2020-02-25

Christopher Gibbons

IBM Data Science Capstone Project

Coursera

**Introduction**

This is the report for the IBM Data Science Professional Certificate capstone project. Thus far we have completed eight courses covering the fundamentals of data science, a range of core algorithms (such as K means, support vector machines, linear and logistic regression), and to how to use Python and SQL to visualize and solve specific problems. In this assignment we have been tasked with using Foursquare data to solve a real-word problem of our choosing.

We have all seen countless failed restaurants, coffee shops, and the like in our travels. Most can probably name a few that are likely to fall away in the coming years. While there are many factors that go into a successful business, and restaurants are no different, without an ideal location a restaurant will not be successful. If there are too many well-established competitors a new coffeeshop will fail to draw business and will not stand a chance. In this project we identify areas around New York City that are deficient in coffee (and/or other establishments if need be) to find low competition areas a coffeeshop might thrive in.

The target audience is those who wish to identify a good location to start a coffeeshop business.

**Data**

As mandated by the guidelines given for the assignment, Foursquare was the primary source of data. It was straightforward to pull zip code-associated venue category data from foursquare, as shown in the following image.

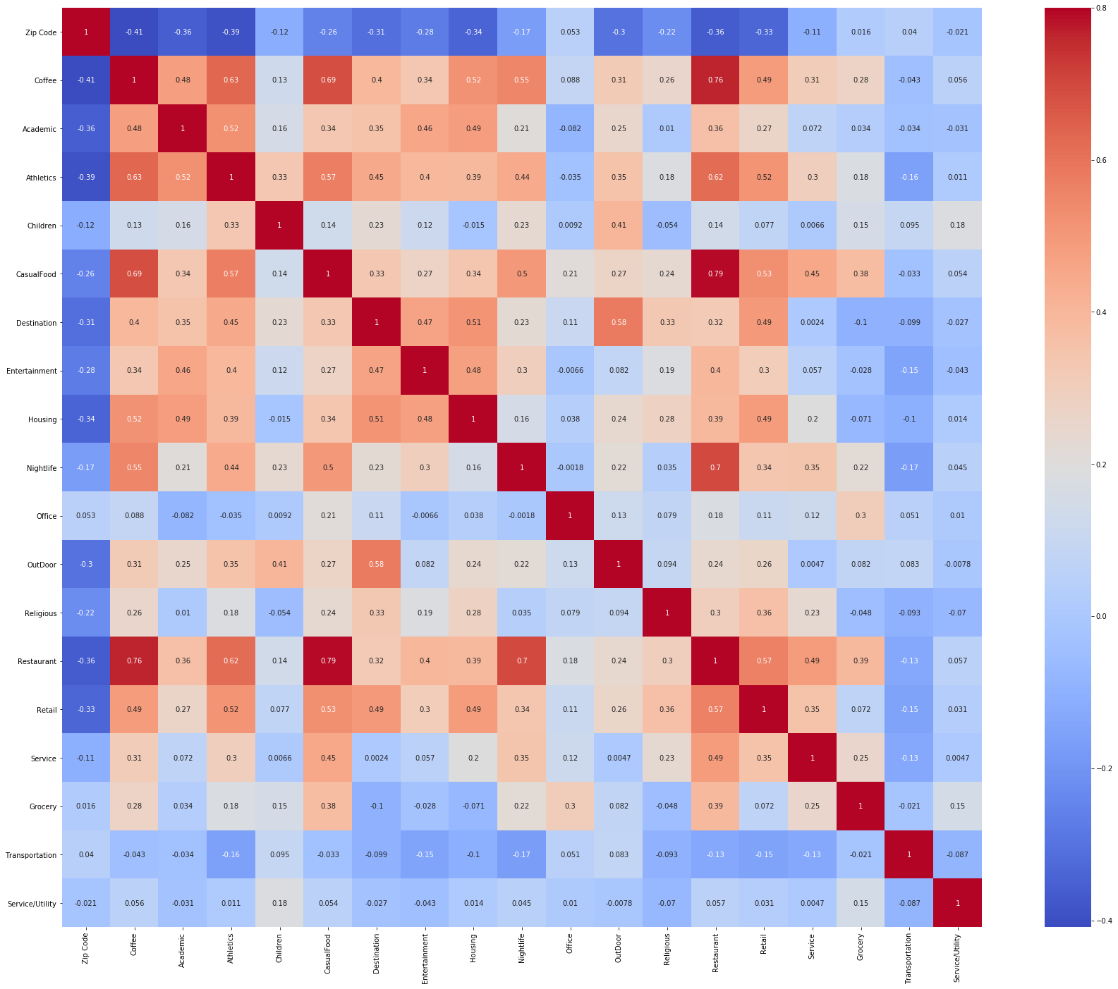


*Fig. 1. An example of data pulled from Foursquare.*

**Methodology**

First, establishment types were grouped based on a very limited number of types of institution. This was done to improve the significance of each column, as many of the raw categories were empty for the vast majority of locations in some cases. The neighborhoods were grouped using K-Means, again, to improve the amount of data under each category.

