Yelp

# Yelp Analysis using Watson IBM

*Paige Young, Tobiah Rothlingshofer, Neeharika Kaja, Raajita Koneru*

December 2nd, 2016

# Introduction

Our IBM Watson Presentation involves analyzing data from Yelp. Yelp is known to be a great source of information for customers, but it can also be extremely helpful for restaurant owners. It is a way to analyze real customer preferences and adapt to them.

For this project, we wanted to help restaurant owners locate prime areas for specific types of cuisines. We started our research with a simple question: **How can we use yelp data to help potential restaurant owner identify ideal locations and key drivers for success?**

Our project will contribute by giving more information to people looking to open successful restaurants, specifically in the Athens areas. This project has also taught us all more about IBM Watson and dealing with obstacles. We constantly ran into issues while utilizing Watson, such as the inability to incorporate latitude and longitude data as planned or teaching Watson to understand the relevance between key variables Review Count and Rating, but we were able to find creative ways to circumvent the errors or to fix them. In performing our analysis we hope to derive valuable data for restaurant owners in Athens and incite strategic decisions.

# Methodology

| Key Variables | |
|---|---|
| **Latitude** | A measure that takes into account a location’s angular presence between East and West parts the earth’s equator. Expressed in degrees. |
| **Longitude** | A measure that takes into account a location’s angular presence between North and South parts the earth’s equator. Expressed in degrees. |
| **Distance from UGA** | Minimum distance between each restaurant and UGA.<br>**Formula is as follows:**<br>=ACOS(COS(RADIANS(90-latitude1)) *COS(RADIANS(90- latitude2)) +SIN(RADIANS(90- latitude1)) *SIN(RADIANS(90- latitude2)) *COS(RADIANS(longitude1- longitude2))) *3958.756 |
| **Distance from Shopping Center** | Minimum distance between each restaurant and closest shopping center. The distance was obtained by taking the minimum distance between each restaurant and 16 shopping centers as outlined below.<br>Georgia Square Mall<br>Athens Plaza Shopping Center<br>Homewood Shopping Center<br>Athens West Shopping Center<br>Epps Village Shopping Center<br>Clarke Crossing Shopping Center<br>College Station Shopping Center<br>Athens Promenade<br>Public Shopping Center<br>Colonial Promenade Beechwood<br>Downtown Athens<br>Walmart Lexington<br>Walmart Epps Bridge<br>Five Points<br>Prince Avenue<br>Beechwood<br>**Formula is as follows:**<br>=ACOS(COS(RADIANS(90-latitude1)) *COS(RADIANS(90- latitude2)) +SIN(RADIANS(90- latitude1)) *SIN(RADIANS(90- latitude2)) *COS(RADIANS(longitude1- longitude2))) *3958.756 |
| **Distance from Apartments** | Minimum distance between each restaurant and closest apartment complex. The distance was obtained by taking the minimum distance between each restaurant and 16 apartment locations as outlined below.<br>Building 1516<br>High Ridge<br>Georgia Heights<br>Eclipse on Broad<br>Summit<br>Uncommon Athens<br>Polo Club<br>Abbey Wes<br>The Reserve<br>Standard<br>Tri Delta House<br>Lakeside<br>The Lodge<br>Whistlebury<br>**Formula is as follows:**<br>=ACOS(COS(RADIANS(90-latitude1)) *COS(RADIANS(90- latitude2)) +SIN(RADIANS(90- latitude1)) *SIN(RADIANS(90- latitude2)) *COS(RADIANS(longitude1- longitude2))) *3958.756 |

