IBM Applied Data Science Capstone

Researching a Neighborhood Before Buying a House in the Southwest Houston area

By: Dante A. Zorrilla

June 2020

1. Introduction

1.1. Background

Houston is one of the fastest growing and most diverse places in the United States. Wealth of job opportunities is bringing people to move in Houston, and thanks to the influx of new residents and median income growth rate that exceeds the national rate, newer and cheaper houses are being built.

1.2. Description of the Problem

Buying a house in Houston seems like a no brainer. However, for most people the whole process of buying a house is a hassle; it doesn't matter if the house will be for personal use or rental purposes, every person should conduct an extensive research before making such an important decision. Researching the market as per preference and budget along with looking into the neighborhoods is time consuming, therefore this project is looking forward to easing the process by applying data science methodology.

1.3. Target Audience

This project is useful for people looking to buy a house in Southwest Houston area. Python language will be used to gather real estate information, crime index per neighborhood and venues from different sources and present them to you to help you buying a new house in Houston.

2. Data

For this project we will use data from the following sources:

- Zillow: This website will be used to get information of real estate properties in Southwest Houston; including address, that will give us the latitudes and longitudes to map each property.
- Neighborhood Scout: This website will be used to extract crime index information for each neighborhood in Southwest Houston. This information will be used to see how prices change depending on the crime index of the neighborhood in which the property is located.
- Geojson.io: This website will be used to create the geojson file from scratch containing the map of the neighborhoods in Southwest Houston that will be used to locate the real estate properties.
- Foursquare API: Foursquare location database will give me a list of venues that will be used to explore the neighborhood for the property in interest.

3. Methodology

3.1. Extracting and Processing Data from Zillow

The zillow.py script was used to extract real estate data from Zillow. To run it, open a command prompt and type “python Zillow.py 77489 newest”, where 77489 is the zip code from which we want the data and newest is the sort argument for new listings.

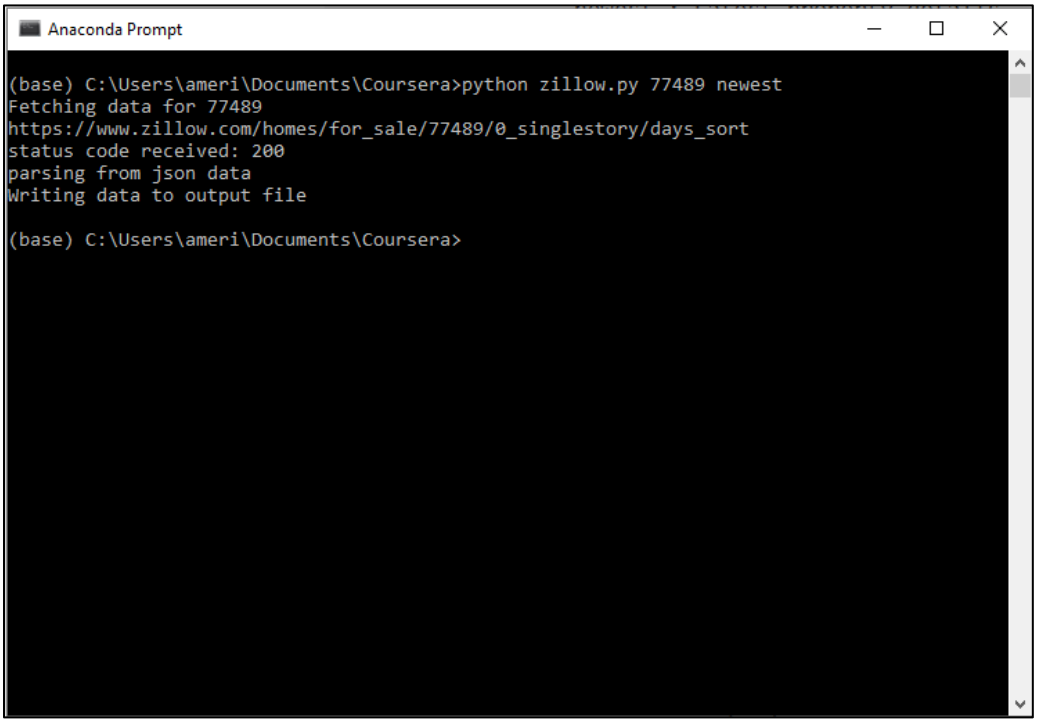

Figure 1: Running zillow.py script

The script will return a csv file named “properties-77489” with the houses for sale data in zip code 77489. To get all the needed real estate data, the script ran for all the zip codes in Southwest Houston.

