Exploring New York City and its Housing Prices

Data Science Capstone Project

Rui Huang

New York City

- New York City, a.k.a. the Big Apple, is situated in northeastern U.S.
- It is the financial capital of the country
- It hosts many neighborhoods within its five boroughs - Manhattan, Brooklyn, the Bronx, Queens, and Staten Island - and each of these boroughs exhibits its own lifestyle
- New York City is heavily populated and ethnically diverse
- It is one of the most international cities in the country.

Business Problem

* New York City boosts tremendous business opportunities, and is a global hub of business and commerce
* Much of New York's business success is due to the real estate market
* The highly competitive business market in New York City has driven up its cost of living
* New comers who wish to move into the city need to carefully analyze the costs and benefits associated with living in New York before making a serious financial commitment
* Particularly, people who are interested in moving to New York City should understand the culturally diverse neighborhoods and their identifying characteristics, as well as the housing prices in each neighborhood

Research Questions

- How could we cluster neighborhoods in New York City using surrounding venues?
- Is there a relationship between the characterized clusters and housing prices in New York City?
- Is there a relationship between the surrounding venues and housing prices in New York City? If yes, which aspects of the venues are most likely to influence housing prices, and whether we can use venues to predict housing prices in NYC?

Target Audience

- The objective of this study is to understand how neighborhood venues may impact housing prices in New York City
- Real estate professional may be interested in our findings
- People who plan to move to New York would find the findings useful too
- The study would also interest anyone who is taking the data science course

Data Description - 1

* Coordinates data for each borough and neighborhood from Geocoder (https://geocoder.readthedocs.io/index.html) - coordinates data will allow us to locate venues on FourSquare API and map the neighborhood segments on a map

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Brooklyn	Bay Ridge	40.625801	-74.030621
2	Brooklyn	Bensonhurst	40.611009	-73.995180
3	Brooklyn	Sunset Park	40.645103	-74.010316
4	Brooklyn	Greenpoint	40.730201	-73.954241

Data Description - 2

* Venues in each neighborhood from FourSquare API - venue data from FourSquare can be used to explore and segment neighborhoods in NYC
* For example, FourSquare APIs offer rich location-based experiences and enable access to places data in real time and venue-related tips, tastes, photos & attributes from the Foursquare community

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueType
0	Bedford-Stuyvesant	40.687232	-73.941785	Sincerely Tommy	40.686066	-73.944294	Boutique
1	Bedford-Stuyvesant	40.687232	-73.941785	Bed-Vyne Brew	40.684751	-73.944319	Bar
2	Bedford-Stuyvesant	40.687232	-73.941785	Bed-Vyne Wine & Spirits	40.684668	-73.944363	Wine Shop
3	Bedford-Stuyvesant	40.687232	-73.941785	Anchor Coffee	40.684145	-73.941015	Coffee Shop
4	Bedford-Stuyvesant	40.687232	-73.941785	Peaches HotHouse	40.683331	-73.943853	Fried Chicken Joint

Data Description - 3

* Average NYC condo prices by neighborhoods from CITYREALTY (https://en.wikipedia.org/wiki/New_York_City#Population_density) - data retrieved from CityRealty lists average prices for different types of condos (ranging from studio to 3-bedroom) aggregated by neighborhoods in NYC
* This project will use the average price of 2-bedroom condos, which is a common type selected by an average family

