“A cup of coffee, please!”

— A Data-Oriented Guide to Coffee Establishment in Manhattan

Steven Du

Cap Stone Project Report, Nov 2019

Coffee in New York

- A recent report (Wallethub, Sep, 2019) shows that New York is the **number one city** in American in term of the number of Coffee Shop, Coffee House and Café per capita.

- A company named Massive Health also reported that, in 2012, New Yorkers drink **6.7 times more coffee** than people in other cities.

- Needless to say, coffee plays an indispensable role in the life of people who lives and visits New York.

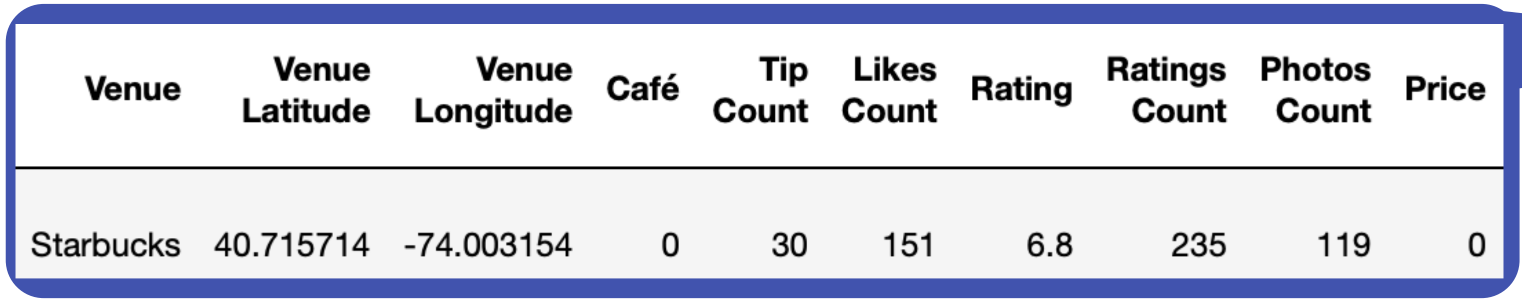
- ***But how much do we know about the coffee establishments— places where one can purchase coffee— in this city (Manhattan Island in particular)?***

What makes a high-quality coffee shop and where shall I go for a good cup of coffee on Manhattan?

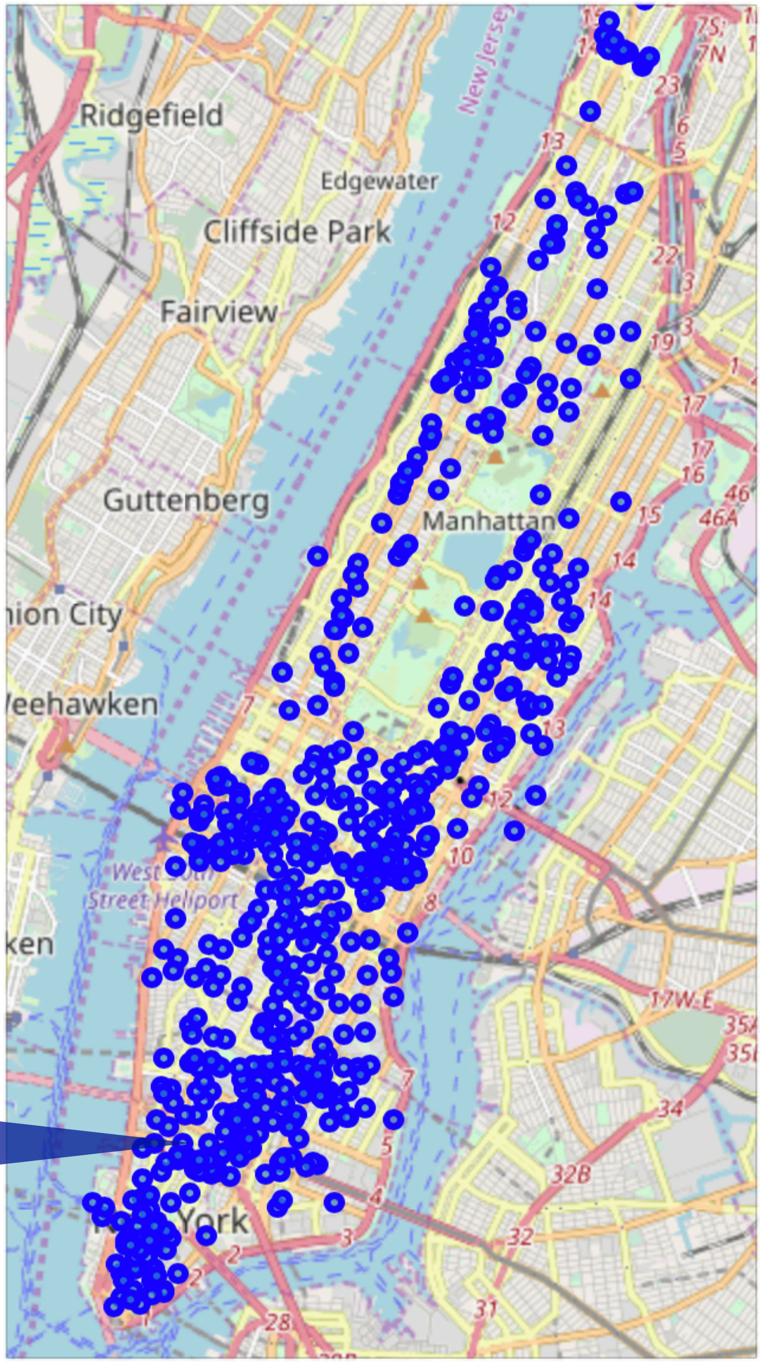
Data Collection

Locating Coffee Establishments in Manhattan

- Define coffee establishments as venues that are categorized as either **'Coffee Shop' or 'Café'** by Foursquare.
- Use 'Search' endpoint to search for coffee establishments around the centers of **40 neighborhoods in Manhattan**.
- Use 'Detail' endpoint to retrieve **the rating, number of likes/ratings/tips/photos, price tier and category** of each coffee establishments.