Once all the csv files “properties-77XXX” were created, “merging_csv.ipynb” script was used to merge all the csv in the current directory into a single csv file

named “data.csv”. To avoid errors, make sure only the csv files from zillow.py script are in the working folder.

title	address	city	state	postal_co	price	facts and	real estat	url
For sale by owner		Missouri C	TX	77489	$249,000	4 bds, 3.0 ba	,2481 sc	https://www.zillow.com/homedetails/931-Crestmont-Place-Loop-Missouri-City-TX-77489/84753224_zpid/
House for sale		Missouri C	TX	77489	$169,900	4 bds, 2.0	Champion	https://www.zillow.com/homedetails/15834-Kenbriar-Dr-Missouri-City-TX-77489/27542788_zpid/
House for sale		Missouri C	TX	77489	$215,000	4 bds, 3.0	New Hom	https://www.zillow.com/homedetails/618-Derby-Ln-Missouri-City-TX-77489/50211160_zpid/
Townhouse for sale		Missouri C	TX	77489	$145,900	3 bds, 3.0	Alliance P	https://www.zillow.com/homedetails/8031-Chasewood-Dr-Missouri-City-TX-77489/72053629_zpid/
House for sale		Missouri C	TX	77489	$210,000	4 bds, 3.0	Realty Un	https://www.zillow.com/homedetails/2823-Eaglerock-Cir-Missouri-City-TX-77489/27594900_zpid/
House for sale		Missouri C	TX	77489	$120,000	4 bds, 2.0	Midland R	https://www.zillow.com/homedetails/7711-Sign-St-Missouri-City-TX-77489/27625026_zpid/
House for sale		Missouri C	TX	77489	$215,000	4 bds, 3.0	Mark Dim	https://www.zillow.com/homedetails/2330-Bright-Meadows-Dr-Missouri-City-TX-77489/27577301_zpid/
House for sale		Missouri C	TX	77489	$179,000	3 bds, 2.0	REALM Re	https://www.zillow.com/homedetails/1838-Ripple-Creek-Dr-Missouri-City-TX-77489/27591248_zpid/
House for sale		Missouri C	TX	77489	$199,900	3 bds, 2.0	MIH REAL	https://www.zillow.com/homedetails/1907-Robcrest-Way-Missouri-City-TX-77489/72057912_zpid/
New construction		Missouri C	TX	77489	$251,190	4 bds, 3.0	Centex Ho	https://www.zillow.com/homedetails/2718-Ibis-Way-Missouri-City-TX-77489/2080924718_zpid/
New construction		Missouri C	TX	77489	$227,190	3 bds, 3.0	Centex Ho	https://www.zillow.com/homedetails/1718-Avocet-Way-Missouri-City-TX-77489/2081566759_zpid/
House for sale		Missouri C	TX	77489	$184,900	3 bds, 2.0	Champion	https://www.zillow.com/homedetails/1102-Mossridge-Dr-Missouri-City-TX-77489/27565843_zpid/
House for sale		Missouri C	TX	77489	$195,000	3 bds, 2.0	REALM Re	https://www.zillow.com/homedetails/707-Stephanie-Dr-Missouri-City-TX-77489/27565222_zpid/
House for sale		Missouri C	TX	77489	$179,000	3 bds, 2.0	Newhous	https://www.zillow.com/homedetails/16403-Quail-Call-Dr-Missouri-City-TX-77489/27592940_zpid/
New construction		Missouri C	TX	77489	$239,390	4 bds, 3.0	Centex Ho	https://www.zillow.com/homedetails/1719-Avocet-Way-Missouri-City-TX-77489/2081566757_zpid/
New construction		Missouri C	TX	77489	$238,850	3 bds, 3.0	Centex Ho	https://www.zillow.com/homedetails/2610-Ibis-Way-Missouri-City-TX-77489/2080629252_zpid/
New construction		Missouri C	TX	77489	$271,830	4 bds, 3.0	Centex Ho	https://www.zillow.com/homedetails/1703-Avocet-Way-Missouri-City-TX-77489/2081299550_zpid/
House for sale		Missouri C	TX	77489	$185,000	4 bds, 2.0	RE/MAX A	https://www.zillow.com/homedetails/14703-Chasemont-Dr-Missouri-City-TX-77489/27626781_zpid/
House for sale		Missouri C	TX	77489	$215,000	3 bds, 2.0	Sharcom F	https://www.zillow.com/homedetails/1818-Skipwood-Dr-Missouri-City-TX-77489/72057859_zpid/

