IBM Applied Data Science
Capstone Project - The Battle of Neighborhoods

Project presentation

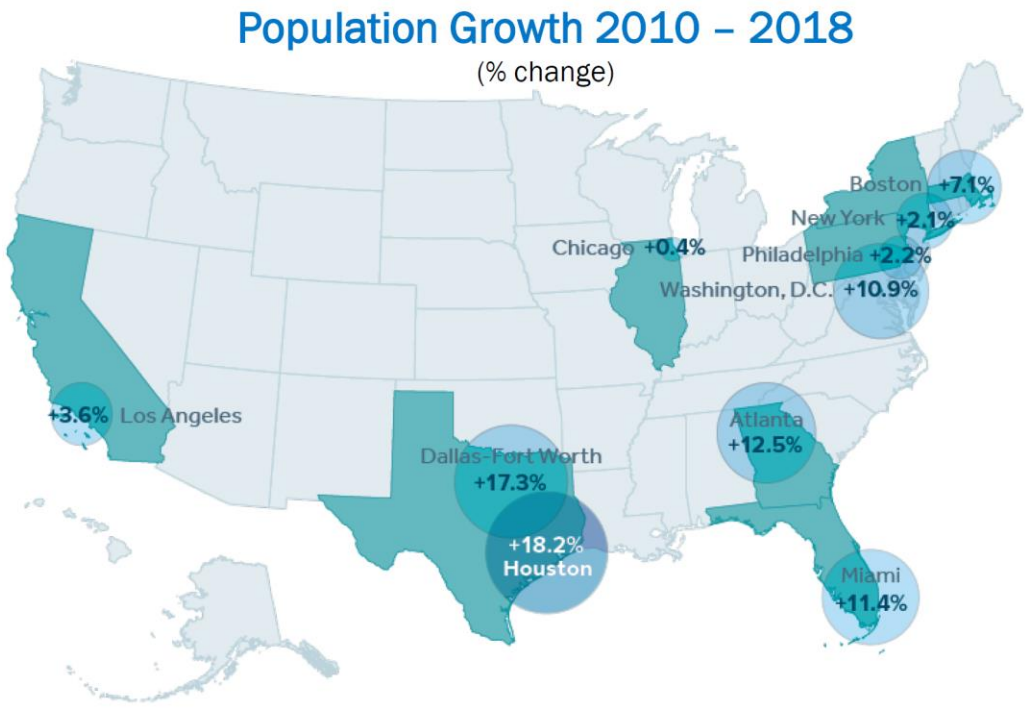
Segmenting and clustering super neighborhoods in Houston, Texas

Samarjit Chakraborty

July 2020

Introduction

- Houston, Texas is the fourth-largest city in the United States with fastest growing consumer market.
- City of Houston has more than 2.3 million residents and covers 634 square miles.
- Comprehensive, community-based efforts is required to promote healthy living in low-income neighborhoods in Houston.
- Geospatial analysis of census data with machine learning methods provides decision makers access to tools necessary for resource planning.
- The goal of this project was to analyze data available in public domain to identify zones of opportunity.
- This in turn would help to attract both practical and innovative investment into underinvested communities while leveraging local and state resources.
- The super neighborhood elects a council comprised of area residents and stakeholders that serves as a forum to discuss issues and identify and implement priority projects for the area.

