Explaining Housing Price in Manhattan using nearby venues

A COURSERA CAPSTONE PROJECT

ESTELLA YU

MAY 2019

Introduction

- New York is such a unique place full of attractions from the financial capital, fashion trends, artistic and historic atmosphere.

- YET! city life comes with a lot of dollar signs -- especially in Manhattan

- Questions to answer:
  - **Why** do we want to analyze the housing price in **New York**?
  - **How** is the apartment rental price vary by neighborhood?
  - If you are planning to move to a new neighborhood, **what** typical venues will you be looking for?
  - Do the popular venue & higher end apartment price align?

Data Description

- Data Source:
  - Zumper (National Rent Report: February 2019)
  - Rentcafe (Manhattan, NY Rental Market Trends)
  - FourSquare API (venue data around each neighborhood)

Data Description

- Data collection
  - **Manhattan**, which is very well tabulated on the website Rentcafe. I will first scrape and name of each neighborhood and the corresponding rental price from the website using the *Beautiful Soup API*, and then translate the data into *pandas* dataframe.
  - Based on the name of each neighborhood, the latitude and longitude will be obtained through *GeoPy API*, which is then plotted on a map using *folium*.
  - The *FourSquare API* was used to get the most common venues around each neighborhood, where the venue data is also added to the folium map.
  - Using the average rental price, a heat map is plotted showing the distribution of average rent demographically.

Problem solving goals

- 1. Visualize the national rent price across nation (folium, heat map)
- 2. Cluster the neighborhood based on rental price, close by venues, etc.
- 3. Plot rental price vs. popular venue across all neighborhood, and explore correlations

Methodology / Analysis

	Neighborhood	Average Rent	loc_lat	loc_lon
0	Marble Hill	1694.0	40.8762983	-73.9104292
1	Inwood	2225.0	40.8692579	-73.9204949
2	Washington Heights	2243.0	40.8401984	-73.9402214
3	Randalls and Wards Islands	2336.0	40.79144785	-73.921023713881
4	East Harlem	3334.0	40.7947222	-73.9425

Fig. 1. Dataframe from web scraping – the name of each neighborhood and the corresponding average rent are scraped from the internet, and the latitude & longitude location are obtained from the Geopy API.