Figure 2: Merged csv file

Since the resulted address column is empty, the url column was used to get them. Used Excel formulas to removed https://www.zillow.com/homedetails/ and everything that follows the slash after the zipcode, then “find and replace” was used to replace “-” for a blank space to finally extract the addresses from the url.

Now the csv can be loaded into Python as a dataframe, using pandas library. The address column was extracted and “Nominatim” module used to locate them into “coord” column.

```
In [2]: house_df=pd.read_csv('data.csv')
        address_df=house_df['address']
        np.size(house_df['address'])

Out[2]: 813

In [3]: # Using geolocator to get latitude and longitude from address
        geolocator = Nominatim(timeout=3)
        house_df['coord'] = house_df['address'].apply(geolocator.geocode)
```

Figure 3: Extracting location from the address

Some addresses will return “None” in the “coord” column; this might happen because the address is incorrect, or the format address used by Zillow cant be processed by the module, so trying to get the latitudes and longitudes will lead to an error. To fix this issue all the None values were deleted, the “APT’ text were changed to “#”, and addresses including “Unit X” were removed. Once we fixed all the addresses that might give an error, the location was used to get the latitudes and longitudes.

```
In [4]: house_df['latitude'] = house_df['coord'].apply(lambda x: (x.latitude))
        house_df['longitude'] = house_df['coord'].apply(lambda x: (x.longitude))
        house_df.head(10)
```

Out[4]:

	title	city	state	postal_code	price	features	provider	url	address	coord	latitude	longitude
0	House for sale	Houston	TX	77096	$255,000	0 bds, 2.0 ba ,2304 sqft	The Kuper Group	9906 Cliffwood Dr Houston TX 77096/28041595_z...	9906 Cliffwood Dr Houston TX	(9906, Cliffwood Drive, Willow Meadows, Westwo...	29.675033	-95.453459
1	Townhouse for sale	Houston	TX	77042	$149,900	2 bds, 3.0 ba ,1520 sqft	Century 21 Olympian Galleria	9850 Pagewood Ln APT 1906 Houston TX 77042/28...	9850 Pagewood Ln #1906 Houston TX	(9850, Pagewood Lane, Jeanetta, Houston, Harri...	29.725298	-95.541977
2	Townhouse for sale	Houston	TX	77042	$163,000	2 bds, 3.0 ba ,1520 sqft	REALM Real Estate Professional	9800 Pagewood Ln APT 3505 Houston TX 77042/28...	9800 Pagewood Ln #3505 Houston TX	(9800, Pagewood Lane, Jeanetta, Houston, Harri...	29.725297	-95.541900
3	Townhouse for sale	Houston	TX	77042	$155,000	2 bds, 3.0 ba ,1520 sqft	Coldwell Banker United,	9800 Pagewood Ln APT 2704 Houston TX 77042/28...	9800 Pagewood Ln #2704 Houston TX	(9800, Pagewood Lane, Jeanetta, Houston, Harri...	29.725297	-95.541900
4	Townhouse for sale	Houston	TX	77099	$65,000	2 bds, 2.0 ba ,1068 sqft	Champions Real Estate Group	9797 Leawood Blvd APT 902 Houston TX 77099/28...	9797 Leawood Blvd #902 Houston TX	(9797, Leawood Boulevard, Parkglen West, Bella...	29.672342	-95.577324