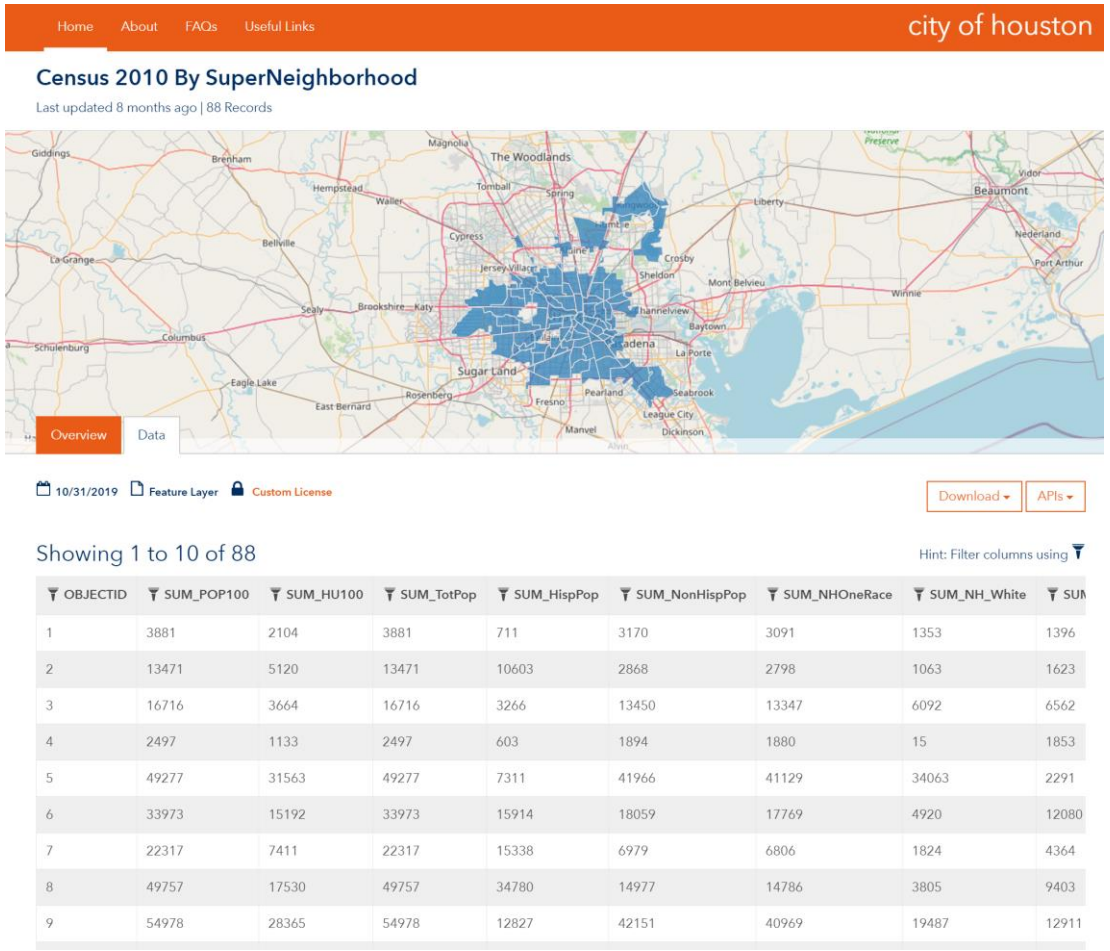


Source: https://www.houstontx.gov/

Data

- Houston is divided into 88 geographically designated areas, referred to as super neighborhoods.
- Residents, civic organizations, institutions, and businesses in these super neighborhoods are encouraged to work together to identify, plan, and set priorities to address the needs and concerns of the community.
- 2010 census data were obtained, free of charge, from City of Houston GIS (COHGIS) Open Data portal - *https://cohgis-mycity.opendata.arcgis.com*.
- This dataset includes demographic information by super neighborhood.
- The boundaries of each super neighborhood rely on major physical features (bayous, freeways, etc.) to group together contiguous communities that share common physical characteristics, identity or infrastructure.
- Demographic data includes, but is not limited to: population, housing, and other social characteristics.

https://cohgis-mycity.opendata.arcgis.com/datasets/census-2010-by-superneighborhood



OBJECTID	SUM_POP100	SUM_HU100	SUM_TotPop	SUM_HispPop	SUM_NonHispPop	SUM_NHOneRace	SUM_NH_White	SUM
1	3881	2104	3881	711	3170	3091	1353	1396
2	13471	5120	13471	10603	2868	2798	1063	1623
3	16716	3664	16716	3266	13450	13347	6092	6562
4	2497	1133	2497	603	1894	1880	15	1853
5	49277	31563	49277	7311	41966	41129	34063	2291
6	33973	15192	33973	15914	18059	17769	4920	12080
7	22317	7411	22317	15338	6979	6806	1824	4364
8	49757	17530	49757	34780	14977	14786	3805	9403
9	54978	28365	54978	12827	42151	40969	19487	12911

Census data

- Downloaded spreadsheet from the COHGIS portal
- Converted spreadsheet data to Pandas dataframes
- Pandas dataframe used for subsequent analysis

[14]:
```
# Read census spreadsheet
# https://cohgis-mycity.opendata.arcgis.com/datasets/census-2010-by-superneighborhood?geometry=-98.408%2C29.406%2C-92.423%2C30.240
df_data_0 = pd.read_csv('2010_Census_COH_DEMOGRAPHICS_MIL.csv')
df_data_0.rename(columns={'Name':'Neighborhood'},inplace=True)
df_data_0.head()
```

(88, 32)

[15]:

	Neighborhood	OBJECTID	SUM_POP100	SUM_HU100	SUM_TotPop	SUM_HispPop	SUM_NonHispPop	SUM_NHOneRace	SUM_NH_White	SUM_NH_Black	...	SUM_VAP_NH_Asi	SUM_VAP_HawPac	SUM_VAP_NH_Oth	SUM_VAP_NH_2or
32	ACRES HOME	33	24465	9288	24465	4782	19683	19525	595	18783	...	71	2	18	96
87	ADDICKS PARK TEN	88	7323	4015	7323	2121	5202	5018	2849	1471	...	511	7	15	118
82	AFTON OAKS / RIVER OAKS AREA	83	14007	8069	14007	1352	12655	12520	11271	457	...	621	4	8	80
52	ALIEF	53	102235	35498	102235	47966	54269	52936	8596	25589	...	14842	25	122	847
60	ASTRODOME AREA	61	17697	11311	17697	1840	15857	15430	5327	3670	...	5670	12	41	356

GeoJSON data

- Downloaded GeoJSON data from the COHGIS portal
- Extracted neighborhood names and geographical coordinates from GeoJSON file
- Converted polygon coordinates into their centroid latitude and longitude values

[25]:
```
# Read GeoJSON file from City of Houston GIS portal
# https://opendata.arcgis.com/datasets/f50cc53c15bf4fb3940ab6e8c2534f3d_2.geojson

!wget --quiet https://opendata.arcgis.com/datasets/f50cc53c15bf4fb3940ab6e8c2534f3d_2.geojson -O ./houston_demo_data.json
```

[49]:
```
houston_data.sort_values(by=['Neighborhood'],inplace=True)
houston_data
```

[49]:

	Neighborhood	Latitude	Longitude
32	ACRES HOME	29.864097236400905	-95.43626668096061
87	ADDICKS PARK TEN	29.813149745221093	-95.62979046456292
82	AFTON OAKS / RIVER OAKS AREA	29.74792116154408	-95.43563576038133
52	ALIEF	29.686317005174633	-95.58639242232414
60	ASTRODOME AREA	29.687929792715206	-95.39496596241122


```
{'type': 'FeatureCollection',
 'name': 'COH_DEMOGRAPHICS_-_MIL',
 'crs': {'type': 'name',
  'properties': {'name': 'urn:ogc:def:crs:OGC:1.3:CRS84'}},
 'features': [{'type': 'Feature',
   'properties': {'OBJECTID': 1,
    'SUM_POP100': 3881,
    'SUM_HU100': 2104,
    'SUM_TotPop': 3881,
    'SUM_HispPop': 711,
    'SUM_NonHispPop': 3170,
    'SUM_NHOneRace': 3091,
    'SUM_NH_White': 1353,
    'SUM_NH_Black': 1396,
    'SUM_NH_AmInd': 11,
    'SUM_NH_Asian': 321,
    'SUM_NH_HawPacI': 1,
    'SUM_NH_Other': 9,
    'SUM_NH_2orMore': 79,
    'SUM_VAP_TotPop': 3103,
    'SUM_VAP_HispPo': 544,
    'SUM_VAP_NonHis': 2559,
    'SUM_VAP_NHOneR': 2510,
    'SUM_VAP_NH_Whi': 1310,
    'SUM_VAP_NH_Bla': 885,
    'SUM_VAP_NH_AmI': 11,
    'SUM_VAP_NH_Asi': 296,
    'SUM_VAP_HawPac': 1,
    'SUM_VAP_NH_Oth': 7,
    'SUM_VAP_NH_2or': 49,
    'SUM_TotHousing': 2104,
    'SUM_OccHU': 1978,
    'SUM_VacantHU': 126,
    'POLYID': 60,
    'Name': 'FOURTH WARD',
    'Shapearea': 12969824.766383596,
    'Shapelen': 16572.0260242156},
   'geometry': {'type': 'Polygon',
    'coordinates': [[[-95.3858120191703, 29.76157948522655],
      [-95.3857510351169, 29.759096134455724],
      [-95.38575701432947, 29.7589274318535],
      [-95.38581341906735, 29.75809220348262],
      [-95.38571553175525, 29.757628886192457],
      [-95.38567397795319, 29.753574178823797],
      [-95.38575841615513, 29.753438202654888],
      [-95.3857221276755, 29.750280400839365],
      [-95.38400247507784, 29.750315844530196],
      [-95.38376001290212, 29.750262353841276],
```