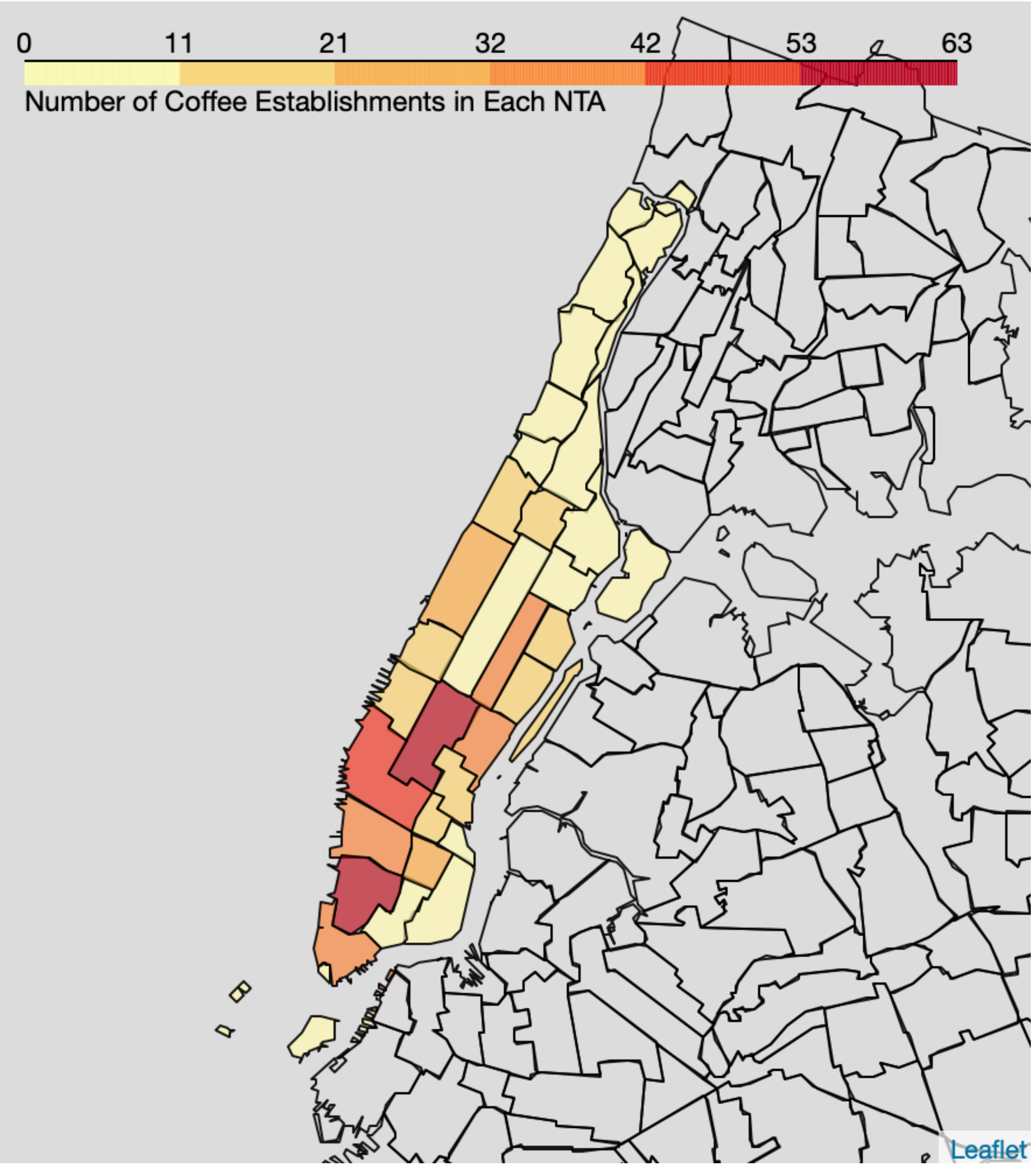
Venue	Venue Latitude	Venue Longitude	Café	Tip Count	Likes Count	Rating	Ratings Count	Photos Count	Price
Starbucks	40.715714	-74.003154	0	30	151	6.8	235	119	0



Data Collection

Coffee Establishments in Neighborhood Tabulation Areas

- New York city is divided into many neighborhood tabulation areas (NTAs), and each NTA may include one or several neighborhood.
- Manhattan is divided into **29 NTAs**.
- We find **633 coffee establishments** in Manhattan!
- The distribution of coffee establishments in Manhattan is far from homogeneous. Regions located in the south and the west part of Manhattan tend to host more coffee establishments.



Brief Overview of Coffee Establishments in Manhattan

- Among all 633 coffee establishments, **196 are ‘Cafe’, 157 are Starbucks** (all categorized as ‘Coffee Shop’)

- Over **83%** of coffee establishments are **low-priced** (price tier 1) and the rest are high-priced (price tier 2 and 3).

- **82** coffee establishments do not have a rating score (0-10). For the rest, the rating ranges **between 5 and 9.3**. The **median rating is 7.7**, and the 25% and 75% quantiles are 6.0 and 8.45.

- Many coffee establishments receive a numbers of likes/ratings/tips/photos. The **distributions of such counts are highly skewed.** While most coffee establishments only accumulate dozens or at most a few hundreds counts, a small number of coffee establishments can receive thousands of likes and photos from users.

