What Makes a Good Starbucks in Los Angeles?

Applied Data Science Capstone project -- The Battle of Neighborhoods

by Andrew Turczyn

January 29, 2020

1. Introduction
   1. Background

LA has an extremely large number of coffee shops, particularly Starbucks. In my West Hollywood neighborhood, there are three Starbucks within a couple blocks of each other. Even though all three have the same core offerings on the menu, I have a definite preference for one of them over the others. I’m sure another person might rate them completely differently based on factors that are important to them. Foursquare has a rating for each Starbucks that is an average of all customer ratings logged on the app.

* 1. The problem

What are the factors that give one Starbucks a higher or lower Foursquare rating than another? In this report, I attempt to answer that question by considering some factors that might be relevant including:

* Location-specific attributes such as outdoor seating, wi-fi availability, etc.
* Neighboring business categories
* Distance to the next nearest Starbucks

Since all the venues in this study are Starbucks, I am not considering factors that might differentiate coffee shops from other brands such as marketing/advertising, distinct menu offerings, aesthetics, etc.

* 1. Interest

The results of this report would be very valuable for the following:

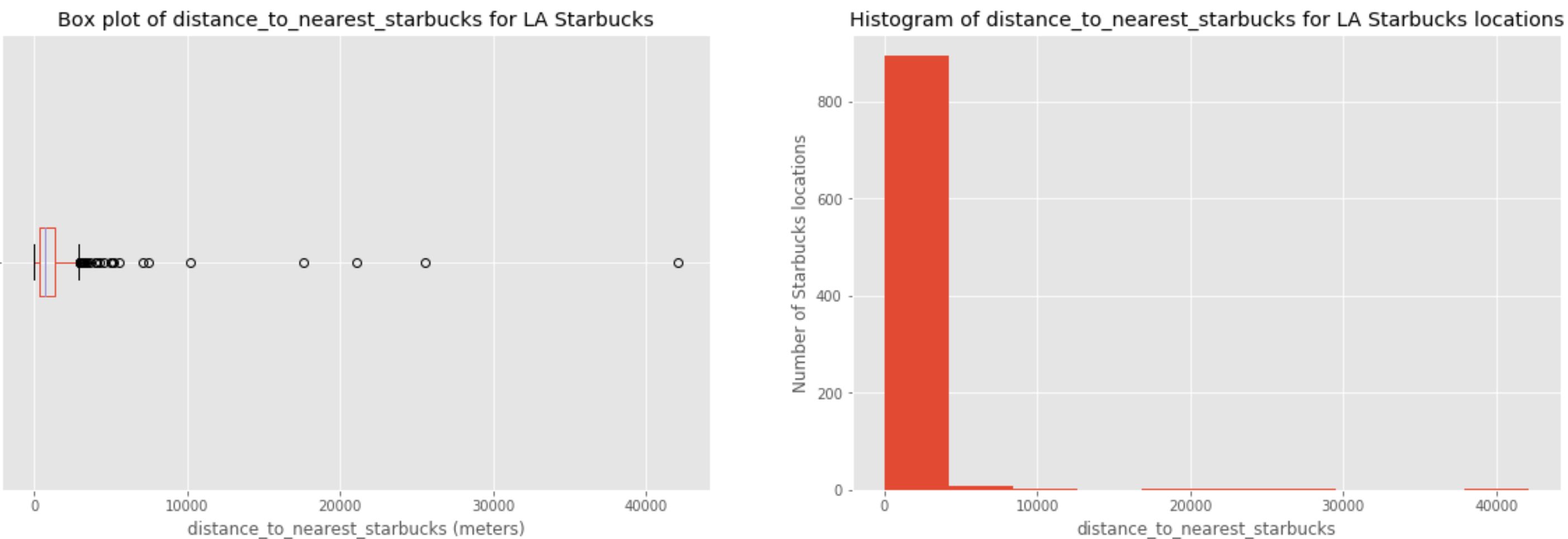
* Owners of existing Starbucks stores in Los Angeles
* Someone who is considering opening a new Starbucks location in Los Angeles

For existing owners, having a higher Foursquare rating is likely to increase sales by influencing new customers to choose their location over another nearby Starbucks or coffee shop with a lower rating. For a potential new Starbucks owner, it would be helpful to understand the features and neighborhood aspects that lead to higher or lower ratings to maximize potential revenue.

1. Data

The following data sources were used to gather information about Starbucks venues and the potential explanatory factors considered in this report:

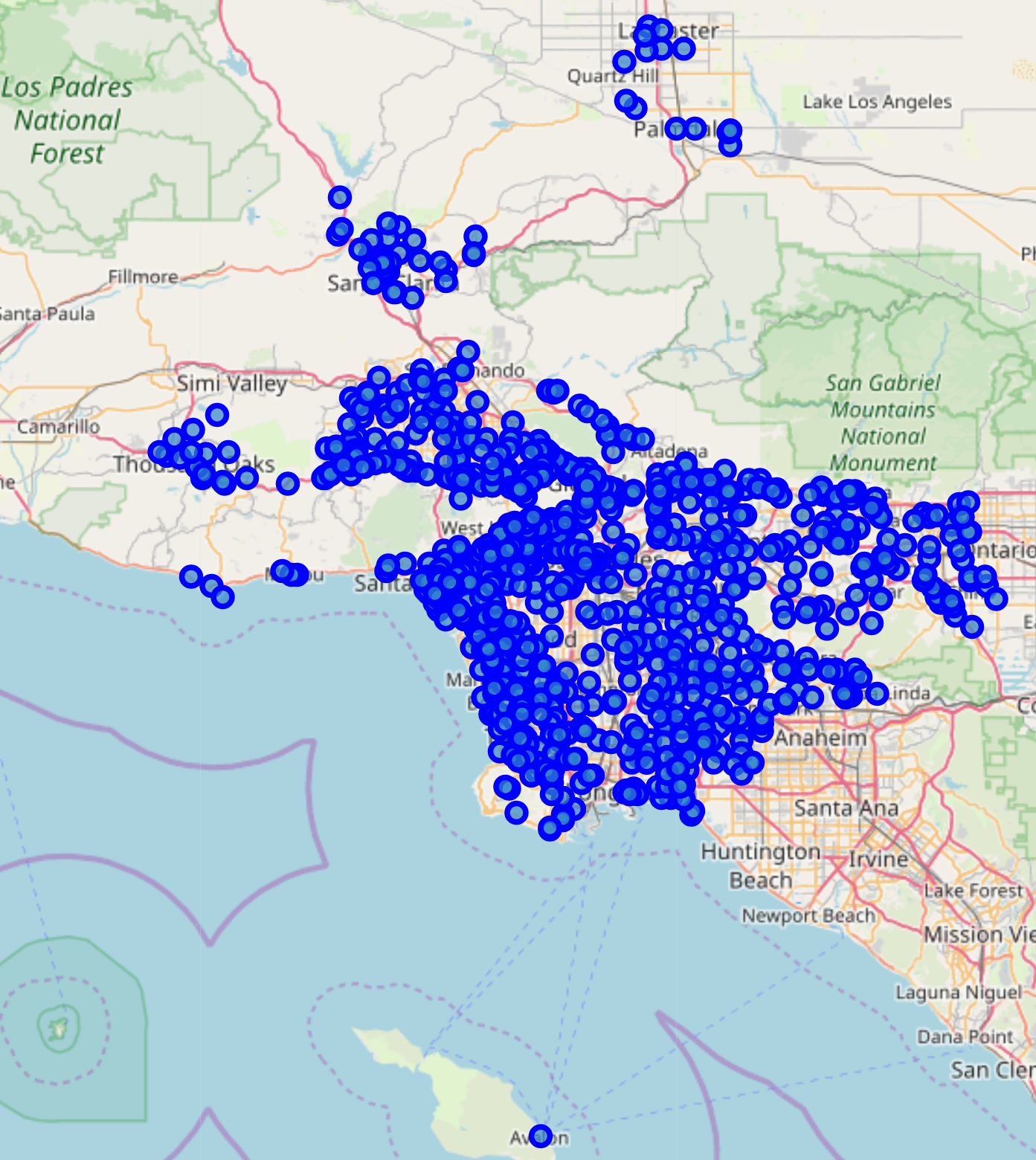
* List of all Starbucks locations in Los Angeles County
  + Get a list of LA zip codes from the LA county website:
    - <http://file.lacounty.gov/SDSInter/lac/1031552_MasterZipCodes.pdf>
    - 487 zip codes found
  + Use geocoder to map postal codes to latitude/longitude
  + Foursquare ‘search’ query by latitude/longitude to find all Starbucks in each postal code
    - Search within 2.5 mile radius of each postal code
    - Remove duplicates found in overlapping search areas
    - Capture each store ID, latitude/longitude, and postal code
      * 966 Starbucks locations identified in LA County
* Starbucks location info
  + Foursquare ‘venues’ query by store ID
  + Capture store rating, list of attributes (e.g. likes count, photo count, tips count, etc.)
  + One-hot encoding to convert attributes to indicator variables
  + Missing values:
    - Drop locations with missing rating after data exploration but before modeling
* Immediate neighbor category types for each Starbucks
  + Foursquare ‘search’ query by latitude/longitude to find top 5 venues nearest to each Starbucks within 0.1 miles
  + Capture category ID and category name
  + One-hot encoding to convert neighbor categories to indicator variables
* Distance to the nearest Starbucks
  + Use geodesic function from geopy.distance
  + Calculate straight-line distance between each pair of Starbucks located within zip codes with the same first three digits based on latitude/longitude of origin and destination
  + Missing values replaced with average distance
  + Correct outliers:
    - Replaced values > 5000 meters with distance = 5000 meters