Figure 4: Processed dataframe

3.2. Getting Crime Index and Creating the Geojson File

NeighborhoodScout provides analytic data for different features of a neighborhood, including crime statistics. Since this site charges a fee to get detailed data for crime statistics, the choropleth map was used to create a csv file with the crime index for each neighborhood, with a range of 1 being the safest neighborhood and 10 the most dangerous.

The same choropleth map was used to create the geojson file defining the borders of each neighborhood in Southwest Houston in the website geojson.io. Once we have the csv file and the geojson file we are ready to create the choropleth map.

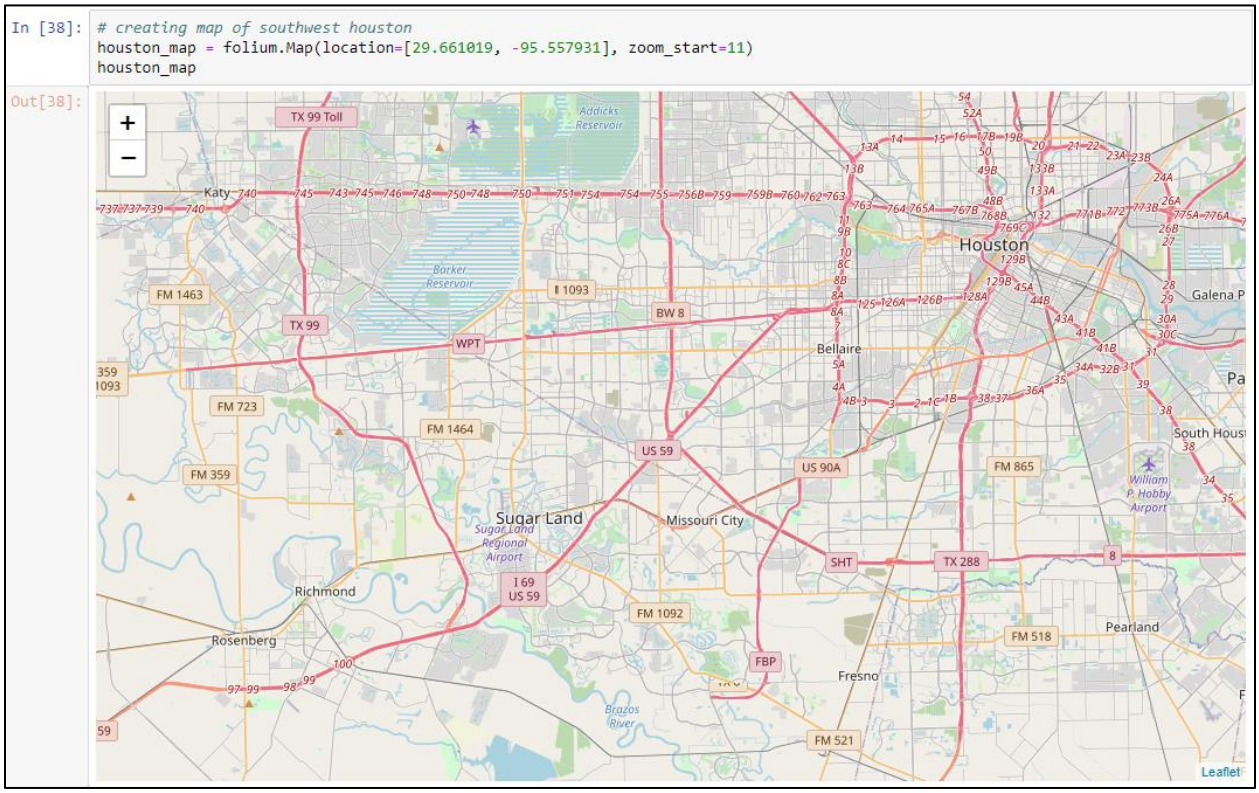

Figure 5: Map of Southwest Houston

Started by using folium to create a map of Southwest Houston area, then the csv and geojson file were loaded to add the choropleth feature to the map.

