Battle of Neighborhoods

CHICAGO, ILLINOIS

BY
Ajay Kumar Mukhi

Problem Statement

- To recommend the best neighborhood to live, to buy a house, to rent an apartment or build a restaurant etc in Chicago, Illinois .
- To understand the similarities and differences between the neighborhoods using Unsupervised K-Mean Clustering Algorithm.

Objective

- Collecting the top trending venues in the using Foursquare API(Beautiful Soup, http request)
- Forming neighborhood clusters based on venue categories using unsupervised k-mean clustering algorithm(sklearn)
- Identifying and understanding the similarities and differences between two chosen neighborhoods to retrieve more insights and to conclude which neighborhood wins over other.

Python packages and Dependencies:

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library

Work flow

- Web Scraping and Data Wrangling
- Top Trending Places Extraction and Clustering
- Decision Making based on the clustered neighborhoods, Population Distribution, School Ratings, Median House Price Analysis

Web Scraping and Data Wrangling

Beautiful Soup

Collecting Neighborhood/Postal code

http://seattlearea.com/zip-codes/

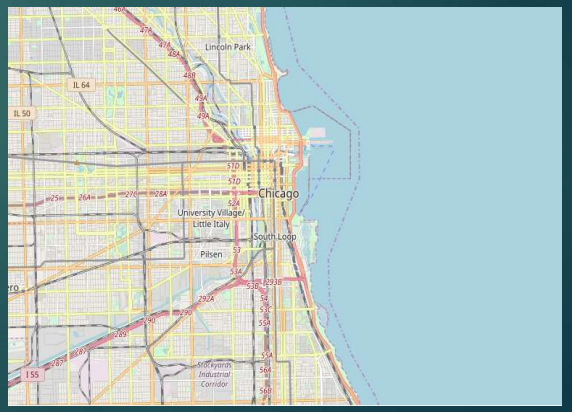
Google Maps API

To Collect Geographical Data

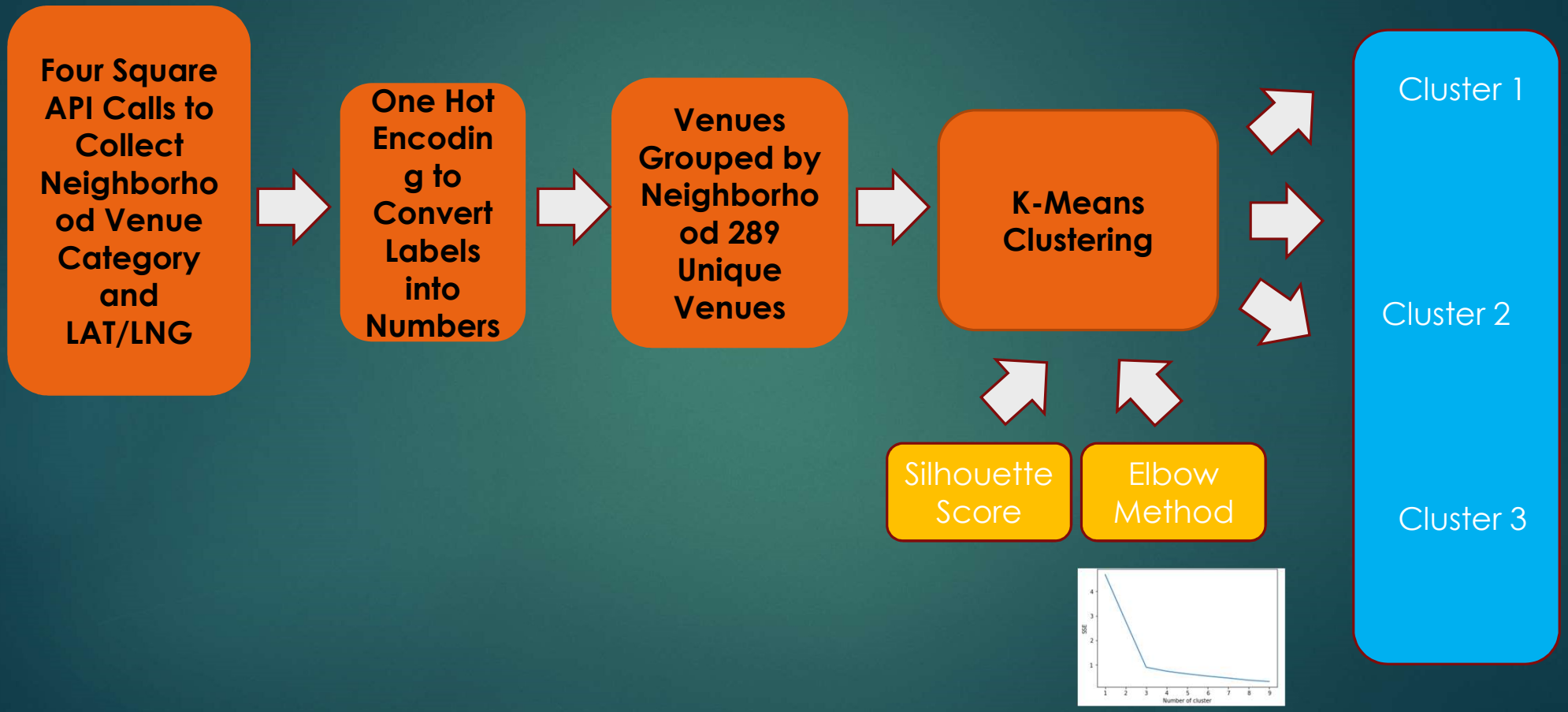
	PostalCode	Neighborhood
0	60625	Albany Park
1	60630	Mayfair
2	60630	North Mayfair
3	60625	Ravenswood Manor
4	60632	Archer Heights

	PostalCode	Neighborhood	Latitude	Longitude
0	60625	Albany Park	41.971107	-87.702482
1	60630	Mayfair	41.972079	-87.751729
2	60630	North Mayfair	41.972079	-87.751729
3	60625	Ravenswood Manor	41.971107	-87.702482
4	60632	Archer Heights	41.807469	-87.707409

