Exploring New York City and its Housing Prices
Final Report

1. Introduction & Business Problem

Problem Background:

New York City, a.k.a. the Big Apple, is situated in northeastern U.S. and 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. In 2017, the city had an estimated population density of 28,491 inhabitants per square mile (11,000/km2) [1]. It is one of the most international cities in the country.

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 [2], although the city is also a major center for many other types of business. The highly competitive business market in New York City has driven up its cost of living. Thus, 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. In addition, what features of a neighborhood are mostly likely to affect the housing price? Answers to these questions will help new comers make a more informed decision about where they should choose to live, given the tradeoffs between the lifestyle they are seeking for and the prices they have to pay for expensive living in the city.

Problem Description:

In this project, we will explore the neighborhoods in New York City, and explore how surrounding venues in each neighborhood may affect its housing prices. Specifically, we will ask the following questions.

1. How could we cluster neighborhoods in New York City using surrounding venues?
2. Is there a relationship between the characterized clusters and housing prices in New York City?
3. 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.

2. Description of the data

This study will use the following data to conduct empirical analysis and seek answers to the business questions identified in Section 2.

- 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

- 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

- 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:

We will go through the following procedure to collect the necessary data.

- 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.

The collected data will be put into 2 dimensional Pandas dataframes, with each row represents a neighborhood and each column describing the surrounding venues. We will combine data collected from different sources so that the average 2-bedroom condo price data is joined with neighborhood and venues data for further analysis.

Using Data to Answer Business Problems:

The following analyses will be performed to answer the questions listed in Section 2.

- Cluster Analysis - used to explore and segment NYC neighborhoods
- 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

3. Methodology

Analytic Approach:

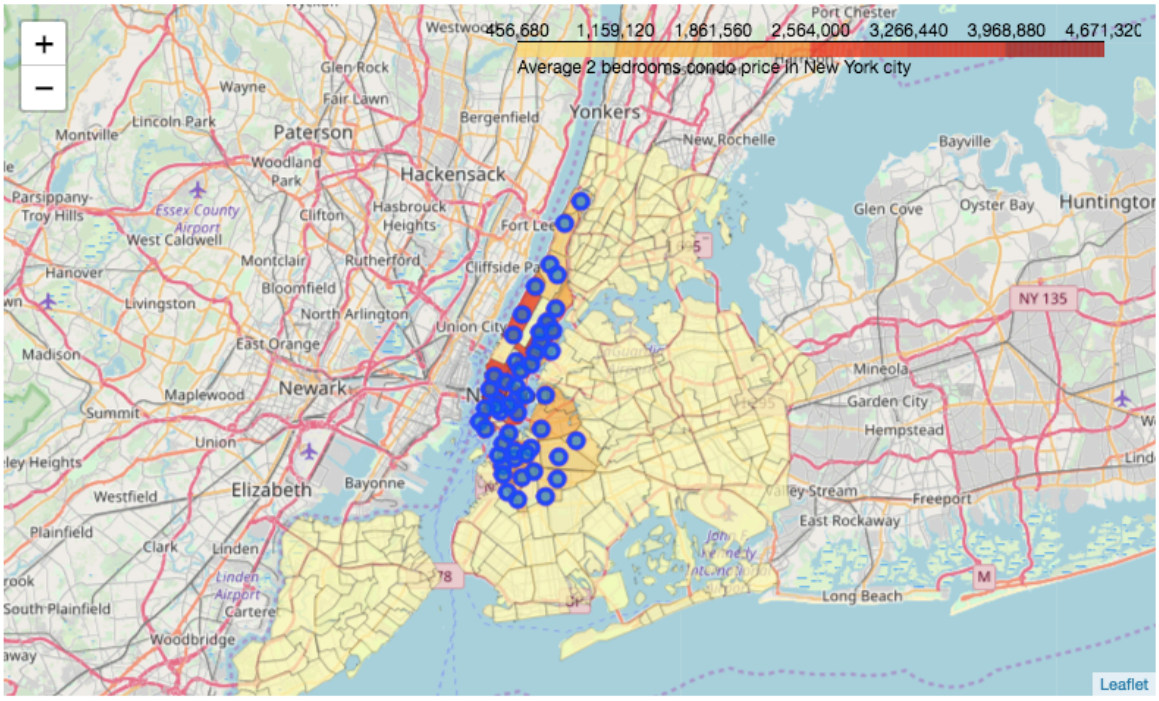
New York city neighborhood has a total of 5 boroughs and 306 neighborhoods. In this project, we will first cluster the neighborhoods based on venue types. We then will examine the association between clusters and average condo prices, as well as

between venue features and average condo prices.

Exploratory Data Analysis:

Data 1 - New York City Geographical Coordinates Data.

1. Load the data and explore data from nyc_geo.json file.
2. Transform the data of nested python dictionaries into a Pandas dataframe.
3. This dataframe contains the geographical coordinates of New York City neighborhoods.
4. This data will used to get Venues data from Fouresquare.
5. We used geopy and folium libraries to create a map of New York city with neighborhoods superimposed on top.