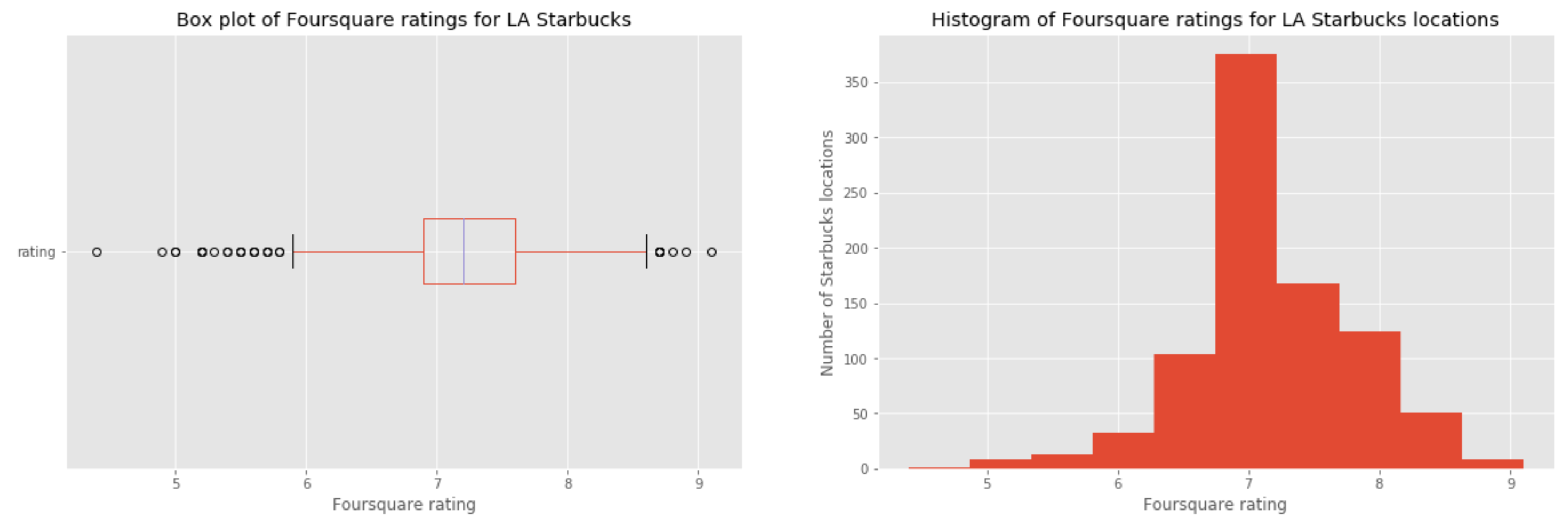
1. Methodology

* Data exploration and visualization to determine appropriate analysis
* Feature selection: Use correlation analysis to select features for modeling
* Machine learning: Use linear regression to predict rating based on proposed features
  1. Exploratory data analysis

Map of 966 Starbucks in LA County:



Examine the distribution of ratings across all Starbucks in LA:



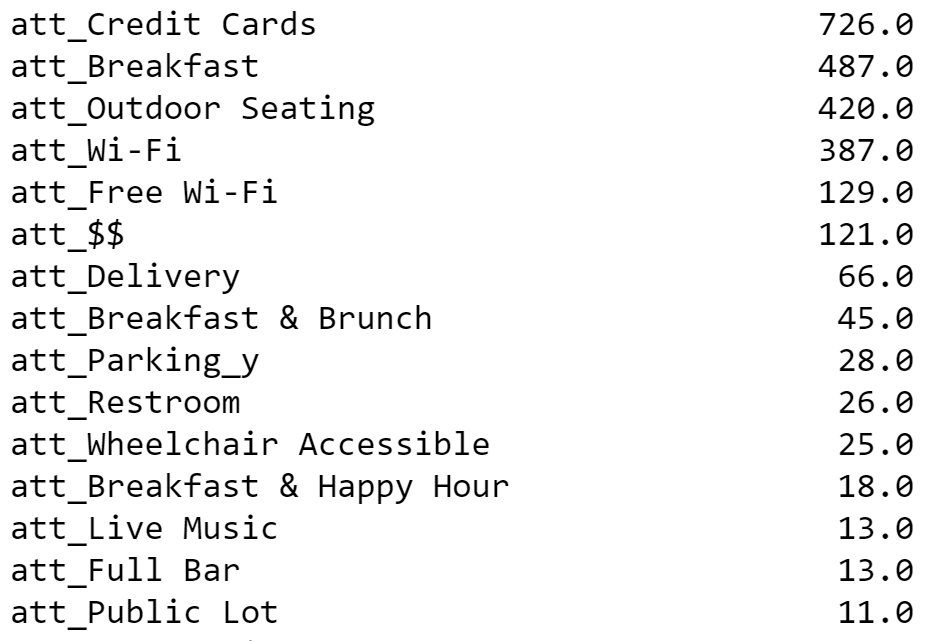
* Foursquare rating stats:
  + Average = 7.2
  + Standard deviation = 0.62
  + Min = 4.4
  + Max = 9.1
* Distribution is fairly normal with slightly positive skew with more outliers on the low side
  1. Feature selection

Potential features:

* **Immediate neighbor categories**
  + 346 venue category types identified across LA Starbucks
  + Most common category counts:



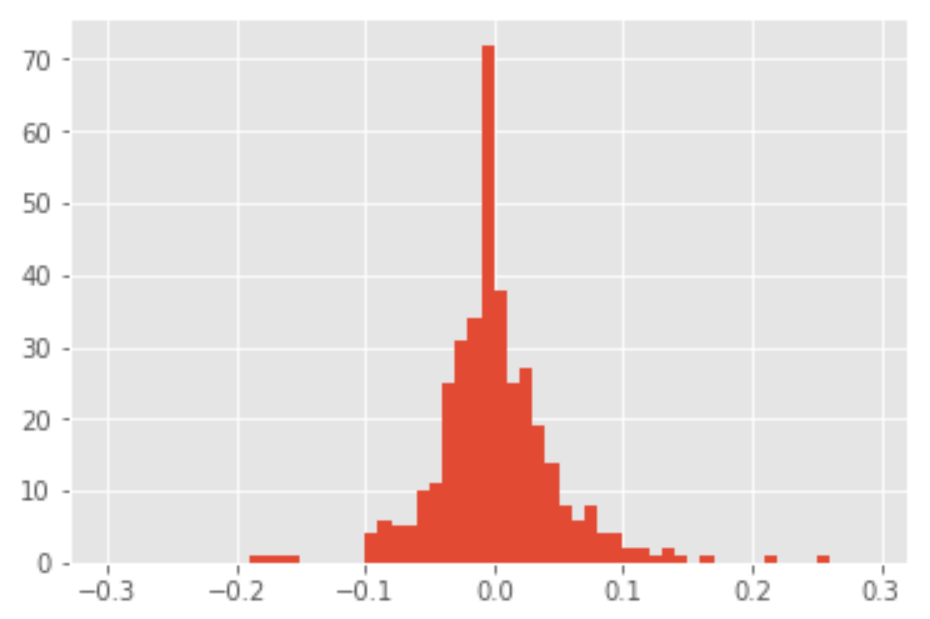
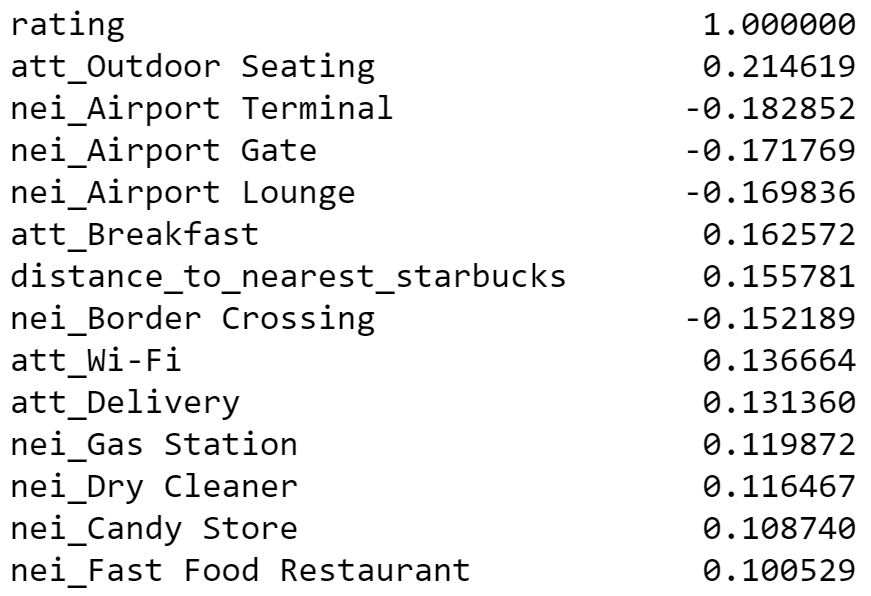
* **Venue attributes**
  + 49 attributes identified across all LA Starbucks
  + Most common attribute counts:



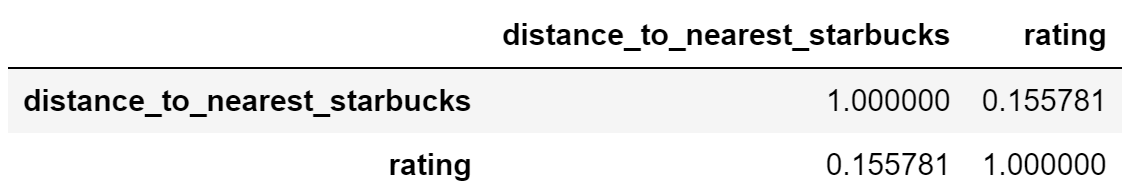
* **Distance to nearest Starbucks** (meters)