Folium Visualization for Chicago Neighborhood

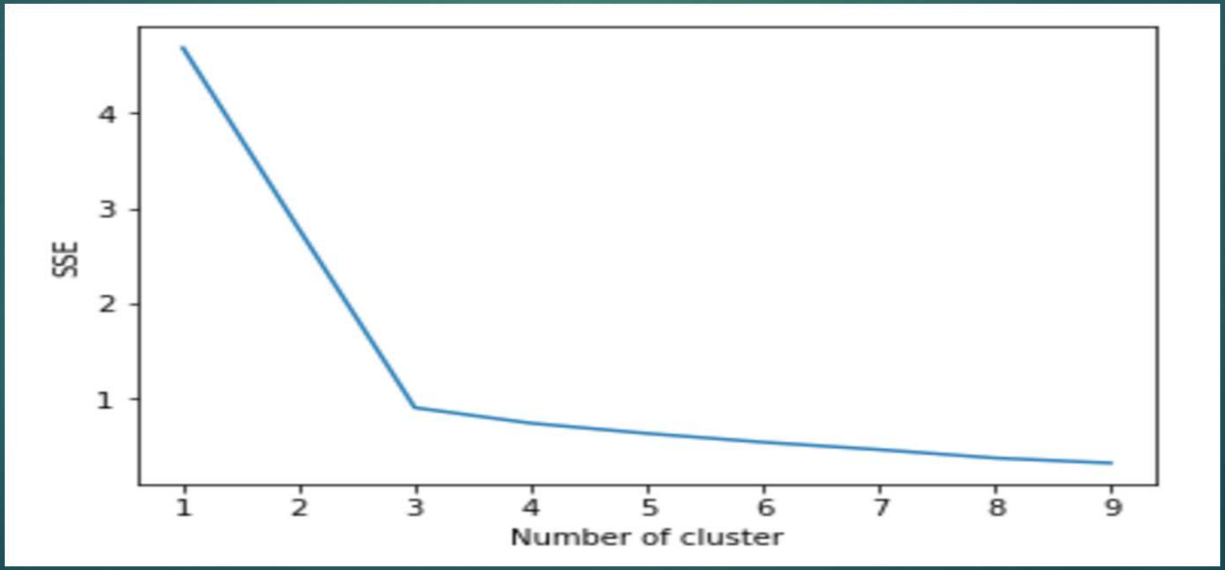


Venues Extraction using Four Square API and Clustering



Elbow Criterion Method

Elbow method is to run k-means clustering on a given dataset for a range of values of k and for each value of k and calculate sum of squared errors (SSE).



sklearn.metrics.silhouette_score

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample.

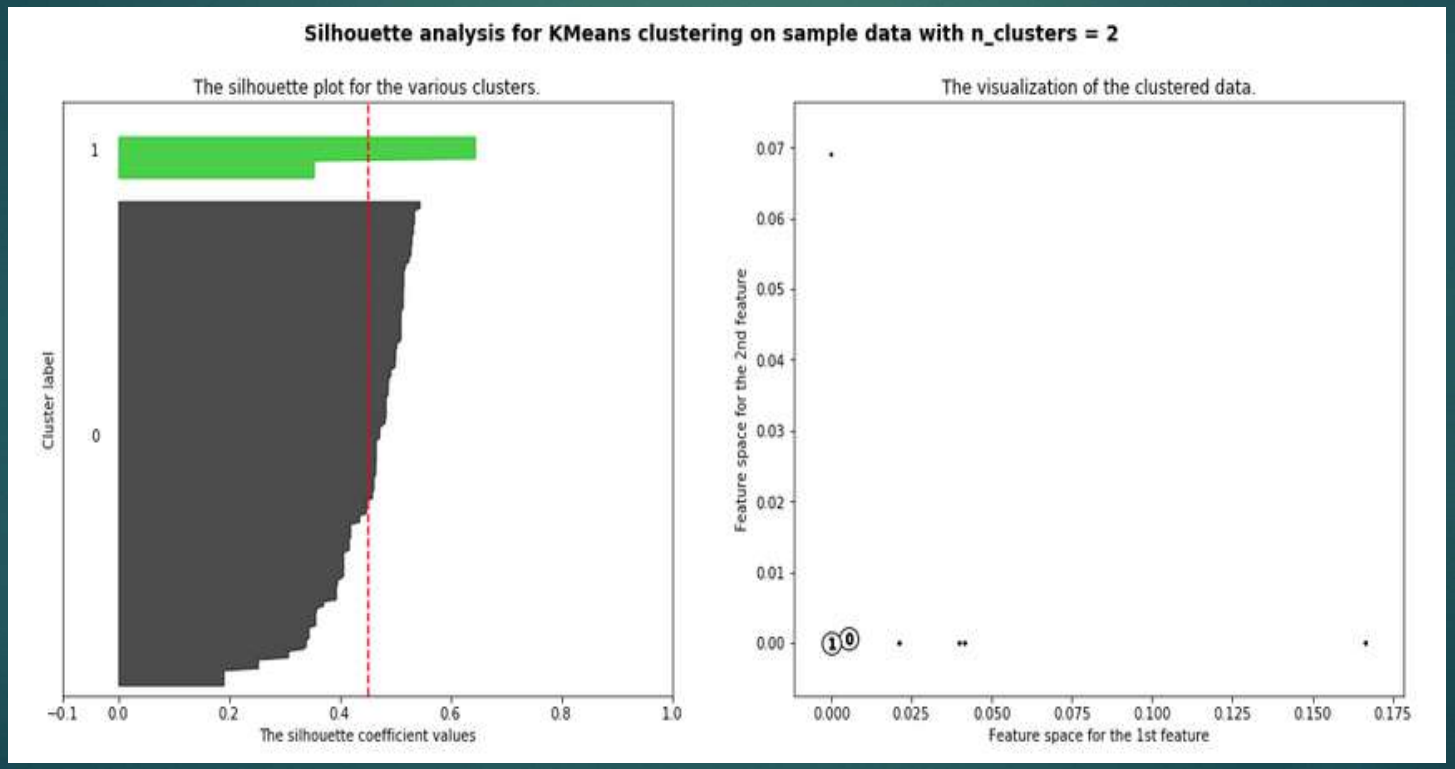
The formula for the Silhouette Coefficient of a sample is

(b - a) / max(a, b).

