**IBM Capstone Project**

**Introduction**

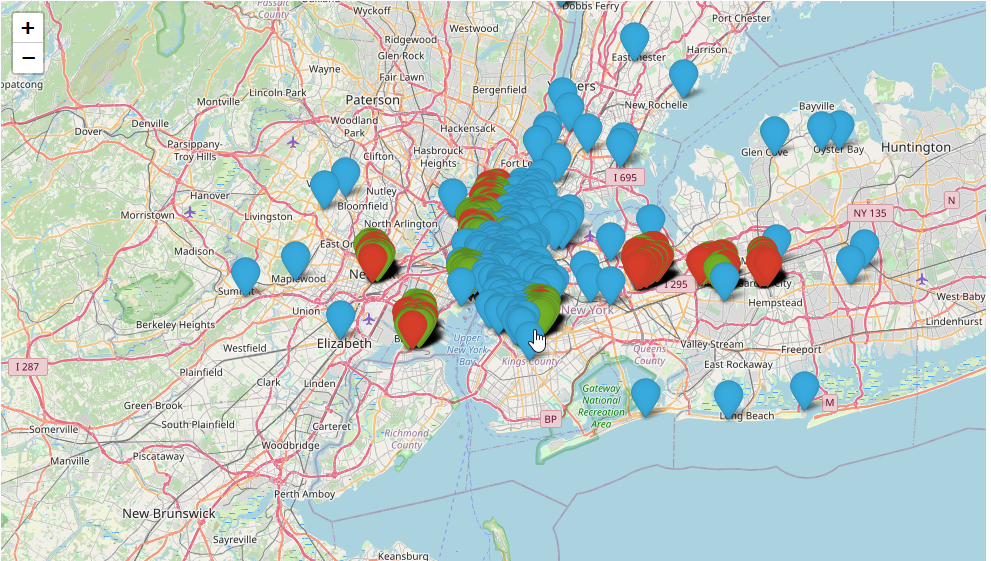
Most of us in the business world have heard the phrase at some point ‘location, location, location’. Considering the potential location of a business is one of the more important parts of starting one. After all, almost any firm based in Times Square is going to do better overall than if that same firm were to be placed in a corn field in South Dakota. Although no one questions the importance of location, there are other factors that go into whether a business succeeds or fails. If a firm were to neglect say, customer service in the name of a high value location, then cultivating a core clientele would be a challenge regardless of how many people pass by your front door. Would it not be handy if there was a way to quantify exactly how important location is? That’s what this capstone aims to find out. This project is aimed to interest any business owner that is either opening a new location, or even just starting out.

**The Data**

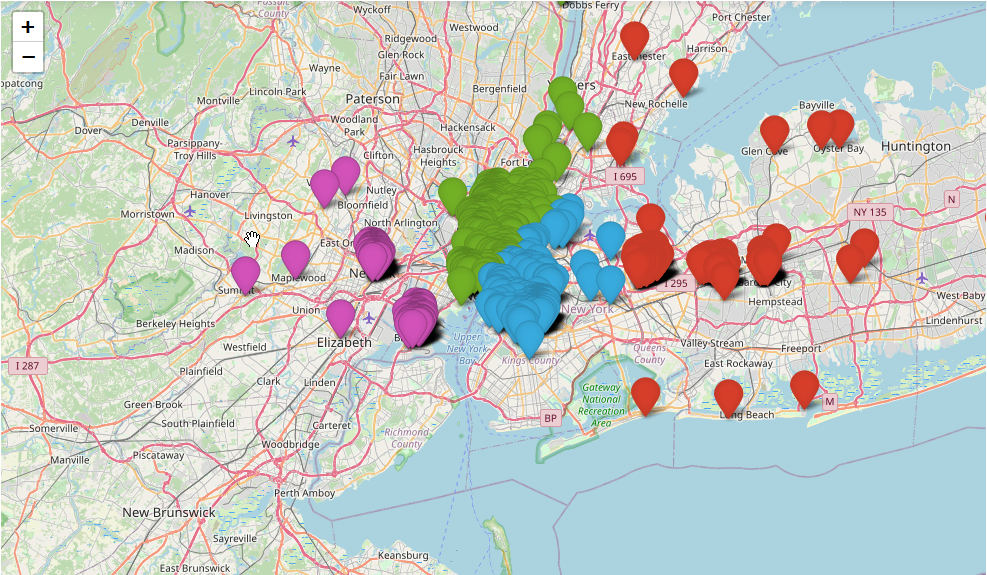
Since it is not possible for an individual such as myself to know the intimate financial details of various businesses in order to deem said business a success or failure, a creative way to determine success of a business must be found. It just so happens that Foursquare’s API has a attribute in their json that returns a rating of the venue. Using this field, it is possible to determine whether or not a business has a strong core clientele or not, so this is the metric that will be used for the success of a business. Because of its density and diversity of venues, New York City and some of its surrounding areas will be the subject of the explore query. All of the venues that result from the initial query will require an additional query to get their respective ratings, since that information is not available in the explore query’s results. Then all this information will be placed into a SQL database. Although the data will essentially be pre-cleaned using this method the downsides are the query limits in Foursquare’s API, so the data will have to be gathered over the course of a multiple days as the limits refresh (thus the reason a SQL database is required as opposed to importing the data directly into a pandas dataframe).

**Methodology**

Out of a little over 1000 venues gathered from New York and some surrounding areas, the average rating is almost exactly 8. Based on some of the other information provided in the describe output of the dataframe, it is possible designate a venue 'average', 'below average', and 'above average' based on where most of the ratings end up being. A look at the 25% mark of the data reveals that the first quarter of the venues are at or below 7.3. This line will be used to delineate the 'below average' label. At the 75% mark, the last quarter of the data has a rating of 9 or above. This rating will dlineate the 'above average' label. On the map shown below, 'above average' venues will have a blue marker, 'average' venues will have a green marker, and the 'below average' venues will have a red marker.