Methodology: Folium map

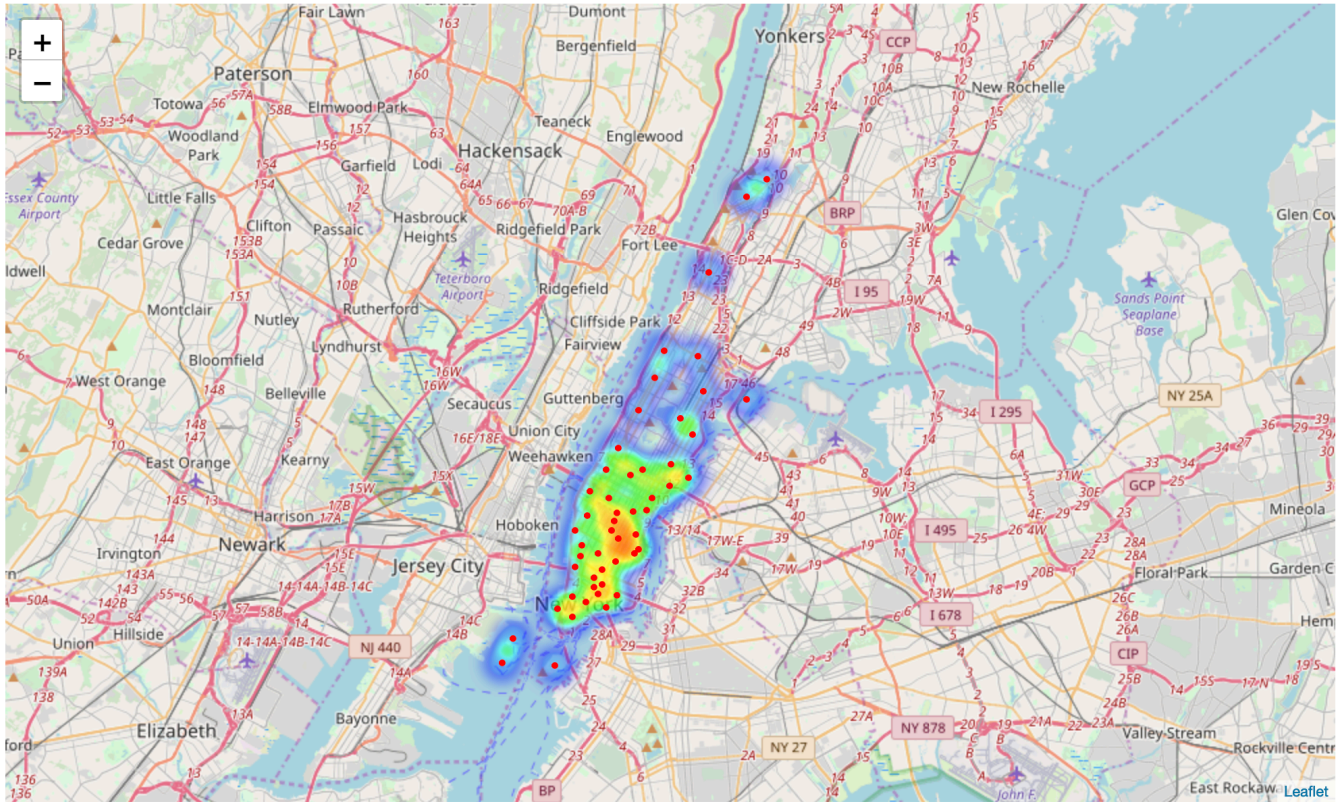

Fig. 2. Folium map plotted based on the dataframe. In addition, a heat map is generated, showing the regions with higher rental price in red and lower price in blue.

Methodology: fetching venues

FourSquare

	Neighborhood	Average_Rent	Venue	lat	lon	Category
0	Marble Hill	1694.0	108 Marblehill	40.876605	-73.909430	Housing Development
1	Marble Hill	1694.0	Marble Hill	40.876111	-73.911111	Neighborhood
2	Marble Hill	1694.0	St. Johns Roman Catholic Church	40.876174	-73.909795	Church
3	Marble Hill	1694.0	CTown Supermarkets	40.876218	-73.908541	Supermarket
4	Marble Hill	1694.0	Wine & Liquors	40.874535	-73.909832	Wine Shop

Fig. 3. Using the FourSquare API, the venues in each neighborhood are fetched. The name of each venue, geological location and its category are cleaned and tabulated.

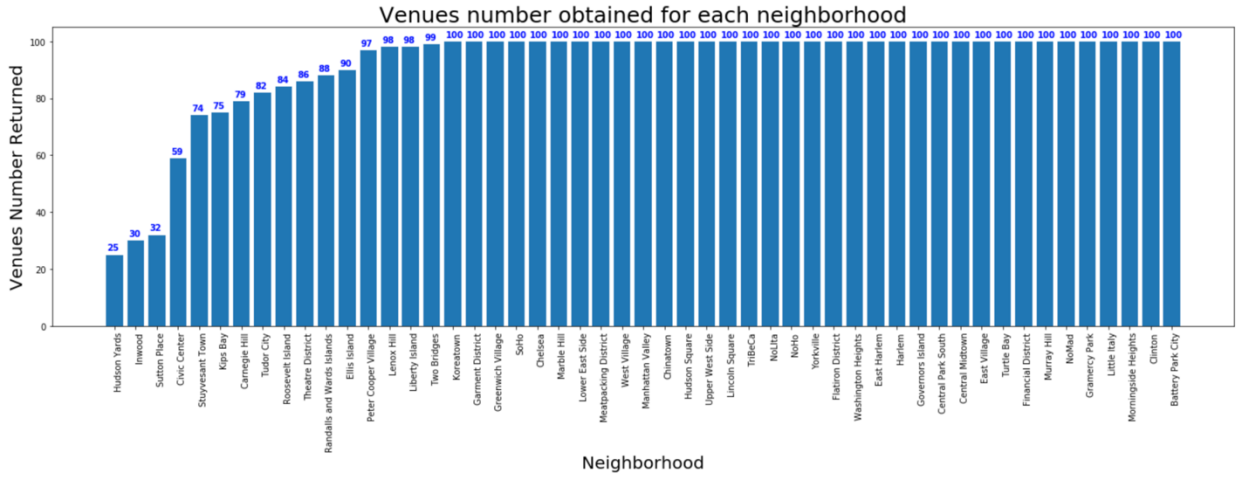

Fig. 4. Total number of venues returned in each neighborhood varies. With a limit of 100 venues set across all neighborhoods, some of the neighborhoods returns venues smaller than the target limit.