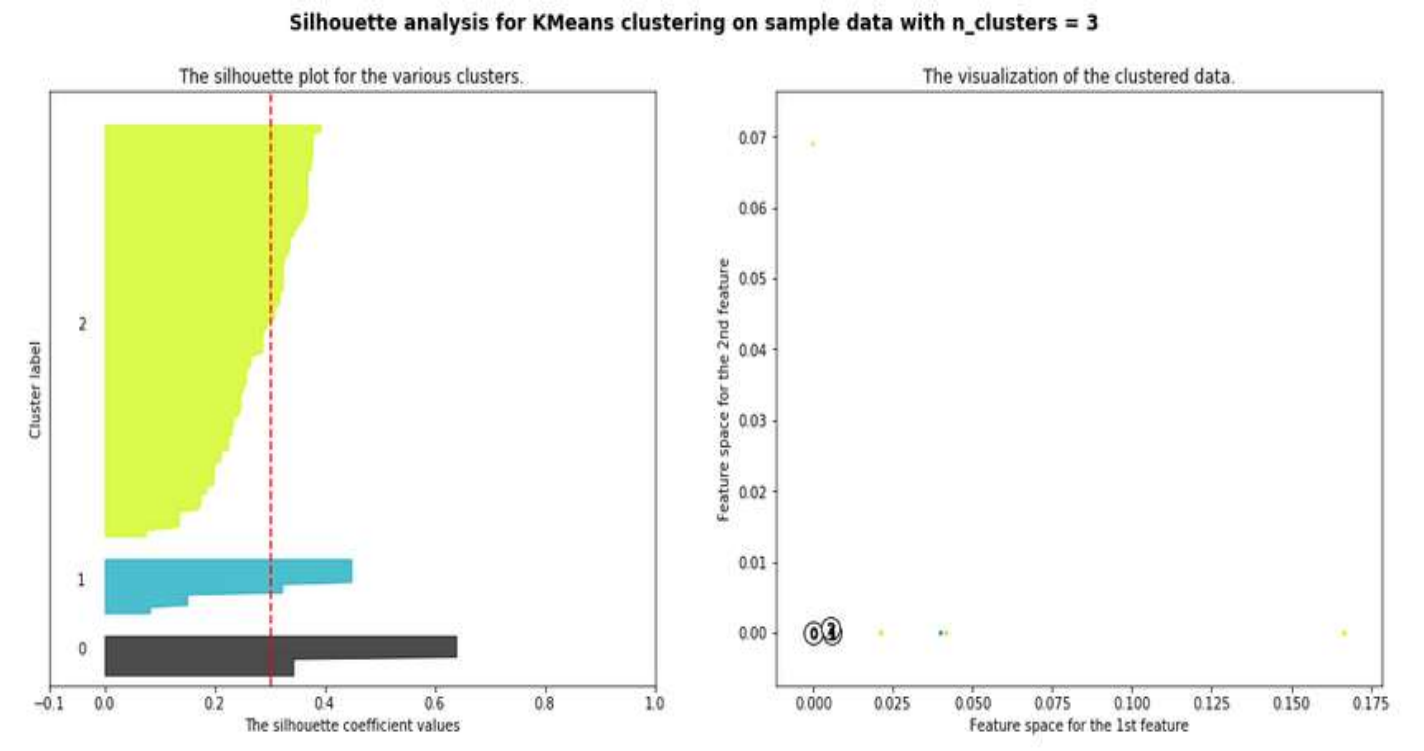
The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster.

Siluhouette Score and Cluster Visualizations

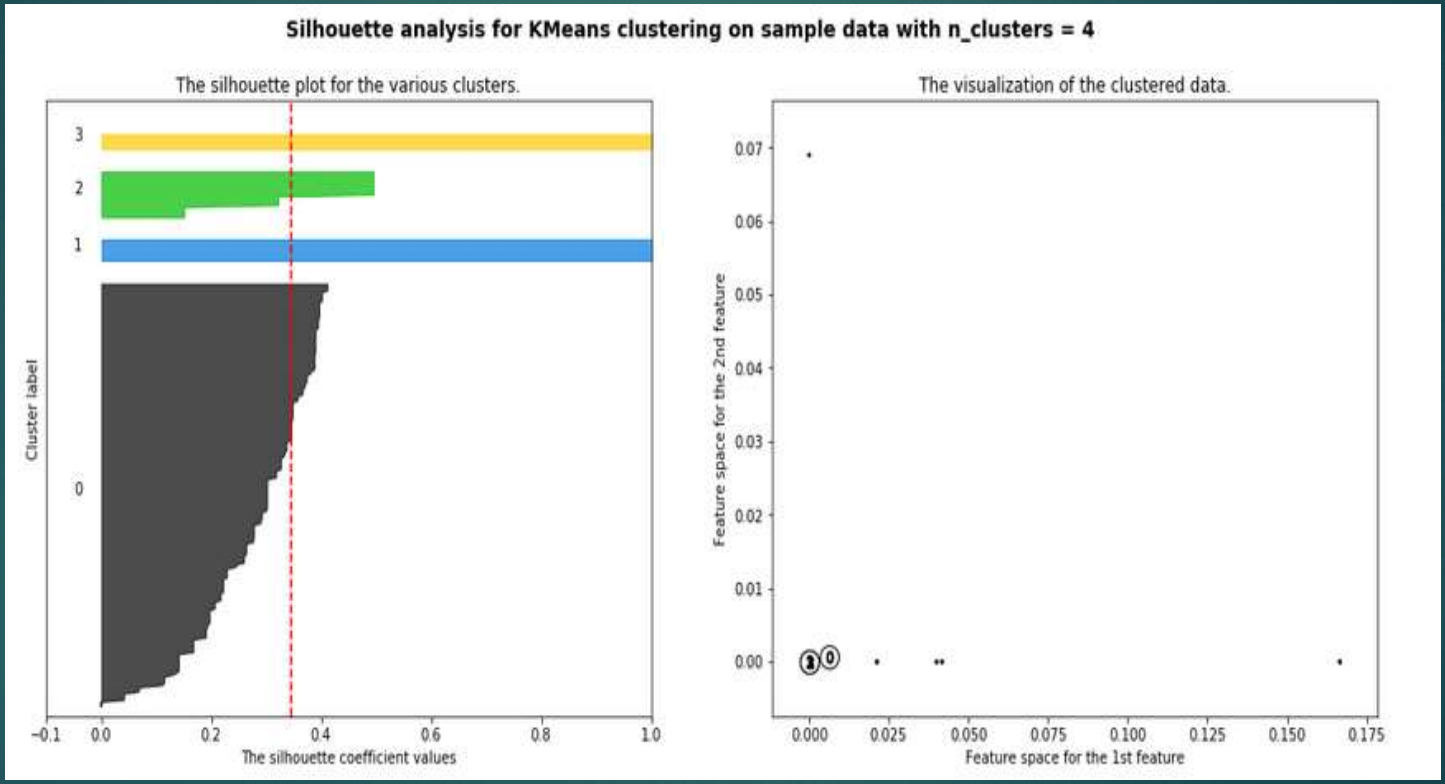
```
For 2  Clusters  the average silhouette_score is : 0.45129474688303467
For 3  Clusters  the average silhouette_score is : 0.30224267891889856
For 4  Clusters  the average silhouette_score is : 0.3449303029949109
For 5  Clusters  the average silhouette_score is : 0.3858319904775224
```



