
Clustering of
New York City
Neighborhoods

Introduction

Segmentation of NYC neighborhoods based on **real estate data** and **location data**

Why is this study valuable?

1

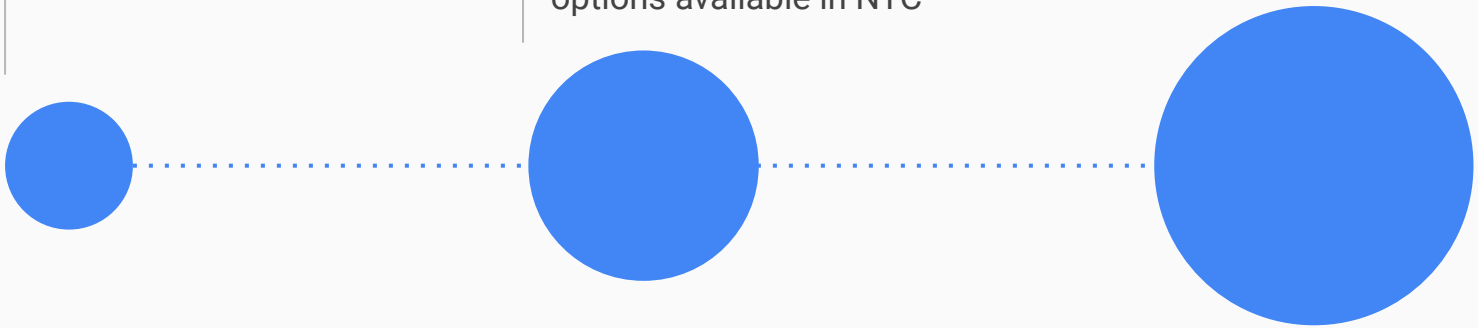
Millennials are becoming first time homeowners

2

Hard to pick as various factors to take into account and many options available in NYC

3

Help real estate companies to find new neighborhoods in NYC to develop new projects



Data

Location Data

- Coordinates of NYC neighborhoods (IBM)
- Foursquare API data for venues
  - Food & Drink Shop
  - Medical Center
  - Residence
  - School
  - Outdoors & Recreation