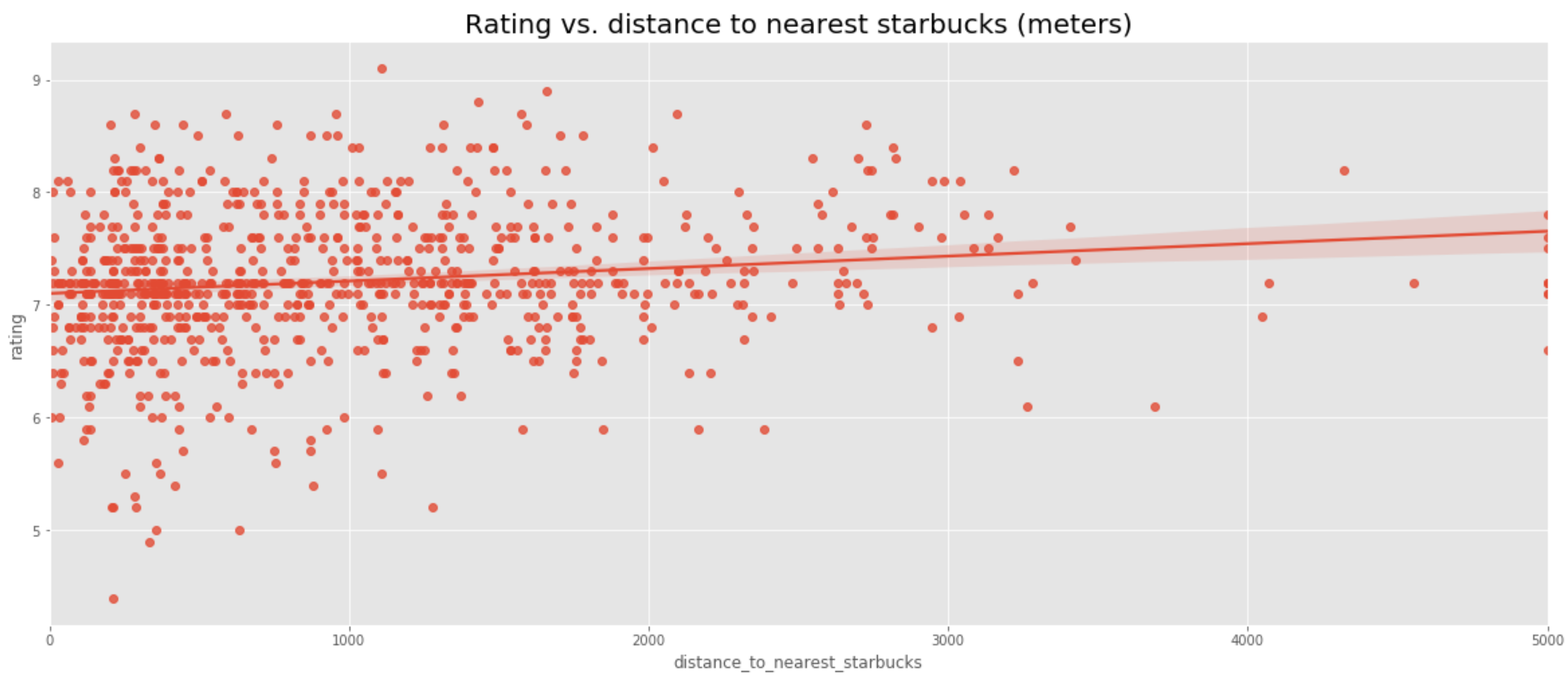
Pearson correlation with Foursquare rating:

* Rating vs. immediate neighbor categories & venue attributes
  + Generally weak, with only a few having absolute correlation > 0.1

* Rating vs. distance to nearest Starbucks
  + Slightly positive (0.16)
  + Starbucks that are located furthest from another Starbucks are generally rated better





* 1. Machine learnings used

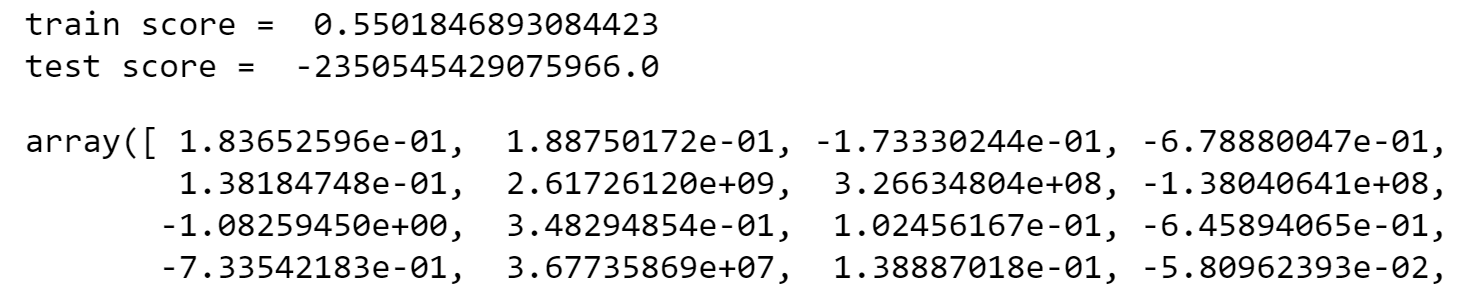
I used a multiple linear regression modeling approach to predict Foursquare rating for LA Starbucks based on the potential features identified previously.