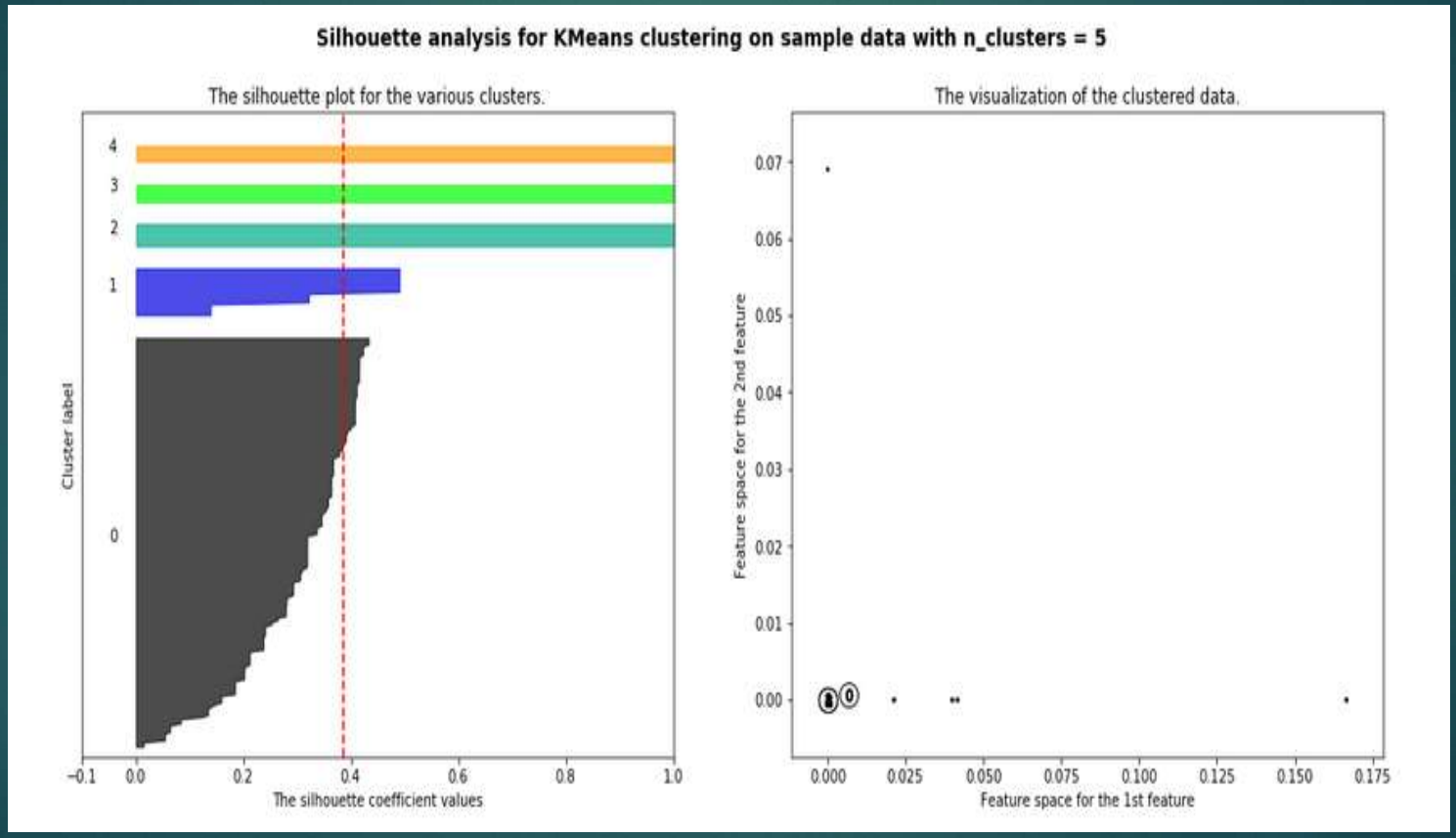
Siluhouette Score and Cluster Visualizations