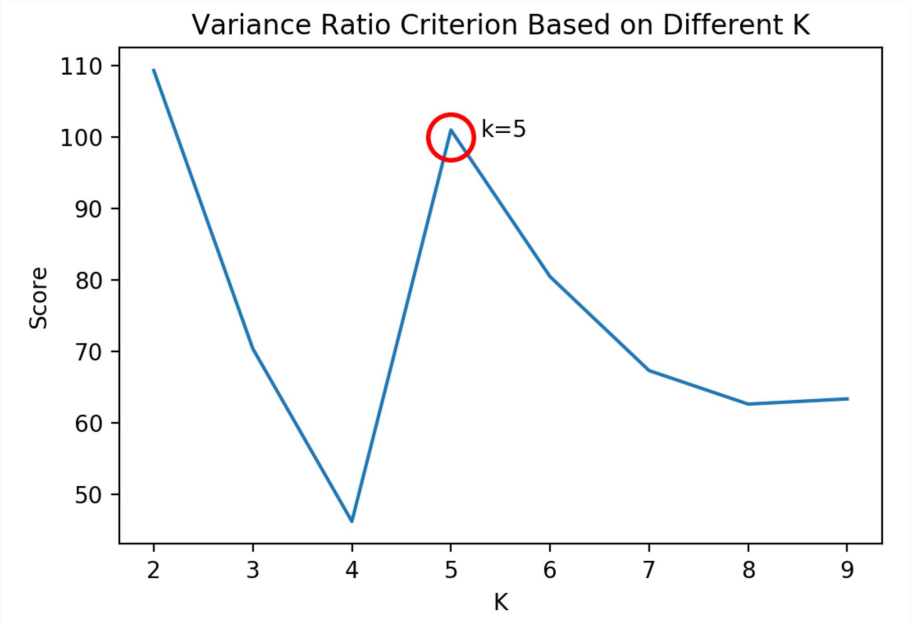
Real Estate Data

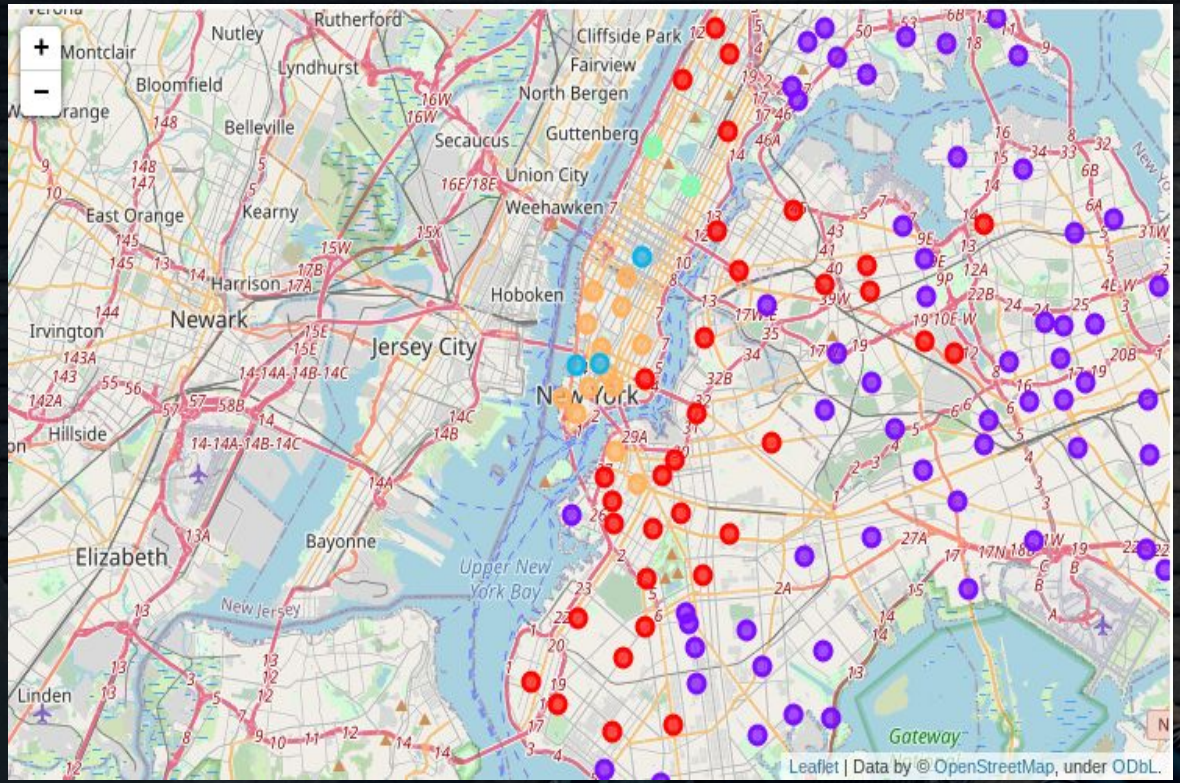
- Q1 2020 average total inventory (StreetEasy)
- Q1 2020 Median price of houses for sale (StreetEasy)

Methodology

K-means clustering algorithm is selected for segmenting NYC houses for sale.

Based on the variance ratio criterion, k=5 is picked for K-means.





5 Clusters determined by K-means algorithm

Basic Characteristics of the Five Clusters

Cluster 1

- Close to Manhattan
- Medium prices
- 20-30 minutes commute to Manhattan
- Various venues

Cluster 2

- Far from Manhattan
- Low prices
- Mainly in Bronx and Queens
- LImited venues

Cluster 3

- Midtown
- Very high prices
- Abundant venues

Cluster 4

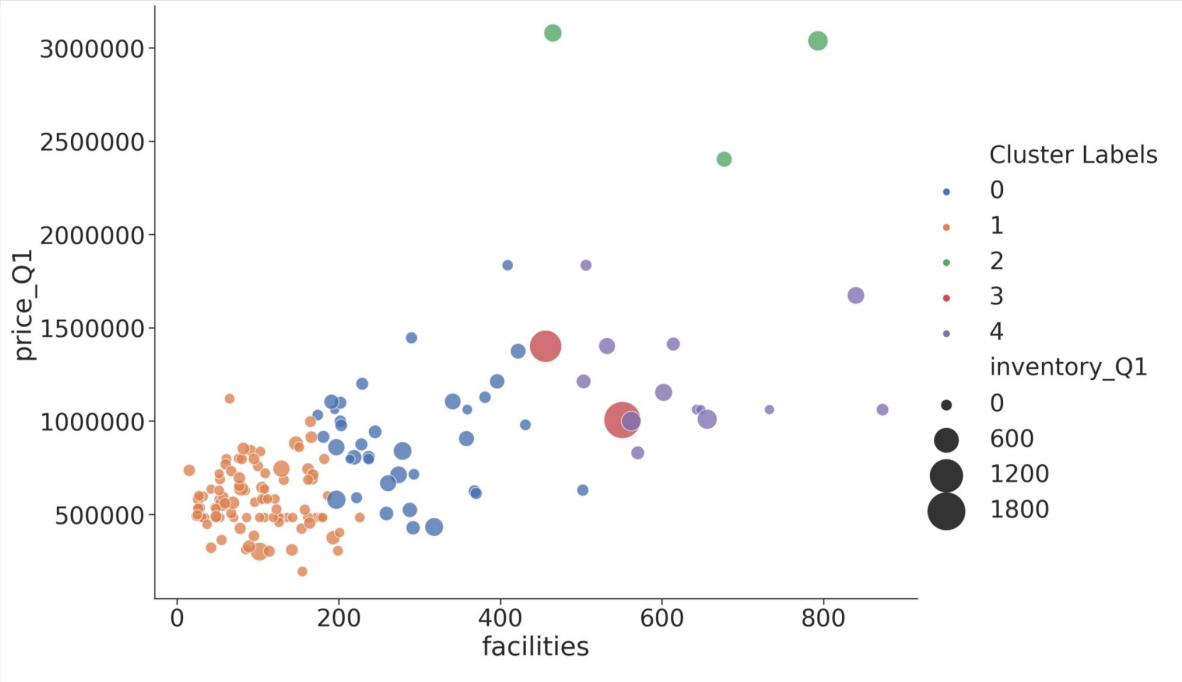
- UES and UWS
- Amble inventory
- Abundant venues

