The Battle of the Neighborhoods - Week 1

Introduction & Business Problem

Problem Background:

Houston is one of the largest city in the United States .A. For employment and business opportunities, many people from around the globe move to this city. When they move their first important task is to find a good locality with good amenities and safe neighborhood. To find the best location to reside using data science can enhance a person's capability to take the decision of selecting a good locality to reside.

Problem Description:

A person who is just moving to Houston asks Real Estate Firm- XYZ Company Ltd to find the best neighborhoods for him which have good school for education of kids and low rate of crime for safety of family. The choice of neighborhood should also consider the home prices and the number of amenities/facilities to do activities within that neighborhood.

Target Audience:

The target audience is the average American families which are looking to move to Houston and are in search for the location to reside which best suits their needs. To recommend the best neighborhood, XYZ Company Ltd has appointed me to lead this task. The objective is to locate and recommend the best-suited neighborhood of Houston city to the family. The Family also expects to understand the rationale of the recommendations made. This would interest anyone who comes to XYZ Company Ltd with similar request for Houston City.

Success Criteria:

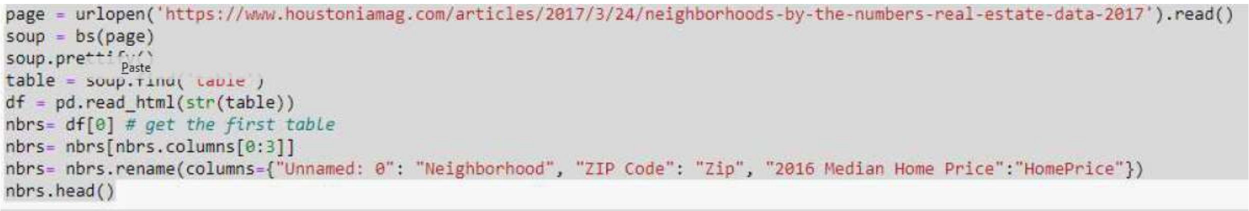
The success criteria of the project will be a suitable recommendation of a Neighborhood/Zip Code choice which satisfy all the basic needs (Schooling,facilities.safety,low price)for a typical family on based on the analysis of Houston area data.

Data

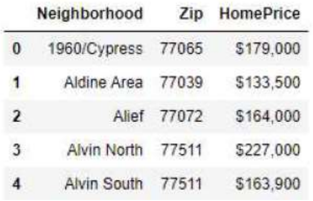
We need below data for our Analysis -
1. All Zip Codes/Neighborhood in Houston
2. Home Price data for each Zip Code
3. Crime data in every Zip Code
4. Average School Rating in every Zip Code
4. Latitude and Longitude of each Zip Code
5. The Number of venues in each Zip Code from Foursquare APIs based on Latitude and Longitude of Zip Code

We can get All Zip Codes/Neighborhood in Houston with average home prices from website of Houstonia- a popular online magazine in Houston. This is available at -
https://www.houstoniamag.com/articles/2017/3/24/neighborhoods-by-the-numbers-real-estate-data-

2017
To get this we would need to scrape a page from website of Houstonia magazine to get the neighborhood data of Houston, TX using BeautifulSuop4 Library. This would return neighborhood name, Zip Code and average home price in that Zip Code.

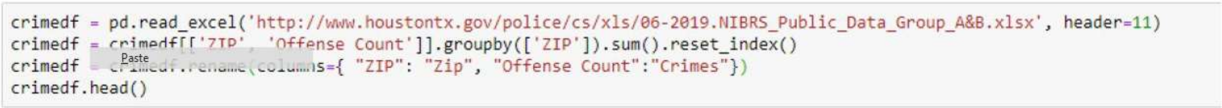


```
page = urlopen('https://www.houstoniamag.com/articles/2017/3/24/neighborhoods-by-the-numbers-real-estate-data-2017').read()
soup = bs(page)
soup.prettify()
table = soup.find('table')
df = pd.read_html(str(table))
nbrs= df[0] # get the first table
nbrs= nbrs[nbrs.columns[0:3]]
nbrs= nbrs.rename(columns={"Unnamed: 0": "Neighborhood", "ZIP Code": "Zip", "2016 Median Home Price":"HomePrice"})
nbrs.head()
```

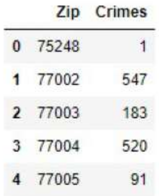


	Neighborhood	Zip	HomePrice
0	1960/Cypress	77065	$179,000
1	Aldine Area	77039	$133,500
2	Alief	77072	$164,000
3	Alvin North	77511	$227,000
4	Alvin South	77511	$163,900

Next, Crime data in every Zip Code can be read from the published data from Houston Police Department at their website -
http://www.houstontx.gov/police/cs/xls/06-2019.NIBRS_Public_Data_Group_A&B.xlsx