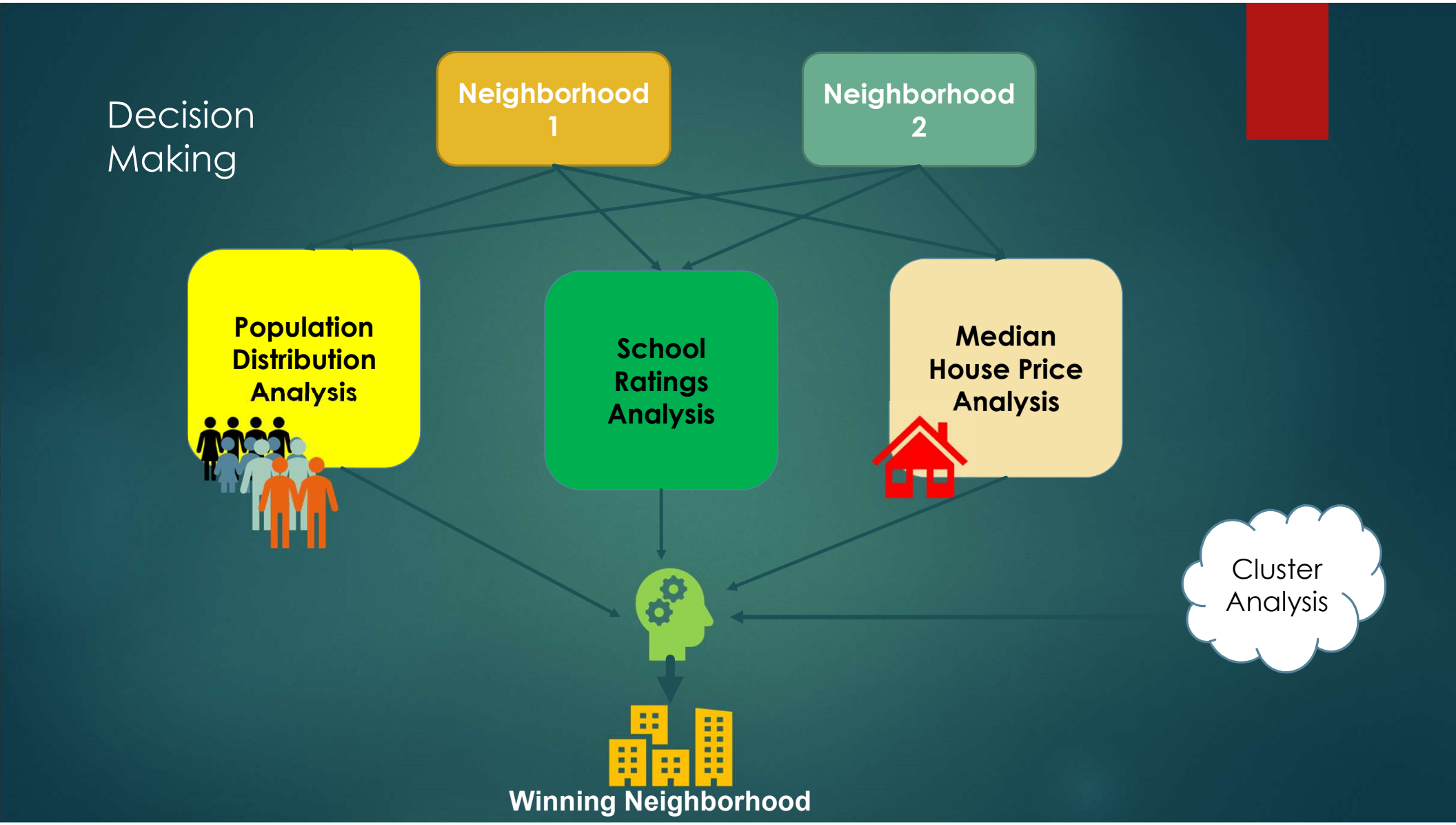
Siluhouette Score and Cluster Visualizations



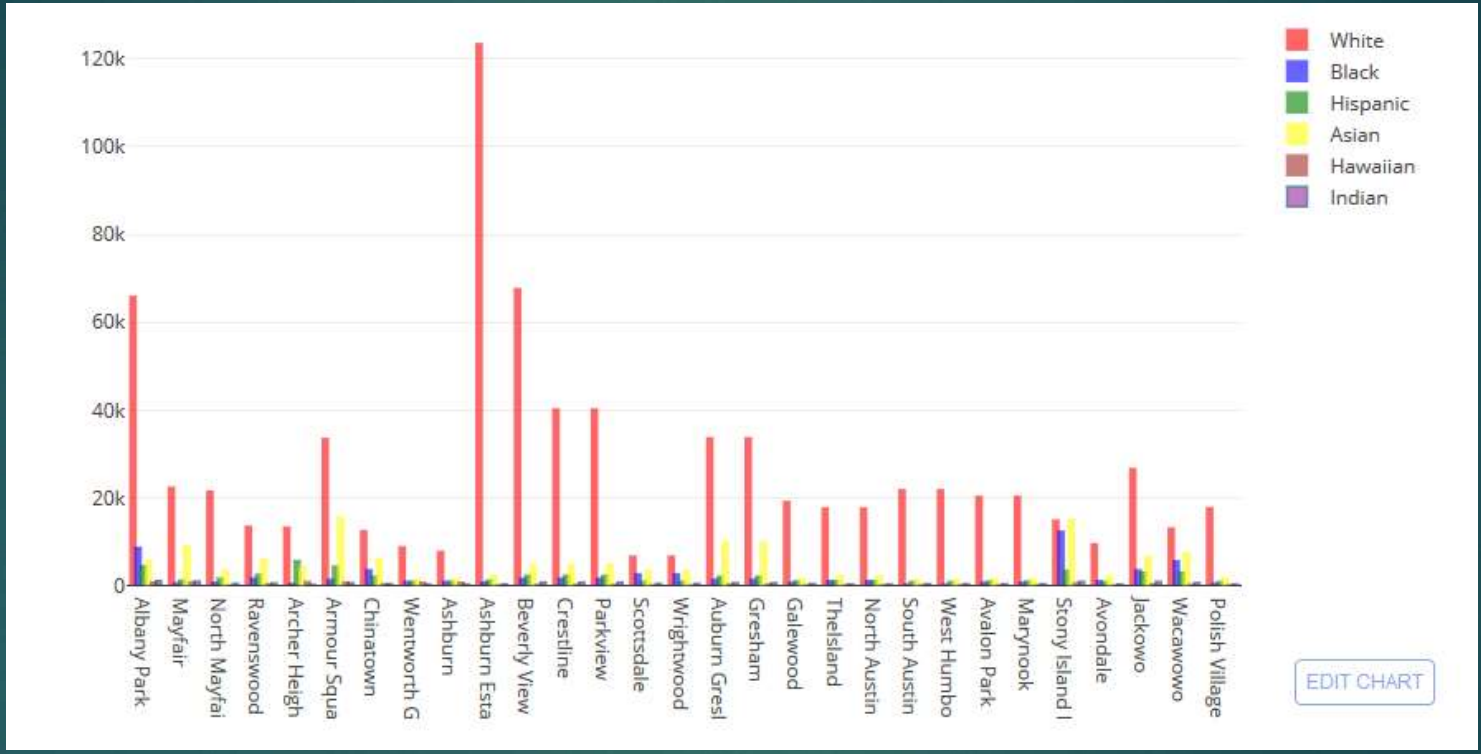
Siluhouette Score and Cluster Visualizations