Methodology: simple regression

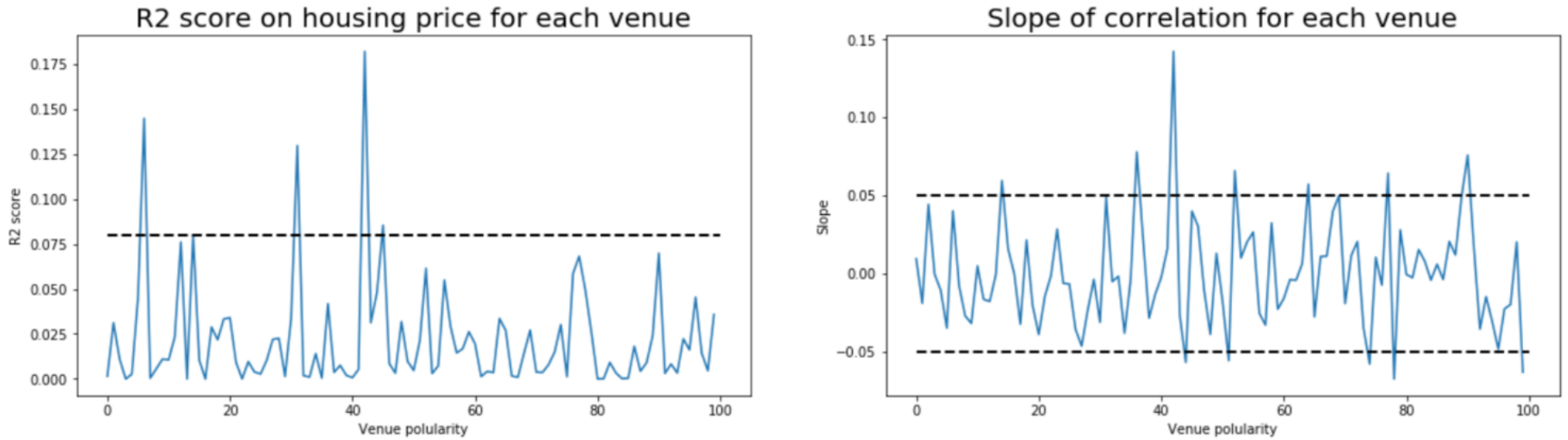

Fig. 8. Simple linear regression on each venue categories across all neighborhood. Although the R^2 score are small, some helpful insights can be obtained. For example, some categories are more likely to contribute to the price variation than other categories, and some positive/negative influences are more impactful than the other.