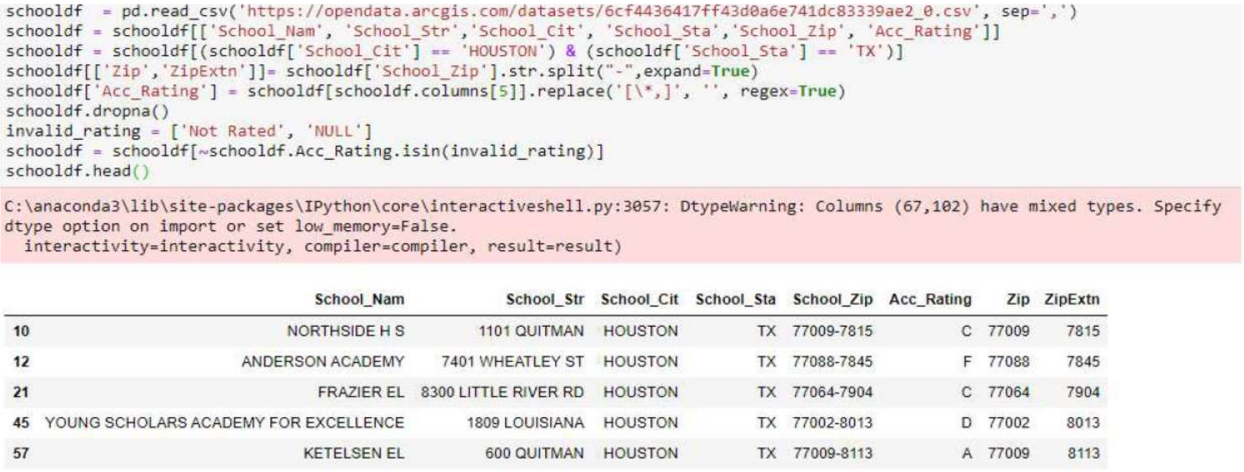
```
crimedf = pd.read_excel('http://www.houstontx.gov/police/cs/xls/06-2019.NIBRS_Public_Data_Group_A&B.xlsx', header=11)
crimedf = crimedf[['ZIP', 'Offense Count']].groupby(['ZIP']).sum().reset_index()
crimedf = crimedf.rename(columns={ "ZIP": "Zip", "Offense Count":"Crimes"})
crimedf.head()
```



	Zip	Crimes
0	75248	1
1	77002	547
2	77003	183
3	77004	520
4	77005	91

School Accountability Ratings are available at
https://opendata.arcgis.com/datasets/6cf4436417ff43d0a6e741dc83339ae2_0.csv

This dataset was made available on ArcGIS website by Texas Education Agency. This Dataset contains the school with its address and its Accountability Rating for whole Texas. We would filter this data set for Houston and get the average rating of all schools in each Zip Code.

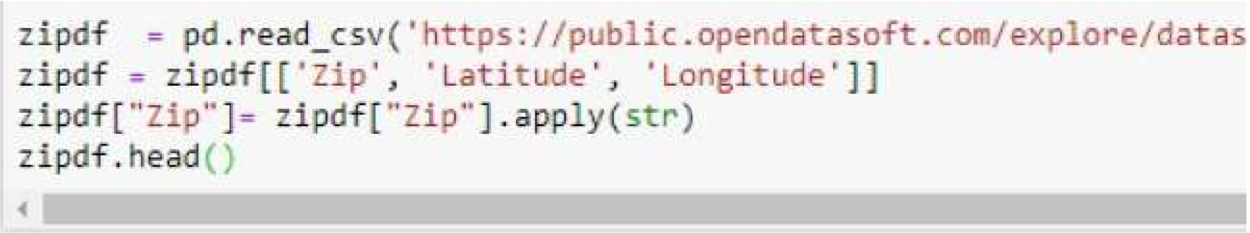


```
schooldf  = pd.read_csv('https://opendata.arcgis.com/datasets/6cf4436417ff43d0a6e741dc83339ae2_0.csv', sep=',')
schooldf = schooldf[['School_Nam', 'School_Str','School_Cit', 'School_Sta','School_Zip', 'Acc_Rating']]
schooldf = schooldf[(schooldf['School_Cit'] == 'HOUSTON') & (schooldf['School_Sta'] == 'TX')]
schooldf[['Zip','ZipExtn']]= schooldf['School_Zip'].str.split("-",expand=True)
schooldf['Acc_Rating'] = schooldf[schooldf.columns[5]].replace('[\*,]', '', regex=True)
schooldf.dropna()
invalid_rating = ['Not Rated', 'NULL']
schooldf = schooldf[~schooldf.Acc_Rating.isin(invalid_rating)]
schooldf.head()
```