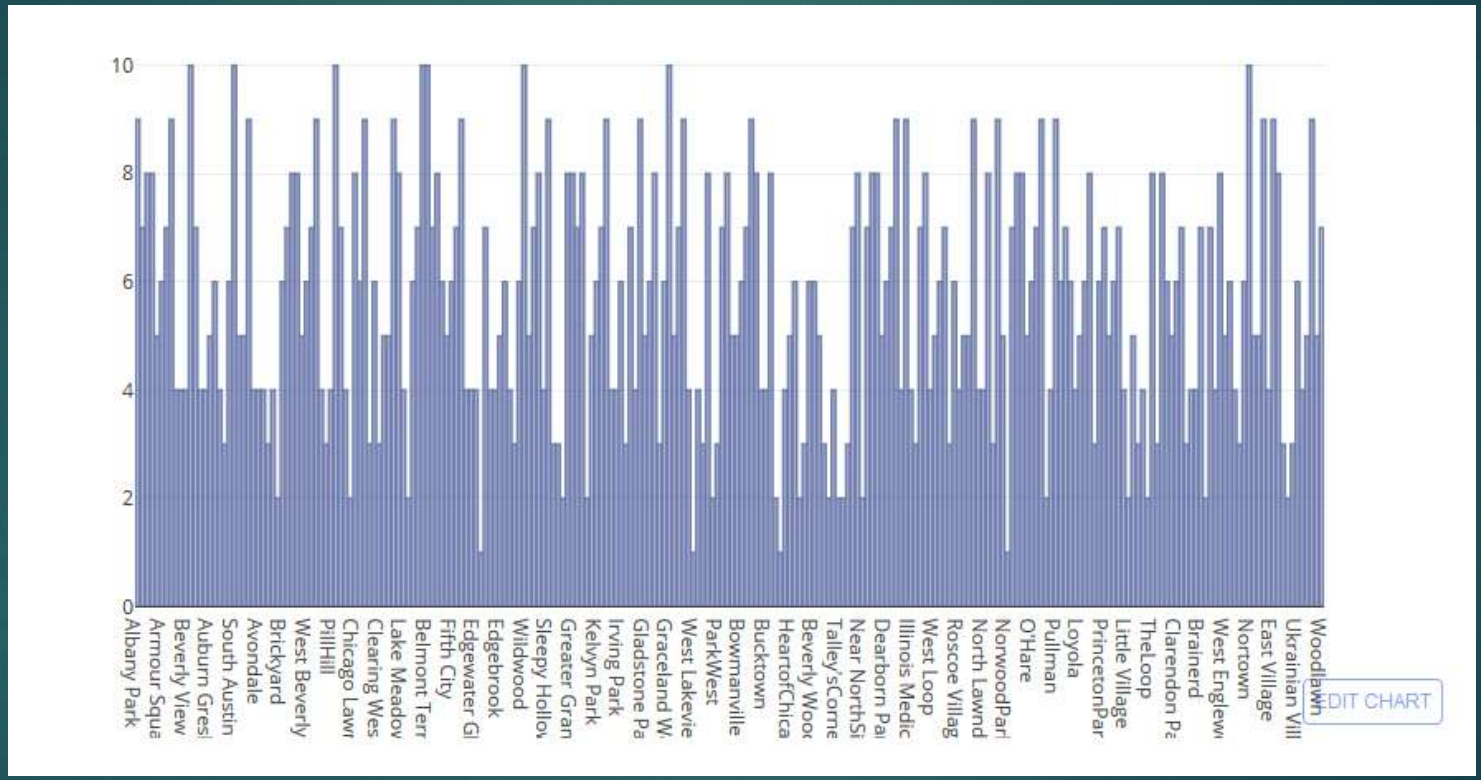
Decision Making



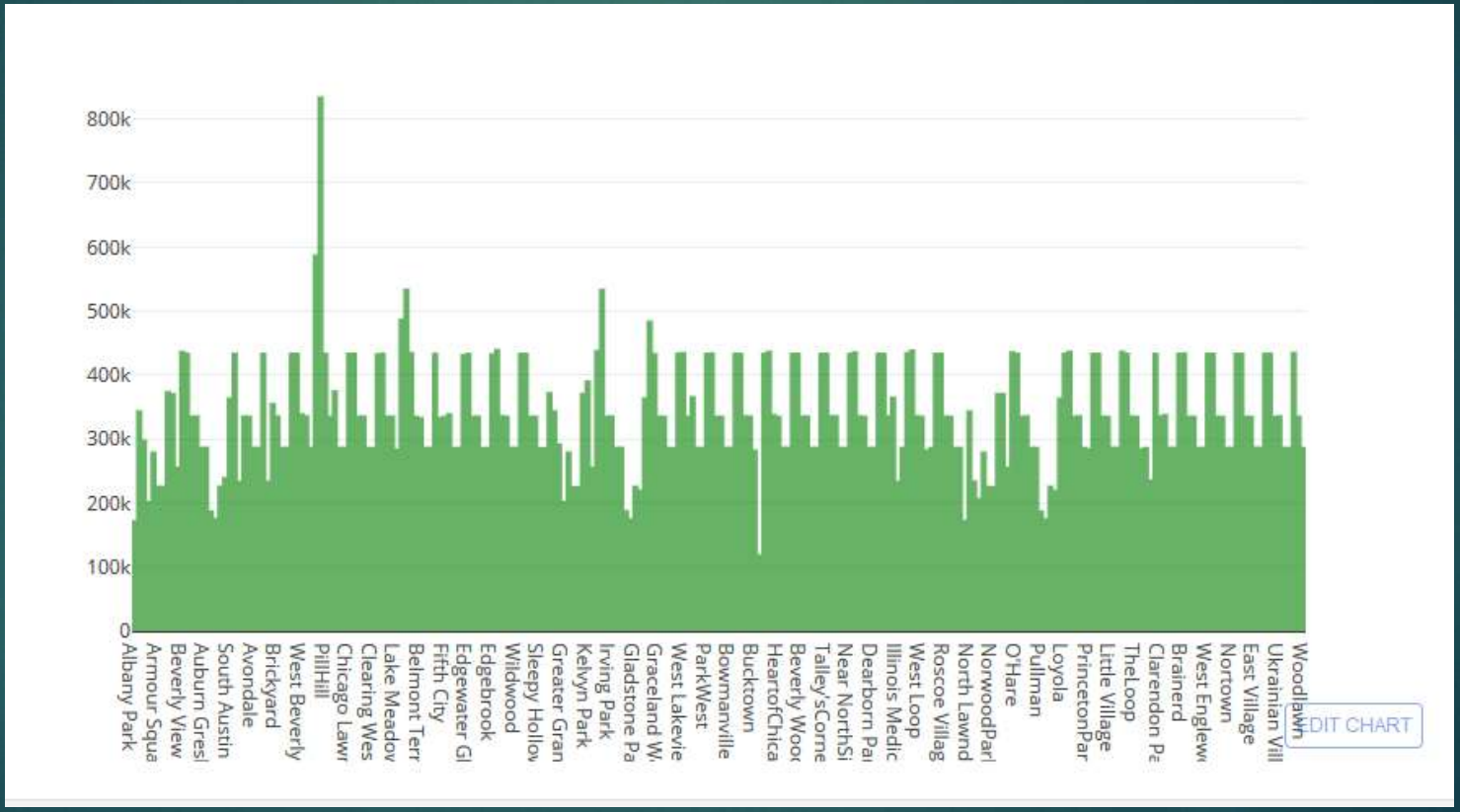
POPULATION DISTRIBUTION



AVERAGE SCHOOL RATINGS



MEDIAN HOUSE PRICE



Comparison between Neighborhoods - Chicago

Armour Square and Parkview

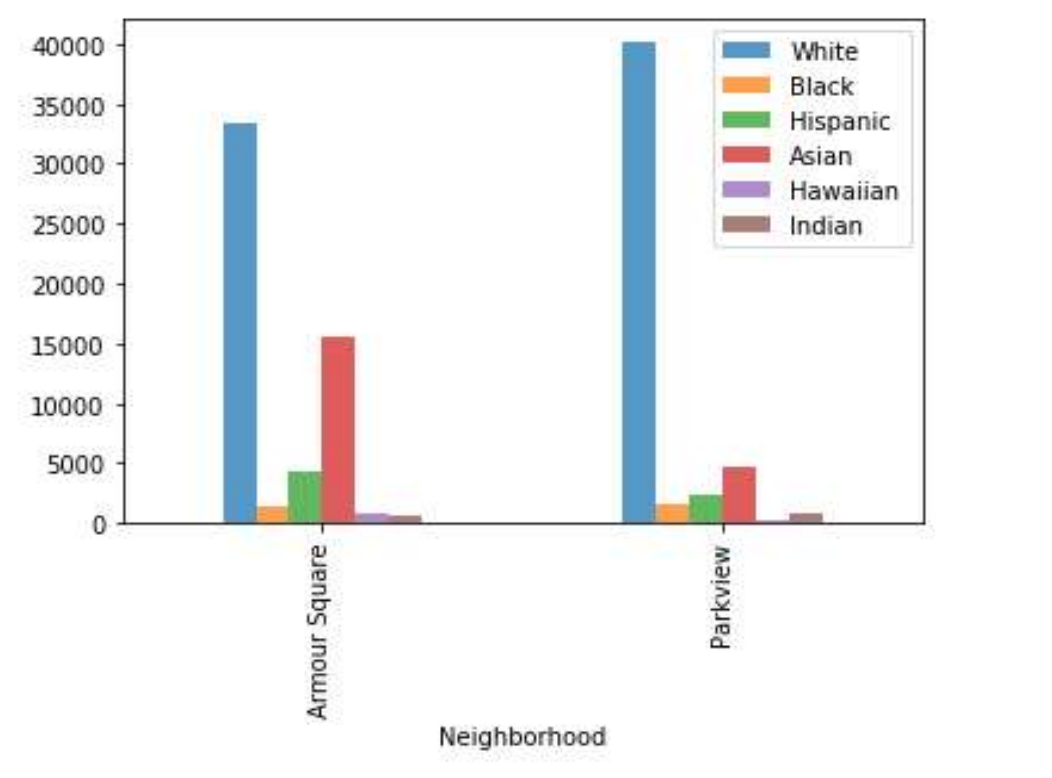
Now lets compare 2 neighborhoods to choose one that best matches our requirements as given below

1. More Indian Population
2. Higher School Rating
3. Reasonable Housing Price in the Range of 300k to 500k
4. Comfortable Neighborhoods

Neighborhood Venues

Neighborhood	Armour Square	Parkview
PostalCode	60609	60652
Latitude	41.8063	41.7429
Longitude	-87.6482	-87.7123
Cluster Labels	1	1
1st Most Common Venue	Discount Store	Gas Station
2nd Most Common Venue	Donut Shop	Italian Restaurant
3rd Most Common Venue	Pharmacy	American Restaurant
4th Most Common Venue	Coffee Shop	Liquor Store
5th Most Common Venue	Supermarket	Clothing Store
6th Most Common Venue	Event Space	Train Station
7th Most Common Venue	Dry Cleaner	Park
8th Most Common Venue	Dumpling Restaurant	Yoga Studio
9th Most Common Venue	Eastern European Restaurant	Electronics Store
10th Most Common Venue	Electronics Store	Donut Shop

