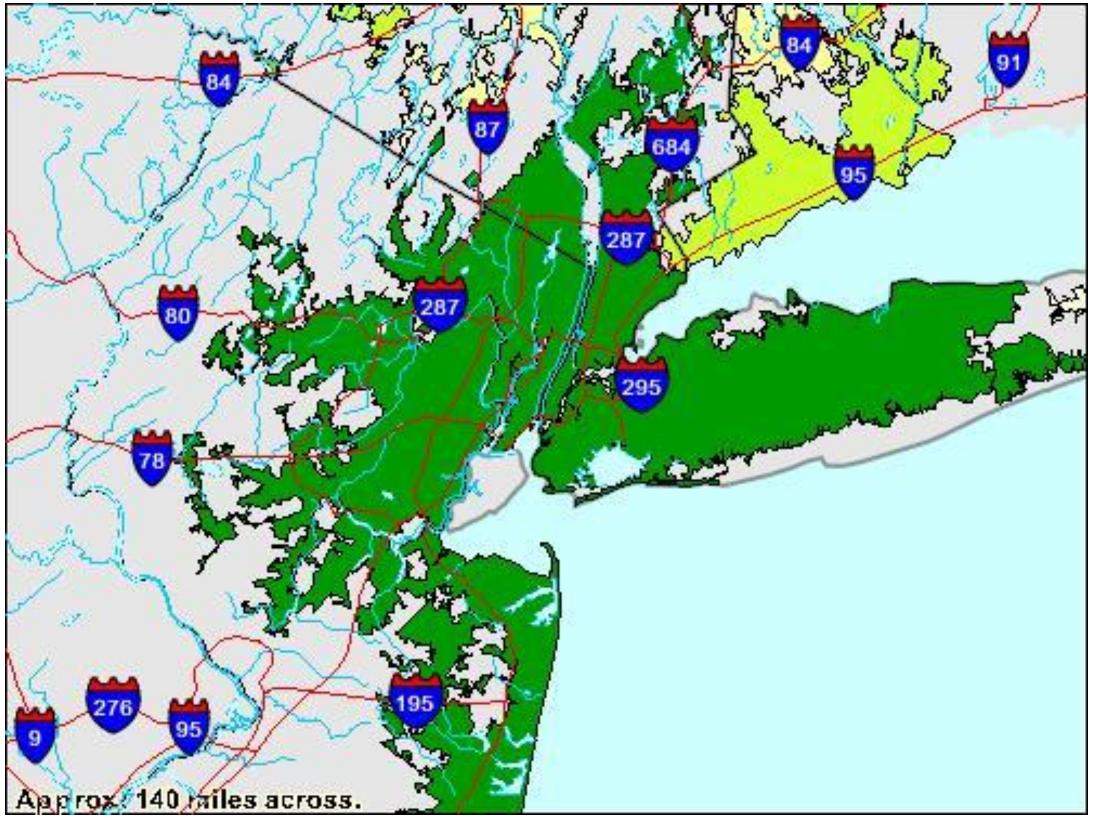
Photo by This file is lacking author information. / Public domain

NEW YORK – REAL ESTATE

APPLIED DATA SCIENCE - CAPSTONE



CONTENTS

- INTRODUCTION
- DATA DESCRIPTION
- RESEARCH METHODS
- VISUALIZATION
- RESULTS
- ANALYSIS
- DESCRIPTION
- CONCLUSION

INTRODUCTION

- Exploring the neighbourhoods of New York city in order to extract the correlation between the real estate value and its surrounding venues.
- So, can the surrounding venues affect the price of a house?
- If so, what types of venues have the most affect, both positively and negatively?
- Potential buyers who can roughly estimate the value of a house based on the surrounding venues and the average price.
- Real estate makers and planners who can decide what kind of venues to put around their products to maximize selling price.
- Houses sellers who can optimize their advertisements etc.,

DATA DESCRIPTION

The availability of real estate prices. Though very limited.

The diversity of prices between neighborhoods. For example, a 2-bedrooms condo in Central Park West, Upper West Side can cost $4.91 million on average; while in Inwood, Upper Manhattan, just 30 minutes away, it's only $498 thousands.

The availability of geo data which can be used to visualize the dataset onto a map.

Scrap the City Realty webpage for a list of New York city neighborhoods and their corresponding 2-bedroom condo average price.