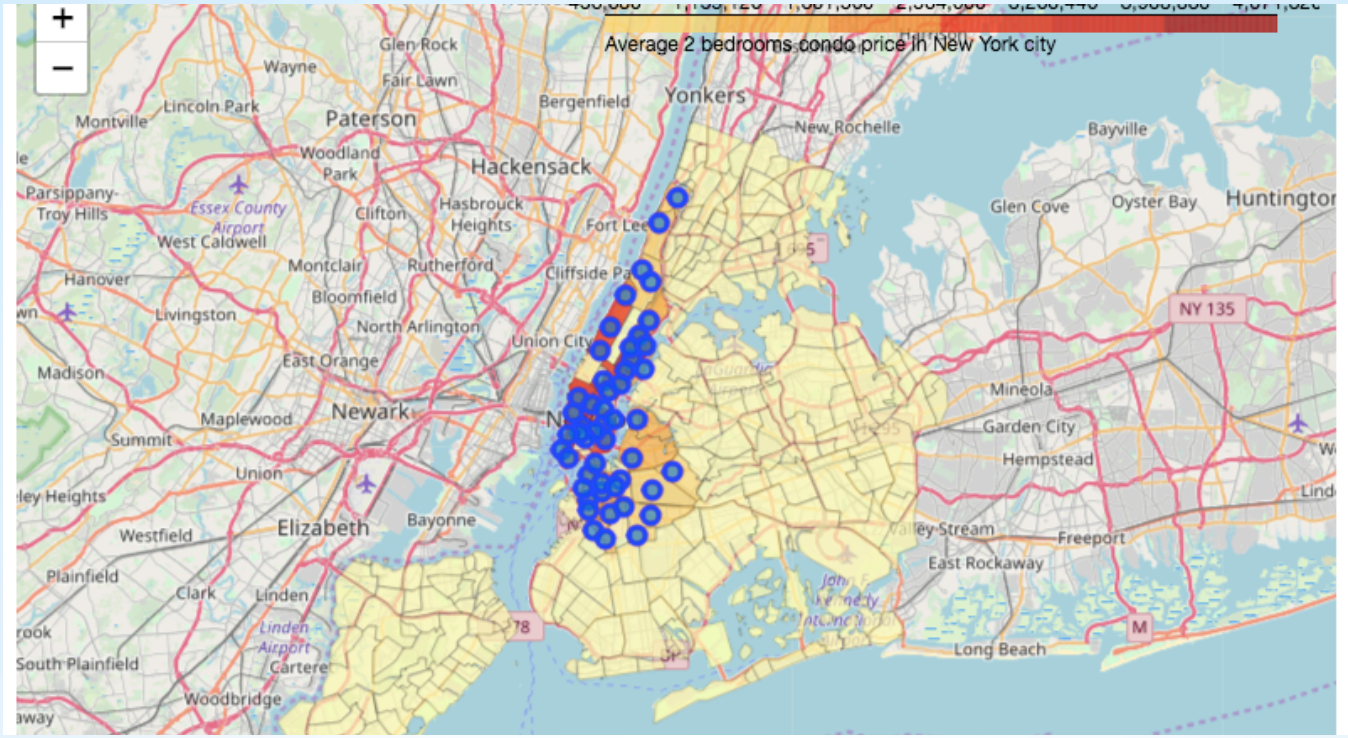
	Area	Neighborhood	AvgPrice
0	Brooklyn	Bedford-Stuyvesant	750000
1	Brooklyn	Boerum Hill	1.69e+06
2	Brooklyn	Brooklyn Heights	2.15e+06
3	Brooklyn	Bushwick	967000
4	Brooklyn	Carroll Gardens	1.51e+06

Data Collection

* For each neighborhood, call Geocoder Python to get its coordinate.
* Based on each neighborhood's coordinate, call FourSquare API to get the surrounding venues.
* We will scrap the average price of 2-bedroom condos for each neighborhood from the CityRealty website

Methodology

- Cluster Analysis - used to explore and segment NYC neighborhoods
- Visualization – used to display data on map
- Correlation and linear regression analysis - used to identify the impacts of neighborhood clusters and venue types on 2-bedroom condo prices
- Machine learning techniques - used to determine how well independent variables can be used to predict condo prices



* New York City Geographical Coordinates Data

Neighborhood	AvgPrice	Latitude	Longitud
Bedford-Stuyvesant	750000	40.687232	-73.94178
Boerum Hill	1.69e+06	40.685683	-73.98374
Brooklyn Heights	2.15e+06	40.695864	-73.99378
Bushwick	967000	40.698116	-73.92525
Carroll Gardens	1.51e+06	40.680540	-73.99465