Foursquare data

- Used Foursquare API to get the most common venues of each neighborhood
- Designed limit to get the top 100 venues within a radius of 1000 meters from the neighborhood’s centroid latitude and longitude values


```
{'reasons': {'count': 0,
  'items': [{'summary': 'This spot is popular',
    'type': 'general',
    'reasonName': 'globalInteractionReason'}]},
 'venue': {'id': '52b81d3111d21ec4eed0021e',
  'name': 'Green Vegetarian Cuisine',
  'location': {'address': '6720 Chimney Rock Rd Ste Y',
   'lat': 29.70629593156019,
   'lng': -95.47655849535282,
   'labeledLatLngs': [{'label': 'display',
     'lat': 29.70629593156019,
     'lng': -95.47655849535282}],
   'distance': 917,
   'postalCode': '77081',
   'cc': 'US',
   'city': 'Houston',
   'state': 'TX',
   'country': 'United States',
   'formattedAddress': ['6720 Chimney Rock Rd Ste Y',
    'Houston, TX 77081',
    'United States']},
  'categories': [{'id': '4bf58dd8d48988d1d3941735',
    'name': 'Vegetarian / Vegan Restaurant',
    'pluralName': 'Vegetarian / Vegan Restaurants',
    'shortName': 'Vegetarian / Vegan',
    'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/vegetarian_',
     'suffix': '.png'},
    'primary': True}],
  'delivery': {'id': '1600387',
   'url': 'https://www.grubhub.com/restaurant/green-vegetarian-cuisine-6720-chimney
mpaign=1131&utm_content=1600387',
   'provider': {'name': 'grubhub',
    'icon': {'prefix': 'https://fastly.4sqi.net/img/general/cap/',
     'sizes': [40, 50],
     'name': '/delivery_provider_grubhub_20180129.png'}}},
  'photos': {'count': 0, 'groups': []}},
 'referralId': 'e-0-52b81d3111d21ec4eed0021e-0'}
```

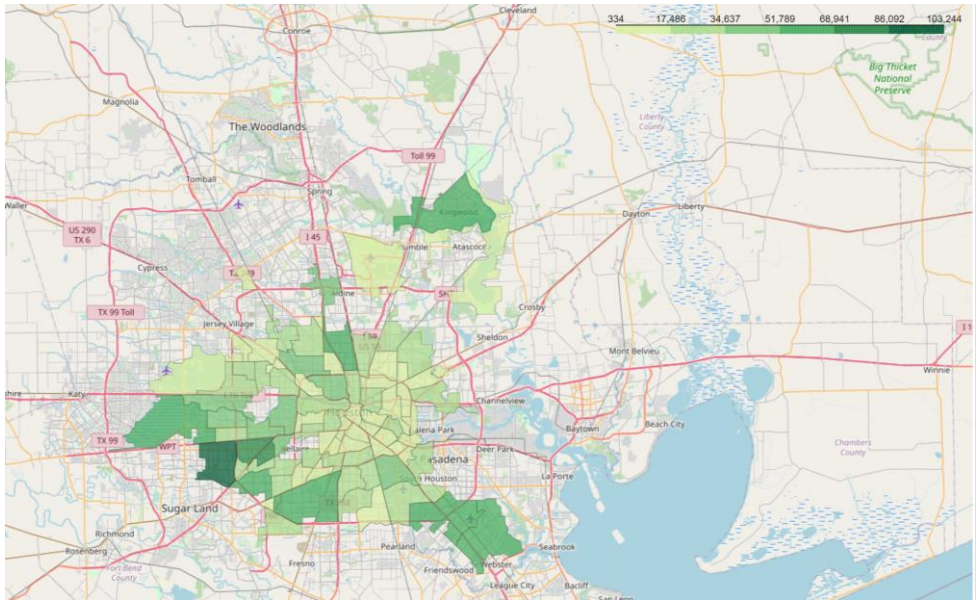
Methodology

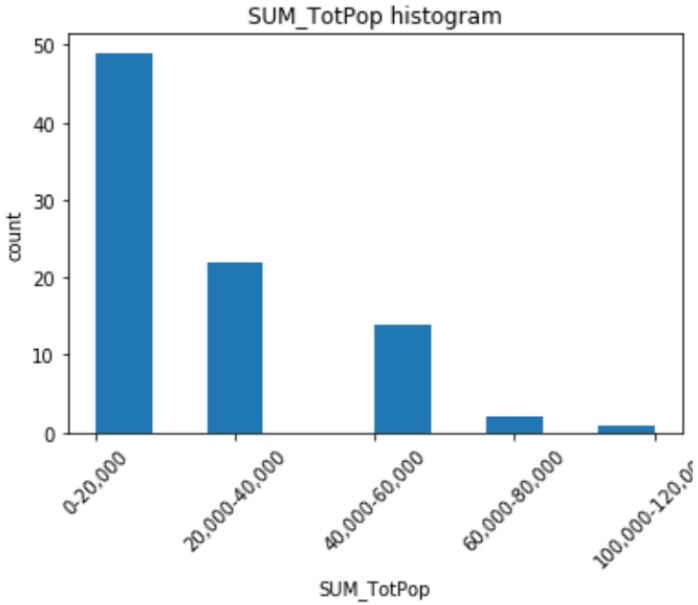
- In the Jupyter notebook, geographical coordinates from polygons (in GeoJSON file) were converted into their equivalent centroid latitude and longitude values.
- Dataframe with latitude and longitude information was merged with census data.
- Census data were used to group the neighborhoods into clusters for initial analysis.
- Foursquare API was used to explore super neighborhoods in Houston.
- The ‘explore’ function was then used to get the most common venue categories in each neighborhood.
- This venue categories feature was then used to group the neighborhoods into clusters.
- The k-means clustering algorithm was used to complete this task.
- Finally, the Folium library was used to visualize the neighborhoods in Houston and their emerging clusters.