In [36]:
```
# Loading json file and mapping the crime index
crime_df = pd.read_csv('sw_houston_crimemap.csv')
houston_geo = r'sw_houston.geojson'
houston_map.choropleth(
    geo_data=houston_geo,
    data=crime_df,
    columns=['Neighborhood','Crime_Index'],
    key_on='feature.properties.snbname',
    fill_color='YlOrRd',
    #bins = [1, 2, 3, 4, 5, 6, 7 ,8 , 9, 10],
    fill_opacity=0.6,
    line_opacity=0.2,
    legend_name='Crime Index in Southwest Houston'
)
houston_map
```

Out[36]:

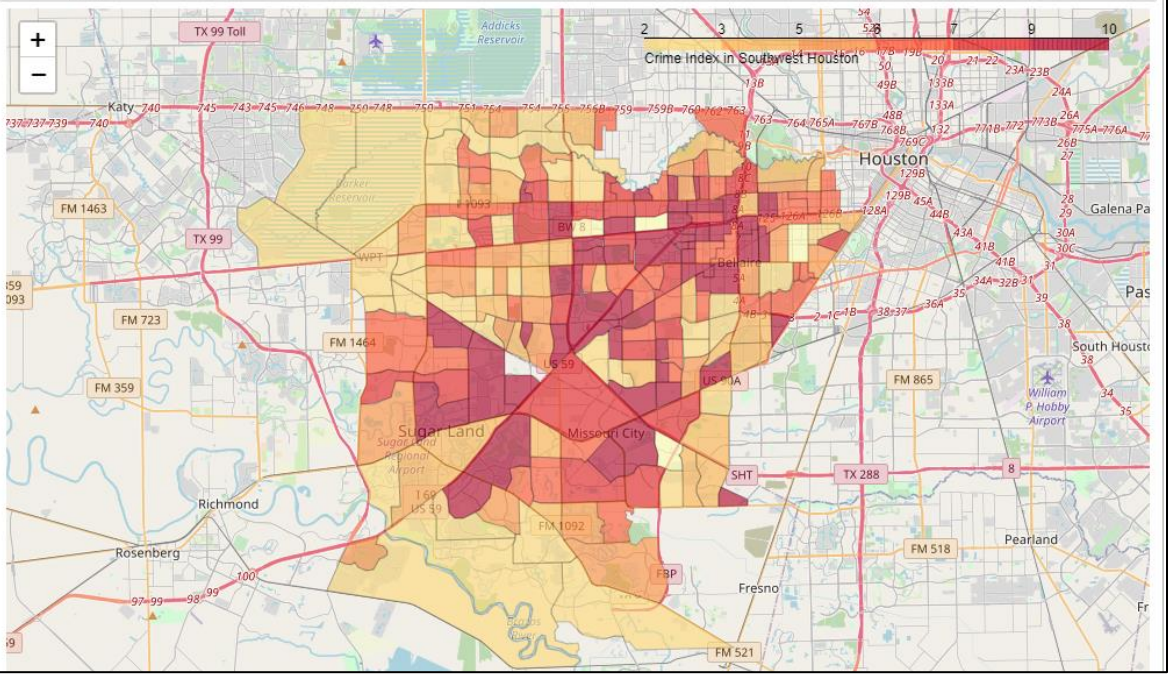

Figure 6: Crime Index Choropleth Map of Southwest Houston

3.4. Locating the Real Estate Properties

Using the previously processed dataframe, the location of the houses for sale in Southwest Houston can be loaded in the choropleth map. A label was added for each marker with the price, features, and address of the property. The result is a choropleth map indicating the crime index per neighborhood, where 1 is the safest and 10 the most dangerous, and markers with popup labels for each property in sale in the area.

```
In [37]: # Mapping location for each house for sale in southwest houston
for lat, lng, label in zip(house_df.latitude, house_df.longitude, 'Price:' + house_df.price + ', ' + 'Features:' + house_df.featu
    folium.features.CircleMarker(
            [lat, lng],
            radius=5, # define how big you want the circle markers to be
            color='yellow',
            fill=True,
            popup=label,
            fill_color='blue',