Data 2 - Average 2-bedroom condo price data from CityRealty.

1. There are 54 neighborhoods captured in this dataset.
2. We combined this data with the coordinates data.

	Neighborhood	AvgPrice	Latitude	Longitude
0	Bedford-Stuyvesant	750000	40.687232	-73.941785
1	Boerum Hill	1.69e+06	40.685683	-73.983748
2	Brooklyn Heights	2.15e+06	40.695864	-73.993782
3	Bushwick	967000	40.698116	-73.925258
4	Carroll Gardens	1.51e+06	40.680540	-73.994654

Data 3 - FourSquare APIs data.

1. For each neighborhood, pass the obtained coordinates to FourSquare API. The “explore” endpoint will return a list of surrounding venues in a pre-defined radius.
2. Count the occurrence of each venue type in a neighborhood. Then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
3. Standardize the average price by removing the mean and scaling to unit variance.
4. The resulting dataset is a 2 dimensions data frame, with each row representing a neighborhood, and each column (except the last one) representing the occurrence of a venue type. The last column has the standardized average price.
5. The dataset has 50 samples and more than 300 features. The number of features may vary for different runs due to FourSquare API may returns different recommended venues at different points in time.
6. List 10 most common venues in each neighborhood

	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

	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

Analysis 1 – Cluster analysis

- To better understand the neighborhoods, k-means was used. First, Elbow analysis was performed to identify the optimal value of k. Then, the suggested k value was plugged into the analysis to segment neighborhoods into 5 clusters.

Analysis 2 – Visualization and interpretation

- To visualize the neighborhood clusters, folium map with markers was created to indicate the locations of each cluster
- Clusters were then examined and interpreted with a meaningful label

Analysis 3 – Linear Regression

- A simple technique of linear regression from sklearn library was chosen.
- Clusters were regressed on average housing price to determine whether neighborhood clusters would be a good predicting variable of condo prices.
- Venue types were also regressed on average housing price to determine whether venue types would be a good predicting variable of condo prices.

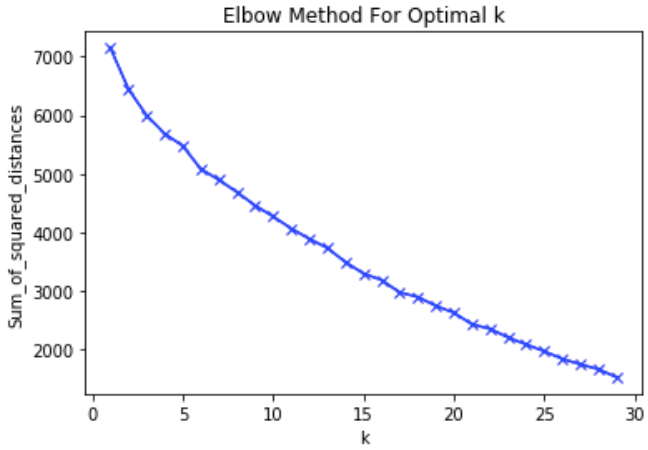
Analysis 4 – Principal Component Regression (PCR)

- PCR can be explained simply as the combination of Principal Component Analysis (PCA) with Linear Regression.
- Since the number of features in the dataset is much bigger than the number of samples. This will cause problem for the analysis process.
- To address this problem, PCR employs the power of PCA, which can convert a set of values of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. As the result, the number of features is reduced while keeping most of the characteristic of the dataset.
- Then PCR use Linear Regression on the converted set to return a coefficient list, just like in normal Regression techniques.

4. Result

Neighborhood K-Means Clustering:

- To cluster the neighborhoods into clusters we used the elbow method and the K-Means clustering Algorithm. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It uses iterative refinement approach.
- First, results from the elbow method were not perfect, but it implies the “elbow” when k=5.

