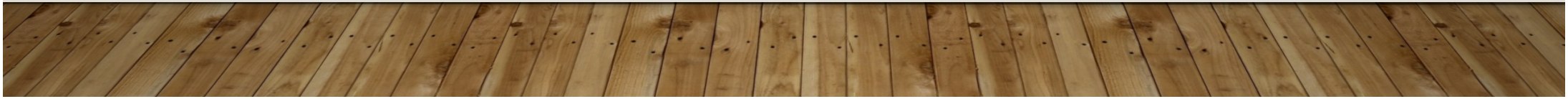
INVESTIGATION OF HOUSE SALE PRICE AND NEIGHBORHOOD VENUES

ZHI YUAN



INTRODUCTION

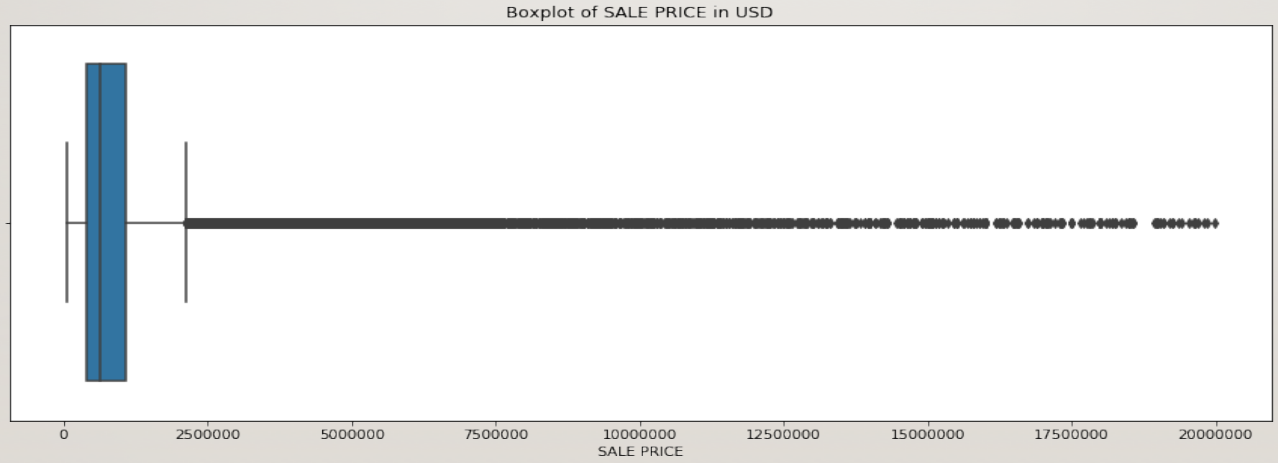
- New York City (NYC) one of the world's most expensive city in house sale price
- House sale price prediction important for New Yorker as well as worldwide investor
- House price is neighborhood dependent; neighbourhood can be grouped by venue categories
- Project idea: explore venue categories in a neighbourhood and investigate which venue categories are correlated with the neighbourhood house sale price

DATA

- Kaggle NYC property sales dataset:
  - One year's property sales records from Sept. 2016 to Sept. 2017
  - 84,548 house sale transactions with a total amount of 89,335,360,909 USD in NYC
- Geocoder Nominatim API to get the neighbourhood latitude and longitude location
- Foursquare API to explore venues in each neighborhood