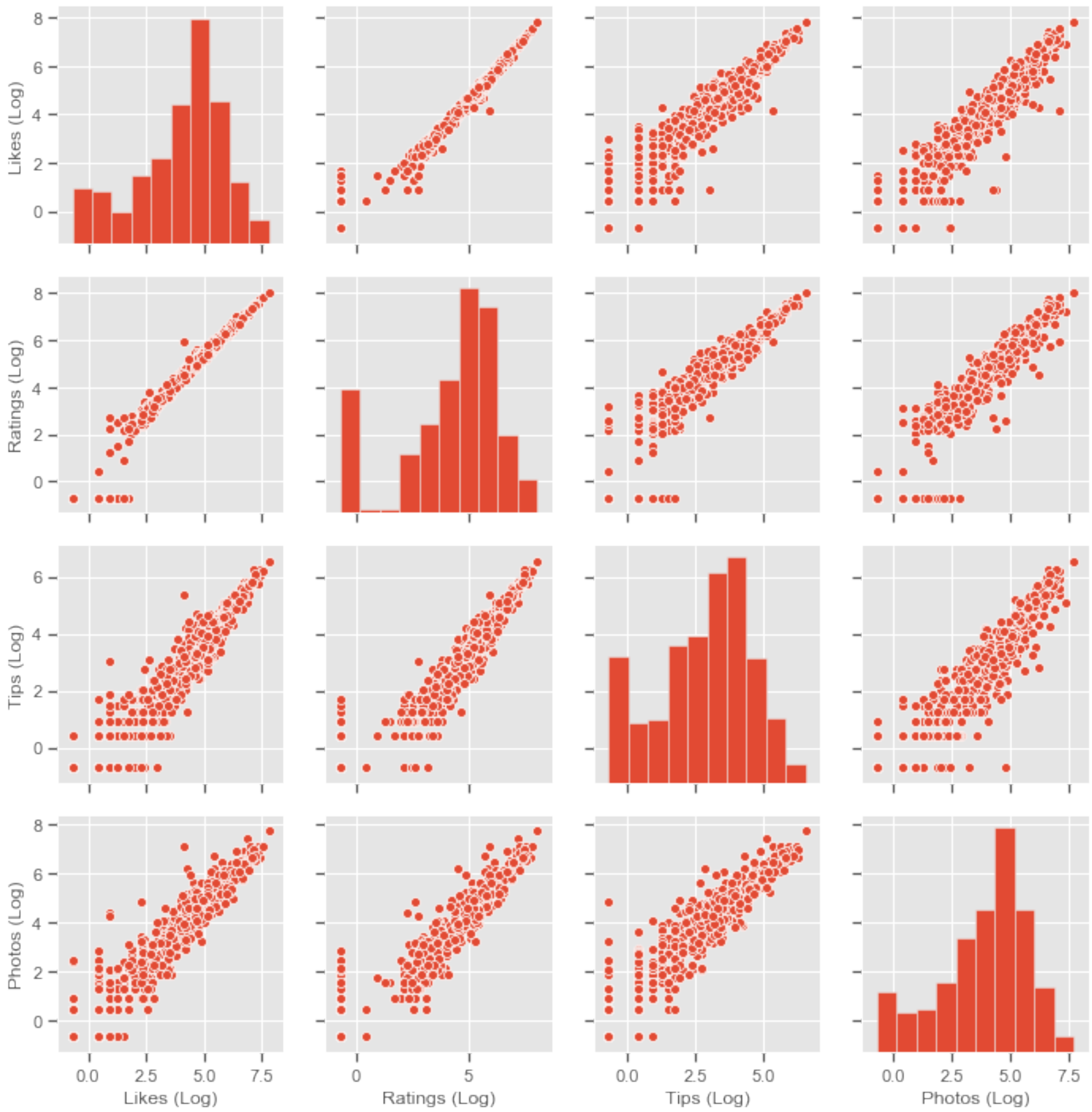
Exploratory Analysis

Number of Likes: A Measure of Popularity

- We conduct **logarithm transformation** on the number of likes/ratings/tips/photos to **reduce the impact of extreme values**.

- Except a number of zero counts, the distributions of logarithms of these four counts are relatively non-skewed. Moreover, **these different counts are highly corrected**.