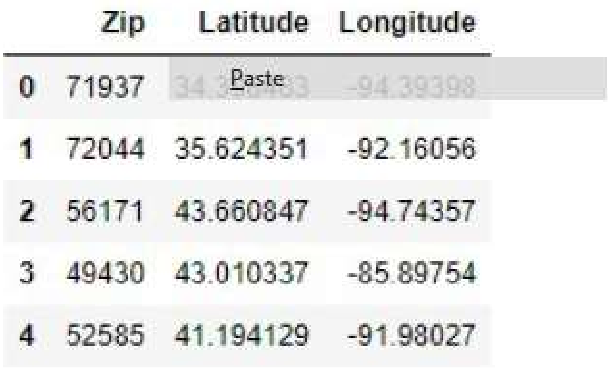
```
C:\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3057: DtypeWarning: Columns (67,102) have mixed types. Specify
dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

	School_Nam	School_Str	School_Cit	School_Sta	School_Zip	Acc_Rating	Zip	ZipExtn
10	NORTHSIDE H S	1101 QUITMAN	HOUSTON	TX	77009-7815	C	77009	7815
12	ANDERSON ACADEMY	7401 WHEATLEY ST	HOUSTON	TX	77088-7845	F	77088	7845
21	FRAZIER EL	8300 LITTLE RIVER RD	HOUSTON	TX	77064-7904	C	77064	7904
45	YOUNG SCHOLARS ACADEMY FOR EXCELLENCE	1809 LOUISIANA	HOUSTON	TX	77002-8013	D	77002	8013
57	KETELSEN EL	600 QUITMAN	HOUSTON	TX	77009-8113	A	77009	8113

Further we get the latitude and longitude data of all Zip Codes in US from open data available at https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-andlongitude/download/?format=csv&timezone=America/Chicago&use_labels_for_header=true&csv_separator=%3B.
Again we would filter this dataset for Houston only.

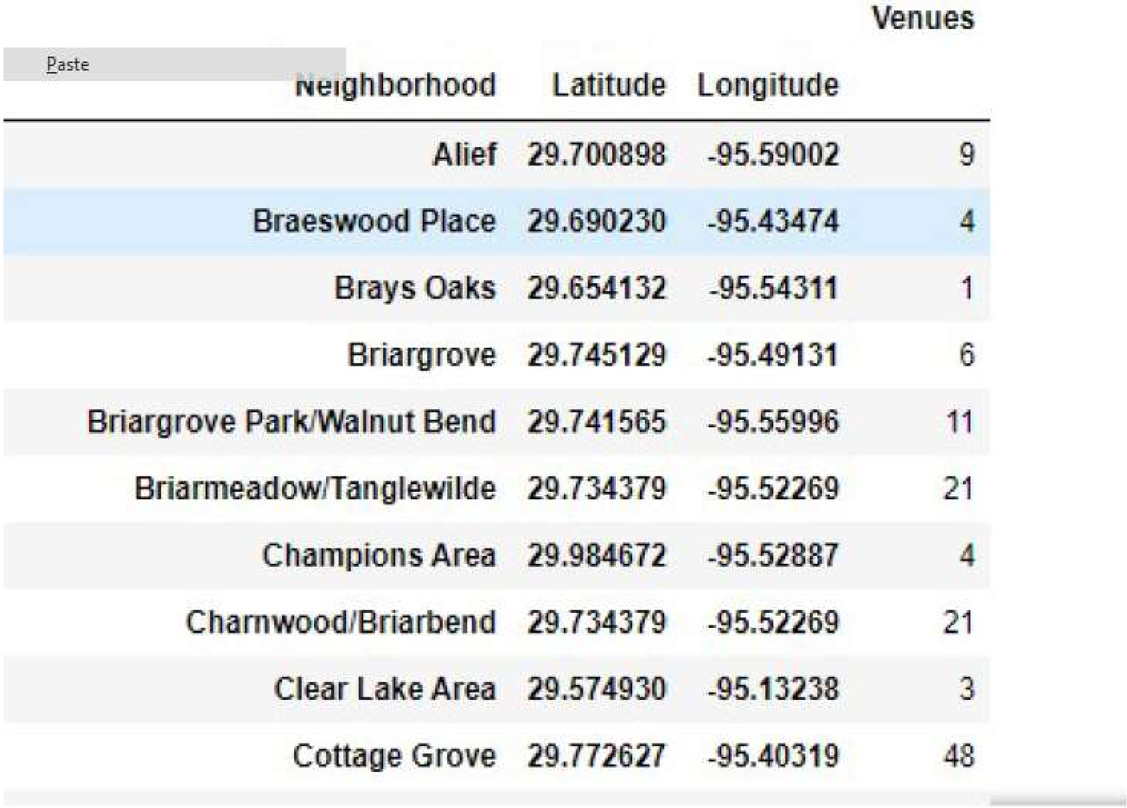