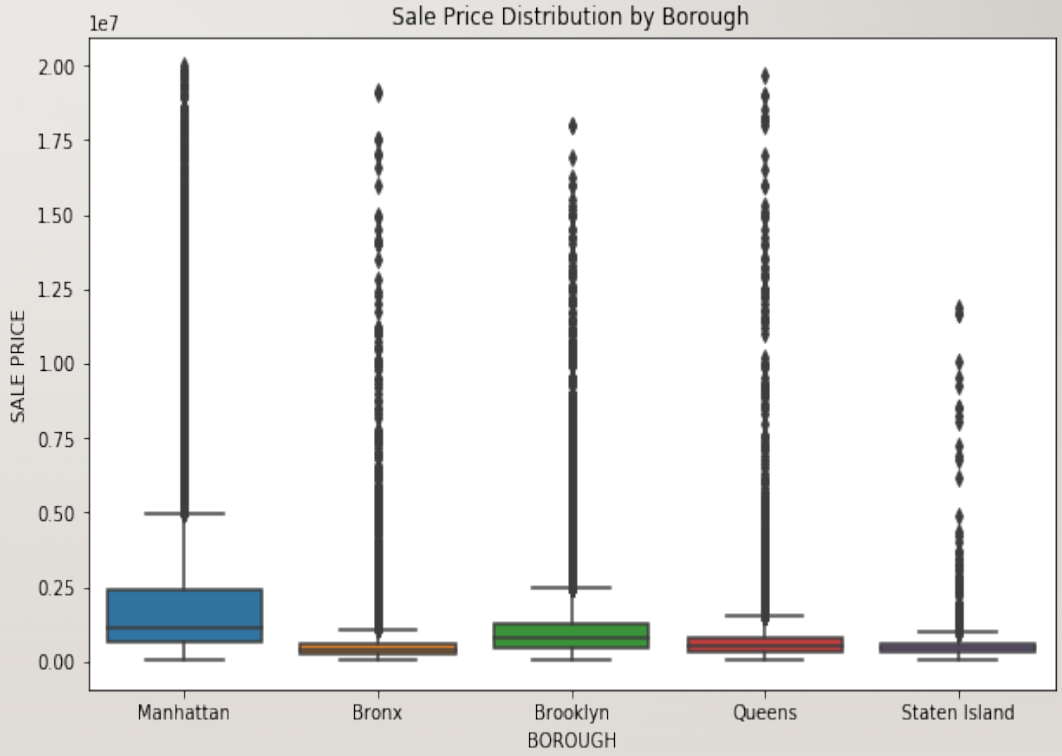
DATA PREPARATION

- Remove duplicated records in dataset
- Remove unrealistically low price entries - mostly transfers e.g. from parents to their child
- Focus on house sale price between 50,000 to 20,000,000 USD



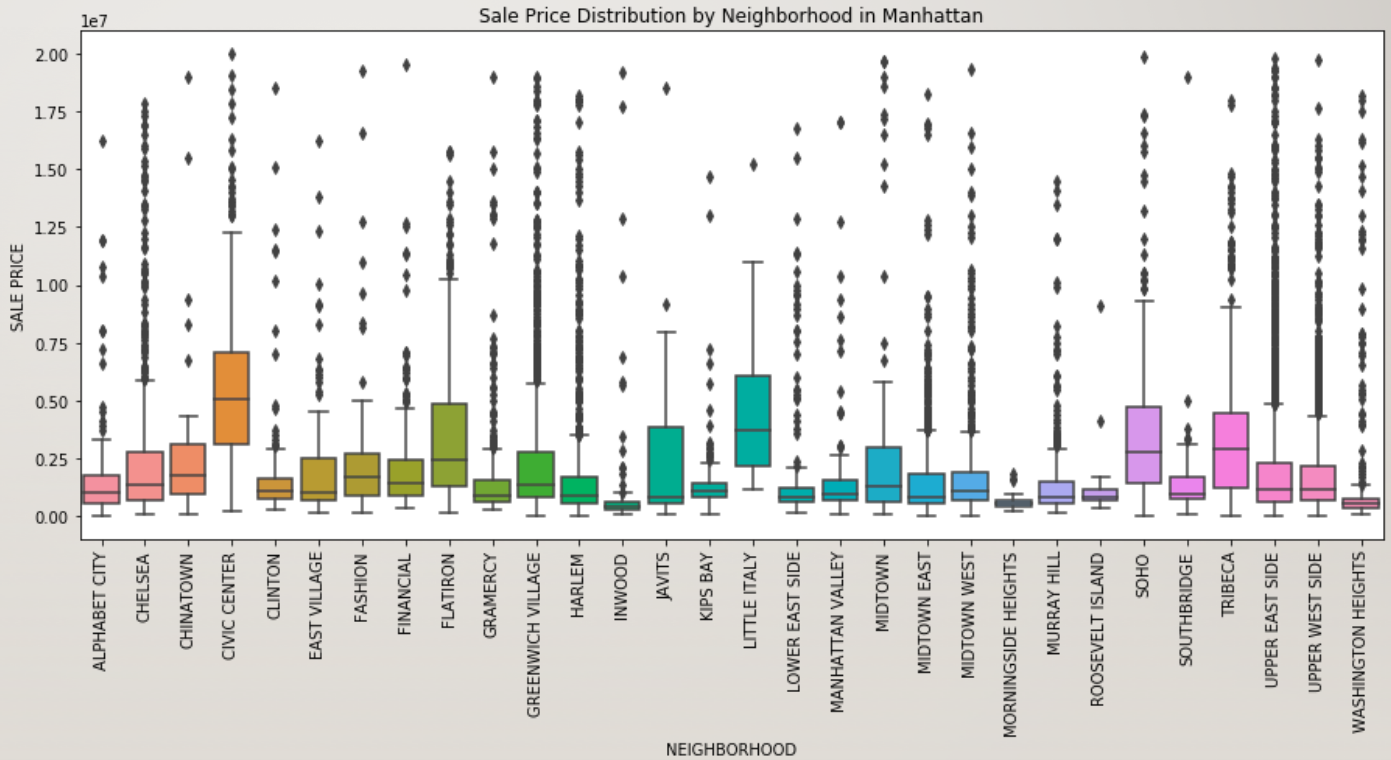
PROPERTY PRICE BY BOROUGH

- Very different price between boroughs
- We focus on Manhattan borough
- 18306 house sale transactions with a total amount of 48196678399.0 in Manhattan