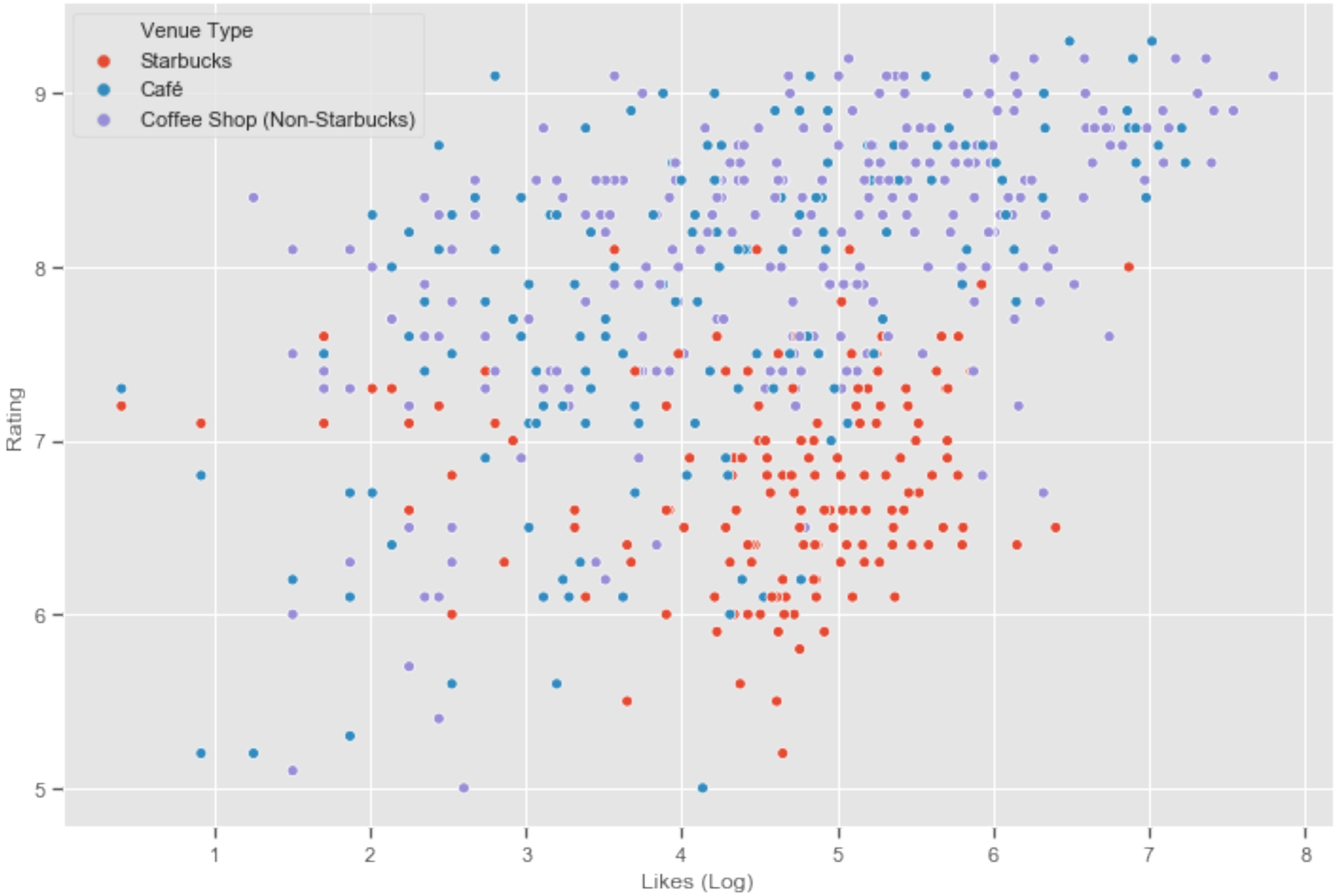
- To avoid multicollinearity, we will choose **the logarithm of the number of likes as the sole measure of popularity** for coffee establishments.



Exploratory Analysis

Rating: A Measure of Quality

- We use **Rating** received by the coffee establishments as **measurement of quality**.
- Our inspection shows that **Rating is positively related to the logarithm of the number of likes.**
- However, such relationship varies between different types of coffee establishments, especially **between Starbucks and non-Starbucks.**