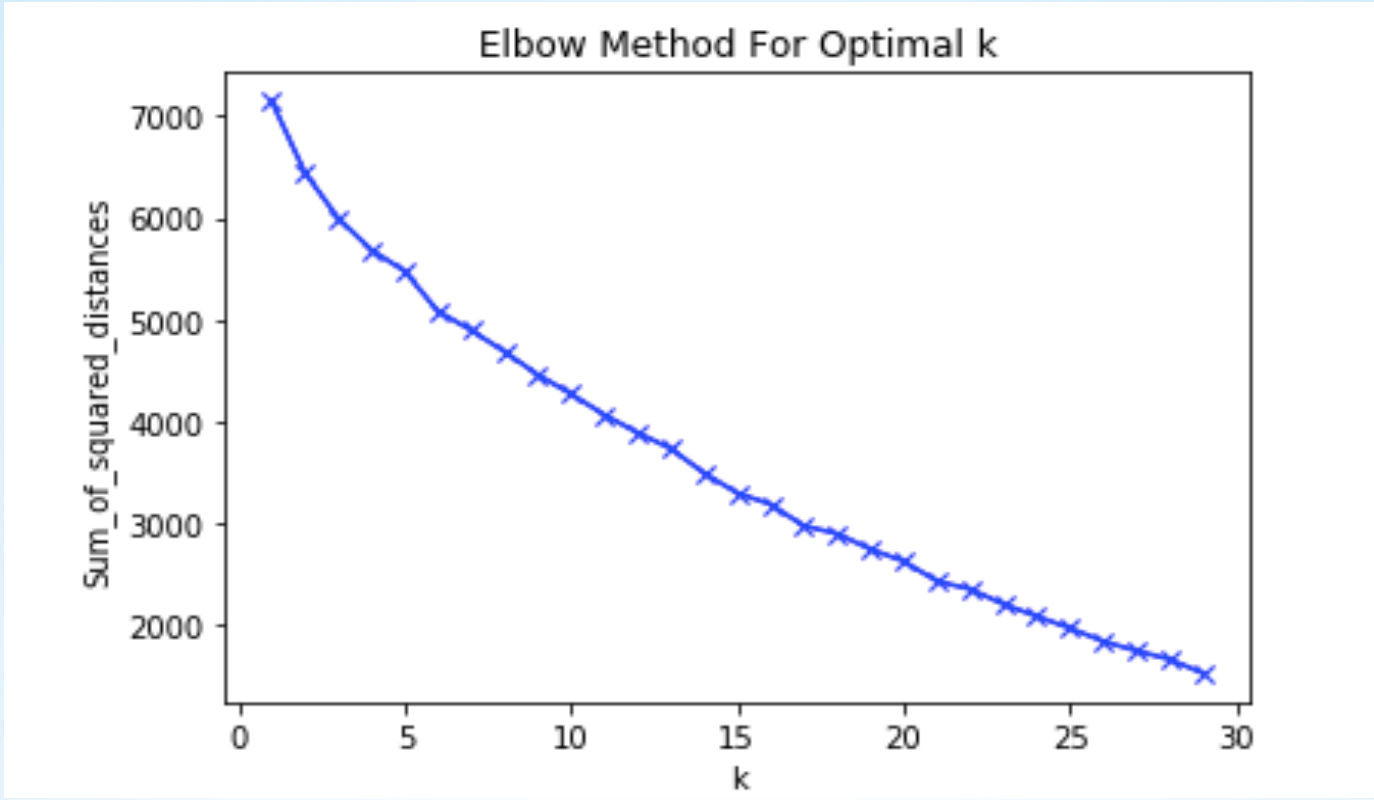
* Average 2-bedroom Condo Price Data

	Neighborhood	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant
0	Battery Park City	0	0	0	1	0	0	0	0	0
1	Bedford-Stuyvesant	0	0	0	0	0	0	0	1	0
2	Boerum Hill	0	0	0	1	0	0	0	0	0
3	Brooklyn Heights	0	0	0	3	0	0	0	0	0
4	Bushwick	0	0	0	1	0	0	0	1	0

*FourSquare APIs Data

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Battery Park City	Park	Coffee Shop	Wine Shop	Plaza	Gym	Hotel	Gym / Fitness Center	Sandwich Place	BBQ Joint
1	Bedford-Stuyvesant	Coffee Shop	Café	Wine Shop	Bar	Pizza Place	Caribbean Restaurant	Juice Bar	Deli / Bodega	Mexican Restaurant
2	Boerum Hill	Coffee Shop	Bar	Cocktail Bar	Bakery	Burger Joint	Dance Studio	Flower Shop	Opera House	French Restaurant
3	Brooklyn Heights	Park	Yoga Studio	Coffee Shop	Italian Restaurant	Wine Shop	Grocery Store	Bar	American Restaurant	Pizza Place
4	Bushwick	Bar	Mexican Restaurant	Pizza Place	Coffee Shop	Dive Bar	Deli / Bodega	Cocktail Bar	Café	Italian Restaurant

