Data Science Capstone Report

# Introduction

The common saying in real estate is the three most important things when buying property are location, location, location. In my capstone project I will attempt to quantify the impact that the venues in a given location have on the price of both buying and short-term renting there.

There were a few different lenses I wanted to use to look at the effect of venues on prices:

* Total number of venues in a neighborhood
* Variety of venues
* Density of venues
* Types of venues - is there one type of venue that affects price more drastically than others?

# Data

I used three different sources to gather my data for this project:

* Foursquare
* Airbnb
* Zillow

I used the Foursquare API to gather information about the venues in each neighborhood. I downloaded summary information and metrics for all rental listings in New York City from Airbnb. I downloaded the average sale price of homes for each neighborhood in NY from Zillow.

# Methodology & Analysis

My first step was downloading all the data from the various sources.

I cleaned the data by getting rid of duplicate and missing values. I also filtered out the information that I didn’t need/wasn’t using. I used maps and charts to visualize the initial data. I manipulated the data to get the attributes I was interested in.

I then split the data into test and train and performed regression analysis on it.

# Results

I looked at three potential predictor variables:

* Total number of venues in a neighborhood
* Number of unique types of venues in a neighborhood
* How expensive (on a scale of 1 – 4) venues in a neighborhood are

I looked at each variable individually then also multiple variables at a time.

|  |  |
| --- | --- |
| Variable(s) | R2 |
| Total Venues | 0.2727 |
| Unique Venue Types | 0.1010 |
| Average Venue Price | 0.0218 |
| Total Venues + Unique Venue Types | 0.1035 |
| Total Venues + Average Venue Price | 0.4620 |
| Unique Venue Types + Average Venue Price | 0.5053 |
| All Three Variables | 0.5081 |

# Discussion

None of the single variables we looked at were good at predicting rental prices in a neighborhood. When we started adding looking at multiple variables our results were much better. The best predictor was when we looked at all three variables together.

That being said, even looking at all three variables it didn't give us a great predictor of rental prices.

I suspect part of it is that there was a 100 venue limit per neighborhood. There is a clear line in the single variable regression graph for total number of venues where the limit cuts off. I think we could have had better results if that limit wasn't there.

# Conclusion

I think there is a lot that could still be done with this data. I would be interested to run the same analysis on all the neighborhoods in NYC instead of just Manhattan to see if that changes anything.

In further analysis I would be interested in utilizing all the data we have on specific rental listings. The large number of listing made it impractical for this project, however, I think looking at the proximity and variety of venues compared to individual listings would be very informative.