Methodology: k-means

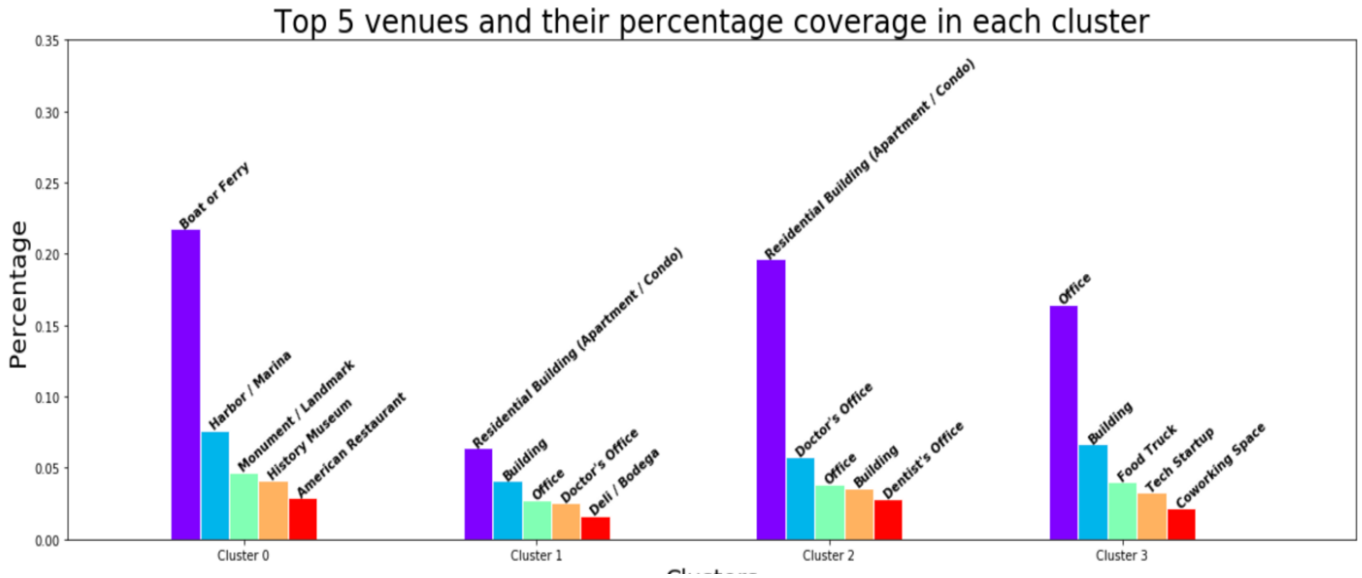

Fig. 12. A summary of the top 5 most common venues in each cluster. We can then provide each cluster with its name based on these characteristics.

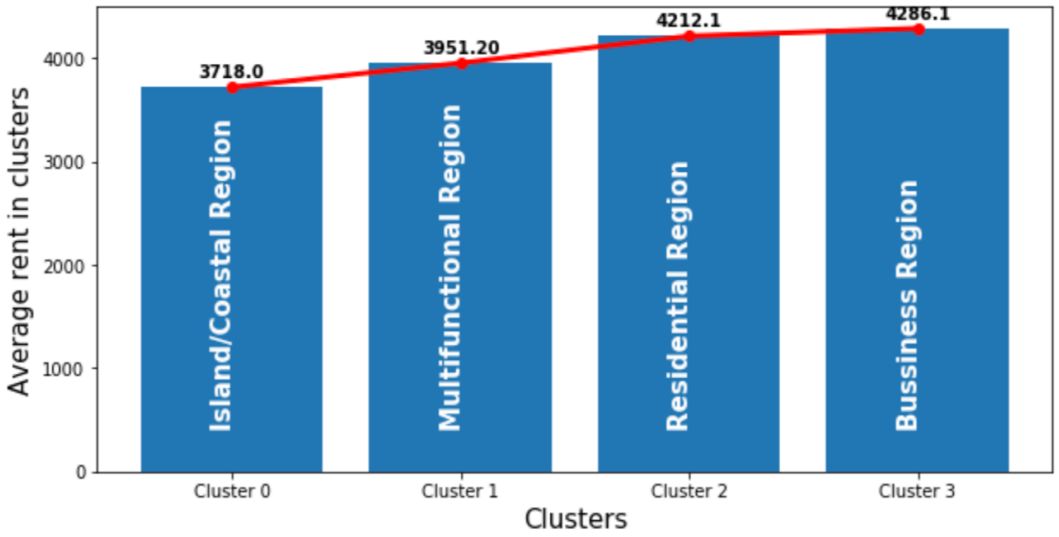

Fig. 13. The clustered region also shows a slight increasing trend of average rental price, with the relatively remote island regions having the lowest rental price across the cluster, and the busy business regions with the highest average rental price.