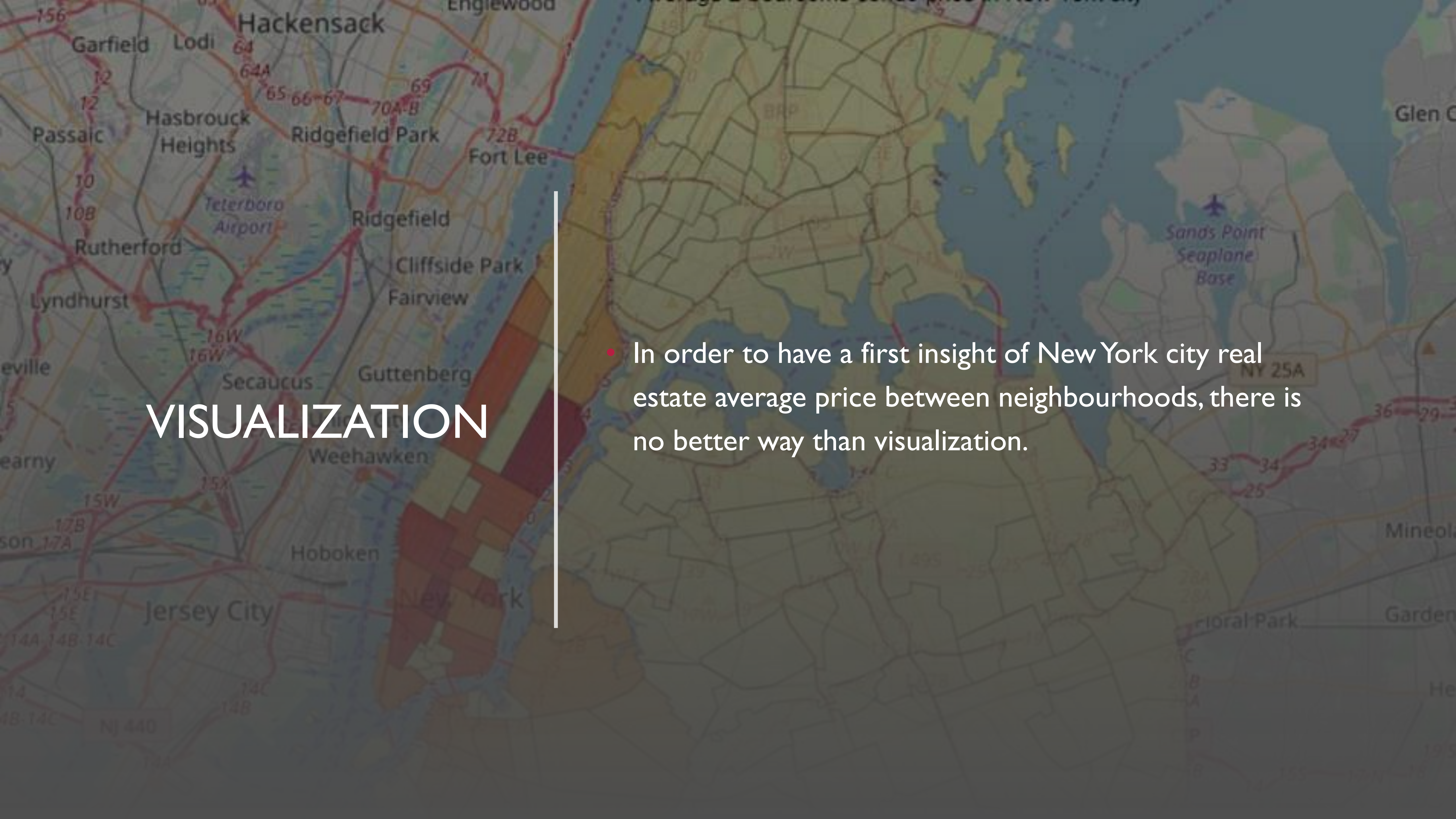
Find the geographic data of the neighborhoods. Both their center coordinates and their border and so on,

RESEARCH METHODS

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.

The number of features is much bigger than the number of samples. This will cause problem for the analysis process. Detail and counter-measurement will be discussed further in the next section.

The assumption is that real estate price is dependent on the surrounding venue. Thus, regression techniques will be used to analyse the dataset. The regressors will be the occurrences of venue types. And the dependent variable will be standardized average prices.