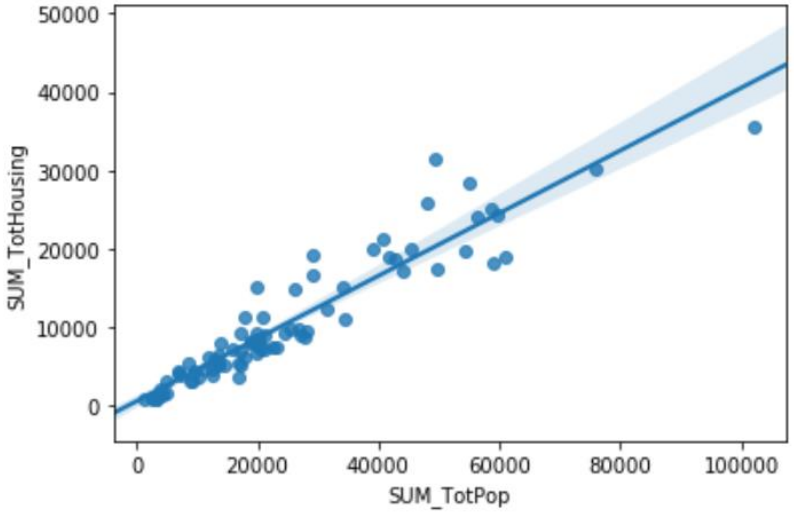
Link to Jupyter notebook in Github

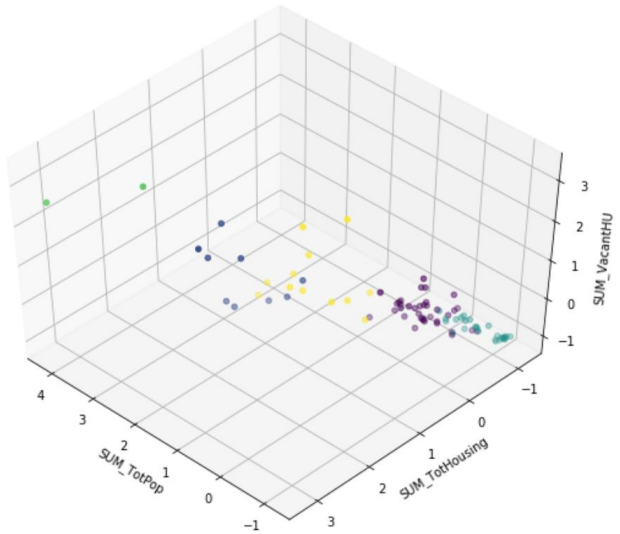
Explore dataset - visualization

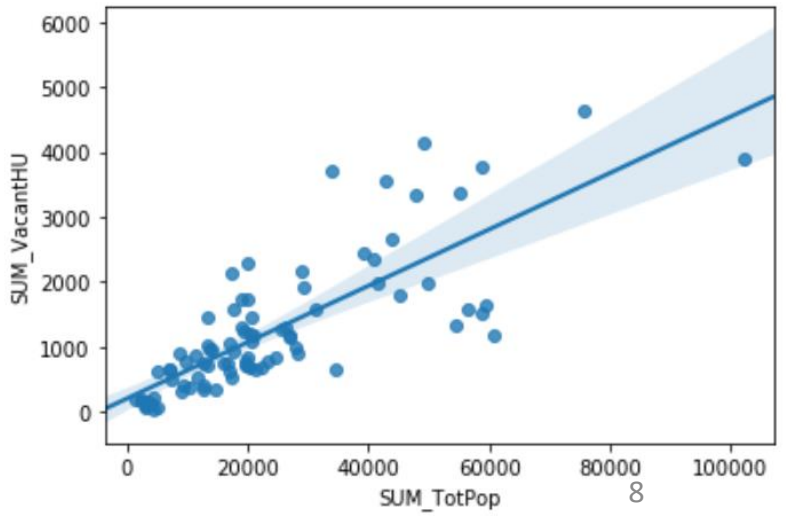
- Visualized data to find correlation between different variables
- Histograms used to understand spread of total population and other variables
- Choropleth maps used to show spatial distribution of census data











Explore dataset – descriptive statistical analysis

- Reviewed variables with a descriptive method
- Computed basic statistics for all continuous variables

```
[17]: df_data_0.columns
```

```
[17]: Index(['Neighborhood', 'OBJECTID', 'SUM_POP100', 'SUM_HU100', 'SUM_TotPop',
       'SUM_HispPop', 'SUM_NonHispPop', 'SUM_NHOneRace', 'SUM_NH_White',
       'SUM_NH_Black', 'SUM_NH_AmInd', 'SUM_NH_Asian', 'SUM_NH_HawPacI',
       'SUM_NH_Other', 'SUM_NH_2orMore', 'SUM_VAP_TotPop', 'SUM_VAP_HispPo',
       'SUM_VAP_NonHis', 'SUM_VAP_NHOneR', 'SUM_VAP_NH_Whi', 'SUM_VAP_NH_Bla',
       'SUM_VAP_NH_AmI', 'SUM_VAP_NH_Asi', 'SUM_VAP_HawPac', 'SUM_VAP_NH_Oth',
       'SUM_VAP_NH_2or', 'SUM_TotHousing', 'SUM_OccHU', 'SUM_VacantHU',
       'POLYID', 'Shapearea', 'Shapelen'],
      dtype='object')
```