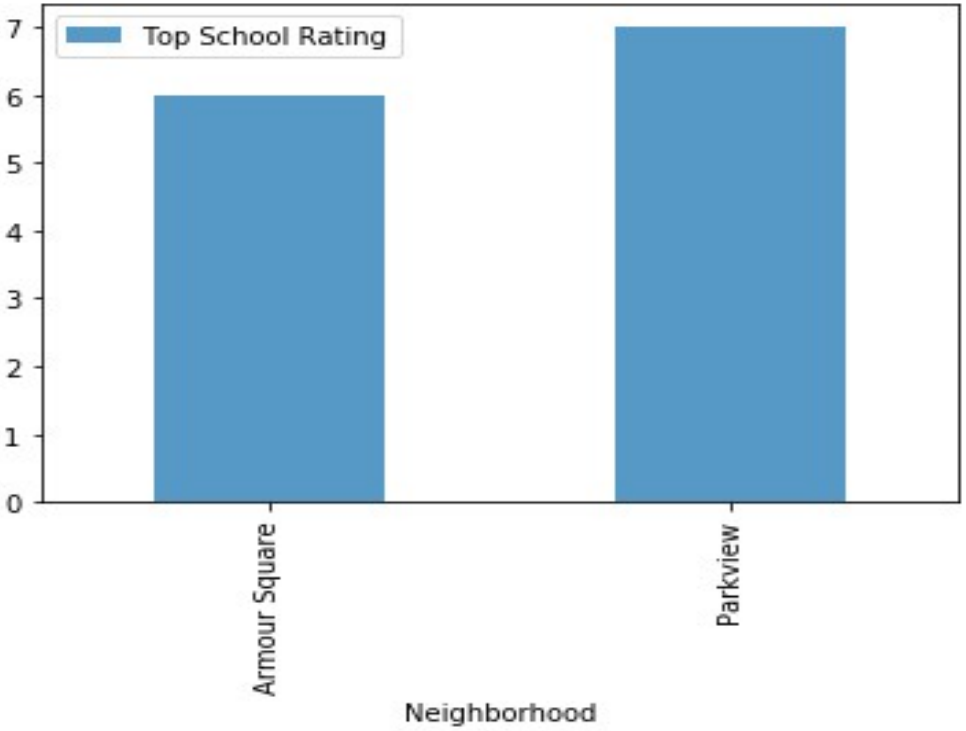
Population distribution




```
Population_Comparison['Indian']
```

```
Neighborhood
Armour Square     618
Parkview          783
```

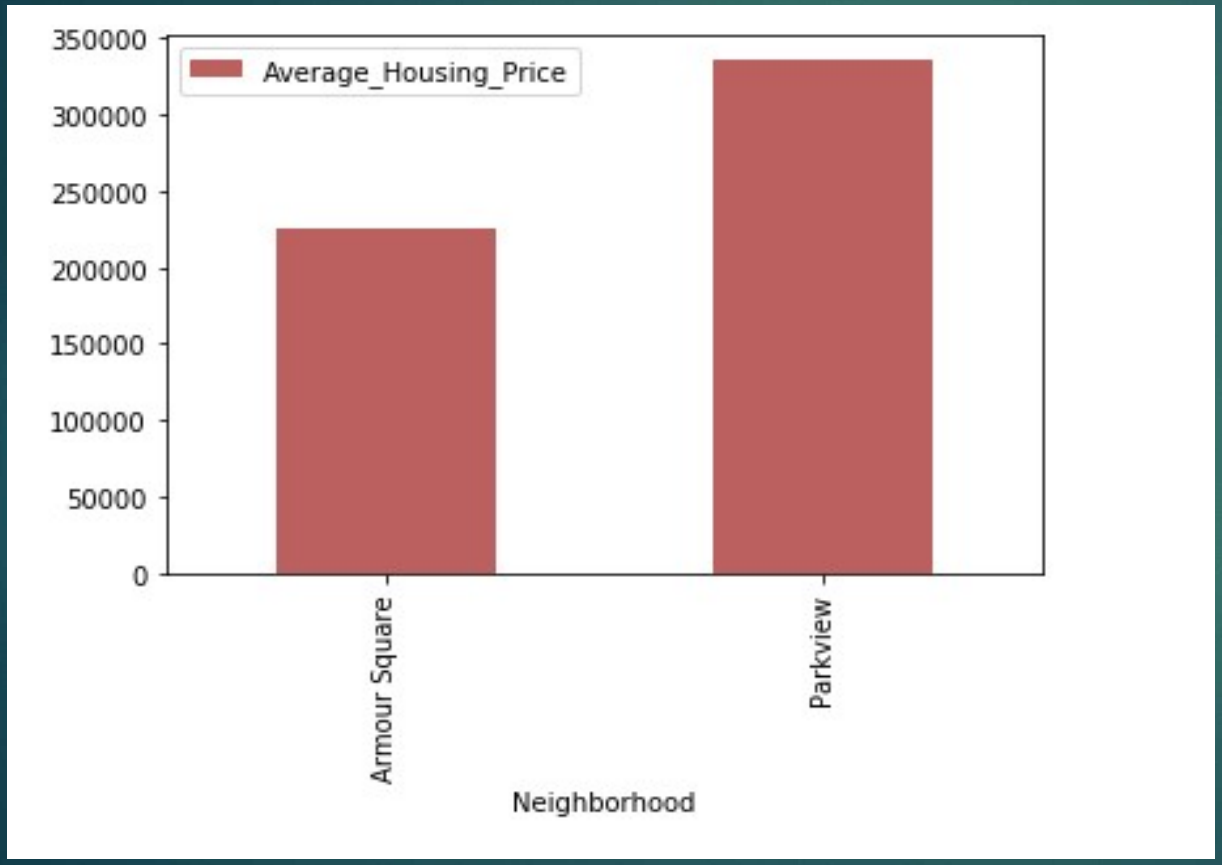
School ratings



School_rating_comparison

	Top School Rating
Neighborhood	
Armour Square	6
Parkview	7

Average housing price



	Average_Housing_Price
Neighborhood	
Armour Square	225740.0
Parkview	335060.0

Conclusion

- This Analysis concludes that compared to Bellevue ,
- Parkview has the higher number of population (including Indians)
- Good school rating of 7
- Reasonable average housing price of approximately 330k
- also top 10 common venues shows Parview has got a good neighborhood with Gas station, Italian and American Restautrant, Train Station, Clothing Store , Park, Donut Shop and many more.

Hence Parview wins over Armour Square!


Thank
you!!