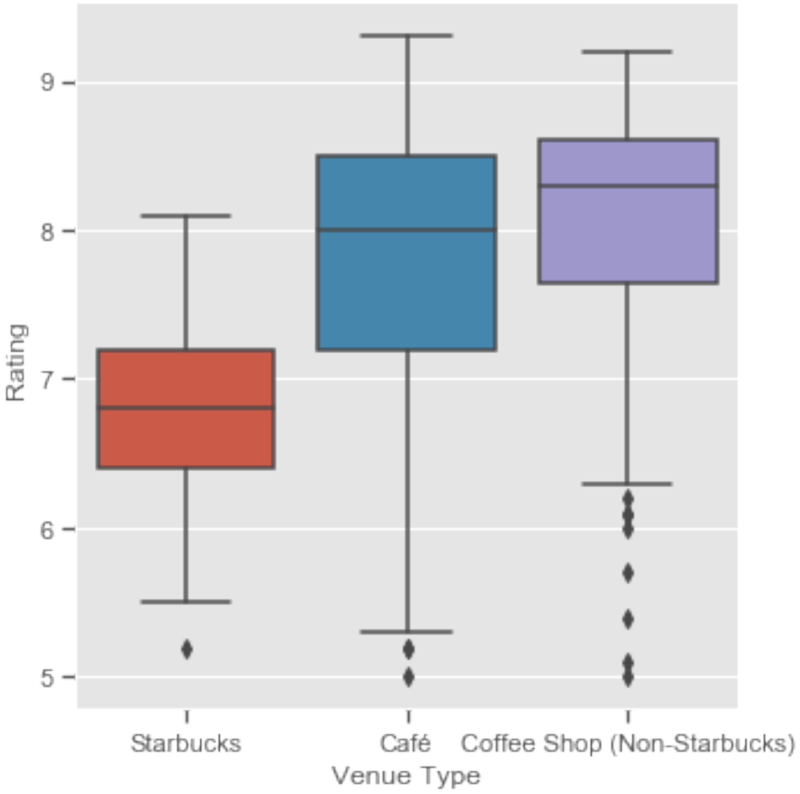


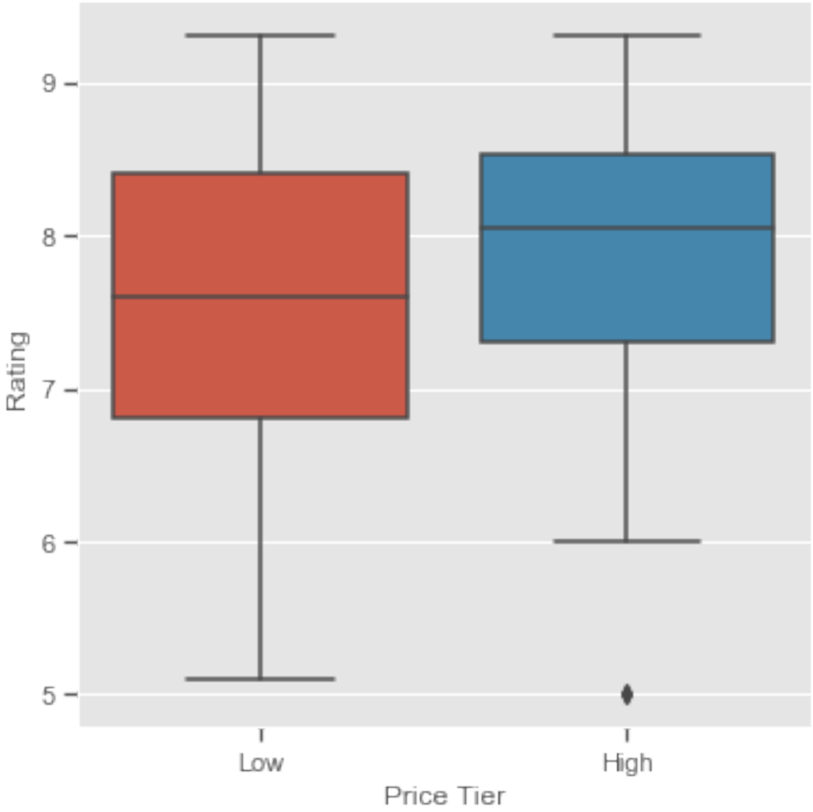
Exploratory Analysis

Rating, Type of Coffee Establishments and Price

- Rating differs between different type of establishments. **Starbucks tend to receive lower rating compared to other establishments.**

- On average, **high-priced coffee establishments receive higher rating** than low-priced coffee establishment.





Multiple Regression Model

How Rating is Related to Other Factors?

- We will use multiple regression model to fit the independent variable **Rating** against the **four dependent variables as well as their interaction terms**:

    - ✦ **Log-Likes:** Logarithm of Number of Likes
    - ✦ **Price:** 0 for low-price and 1 for high-price
    - ✦ **Starbucks:** 0 for being a Starbucks and 1 for not
    - ✦ **Café:** 0 for being categorized as Café, 1 for being categorized as Coffee Shop

- All the coffee establishments that receive no rating (82 in total) will be dropped as missing data.

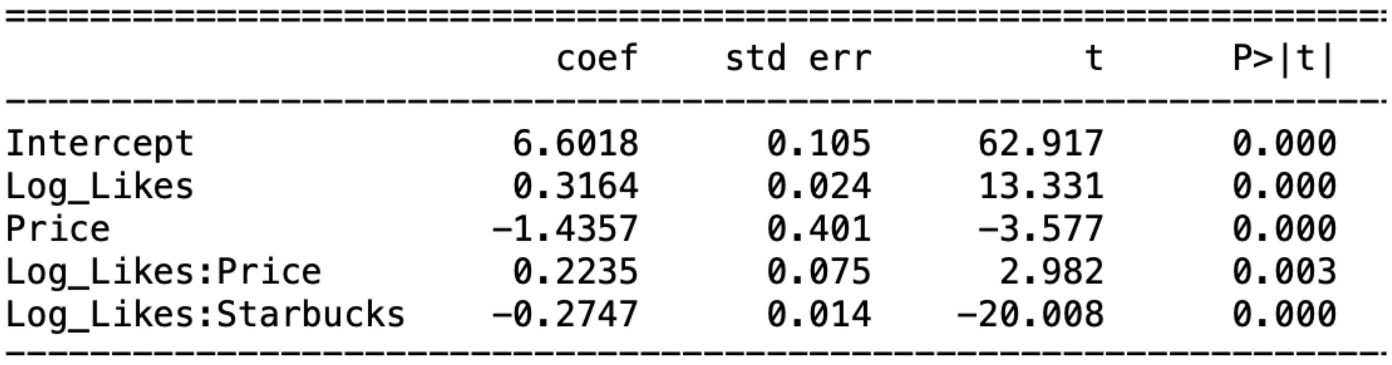
- **BIC** is used as the model selection criterion for dropping insignificant predictors.

Multiple Regression Model

Results

- The following model is selected based on the lowest BIC:

$$
\begin{aligned}
\text{Rating} &= 6.6018 \\
&+ 0.3164 \text{ Log-Likes} \\
&- 1.4357 \text{ Price} \\
&+ 0.2235 \text{ Log-Likes} \times \text{Price} \\
&- 0.2747 \text{ Log-Likes} \times \text{Starbucks} \\
&+ \epsilon
\end{aligned}
$$



```
==============================================================================
                          coef    std err          t      P>|t|
------------------------------------------------------------------------------
Intercept               6.6018      0.105     62.917      0.000
Log_Likes               0.3164      0.024     13.331      0.000
Price                  -1.4357      0.401     -3.577      0.000
Log_Likes:Price         0.2235      0.075      2.982      0.003
Log_Likes:Starbucks    -0.2747      0.014    -20.008      0.000
------------------------------------------------------------------------------
```

- Our analysis confirms the following observations:

☑ The rating is positively related with the logarithm of the number of likes.

☑ This relationship does not depend on whether a coffee establishment is categorized as Café or Coffee Shop.