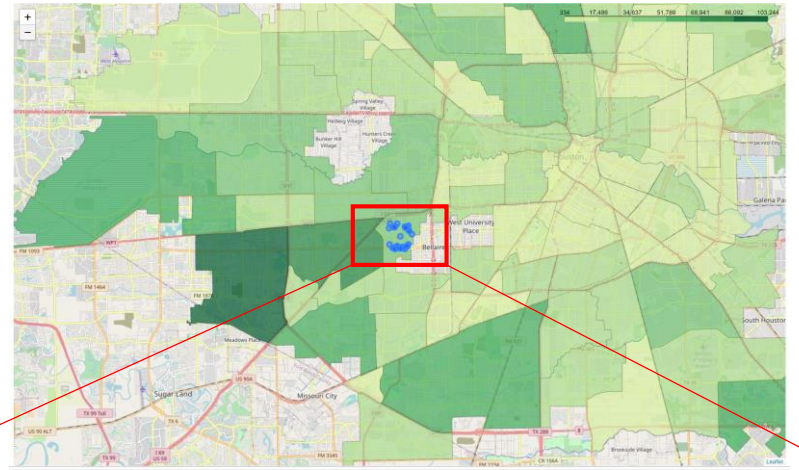
```
[16]: # Descriptive statistical analysis
df_data_0.describe()
```

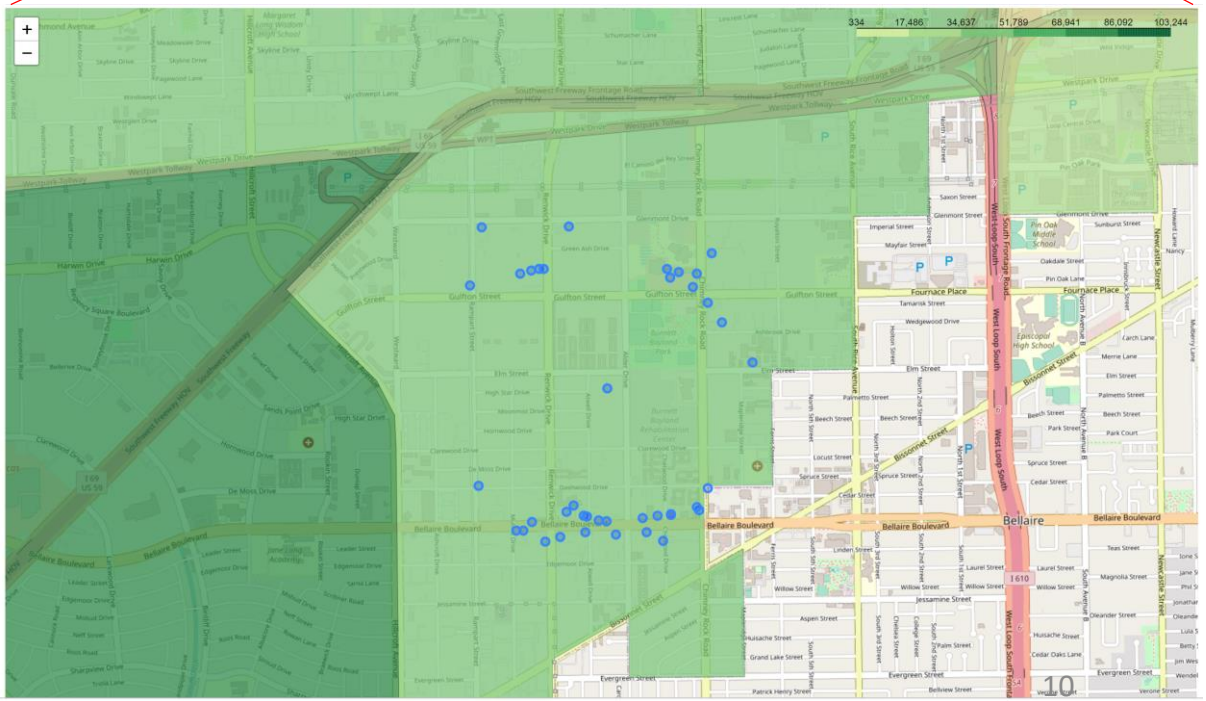
[16]:

	OBJECTID	SUM_POP100	SUM_HU100	SUM_TotPop	SUM_HispPop	SUM_NonHispPop	SUM_NHOneRace	SUM_NH_White	SUM_NH_Black	SUM_NH_AmInd	...	SUM_VAP_NH_Asi	SUM_VAP_HawPac	SUM_VAP_NH_Oth	SUM_VAP_NH_2or
count	88.000000	88.000000	88.000000	88.000000	88.000000	88.000000	88.000000	88.000000	88.000000	88.000000	...	88.000000	88.000000	88.000000	88.000000
mean	44.500000	23485.784091	9971.477273	23485.784091	10312.284091	13173.500000	12922.329545	6015.704545	5424.238636	39.170455	...	1139.022727	6.272727	31.147727	161.636364
std	25.547342	18836.827846	7982.989758	18836.827846	10766.027192	12359.164964	12095.191704	8740.416109	6462.416309	35.533446	...	2165.354055	8.073979	40.989357	183.982144
min	1.000000	1343.000000	729.000000	1343.000000	131.000000	318.000000	316.000000	15.000000	143.000000	0.000000	...	0.000000	0.000000	0.000000	1.000000
25%	22.750000	11009.250000	4451.250000	11009.250000	2238.000000	3837.500000	3773.250000	612.250000	956.500000	16.750000	...	74.750000	1.000000	8.750000	48.000000
50%	44.500000	19520.500000	7836.500000	19520.500000	6976.500000	9779.500000	9571.000000	2269.000000	2287.500000	30.000000	...	414.000000	3.000000	18.000000	84.500000
75%	66.250000	28986.250000	14956.000000	28986.250000	15736.250000	17496.250000	17296.250000	7241.000000	7575.250000	46.250000	...	1024.250000	9.000000	37.250000	211.000000
max	88.000000	102235.000000	35498.000000	102235.000000	48613.000000	54269.000000	52936.000000	47441.000000	26146.000000	189.000000	...	14842.000000	41.000000	306.000000	847.000000

Explore single neighborhood

- Used Foursquare API to get 42 venues around Gulfton
- Designed limit to get the top 100 venues within a radius of 1000 meters from the neighborhood's latitude and longitude values
- Used python **folium** library to visualize venues in 'Gulfton' neighborhood