Cluster 5

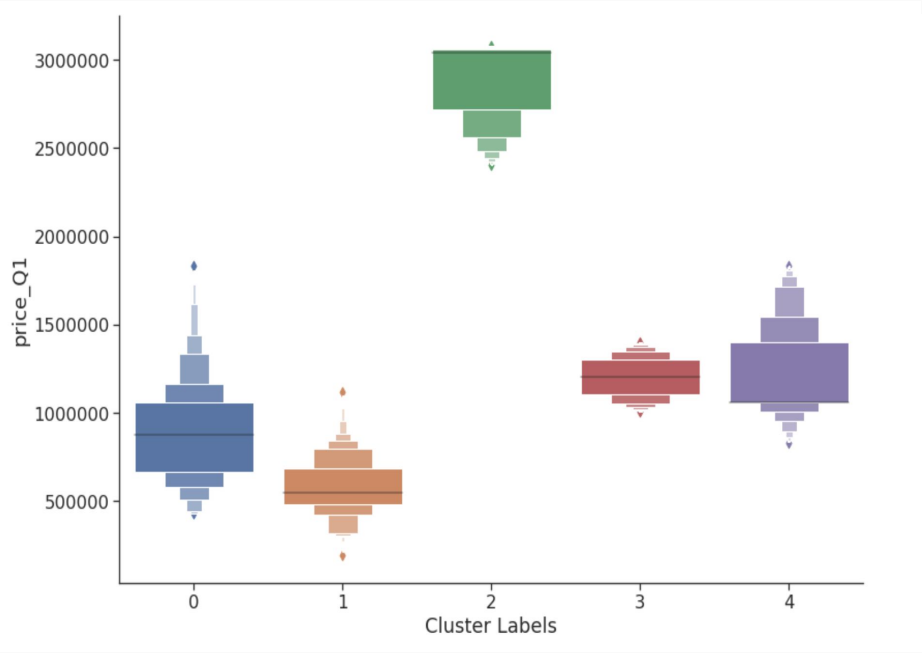
- Very close to Manhattan
- High prices
- Plentiful venues