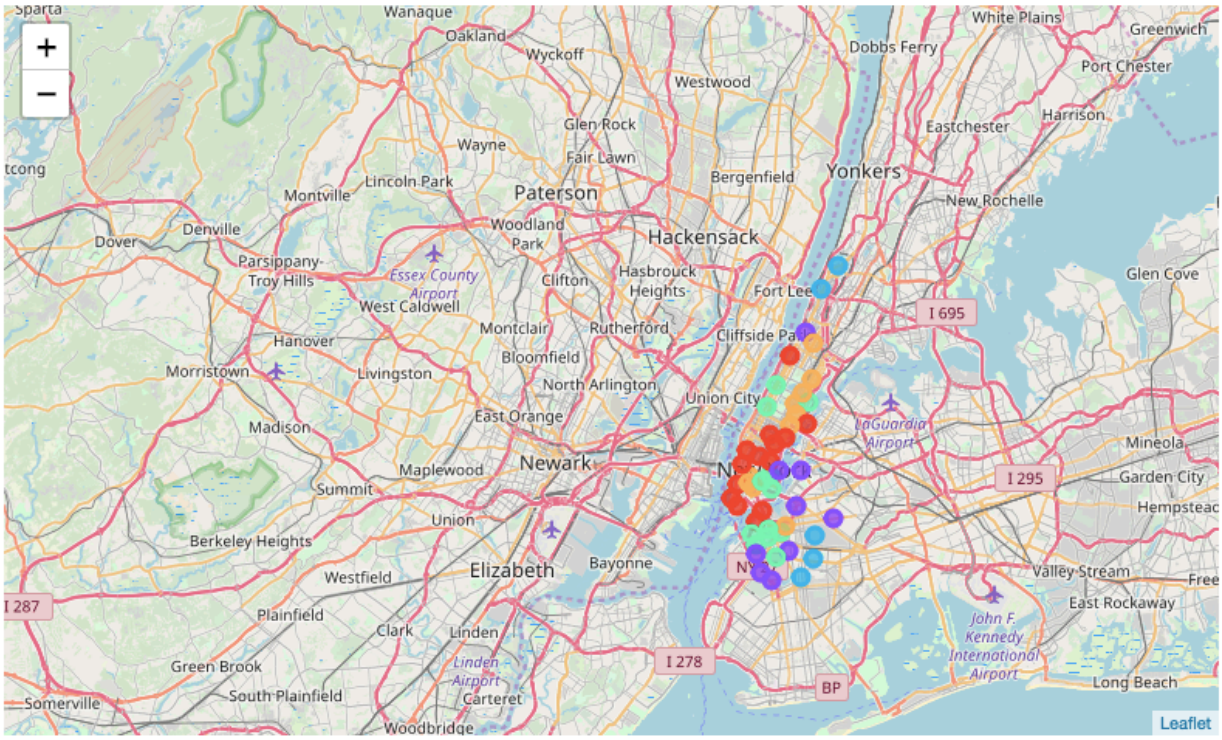


- We then segmented NYC neighborhoods into 5 clusters and added the cluster label into the dataset.

	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 Visualization and Interpretation:

- We visualized the clusters with markers on a map.
- We applied five labels to the five clusters, namely: 1) active lifestyle, 2) food adventurers, 3) simplified lifestyle, 4) art lovers, and 5) have-it-all.



Linear Regression:

- We first regressed neighborhood clusters on average condo price.
- As can be seen from the results below, R2 score is small, which means the model may not be suitable for the data.

```
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']
```

- Next, we regressed venue types on average condo price.
- As can be seen from the results below, R2 was still small but looked better as compared to the regression model of clusters. We also saw some venues with positive coefficients and some venues with negative coefficients, implying restaurant venues are likely to increase the value of a location, whereas recreational and shopping venues may decrease the value of a location.

```
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']
```

PCR:

- PCR is a regression technique involving two steps. First, it obtains the principle components from the feature set and selects a subset for the next step. Second, it uses linear regression on the subset of principal components to get a list of correlation coefficients.
- The result is promising as it shows improvement over the simple Linear Regression.
- The insights are consistent with results from the Linear Regression. That is, restaurant venues are likely to increase the value of a location, whereas recreational and shopping venues may decrease the value of a location.

```
Best n: 50 R2 score: 0.341178436219
Best n: 50 MSE: 0.230593910934
Max positive coefs: [ 0.09020125  0.05865386  0.05728785  0.05657886  0.05183535  0.05173739
  0.05139713  0.05002992  0.04936853  0.04842364]
Venue types with most positive effect: ['Dumpling Restaurant' 'Pilates Studio' 'Library' 'Ten
nis Court'
 'Sushi Restaurant' 'Pie Shop' 'Chinese Restaurant'
 'Paper / Office Supplies Store' 'Korean Restaurant' 'Colombian Restaurant']
Max negative coefs: [-0.05523341 -0.0426949  -0.04228755 -0.04187764 -0.04165757 -0.04134184
 -0.0411243  -0.03973267 -0.03710717 -0.03640676]
Venue types with most negative effect: ['Market' 'Golf Driving Range' 'Lingerie Store' 'Food
& Drink Shop'
 'Newsstand' 'New American Restaurant' 'Cosmetics Shop' 'Tapas Restaurant'
 'Bridal Shop' 'Trail']
Min coefs: [ -1.67905975e-06  -4.35235710e-06   6.21873614e-05  -6.54926719e-05
  -6.54926719e-05  -8.80812119e-05   8.87232334e-05   8.87232334e-05
   8.87232334e-05   1.44136814e-04]
Venue types with least effect: ['Cemetery' 'State / Provincial Park' 'Gym Pool' 'Food Stand'
 'Pakistani Restaurant' 'Breakfast Spot' 'General Entertainment'
 'Molecular Gastronomy Restaurant' 'Volleyball Court' 'Bridge']
```

5. Discussion

Findings from this study suggests the following:

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

6. 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.

Reference:

1. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/New_York_City#Population_density
2. "Why New York Is The Best City for Small Business In America" Forbes, 2018. Retrieved from https://www.forbes.com/sites/rohitarora/2018/05/04/why-new-york-is-the-best-city-for-small-business-in-america/#4c8ae2cf56b3