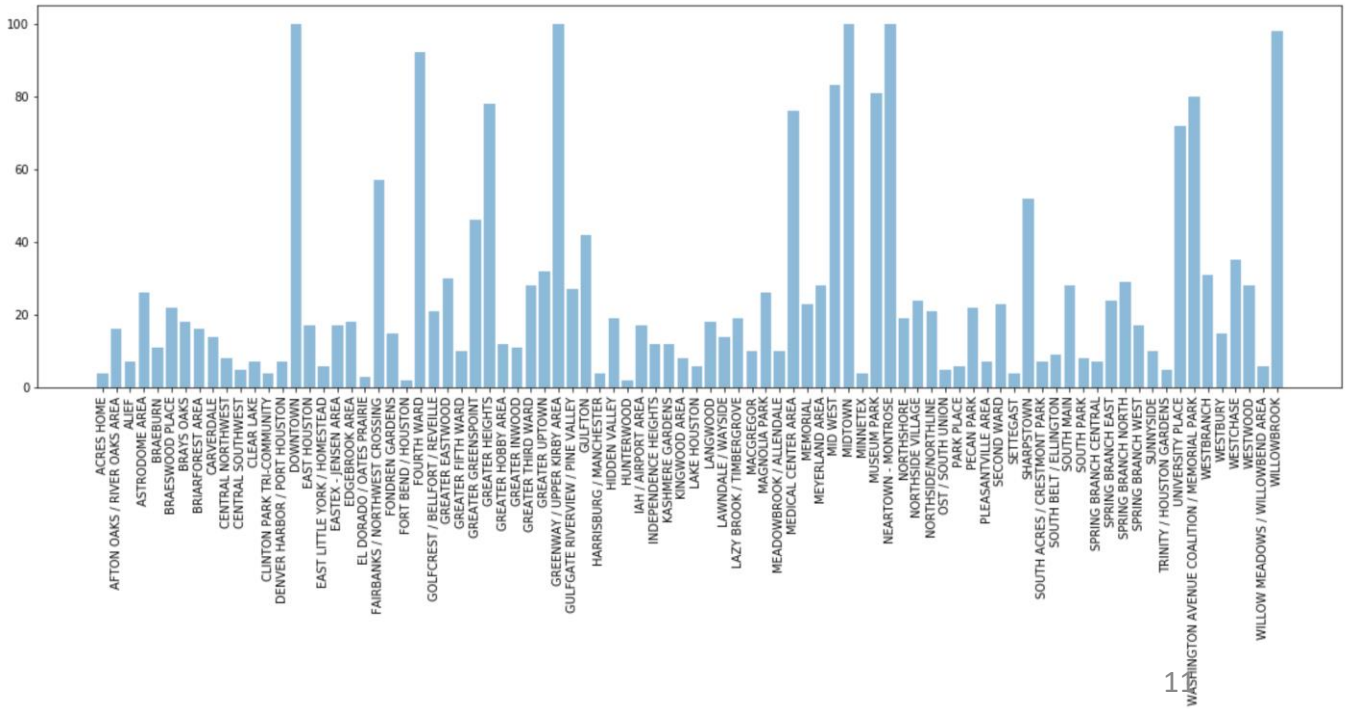
Analyze each neighborhood

- Used Foursquare API calls to get trending venues in each neighborhood
- Designed limit to get the top 100 venues within a radius of 1000 meters from the neighborhood's latitude and longitude values
- Bar plot showed some neighborhoods (Downtown, Midtown, etc.) reached the venue limit of 100
- Created table that shows list of top 10 venue category for each neighborhood
- Foursquare returned 303 unique venue categories
- Results depend on query parameters, and geographical coordinates

```
print(houston_venues.shape)
houston_venues.head(11)
```