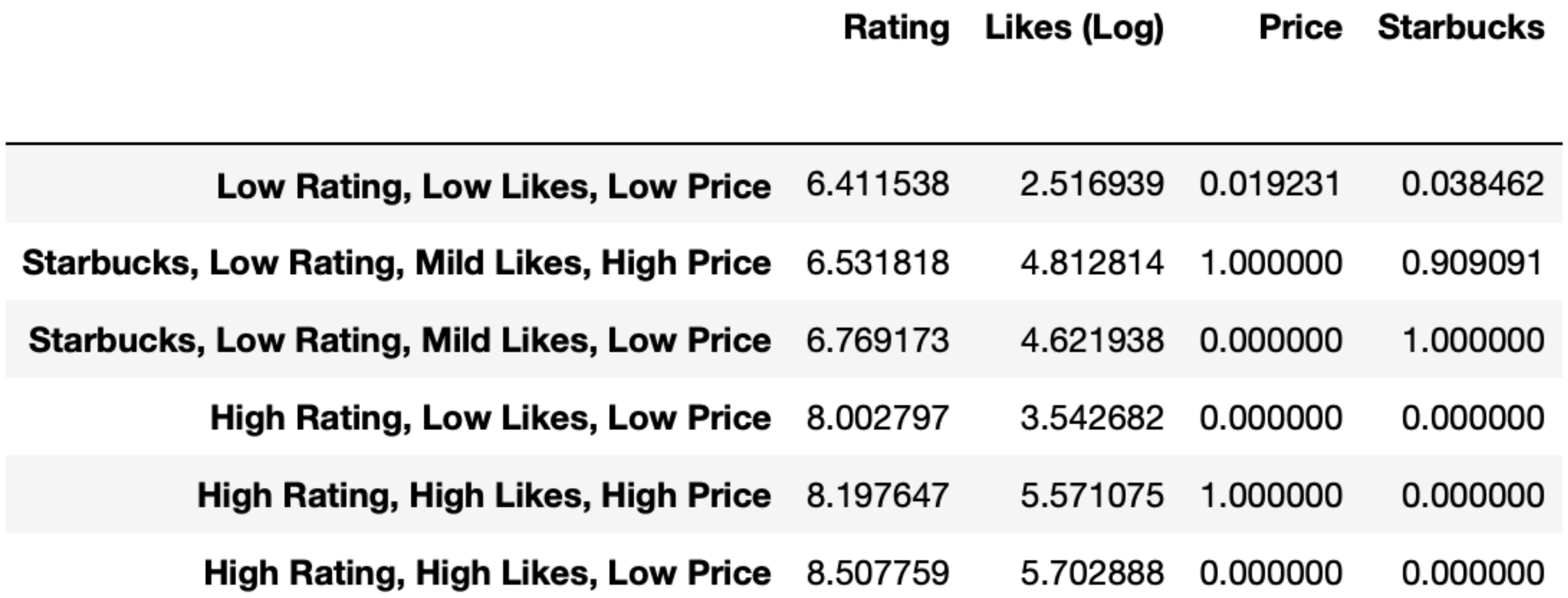
☑ Starbucks on average receive significantly lower rating compared to non-Starbucks coffee shop.

☑With all else being equal, a high-priced coffee shop tends to receive lower rating compared to a low-priced one.

Clustering Coffee Establishments

Find out the Major Archetypes of Coffee Establishments

- To understand how the coffee establishments differ between NTAs in Manhattan, we use the rating, logarithm of likes, price and whether the coffee establishment is Starbucks or not as inputs, and apply **K-Means algorithm** to cluster the coffee establishments into **6 archetypes** with distinctive characteristics:



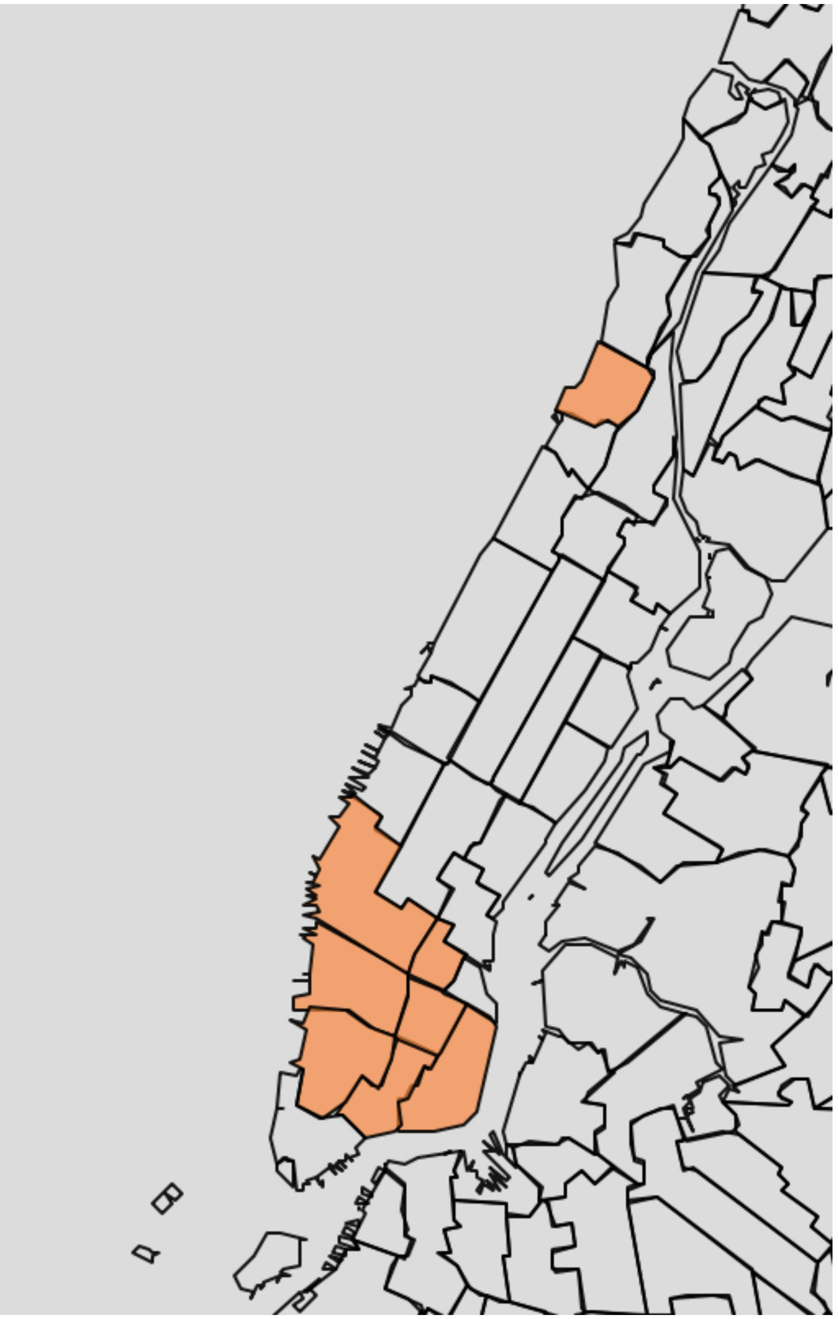
	Rating	Likes (Log)	Price	Starbucks
Low Rating, Low Likes, Low Price	6.411538	2.516939	0.019231	0.038462
Starbucks, Low Rating, Mild Likes, High Price	6.531818	4.812814	1.000000	0.909091
Starbucks, Low Rating, Mild Likes, Low Price	6.769173	4.621938	0.000000	1.000000
High Rating, Low Likes, Low Price	8.002797	3.542682	0.000000	0.000000
High Rating, High Likes, High Price	8.197647	5.571075	1.000000	0.000000
High Rating, High Likes, Low Price	8.507759	5.702888	0.000000	0.000000