            fill_opacity=0.6
        ).add_to(houston_map)

# show map
houston_map
```

Out[37]:

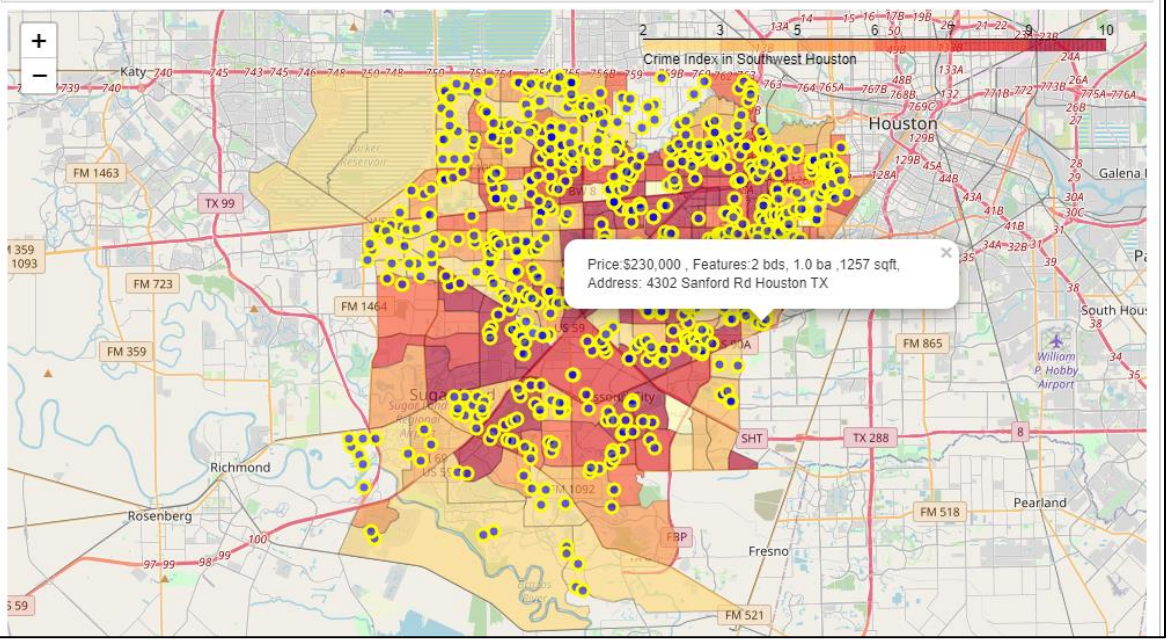

Figure 7: Houses for Sale in Southwest Houston

3.5. Exploring the Neighborhood

The next step in this project is to explore a neighborhood for the property in interest. A call will be sent to the Foursquare API to search for popular places around the property.

We started by defining Foursquare credentials and version and used geocode to locate the property in interest by using the address; for this case a property with a reasonable price and in a safe neighborhood will be used.

```
In [44]: CLIENT_ID = 'QLYEYPO4WPSYWUQ0G5D4KJK42QQMCDW4PIAWW5DLOECPFIFH' # your Foursquare ID
         CLIENT_SECRET = 'MNGMXBHOAEWXZTBYZ4BL4MXOFS2X12N0HCYWAMTMULBO5QRU' # your Foursquare Secret
         VERSION = '20180604'
         LIMIT = 30
         print('Your credentails:')
         print('CLIENT_ID: ' + CLIENT_ID)
         print('CLIENT_SECRET:' + CLIENT_SECRET)

         Your credentails:
         CLIENT_ID: QLYEYPO4WPSYWUQ0G5D4KJK42QQMCDW4PIAWW5DLOECPFIFH
         CLIENT_SECRET:MNGMXBHOAEWXZTBYZ4BL4MXOFS2X12N0HCYWAMTMULBO5QRU

In [45]: address = '13014 Bassford Dr Houston TX'

         geolocator = Nominatim(user_agent="foursquare_agent")
         location = geolocator.geocode(address)
         latitude = location.latitude
         longitude = location.longitude
         print(latitude, longitude)

         29.676706859308176 -95.61027475262581
```