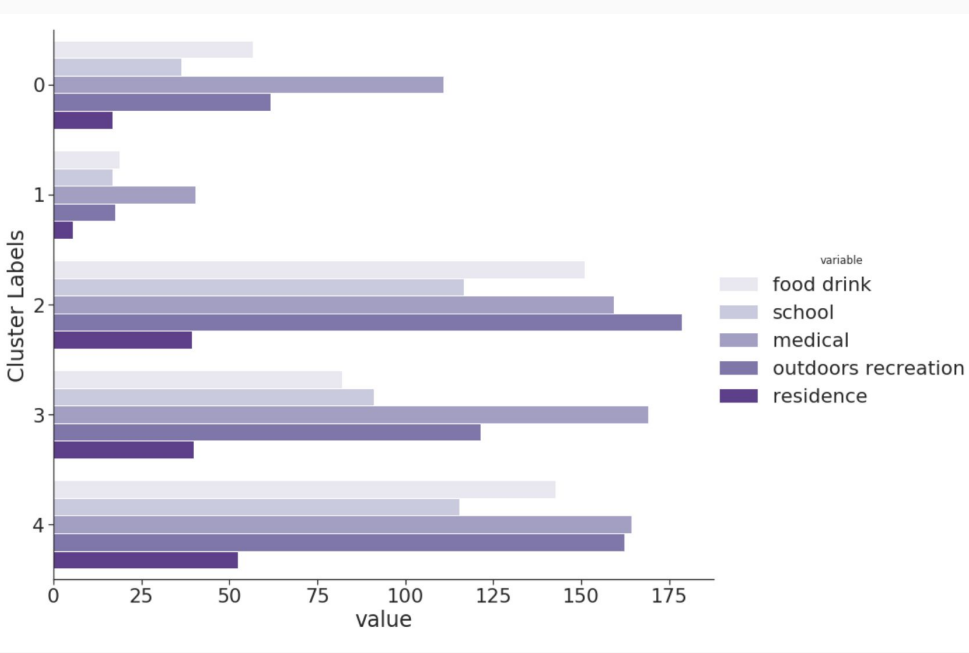
Overall Distribution of the Clusters

Distribution of the clusters regarding housing price, inventory, and total number of venues.

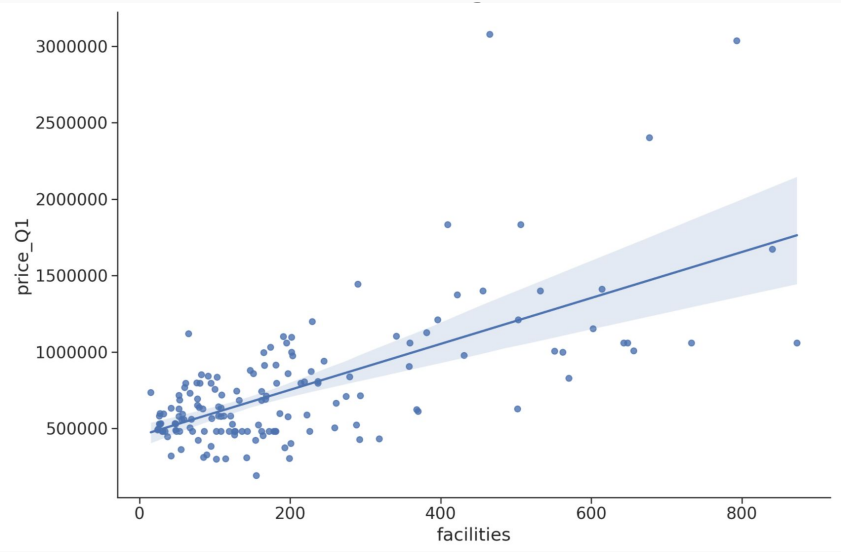


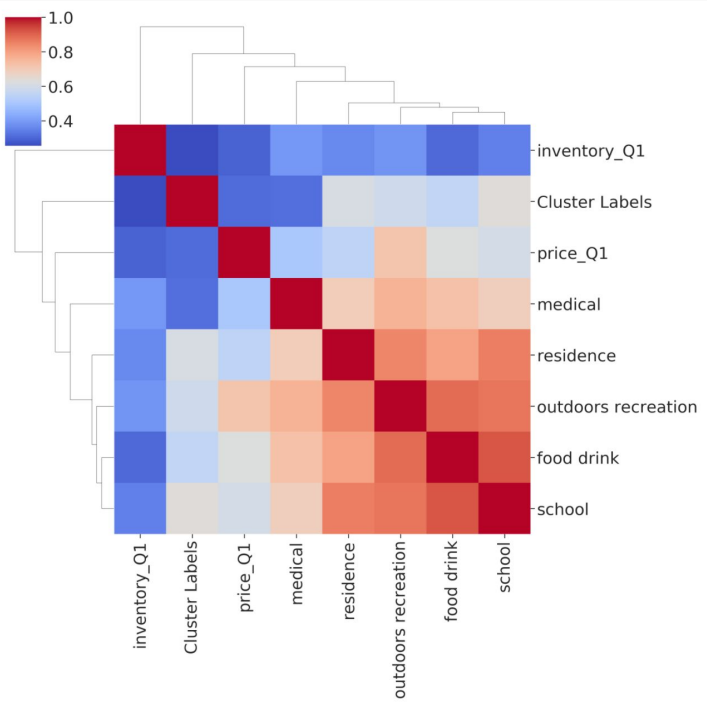
Housing Price & Venues for the Clusters





Correlations





Conclusion

- K-means algorithm was employed
- 5 clusters were discovered based on location and real estate data
- Clusters offer potential homebuyers options to choose houses based on prices/venues nearby
- Limitations: Staten Island data was not available