# Introduction

This report presents the final, culminating project for the Data Science specialization, created in collaboration with IBM. The project requires use of Foursquare location data retrieved through their API but is otherwise left up to the student to come up with a problem to solve and an analysis methodology.

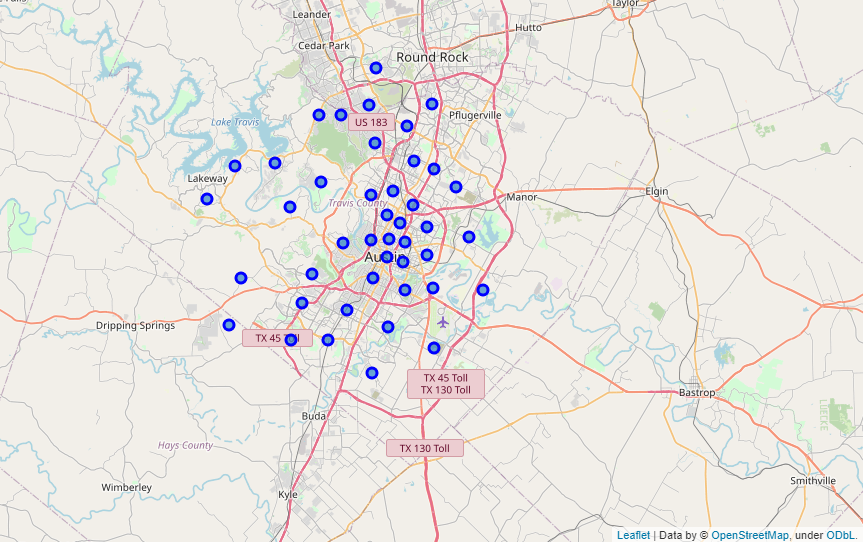
The goal of this work is to find the optimal location for opening a new Barbeque restaurant in Austin, TX. This represents a unique challenge as Austin is widely considered a hotspot for Barbeque, and being a relatively small city, is more than likely very saturated already.

This approach should benefit anyone looking to open a specific type of restaurant or store in a location that might minimize their competition and maximize their potential business.

# Data

To carry out this project, we will need to start with the locations of existing Barbeque restaurants in the city of Austin. We will retrieve this information by making calls to Foursquare’s “venues” API, querying “near Austin, TX”. A variety of data points are returned for each venue; however, we will focus on Venue Name, Latitude, Longitude, and Zip Code.

In analyzing this data, we will compare venues at the Zip Code level with other data points at the Zip Code level, such as Population, Population Density (per sq mile), and Average Income (per person). This data was retrieved by web scraping ZipAtlas.com using the BeautifulSoup Python package.



*Zip Codes for Austin, TX*

# Methodology

The primary assumption for determining the optimal location for a new restaurant is that the number of restaurants in a zip code correlates with either population, population density, or average income, OR a combination of those. To test for correlations, start with visual inspection of each versus venue count, where venue count is the dependent variable.

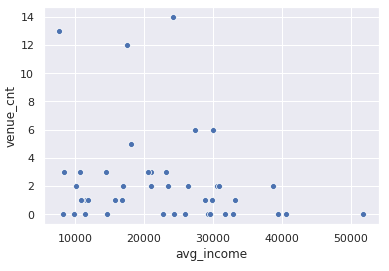
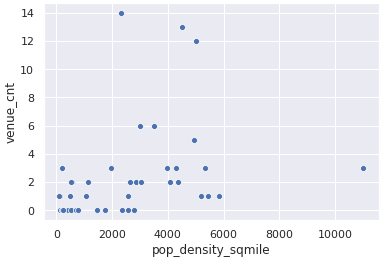
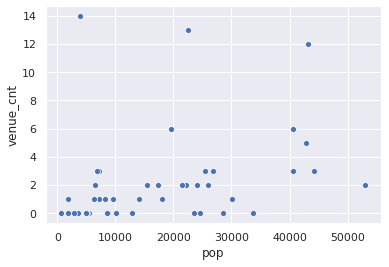
If visual inspection leads to a likely linear or multi-linear relationship, linear regression can be performed using the Scikit-Learn package. A regression model COULD predicate an estimated viable venue count for zip codes without venues, based on similar population, average income, etc. However, the nature of the analysis suggests that a clustering approach might be a better indicator of optimal location. Again, optimal location defined as the zip code without any current venues that would be most viable for a new location based on socioeconomic factors.

Given that this problem is geographic in nature, this is a perfect opportunity to cluster the existing venues using DBSCAN. Venues will be clustered using DBSCAN against latitude and longitude as well as a combination of latitude/longitude and population density. Both clustering methods will be plotted on a map of Austin.