VISUALIZATION

- In order to have a first insight of New York city real estate average price between neighbourhoods, there is no better way than visualization.

ANALYSIS

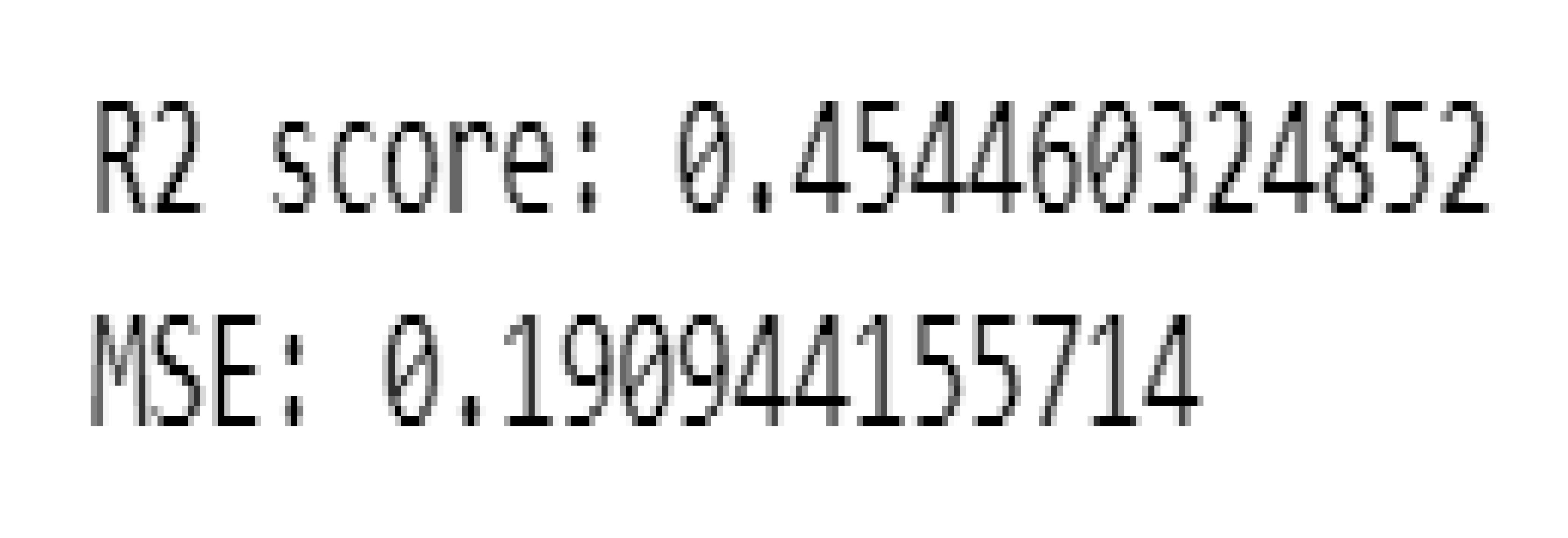
Linear Regression

Linear Regression was chosen because it is a simple technique. And by using Sklenar library, implementing the model is quick and easy. Which is perfect to start the analysing process.

The model will contain a list of coefficients corresponding to venue types. R2 score (or Coefficient of determination) and Mean Squared Error (MSE) will be used to see how well the model fit the data.

ANALYSIS CONT:

- **Principal Component Regression (PCR):**



```
R2 score: 0.454460324852
MSE: 0.190944155714
```

RESULTS

Even though the scores seem to be improved after applying a more sophisticate method, the model is still not suitable for the dataset. Thus, it can’t be used to precisely predict a neighbourhood average price.

Explanations for the poor model can be:

The real estate price is hard to predict.

The data is incomplete (small sample size, missing deciding factors).

The machine learning techniques are chosen or applied poorly.

But again, on the bright side, the insight, gotten from observing the analysis results, seems consistent and logical. And the insight is business venues that can serve the needs of most normal people usually situated in pricy neighbourhoods.

DISCUSSION

The real challenge is constructing the dataset:

Usually the needed data isn't publicly available.

When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.

For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.

Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

CONCLUSION

Some notes on the analysis result:

This project is done by a web developer who only started self-studying Data Science for 4 months. So please take it with a grain of salt.

The coefficients only show correlation, not causation. So, if your neighborhood average price is low, please don't go destroying the surrounding bars and food trucks. There might be another reason.