* Split the modeling data set into train (80%) and test (20%) sets
* Use R-square as a performance metric over both the train and test datasets to compare various model configurations
* Select feature subsets to find a combination that produces the best predictive results over the test dataset

1. Results

Model 1: (All features)

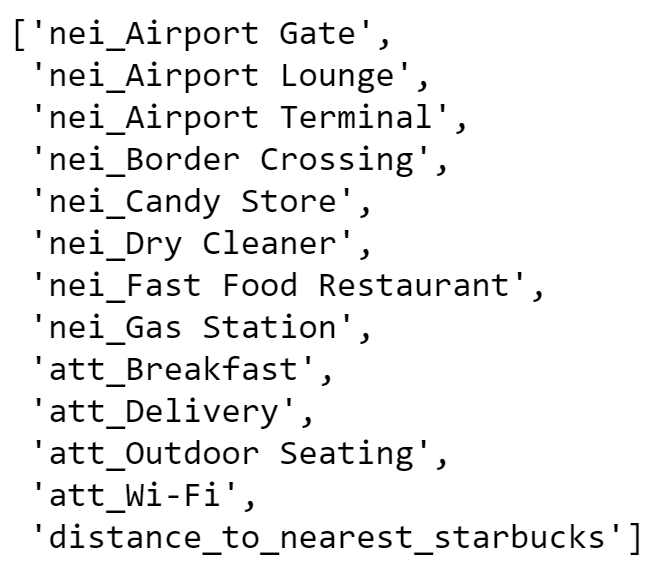
* Input features:
  + All immediate neighbor categories
  + All venue attributes
  + Distance to nearest Starbucks
  + Latitude/longitude of store location (to indicate general geographic desirability)
* Results:



* + R-square is quite good (0.55) over the train dataset but is extremely negative over the test dataset indicating that the model is overfitting the train dataset and is not valuable for prediction at all.
  + The coefficients are extreme in magnitude, frequently to the power of 8 or 9. These are not realistic weights and may be due to overfitting.

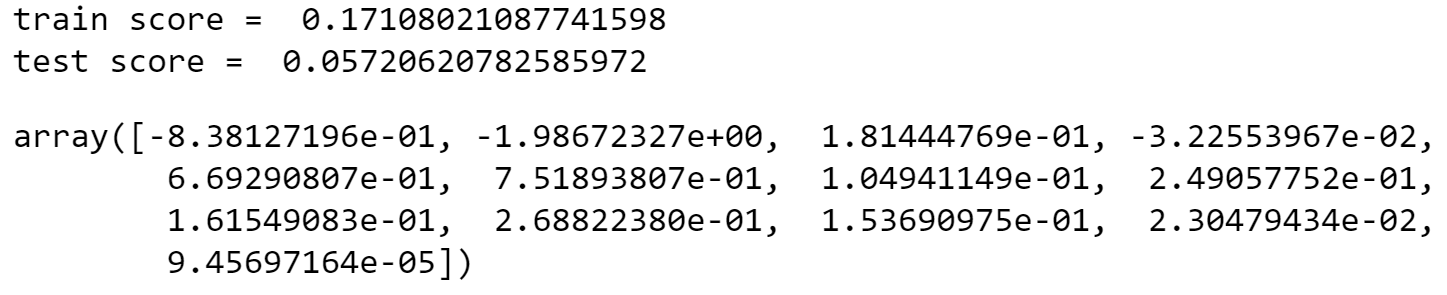
Model 2: (Features with correlation > 0.1)

* Input features:





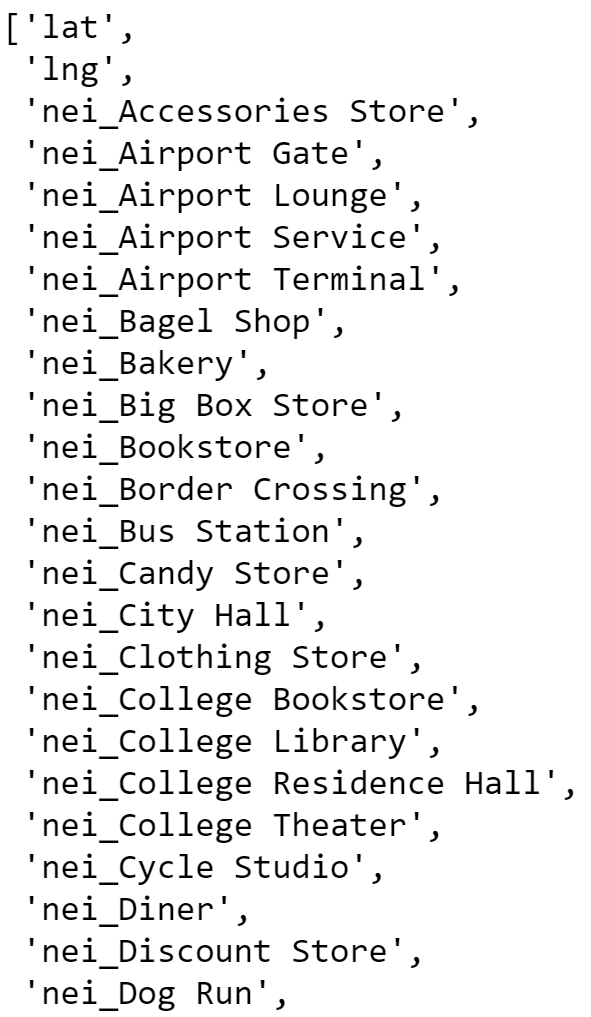
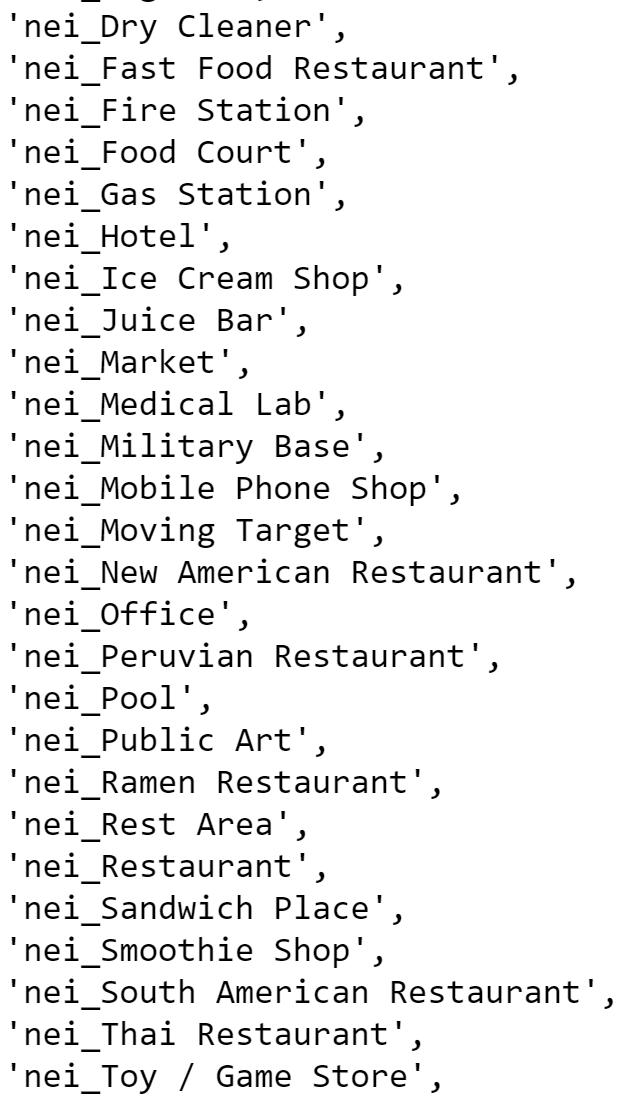
* + Immediate neighbor category types with correlation >= 0.1
  + Venue attributes with correlation >= 0.1
  + Distance to nearest Starbucks
* Results:



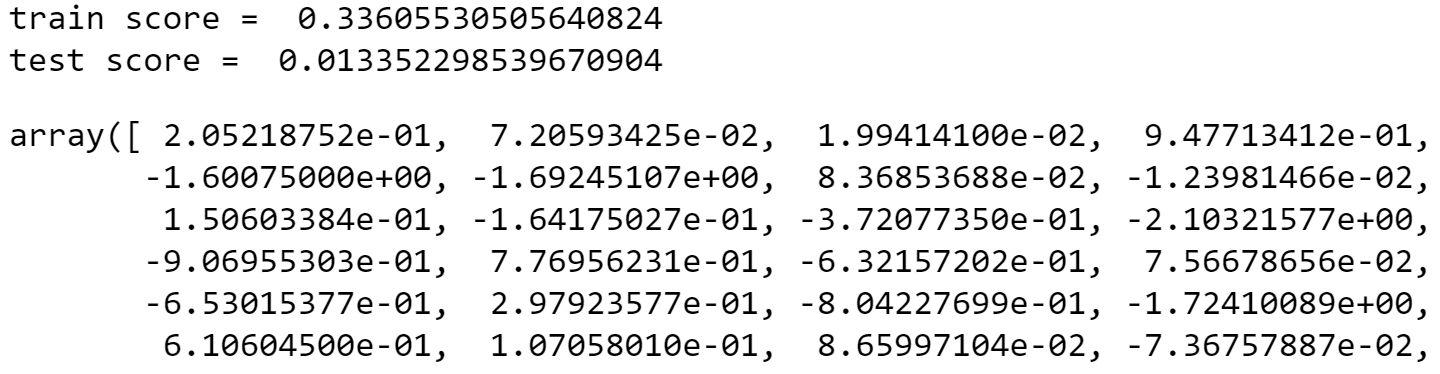
* + R-square over the train dataset (0.17) is much lower vs. model 1 (0.55) indicating that the model is not overfitting the train dataset as much
  + R-square over the test dataset (0.057) is much higher vs. model 1 (-2 x 10~16), however it still has minimal predictive power.
  + The coefficients are much more reasonable in magnitude (all but one < 1) and have a much more realistic impact on Foursquare rating.

Model 3: (Features with correlation > 0.05)

* Input features:

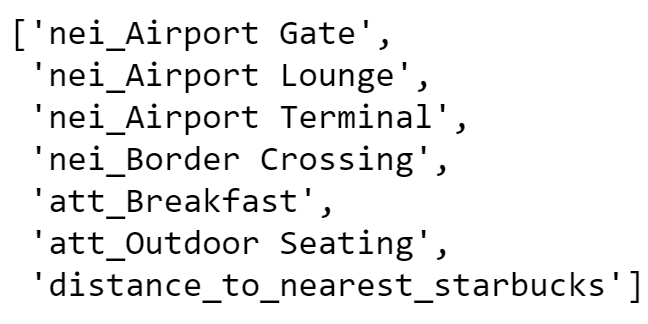
* + Immediate neighbor category types with correlation >= 0.05
  + Venue attributes with correlation >= 0.05
  + Distance to nearest Starbucks
  + Latitute/longitude
* Results:

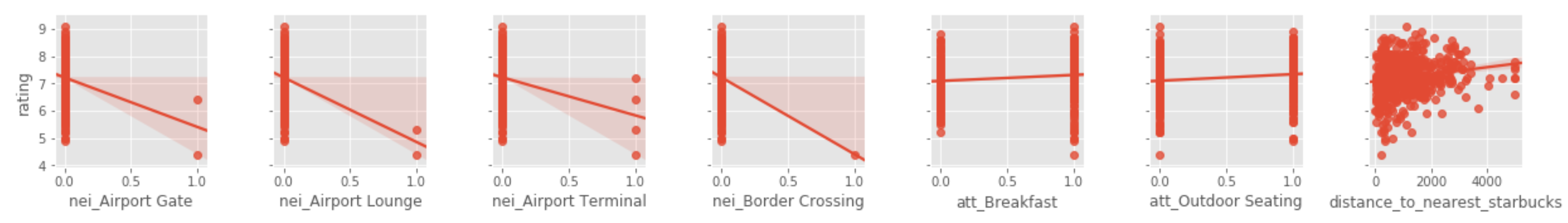


* + R-square over the train dataset (0.33) is higher vs. model 2 (0.17) indicating that the model is overfitting the train dataset a bit more
  + R-square over the test dataset (0.01) is almost zero indicating poor predictive power
  + The coefficients are still reasonable in magnitude

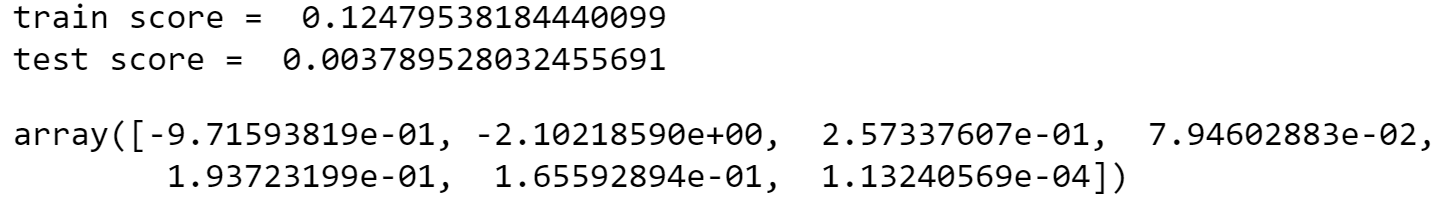
Model 4: (Features with correlation > 0.15)

* Input features:





* + Immediate neighbor category types with correlation >= 0.15
  + Venue attributes with correlation >= 0.15
  + Distance to nearest Starbucks
* Results:



* + R-square over the train dataset (0.12) is not as high as model 2 (0.17)
  + R-square over the test dataset (0.003) is zero indicating that the proposed features are not predictive for Foursquare rating and additional model features would be helpful

Model summary:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Input features** | **R-squared over TRAIN** | **R-squared over TEST** |
| 1 | All features | 0.550 | -2.35 x 1016 |
| 2 | Features with correlation > 0.1 | 0.171 | 0.057 |
| 3 | Features with correlation > 0.05 | 0.336 | 0.013 |
| 4 | Features with correlation > 0.15 | 0.125 | 0.004 |

1. Discussion
   1. Observations

The features explored in this report (store attributes, neighbor venue categories, and distance to nearest Starbucks) turned out to be poor out-of-sample predictors for Foursquare ratings at LA Starbucks. Linear correlation was not very high with respect to Foursquare rating for any of the features individually (all < 0.22), and when tested in combination using linear regression, the R-squared values over the test dataset were all < 0.06 showing weak prediction capability.

* 1. Recommendations

For Starbucks owners:

* Consider adding the following services based on high positive correlation with Foursquare rating:
  + Outdoor seating
  + Breakfast
  + Wi-fi
  + Delivery
  + Private parking lot
* Remove the following attributes due to strong negative correlation with Foursquare rating:
  + ATMs
  + TVs

For a potential new Starbucks owner in LA:

* When choosing a location, avoid locating near transportation sites as they tend to have strong negative correlation with Foursquare rating. Customers near these locations may be more sensitive to service times due to scheduled departures.
  + Airports (terminals, gates, lounges, and border crossings)
  + Bus stations
  + Train stations
* Also avoid locating on college campuses near the following location types as they tend to have lower Foursquare ratings:
  + College theater, library, bookstore, residence halls
  + University

1. Conclusion

A high Foursquare rating is likely to increase revenue at a Starbucks location, but the factors that drive a high or low rating go beyond those considered in this analysis. Some general recommendations can be made based on venue attributes, neighbor store types, and distance to the nearest Starbucks, but none of them are reliably predictive on out-of-sample locations to make strong conclusions.

* 1. Further analysis recommendations

Consider the following list of additional store attributes to find features that may be more predictive of Foursquare rating. Unfortunately, these features may not be easily available in public data sources because they are related to individual store management:

* Store cleanliness
* Staff friendliness
* Average wait times
  + Store staffing levels
  + Customer volume

This analysis could also be extended by analyzing tips and user comments to find common attributes for stores that are rated very high or very low.