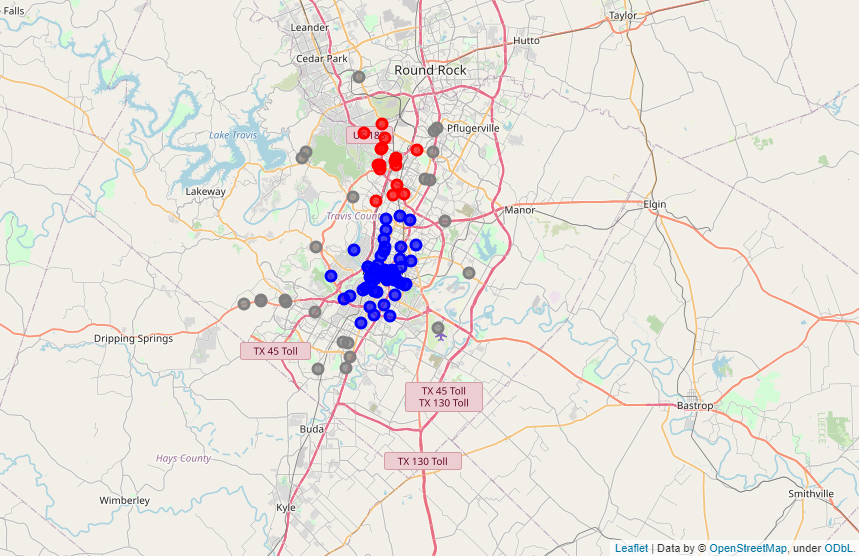
# Results

Austin, TX falls into three separate counties and has 74 separate zip codes (though some of these are strictly for PO boxes). With 98 total Barbeque restaurants in the Foursquare dataset, this averages out to about 1-2 restaurants per zip code and suggests that a zip code level analysis could be appropriate. This also validates that the methodology in determining location by finding zip codes without an existing restaurant. However, all 98 restaurants fell into only 28 of the zip codes, and of those, 39 of the restaurants fell into only three counties. This suggests a high level of clustering, and mapped, shows most of them in the downtown area.

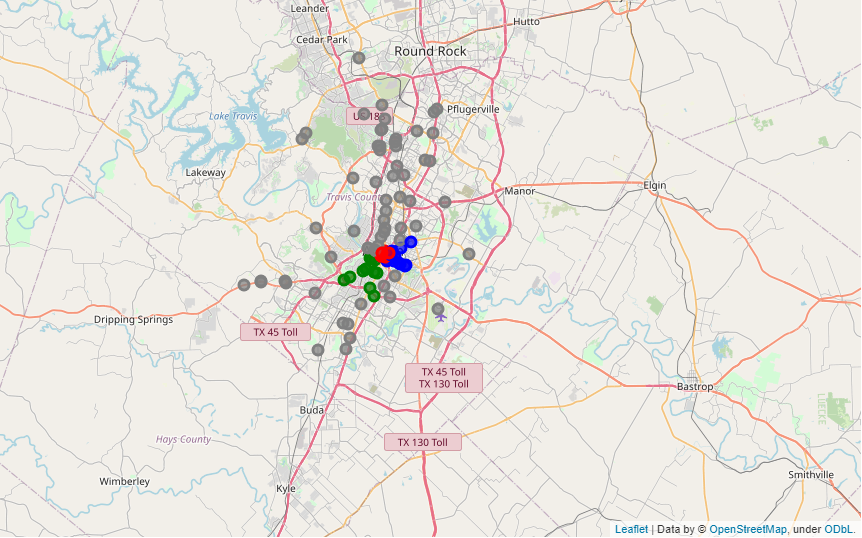
When plotting venue counts against population, population density, and average income, there was no discernable relationship, linear or nonlinear. This suggests that location has much more to do with geography alone than with any of the sampled socioeconomic factors. See figures below.



Despite having no gain insight from correlations in the data, further analysis of the existing restaurants can still be done. DBSCAN, or Density-Based Spatial Clustering of Applications with Noise, which is an unsupervised machine learning algorithm, was used to cluster the restaurants based on both their location and their location combined with population density of their zip code. See below for clusters plotted on a map of the city.



*DBSCAN clustering using latitude and longitude*



*DBSCAN clustering using latitude, longitude, and population density*

# Discussion

Unfortunately, the data gathered does not provide enough insight alone to help find an optimal location. Since most the existing restaurants are clustered downtown, population in their respective zip codes likely does not represent as significant a portion of their customer base as a more rural restaurant might (people might be more willing to drive to a bigger restaurant downtown, commuters may eat there during the work week, etc.).

Since there are only a handful of zip codes that do not have a restaurant, to continue the analysis, we could merely narrow to those areas and further examine them to find a suitable location.

Given more time for completion, I would further explore how to enrich the dataset. This could be new data points or more granularity in population, population density, etc.

# Conclusion

Despite the lack of clear insight from the compiled dataset, the process of gathering, cleaning, and exploring the dataset with still a fun and worthwhile experience. The full experience provided great practice in the art of using python for data analytics and data science. Solving a problem without a known end goal provides a unique experience from structured labs because it forces the learner to overcome challenges throughout the course of the project. Gathering data from multiple sources via web APIs and web scraping, combining those sources together, and cleaning the data can have a daunting array of challenges come up, and solving those provides a wealth of applicable experience.

To anyone that has read this far, thank you for your time!