Clustering Neighborhood Tabulation Areas

Battle Between Neighborhoods for Coffee Establishments

- Different NTA has different preference on the archetypes of coffee shops. By using the **proportions of archetypes** in a NTA as input, we can further **clusters the NTAs in to 3 major regions and 2 single-NTA regions**:

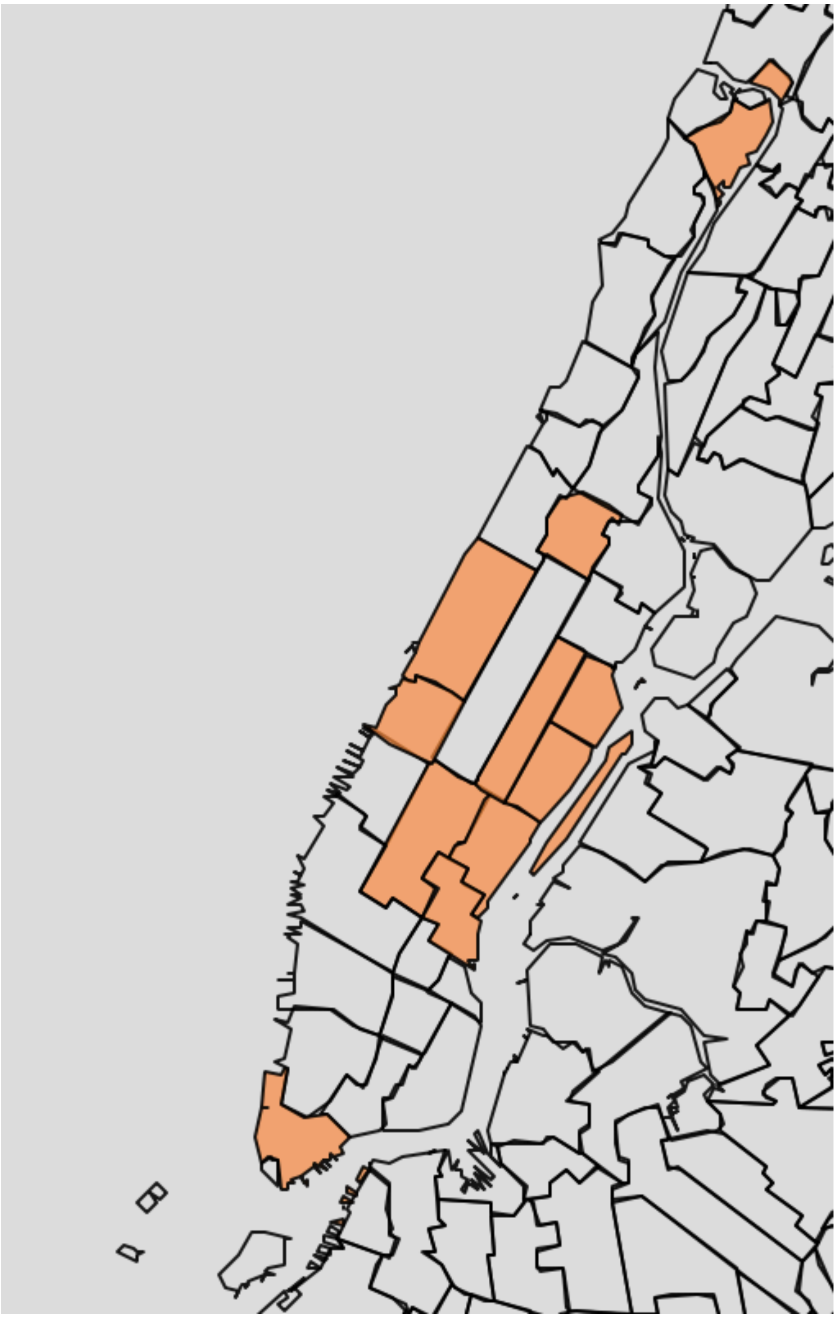
- First Major Region:
  - *Neighborhood*: Most neighborhoods in the southern part of Manhattan, along with Hamilton Height in the north
  - *Feature*: **Mainly host coffee establishment with high ratings**
  - *Recommend for*: **High quality coffee**