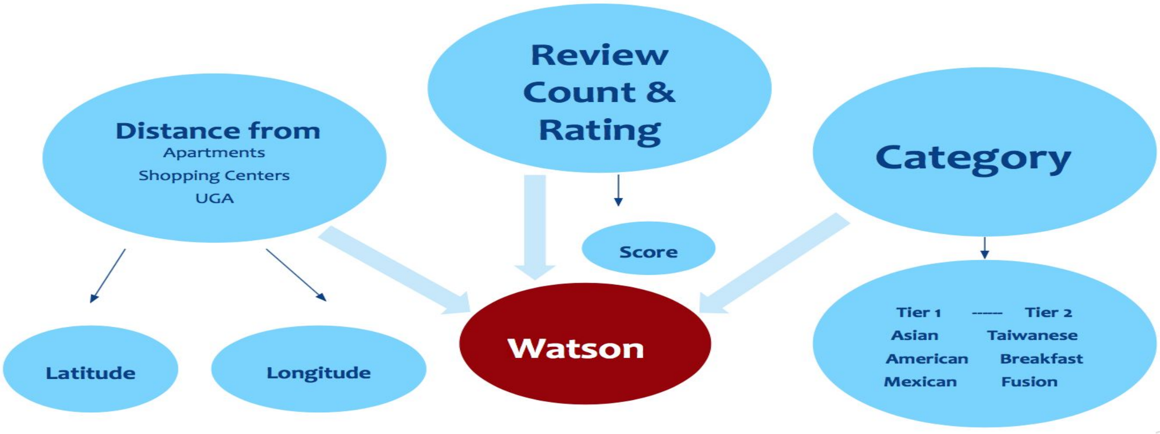
| | |
|---|---|
| **Review Count** | The number (count) of reviews for each particular restaurant provided by Yelp's API. Used to assess the success and significance of a restaurant location. |
| **Rating** | A score between 1-5 assessed by Yelp users detailing their overall satisfaction of a restaurant. Used to assess the success and significance of a restaurant location. |
| **Category** | The type, ethnicity, or categorical significance of a restaurant for results. Categories were split and factored into 2 tiers to allow drill-down analysis. |
| **Tier 1**<br>**Tier 2** | a. Ex. Asian, American, Mexican (ethnic based)<br>b. Ex. Taiwanese, Breakfast, Fusion (niche) |
| **Score** | A proprietary measure that combines and magnifies the positive or negative result of Review Count and Rating variables.<br><br>**Formula is as follows:**<br>If(Rating > 3) {<br>Score = Rating * Review Count<br>} else {<br>Score = (Rating-5) * Review Count<br>} |
| **Reverse Miles** | The inverse result of restaurant miles from key population centers (UGA, closest shopping center, closest apartment complex) that allows Watson to process close restaurant distances as a positive metric instead of negative metric.<br><br>**Formula is as follows:**<br>Max(Distance of Restaurant from Variable Being Analyzed) – Distance of Restaurant from Variable Being Analyzed |

## Description of Methodology

Our sample data consisted of 626 sample restaurants extracted from Yelp's API extraction tool database. Our goal was to utilize these samples to produce viable, cost-effective, and metric-based suggestions to help restaurant owners identify optimal restaurant locations in the Athens area. The Yelp API allowed us to extract 20 results at a time which resulted in numerous JSON files. These files were later cleaned and formatted into CSV files, which we fed into Watson to perform our analysis.

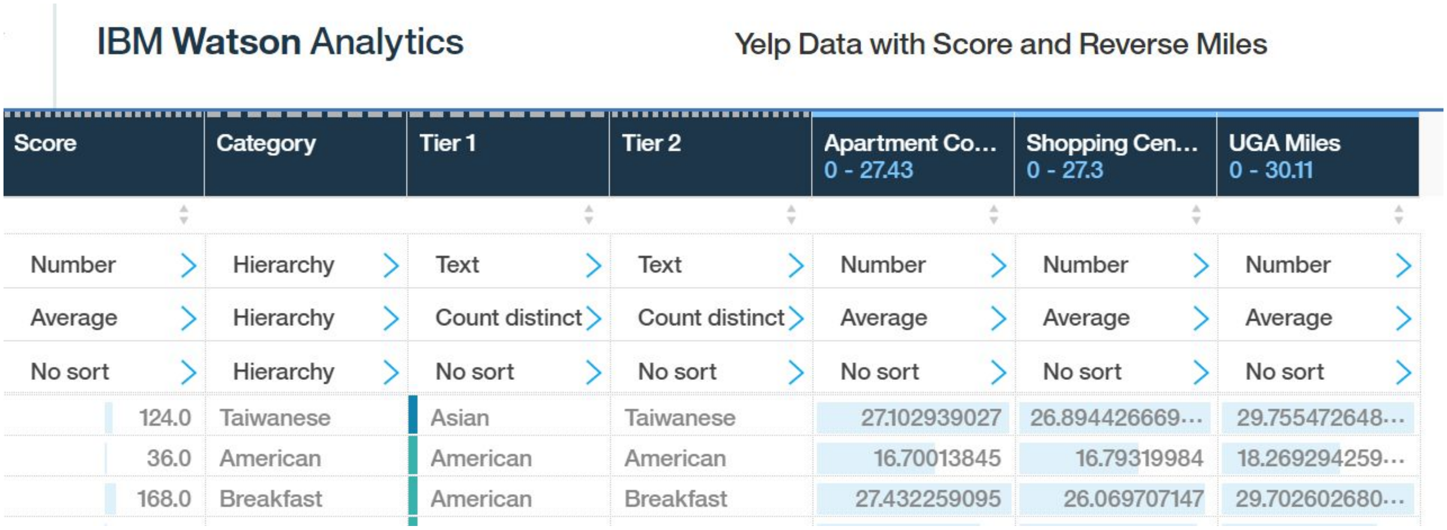
Key variables like Review Count, Rating, Latitude and Longitude, and Category were already provided to us through Yelp. However, we had to process this data further to make it permeable to Watson's interface. A few variables that we created in the process include: Distance of restaurant from UGA, Distance of restaurant from nearest shopping center and Distance of restaurant from nearest apartment complex. Creating these variables allowed us to process the data by radius and distance rather than coordinates, which was a format that Watson could understand. Our analysis consisted of various results that pointed to ideal locations for ethnic restaurants (close to UGA's campus and apartment complexes) and ideal locations for fast food chains (shopping centers).