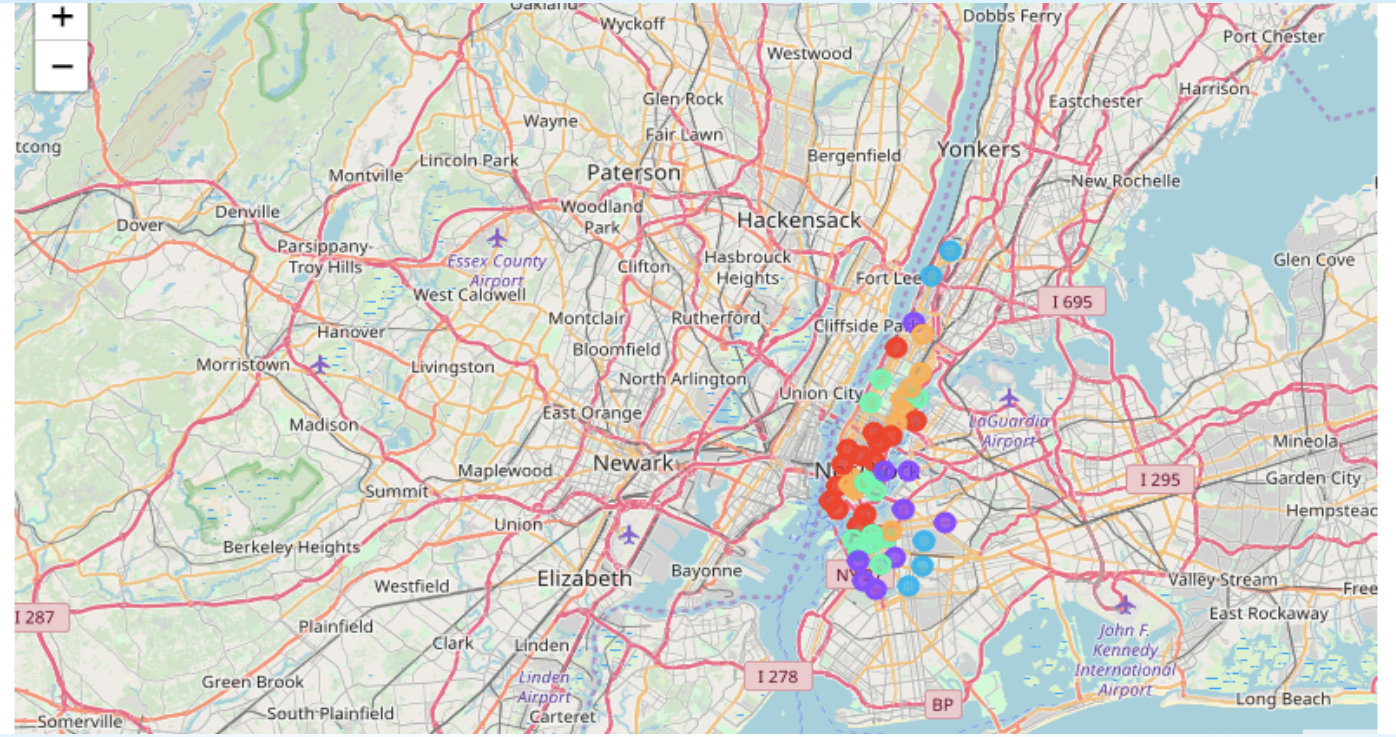
*FourSquare Data - Continued



*Cluster Results

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	
4	Brooklyn	Greenpoint	40.730201	-73.954241	1	Bar	Cocktail Bar	Coffee Shop	Café	Record Shop	F F
10	Brooklyn	Crown Heights	40.670829	-73.943291	2	Café	Caribbean Restaurant	Pizza Place	Coffee Shop	Grocery Store	D E
13	Brooklyn	Windsor Terrace	40.656946	-73.980073	1	Café	Italian Restaurant	Park	Deli / Bodega	Wine Shop	D
14	Brooklyn	Prospect Heights	40.676822	-73.964859	1	Bar	Wine Shop	Sushi Restaurant	Plaza	Cocktail Bar	E
16	Brooklyn	Williamsburg	40.707144	-73.958115	1	Bar	Pizza Place	American Restaurant	Coffee Shop	Yoga Studio	E S