```
zipdf  = pd.read_csv('https://public.opendatasoft.com/explore/datas
zipdf = zipdf[['Zip', 'Latitude', 'Longitude']]
zipdf["Zip"]= zipdf["Zip"].apply(str)
zipdf.head()
```



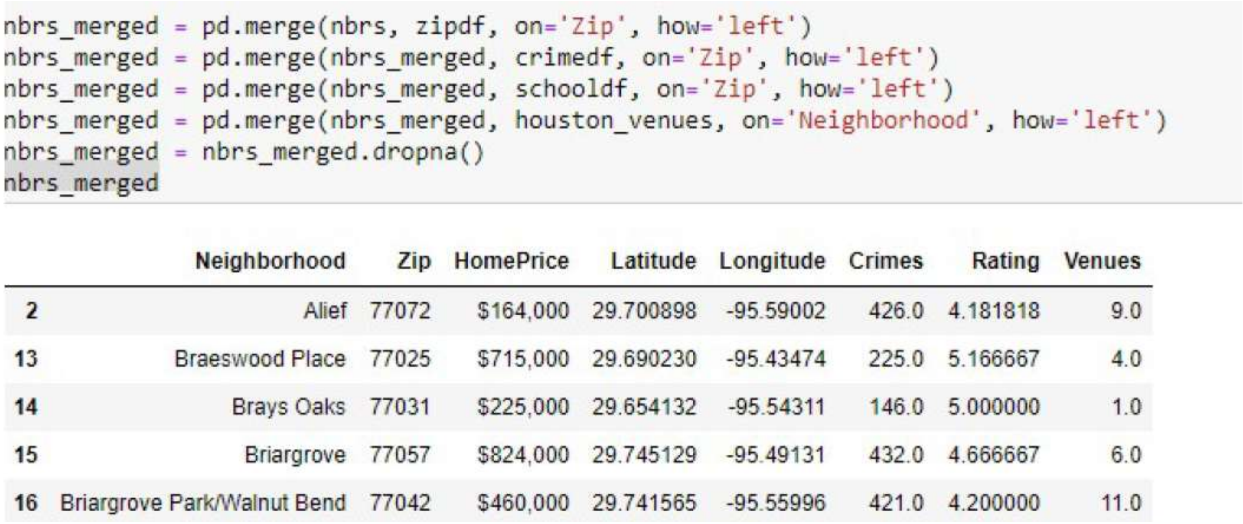
	Zip	Latitude	Longitude
0	71937	34.3[illegible]83	-94.39398
1	72044	35.624351	-92.16056
2	56171	43.660847	-94.74357
3	49430	43.010337	-85.89754
4	52585	41.194129	-91.98027

The Latitude and Longitude of each Zip Code would be used to get all venues in that Zip Code from FourSquare API. We would be interested in only total number of Venues in that zip code.



Neighborhood	Latitude	Longitude	Venues
Alief	29.700898	-95.59002	9
Braeswood Place	29.690230	-95.43474	4
Brays Oaks	29.654132	-95.54311	1
Briargrove	29.745129	-95.49131	6
Briargrove Park/Walnut Bend	29.741565	-95.55996	11
Briarmeadow/Tanglewilde	29.734379	-95.52269	21
Champions Area	29.984672	-95.52887	4
Charnwood/Briarbend	29.734379	-95.52269	21
Clear Lake Area	29.574930	-95.13238	3
Cottage Grove	29.772627	-95.40319	48

For analysis, we would join all above data in single dataset.



```
nbrs_merged = pd.merge(nbrs, zipdf, on='Zip', how='left')
nbrs_merged = pd.merge(nbrs_merged, crimedf, on='Zip', how='left')
nbrs_merged = pd.merge(nbrs_merged, schooldf, on='Zip', how='left')
nbrs_merged = pd.merge(nbrs_merged, houston_venues, on='Neighborhood', how='left')
nbrs_merged = nbrs_merged.dropna()
nbrs_merged
```

	Neighborhood	Zip	HomePrice	Latitude	Longitude	Crimes	Rating	Venues
2	Alief	77072	$164,000	29.700898	-95.59002	426.0	4.181818	9.0
13	Braeswood Place	77025	$715,000	29.690230	-95.43474	225.0	5.166667	4.0
14	Brays Oaks	77031	$225,000	29.654132	-95.54311	146.0	5.000000	1.0
15	Briargrove	77057	$824,000	29.745129	-95.49131	432.0	4.666667	6.0
16	Briargrove Park/Walnut Bend	77042	$460,000	29.741565	-95.55996	421.0	4.200000	11.0