Key Variables (figure a.)



# Findings

Our first question was is it possible to predict score? (Score is a value the combines both rating and number of reviews into one metric). To start we used a simplified subset of our data to allow Watson access to only data points that are relevant to predicting what drives the Score.

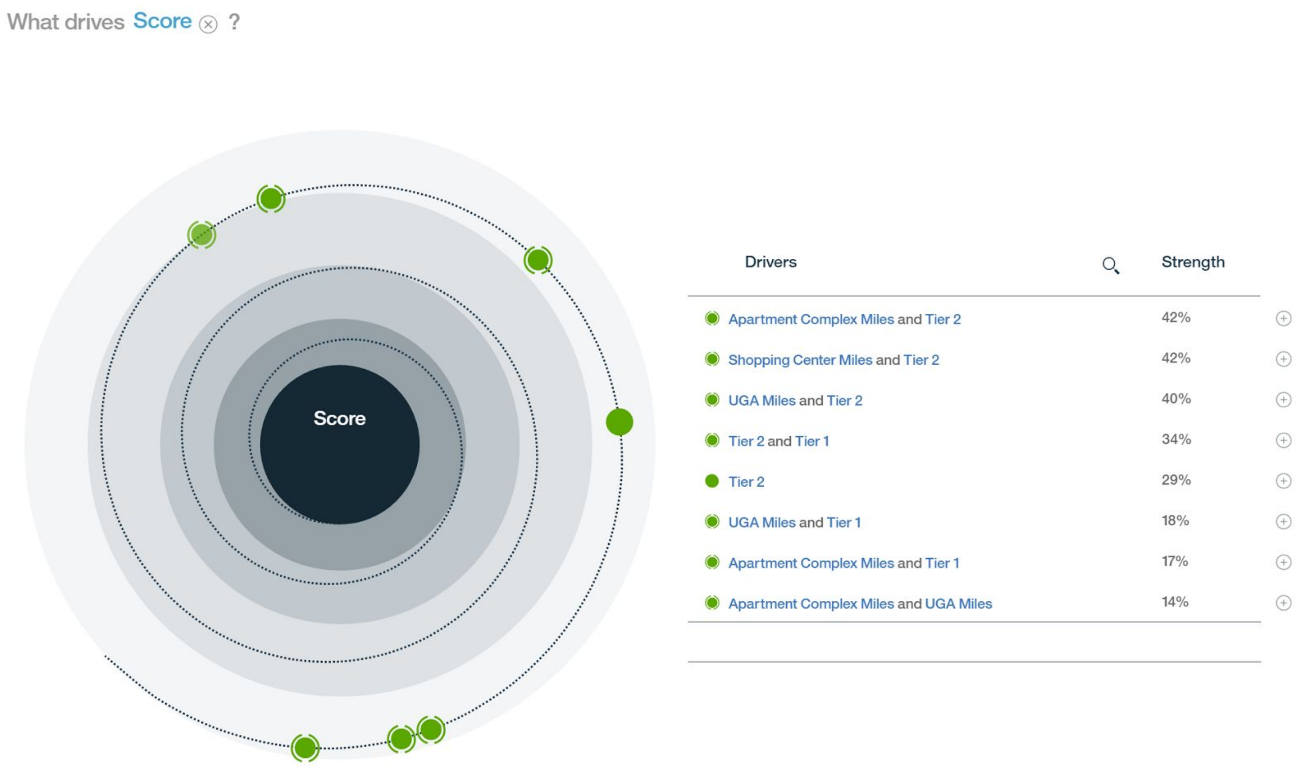


IBM Watson Analytics

Yelp Data with Score and Reverse Miles

| Score | Category | Tier 1 | Tier 2 | Apartment Co… 0 - 27.43 | Shopping Cen… 0 - 27.3 | UGA Miles 0 - 30.11 |
|---|---|---|---|---|---|---|
| Number | Hierarchy | Text | Text | Number | Number | Number |
| Average | Hierarchy | Count distinct | Count distinct | Average | Average | Average |
| No sort | Hierarchy | No sort | No sort | No sort | No sort | No sort |
| 124.0 | Taiwanese | Asian | Taiwanese | 27.102939027 | 26.894426669… | 29.755472648… |
| 36.0 | American | American | American | 16.70013845 | 16.79319984 | 18.269294259… |
| 168.0 | Breakfast | American | Breakfast | 27.432259095 | 26.069707147 | 29.702602680… |

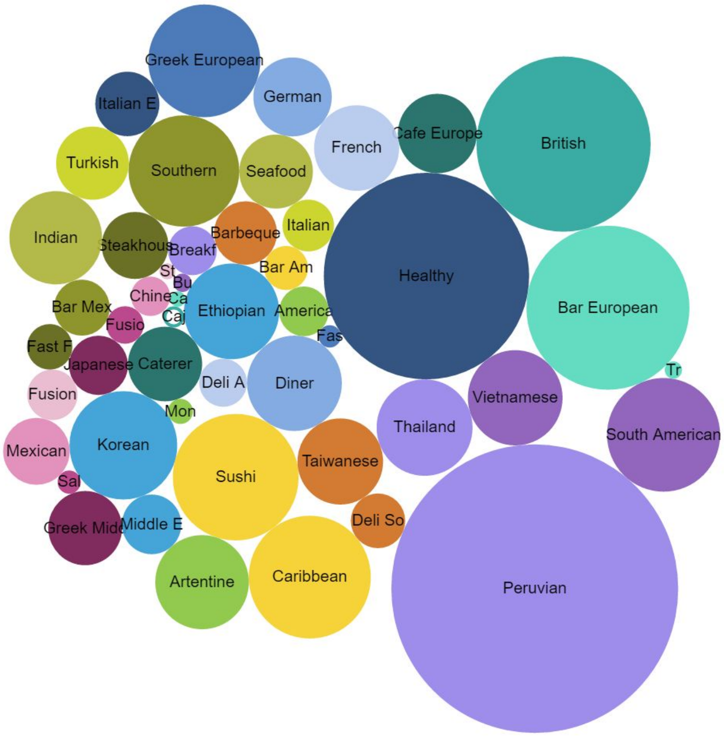