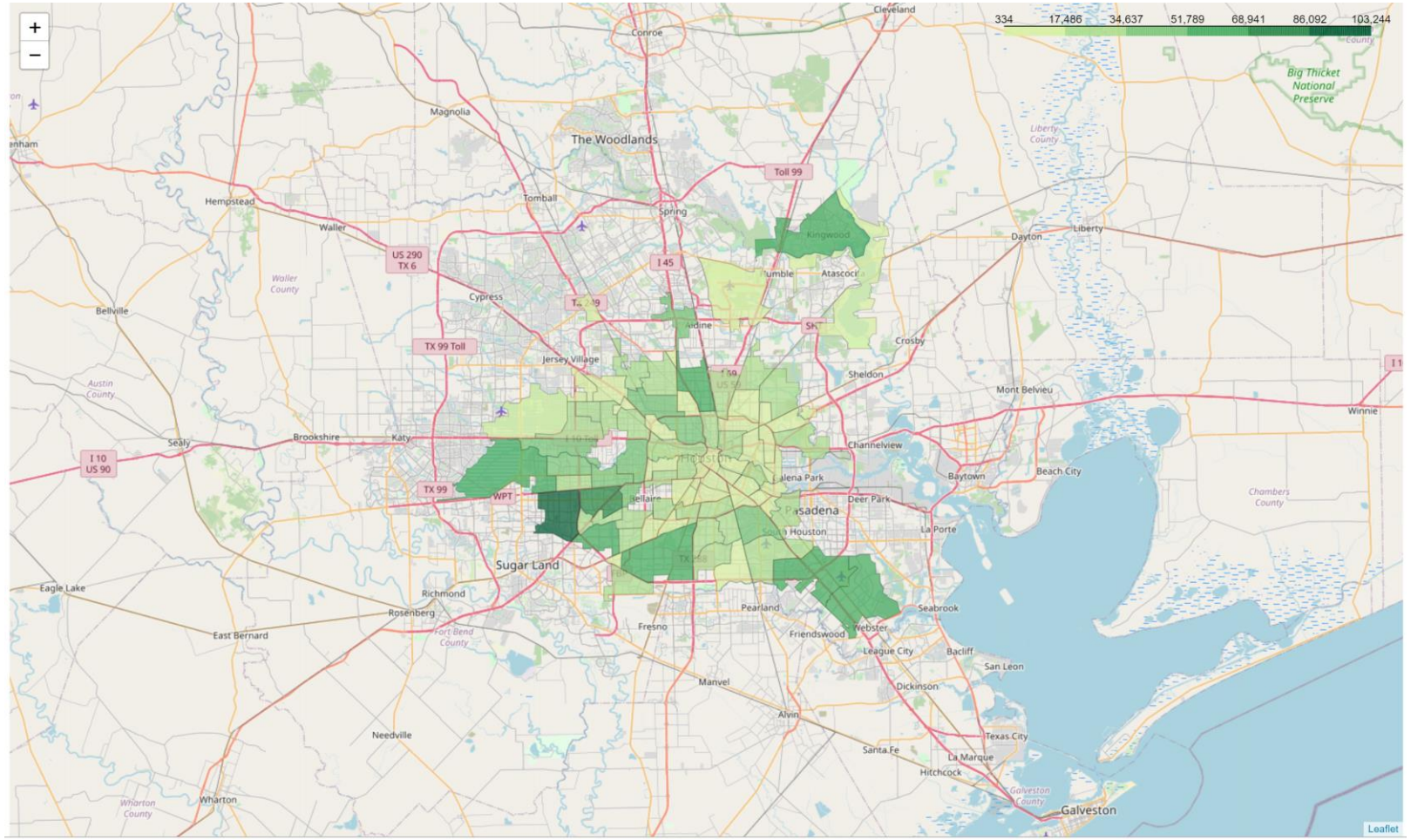
```
(2293, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ACRES HOME	29.864097	-95.436267	Family Dollar	29.872253	-95.437729	Discount Store
1	ACRES HOME	29.864097	-95.436267	METRO Acres Homes Transit Center	29.864708	-95.430969	Bus Station
2	ACRES HOME	29.864097	-95.436267	Vera Steel	29.861840	-95.442030	Construction & Landscaping
3	ACRES HOME	29.864097	-95.436267	Chick Chung	29.863635	-95.428402	Chinese Restaurant
4	AFTON OAKS / RIVER OAKS AREA	29.747921	-95.435636	Frank's Americana Revival	29.742064	-95.439905	New American Restaurant
5	AFTON OAKS / RIVER OAKS AREA	29.747921	-95.435636	The Briar Club	29.740797	-95.436331	American Restaurant
6	AFTON OAKS / RIVER OAKS AREA	29.747921	-95.435636	Bebidas	29.741466	-95.433946	Café
7	AFTON OAKS / RIVER OAKS AREA	29.747921	-95.435636	Whataburger	29.741277	-95.436867	Burger Joint
8	AFTON OAKS / RIVER OAKS AREA	29.747921	-95.435636	Pinkberry	29.742082	-95.442945	Frozen Yogurt Shop
9	AFTON OAKS / RIVER OAKS AREA	29.747921	-95.435636	Ouisie's Table	29.746045	-95.443117	Southern / Soul Food Restaurant
10	AFTON OAKS / RIVER OAKS AREA	29.747921	-95.435636	River Oaks Park	29.741917	-95.434753	Park



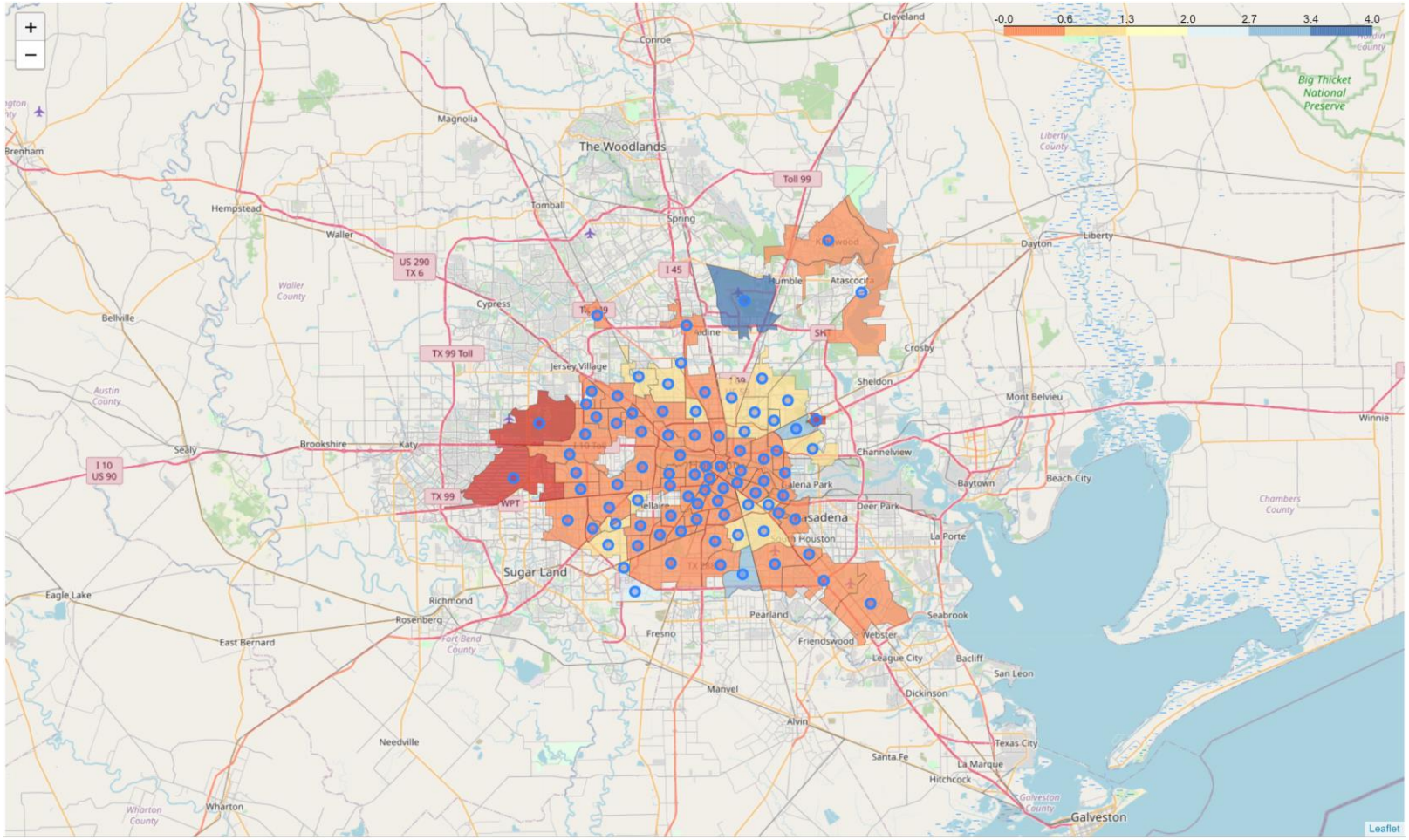
Map of total population

- Choropleth maps used to show total population from census data
- Similar maps could be generated for other variables



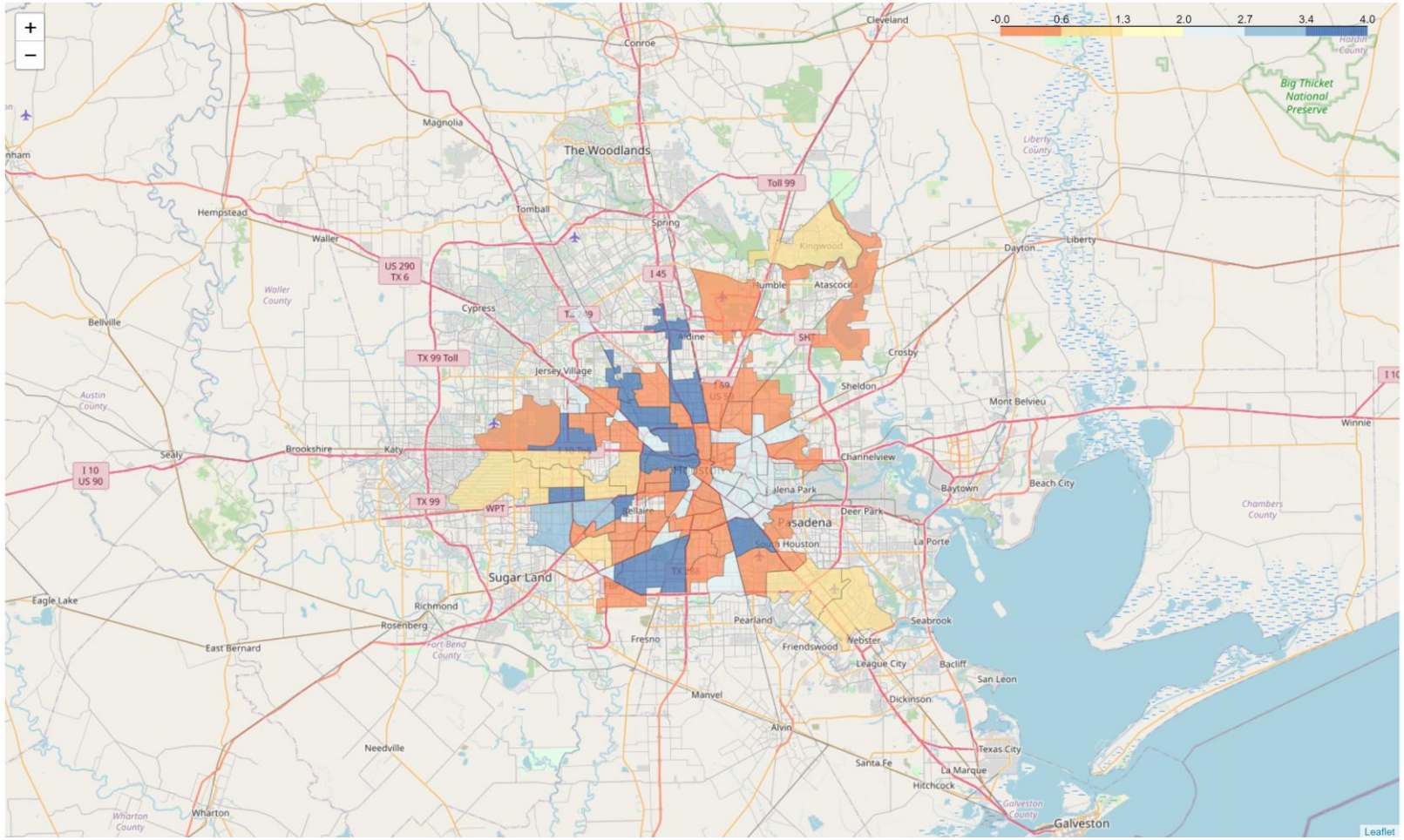
Cluster neighborhood by venue data

- Used unsupervised learning k-means algorithm to cluster the neighborhoods
- Ran k-means to cluster the neighborhood into 5 clusters
- Used census demographics data only for this run



Cluster neighborhood by census data

- Used unsupervised learning k-means algorithm to cluster the neighborhoods
- Ran k-means to cluster the neighborhood into 5 clusters
- Used venues data only for this run



Results

- Descriptive statistical analysis was done to gain insights
- Census and venue data were used to cluster neighborhood
- In future other data (household, income, etc.) can be visualized along with cluster data for further analysis