Figure 8: Defining Foursquare Credential and Address to Explore

The URL to explore the venues was defined and the GET request sent to get JSON file with the venues in the area. The JSON file was processed and converted in a dataframe.

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng	postalCode
0	Timmy Chan's	Chinese Restaurant	9550 Bissonnet St	US	Houston	United States	NaN	612	[9550 Bissonnet St, Houston, TX 77036, United ...	[{'label': 'display', 'lat': 29.68057084332027...	29.680571	-95.614788	77036
1	Old Richmond Farm	Farm	13235 Old Richmond Road Houston, TX	US	Houston	United States	NaN	401	[13235 Old Richmond Road Houston, TX, Houston,...	[{'label': 'display', 'lat': 29.67665621111128...	29.676656	-95.614428	77083
2	Walgreens	Pharmacy	9150 S Dairy Ashford Rd	US	Houston	United States	NaN	724	[9150 S Dairy Ashford Rd, Houston, TX 77099, U...	[{'label': 'display', 'lat': 29.6786494, 'lng'...	29.678649	-95.603121	77099
3	SUBWAY	Sandwich Place	13151 Bissonnet St	US	Houston	United States	NaN	483	[13151 Bissonnet St, Houston, TX 77099, United...	[{'label': 'display', 'lat': 29.68034236500281...	29.680342	-95.613016	77099
4	Redbox	Video Store	9150 S Dairy Ashford Rd	US	Houston	United States	NaN	728	[9150 S Dairy Ashford Rd, Houston, TX 77099, U...	[{'label': 'display', 'lat': 29.67887687683105...	29.678877	-95.603165	77099
5	Family Dollar	Discount Store	12671 Bissonnet St	US	Houston	United States	Dairy Ashford	814	[12671 Bissonnet St (Dairy Ashford), Houston, ...	[{'label': 'display', 'lat': 29.67679023742675...	29.676790	-95.601852	77099
6	Pizza Hut	Pizza Place	12775 Bissonnet St	US	Houston	United States	NaN	586	[12775 Bissonnet St, Houston, TX 77099, United...	[{'label': 'display', 'lat': 29.67825985387720...	29.678260	-95.604481	77099
7	Vinh Chau Food Market	Grocery Store	13151 Bissonnet St Ste 700	US	Houston	United States	NaN	495	[13151 Bissonnet St Ste 700, Houston, TX 77099...	[{'label': 'display', 'lat': 29.680379, 'lng':...	29.680379	-95.613175	77099
8	Pho Com Ga Texas	Vietnamese Restaurant	13151 Bissonnet St Ste 400	US	Houston	United States	Synott	500	[13151 Bissonnet St Ste 400 (Synott), Houston,...	[{'label': 'display', 'lat': 29.68039703369140...	29.680397	-95.613228	77099
9	Mix Karaoke	Karaoke Bar	13151 Bissonnet St #400	US	Houston	United States	NaN	558	[13151 Bissonnet St #400, Houston, TX 77099, U...	[{'label': 'display', 'lat': 29.68091315735761...	29.680913	-95.613417	77099

Figure 9: Dataframe with the Popular Spot in the Area

For the final step in this project, the new dataframe was used to visualize these spots on the map around the property of interest using folium library.