Methodology: clustered folium map

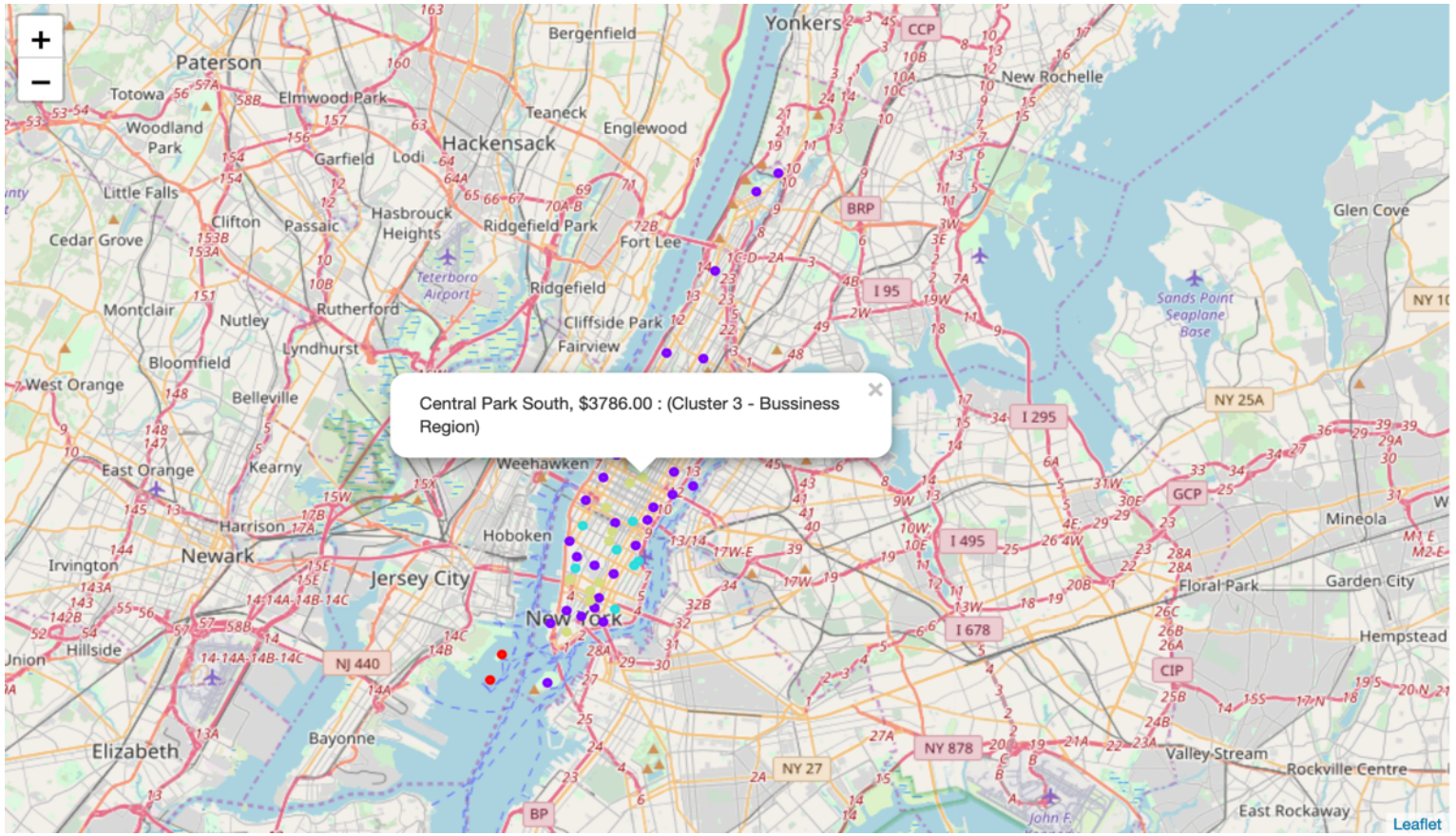

Fig. 14. The clustered neighborhoods are shown on folium map, with labels indicating the neighborhood name, average rental price, and the cluster it belongs to. For example, Central Park South will be considered as a business region (based on the surrounding venues).

Results & discussion

- **New York** is a big city with a **high population density in a narrow area**
- **51** neighborhoods are found, 4000+ venues collected across all neighborhoods
- Simple regression
  - Predicting the “trust-worthiness” of the regression result
  - Output the more influential venue categories
  - Predicts the likelihood for a new venue com

Conclusion

- useful for
  - a person looking for a new rental place: not only the rental price and neighborhood venues have certain correlation, but the nearby venues also determine whether a place is ideal to live in.
  - Investors: find valuable neighborhood to invest in business
- Similar analytic methods can also be generalized to analyze data sets in many different cities across the world.