Clustering Neighborhood Tabulation Areas

Battle Between Neighborhoods for Coffee Establishments

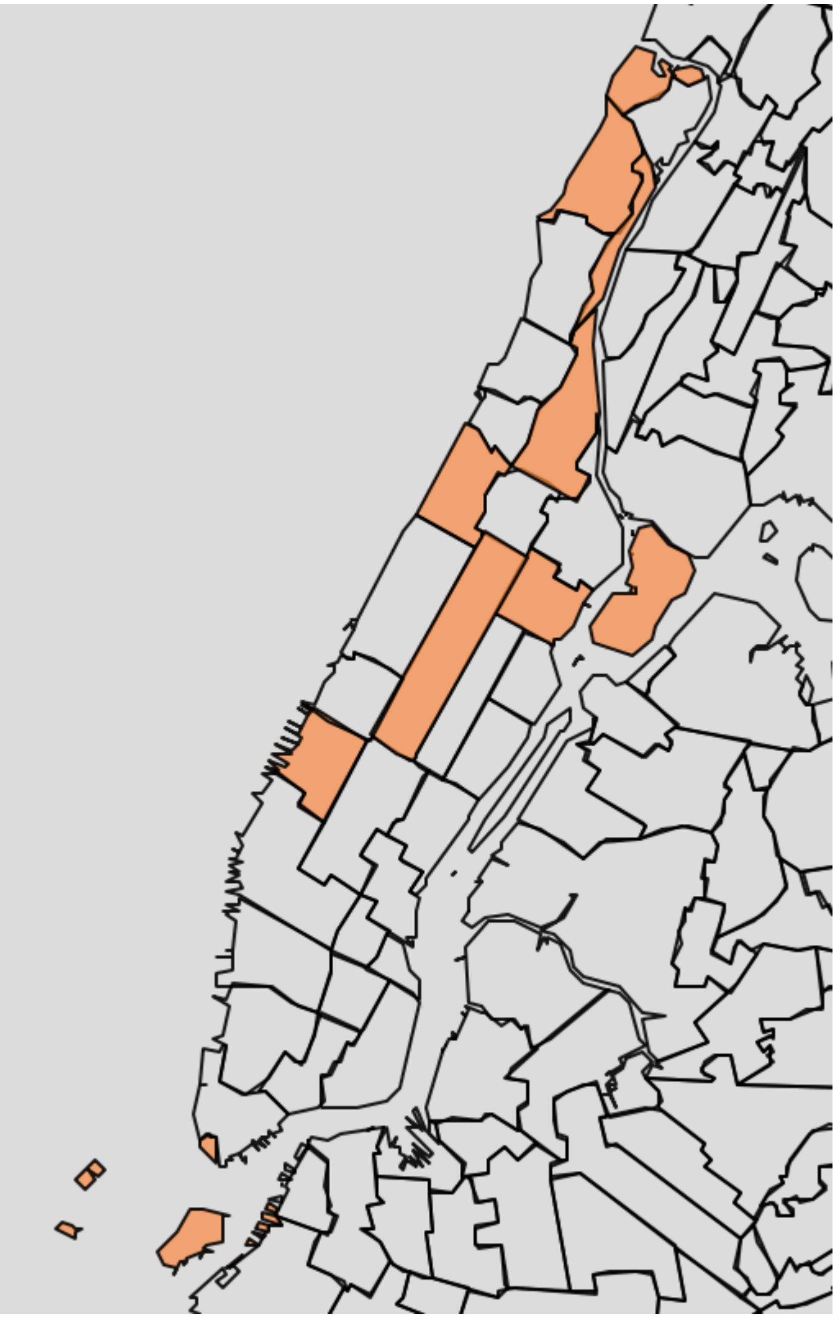
- Second Major Region:
  - *Neighborhood*: Downtown Manhattan, most region in middle Manhattan, Marble Hill and Inwood in the north.
  - *Feature*: **Many high rating, but less popular coffee establishments. The proportion of Starbucks is also significantly higher.**
  - *Recommend for*: **Enjoying good coffee at a more leisure pace**



Clustering Neighborhood Tabulation Areas

Battle Between Neighborhoods for Coffee Establishments

- Third Major Region:
  - *Neighborhood*: Mainly neighborhoods located to the north of central park, along with Clinton neighborhood.
  - *Feature*: **Faire shares of both high rating and low rating coffee establishments with relatively low level of popularity.**
  - *Recommend for*: **Exploration**



Conclusion

A List of Recommendations

- ✓ High numbers of likes/ratings/tips/photos are generally good indicators for high quality.
- ✓ It does not matter whether a coffee establishment is categorized as “Coffee Shop” or ‘Café’.
- ✓ Starbucks can be nice to meet new people (as they tend to attract reasonable level of popularity), but are not idea for enjoying coffee (due to the low rating).
- ✓ Different neighborhoods can provide quite difference experience in coffee establishments, so choose wisely!
- ✓ Nonetheless, no matter where you are, there is always a chance to find a good coffee establishment as long as you keep looking!

Discussion

Myths of Starbucks

- Starbucks tend to receive lower rating but higher number of likes compared to other coffee establishments.

  ➡ Are the Starbucks under-rated? Or Starbucks have a few tricks to boast popularity?

- While Starbucks can be found in almost all areas of Manhattan (10%-15%), the proportion of Starbucks in the second major region is considerably higher (30%).

  ➡ Do the Starbucks hold certain market advantage over the archetype of high rating, low number of likes coffee establishments (the major type of non-Starbucks in the second major region)?

Discussion

Limitation of This Study

- Rating and number of likes do not tell everything about the quality and popularity of coffee establishments.

- Our information on individual coffee establishments are limited. It would be great if we can also analyze the detailed tips and photos posted by the users.

- Neighborhood Tabulation Areas are not necessary the best ways to divide Manhattan for the purpose of our study.

- Our study is based on the observational data. Thus, we can not draw causal conclusion based on the results of our analysis (It is uncertain whether a shop owner can increase the rating of his/her shop by magically double the number of likes the shop received).

MANY THANKS FOR YOUR TIME!
HOPE YOU ENJOY THIS PROJECT!