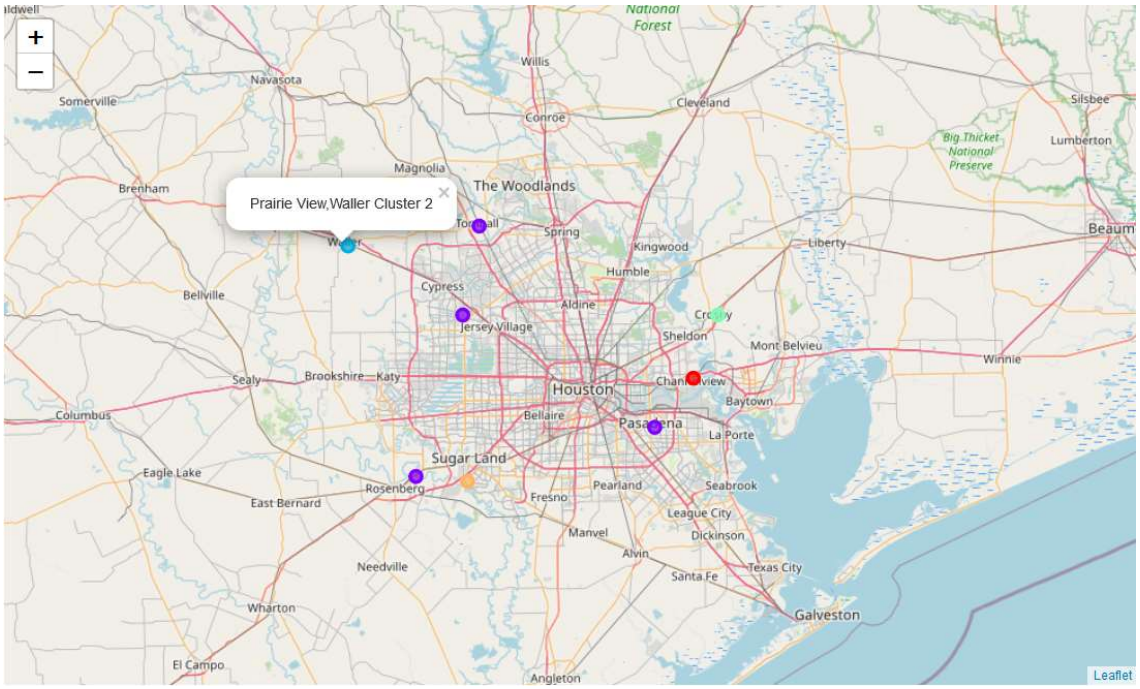
A more technical definition of **k-means clustering** is a method of vector quantization, originally from signal processing, that aims to partition **n** observations into **k** clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

We will use a cluster size of 5 for this project that will cluster the 8 neighbourhoods into 5 groupings. so that people can shortlist the area of their interests based on the venues/amenity's profiles around each neighbourhood.

Results

After running the K-means clustering we can access each cluster created to see which neighbourhoods were assigned to each of the five clusters.

We can visualise these groupings graphically as follows



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

Cluster 0

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

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.

Cluster 1

	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

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.

Cluster 2

	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
3	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

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.

Cluster 3

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

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

Cluster 4

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

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

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.

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 model is also split into 2 major sections namely Crime profiling and Venue / Amenity profiling so they can be used independently depending on the user requirements.

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.