*Cluster Results - Continued



*Cluster Visualization

```
R2-score: -0.00379444311852
Mean Squared Error: 0.351337750824
Max positive coefs: [ 0.00199386]
Venue types with most postive effect: ['Cluster Labels']
Max negative coefs: [ 0.00199386]
Venue types with most negative effect: ['Cluster Labels']
Min coefs: [ 0.00199386]
Venue types with least effect: ['Cluster Labels']
```

*Linear Regression - Clusters as the IV

```
R2-score: 0.235892603046
Mean Squared Error: 0.267444969509
Max positive coefs: [ 0.31137536  0.29670491  0.29670491  0.29173072  0.25605828  0.254013
  0.24977756  0.24977756  0.24977756  0.24977756]
Venue types with most postive effect: ['Dumpling Restaurant' 'Cafeteria' 'Buffet' 'Colombian
Restaurant'
 'Other Nightlife' 'Botanical Garden' 'Jewish Restaurant'
 'Persian Restaurant' 'Resort' 'Train Station']
Max negative coefs: [-0.28102325 -0.24130035 -0.24032501 -0.22989408 -0.22989408 -0.22989408
 -0.19827281 -0.19827281 -0.19115098 -0.18695974]
Venue types with most negative effect: ['Golf Driving Range' 'Newsstand' 'Board Shop' 'Street
Food Gathering'
 'Print Shop' 'Other Repair Shop' 'Lighthouse' 'Rest Area' 'Roof Deck'
 'Drugstore']
Min coefs: [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
Venue types with least effect: ['Science Museum' 'Leather Goods Store' 'Mini Golf' 'Volleybal
l Court'
 'Molecular Gastronomy Restaurant' 'Indoor Play Area'
 'Pakistani Restaurant' 'Shipping Store' 'Bridge' 'TV Station']
```

*Linear Regression - Venue Types as the IV

```
Best n: 50 R2 score: 0.341178436219
Best n: 50 MSE: 0.230593910934
coefs: [ 0.09020125  0.05865386  0.05728785  0.05657886  0.05183535
 0.05002992  0.04936853  0.04842364]
ith most positive effect: ['Dumpling Restaurant' 'Pilates Studio' 'L

urant' 'Pie Shop' 'Chinese Restaurant'
ice Supplies Store' 'Korean Restaurant' 'Colombian Restaurant']
coefs: [-0.05523341 -0.0426949  -0.04228755 -0.04187764 -0.04165757
-0.03973267 -0.03710717 -0.03640676]
ith most negative effect: ['Market' 'Golf Driving Range' 'Lingerie S

'New American Restaurant' 'Cosmetics Shop' 'Tapas Restaurant'
' 'Trail']
-1.67905975e-06  -4.35235710e-06   6.21873614e-05  -6.54926719e-05
e-05  -8.80812119e-05   8.87232334e-05   8.87232334e-05
e-05   1.44136814e-04]
ith least effect: ['Cemetery' 'State / Provincial Park' 'Gym Pool' '
estaurant' 'Breakfast Spot' 'General Entertainment'
astronomy Restaurant' 'Volleyball Court' 'Bridge']
```

*Principal Component Regression (PCR)

Discussions

- NYC neighborhoods may be segmented into five clusters.
- Regression results, seems consistent and logical.
- Neighborhoods with interesting restaurants may be more expensive to live in.
- Despite the variables identified in this study, the real estate price can be hard to predict.
- The machine learning techniques applied in this study may not accurately predict future housing prices in NYC.
- Future studies may be needed to better understand the factors that would help explain housing prices in NYC.

Conclusion

* Based on the analysis, neighborhoods with interesting restaurants seem to boost housing values the most
* There may be a logical endogeneity in this observation though, as neighborhoods with a lot of eateries tend to be those with highly population density and it might simply be the housing demands that drive up the living prices
* To make a more accurate conclusion about how features of neighborhoods may impact housing prices, future studies are needed to explore additional variables