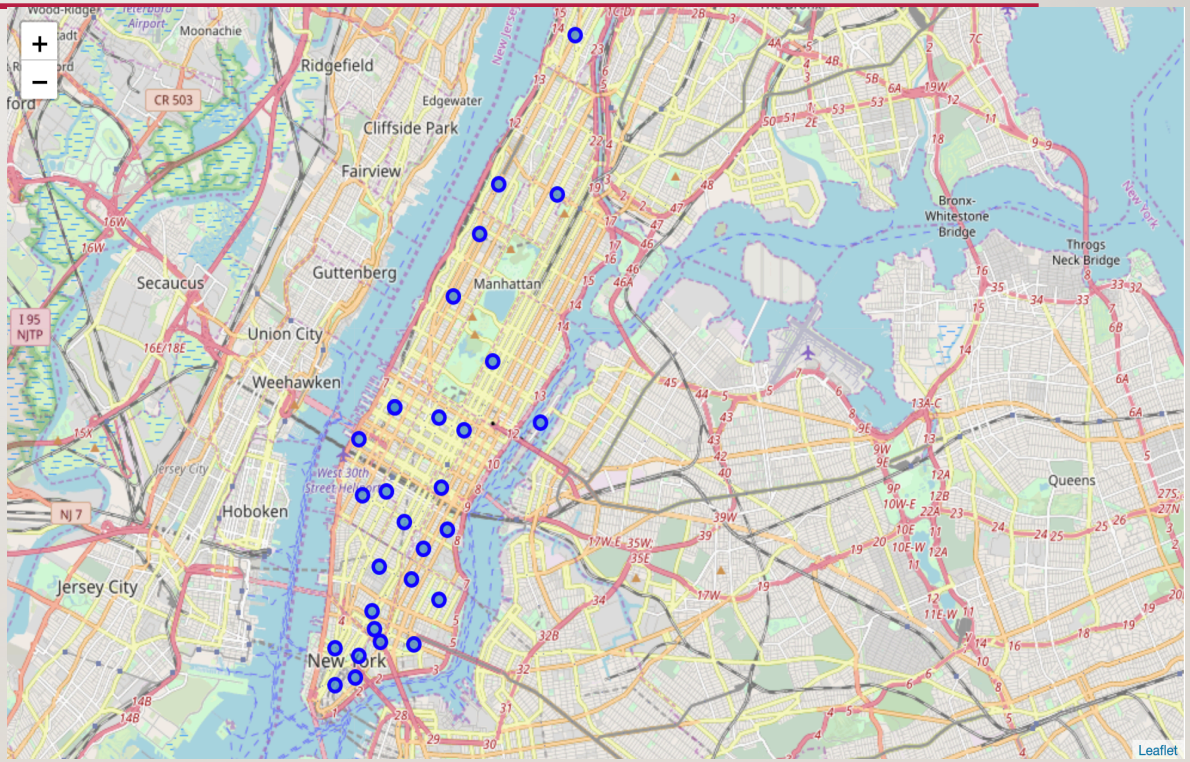
MANHATTAN PRICE BY NEIGHBORHOOD

- Property price differs in neighborhoods



LOCATION OF MANHATTAN NEIGHBORHOOD

- Query the location of each Manhattan neighbourhood by geocoder Nominatim
- Display neighbourhoods by Folium



VENUE CATEGORIES BY FOURSQUARE

- Use Foursquare API to explore the venues in each neighbourhood
- Get the category of the top 100 venues of each neighbourhood
- Calculate the distribution of the venue category in each neighbourhood
- Investigate which venue category of a neighbourhood correlates with the median house sale price of the neighborhood

TOP CORRELATED VENUE CATEGORIES

- Correlation by Pearson's statistics
- 14 venue categories are significantly correlated with neighbourhood property price, i.e., with p-value < 0.05
- 11 venue categories are significantly positively correlated, and 3 are significantly negatively correlated

venue_category	pearson_coef	p_value
Dim Sum Restaurant	0.603476	0.000415
Optical Shop	0.568231	0.001054
Salon / Barbershop	0.511324	0.003879
Martial Arts Dojo	0.453967	0.011740
Falafel Restaurant	0.446707	0.013336
Deli / Bodega	-0.435020	0.016284
Furniture / Home Store	0.420752	0.020598
Arts & Crafts Store	0.412430	0.023521
Dance Studio	0.410312	0.024317
Dessert Shop	0.407086	0.025571
Chinese Restaurant	0.390620	0.032825
Women's Store	0.379428	0.038644
Dog Run	-0.376383	0.040363
Pizza Place	-0.369133	0.044703

DISCUSSION AND CONCLUSION

- Investigate the correlation between neighbourhood real estate property price and neighbourhood venue categories
- Identify some interesting venue categories that are significantly correlated with property price
- Such venue categories should be also useful for predicting individual property prices
- Unfortunately, the limitation of Geocoder Nominatim (frequent service timeout) makes location query for all individual properties listed in the Kaggle dataset not possible