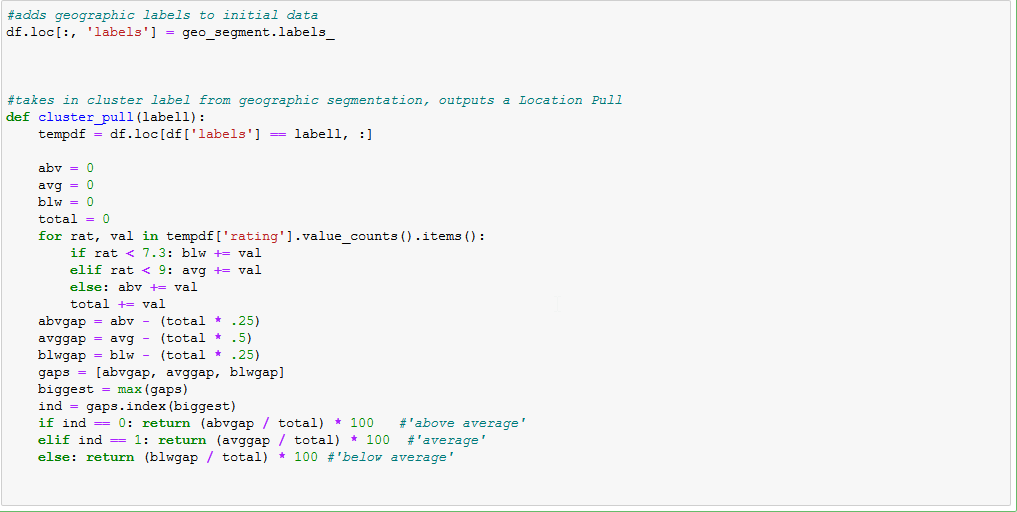
To the surprise of no one, the map indicates a correlation between location and rating. This however was not the objective of this exercise. The challenge is to find a way to quantify the strength of location's effect on a business's success (which we are using rating to measure). Looking at the map through exclusively a geographical lens, it is possible to see roughly three clusters. However, the large mass in the center should be split two ways since it is so large in comparison to the other two clusters. When the venues are segmented using a k-means algorithm based of their coordinates alone, the resulting map looks like this:



Now, in an imaginary world where location is not at all a factor in determining the success of a business, the distribution of ratings would be even in a given cluster. There would be a similar numbers of highly rated businesses and poorly rated businesses. Recalling back to when the venues were labeled 'average', 'below average', and 'above average', it would be reasonable to expect in this imaginary world that 50% of the venues would be average, and the other two categories would each have a 25% presence in given cluster. The strength of the effect of location on a venue's rating can be measured by comparing a cluster's true collective ratings against this imaginary world. An example with nicer numbers would make understanding what will done to measure location's strength a ton easier. Say there are two clusters with 100 venues each. Cluster One has 50 venues that are 'above average', 25 are 'below average', and 25 are 'average'. Cluster Two has 70 venues that are 'below average', 25 venues that are 'average', and 5 that are 'above average'. Since it has been established that location does pull a venue's rating either up or down, it can be assumed that a particular location has an inherent 'above average', 'average', or 'below average' quality to it. This quality can be ascertained by looking at which quality has the largest gap between its actual value and expected value. In Cluster One's case, it would have a 'above average' location, and Cluster Two would have a 'below average' location. The difference between Cluster One's amount of 'above average' venues and its expected value from the imaginary world (25) is 25. This means that the location effectively doubled the amount of venues considered 'above average'. If recorded as a multiplier to the expected number of 'above average' venues, this would be 100% (from now on this number will be referred to as Location Pull). Performing the same operation on Cluster Two yields 180%, indicating a very strong pull. To get a final representative number of the strength of location in general, the two Location Pulls are averaged together to get 140%. Now that the process is hopefully understood, it is time to apply it to the real data.

**Results**

To get the final Location Pull values, all that needs to be done is to define a function to classify the quality of clusters, and then calculate the Location Pull based on that label. Here is what the function looks like:



Now in order to get the final value for the strength of location in general, all the values must be averaged together. Here is the final result:

**15.87% Location Pull**

**Discussion**

Now before getting too excited about this number, there is one limitation to this project that must be discussed, and that is the data itself. Casting a widespread yet diverse net proved difficult during the course of this project, due to the fact that the Foursquare API will present the best venues first, which makes sense from their perspective, but resulted in inaccurate data. In order to get a more accurate representation, more specific queries had to be made to the API so that there was no choice but to present ALL venues, no matter how poorly they were rated. This process consumed the query limit rather quickly, making data gathering more tedious than it truly needed to be. However, in a real business situation, it is more likely that a premium account would be made available for use, lifting those query limits and freeing up precious time to be more systematic and therefore accurate. Does all of this mean that the work done here is useless? Not exactly.

**Conclusion**

The mission of this project was to get a definitive number that can value the strength of location on a business's success. A trustable definitive number was not reached. HOWEVER, like many instances throughout history, something else even better may have been unearthed by accident, and is good reason to not relegate this project to the recycle bin. While the result is untrustworthy, process still is. If approached by a client with a specific location that he/she is uncertain about, the capability to tell them both the quality of the location and the magnitude of its effect on rating is absolutely there once all of the surrounding venues in the location are included in a database of its own. This is far more useful than the initial goal because the process used here is immediately applicable to the real world. In summary, this project looked for copper and potentially found gold.