Discussions

- Datasets were read from the Houston GIS open data portal
- Data wrangling was done to convert spreadsheet and GeoJSON data to Pandas dataframes for analysis
- Histograms were plotted to see the distribution of total population
- Scatterplots (with fitted regression lines) were generated using ‘regplot’ to understand relationship between variables
- Polygon coordinates were converted into their centroid latitude and longitude values
- Foursquare API was used to explore neighborhoods in Houston
- The ‘explore‘ function was used to get the most common venue categories in each neighborhood, and this feature was then used to group the neighborhoods into clusters
- The k-means clustering algorithm was used to complete this task
- Data were normalized so the variable average is 0 and variance is 1 before clustering
- Finally, the Folium library was used to visualize the neighborhoods in Houston and their emerging clusters
- Compared clusters obtained from two datasets (demographics and venues)

Cluster by demographics data

```
df['Clus_km'].value_counts()

0    37
2    28
4    12
1     9
3     2
Name: Clus_km, dtype: int64
```

Cluster by venue data

```
houston_merged['Cluster Labels'].value_counts()

0.0    62
1.0    19
3.0     2
4.0     1
2.0     1
Name: Cluster Labels, dtype: int64
```

Conclusions

- The main goal of this project was to analyze data available in public domain to identify zones of opportunity.
- This in turn would help to attract both practical and innovative investment into underinvested communities while leveraging local and state resources.
- This information would be useful for area residents and stakeholders that serves as a forum to discuss issues and identify and implement priority projects for the area.
- Most of the data (census, venue, etc.) are updated on a regular basis.
- Therefore, it is important to constantly update the underlying model based on newly available data as necessary.
- In future other data (household, income, etc.) can be visualized along with cluster data for further analysis.