After this was done, a correlation heatmap was used to confirm that there exists a correlation between the category variables and coffee shot frequency.

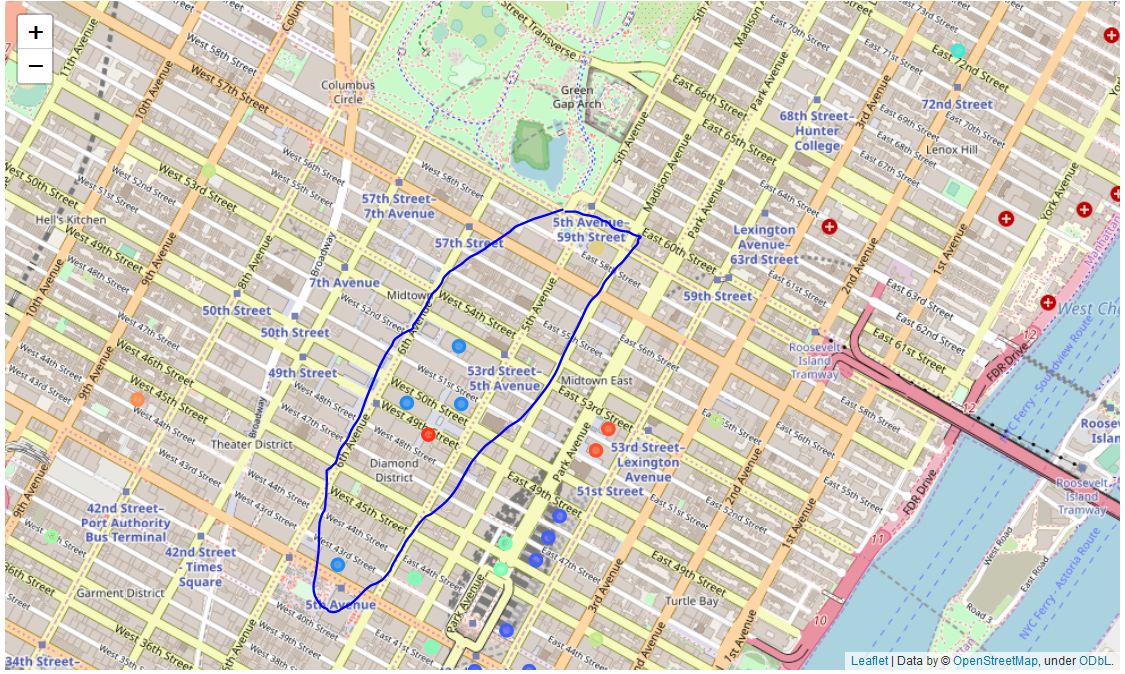


*Fig. 2. The correlation heatmap showing degree of correlation in the grouped data*

A range of regression models was attempted. Ridge Regression with cross validation was found to have the best results (R2 of 0.7-0.8 based on a range of cross validation outcomes). Built-in stats reporting along with cross-validation was used to evaluate models. A prediction of coffeeshop frequency was made using this model, and the results were sorted based on the difference between predicted and actual frequency resulting in a list of potential coffee shop locations that may be best from a standpoint of minimizing competition.

**Results**

Initially, it was hoped to use Atlanta, the author’s home city, as a choice of location. However, the Foursquare data for Atlanta was found to be too sparse to be useful. The data for New York was, as anticipated, much richer and perfectly suitable for the task after grouping.



*Fig. 3. The three most coffee deficient clusters, circled.*

The three most coffee-deficient zones were found to be clustered together in and around the Diamond District. The area looks walkable at a glance and is well positioned by Central Park. There does not appear to be a good reason for there to be few coffee shops in the area. So my recommendation would be to look at the diamond district in Manhattan as a possible site for a new shop.

**Discussion**

The first limitation that comes to mind with this model is the relative scarcity of data. Ideally, the model would be improved by including several cities before grouping by K-Means. Each variable, particularly coffeeshops per cluster, should have a significant number of cases. In some clusters the number of coffeeshops is zero. This is clearly not ideal.

Other factors, such as walkability and property values are important considerations and could be added to a follow-up project.

**Conclusion**

The data is does not appear to be ideal for a linear regression analysis. Nevertheless, A fit with an R2 of 0.7-0.8 (based on a range of cross validation outcomes) was made and found a group of clusters in Manhattan was found which appear to potentially have significantly less competition for an upstart coffee shop. This is too little to justify a business decision on its own but may represent a solid place to start from in the search for locations for a new business.