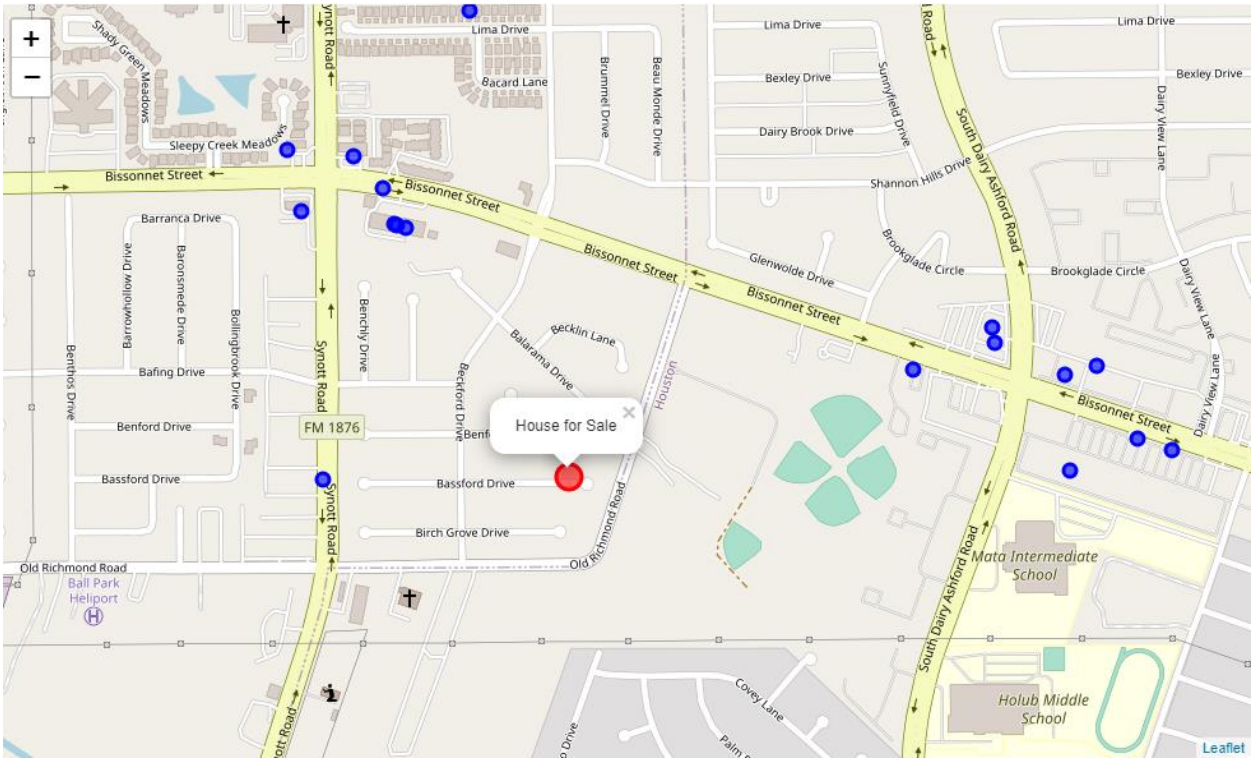

Figure 10: Popular Spots around the Property.

4. Results

This project allowed us to create a Python script that localizes houses for sale in Southwest Houston in a choropleth map indicating crime index per neighborhood. Furthermore, we can explore the area around a house of interest by making Foursquare API calls and extract information about restaurants, hotels, parks and recreations in the neighborhood.

5. Discussion and Recommendation

While working on this project it was noted that the most expensive houses were those closer to Downtown Houston, regardless of the crime index of the area. This was expected because this area is where the major attractions and business buildings are located.

Except for the Center of Houston, all the other areas show a lower cost on houses located in zones considered “most dangerous”. Based on this, my recommendation is to look for properties in areas with a low crime index, even though they might be more expensive, it is worth knowing you are living in a safe neighborhood.

6. Conclusion

Using Python we were able to gather real estate information, neighborhood crime index, and popular spots around an area of interest, creating a script that provides this information useful for people thinking in buying a property in Southwest Houston.