The chart above shows a sample view of the data that was used when Watson was asked to answer the question, “What drives Score?” We created a custom column within Watson labeled Category using the Hierarchy data type. This allowed us to specify that Tier 2 was related to Tier but that it fell under the Tier 1 umbrella category. In addition we changed the mileage columns to return an average as opposed to the default sum aggregation method. Also each distance column as calculated using the reverse milage formulas.



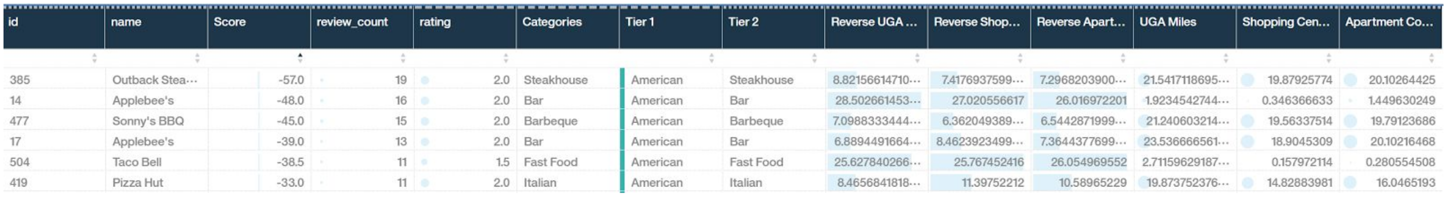
| Drivers | Strength |
|---|---|
| ● Apartment Complex Miles and Tier 2 | 42% |
| ● Shopping Center Miles and Tier 2 | 42% |
| ● UGA Miles and Tier 2 | 40% |
| ● Tier 2 and Tier 1 | 34% |
| ● Tier 2 | 29% |
| ● UGA Miles and Tier 1 | 18% |
| ● Apartment Complex Miles and Tier 1 | 17% |
| ● Apartment Complex Miles and UGA Miles | 14% |

The above visualization, shows that a restaurant's score is primarily driven by its distance from Apartment Complexes, Shopping Centers, and from UGA. And that its Tier 2 category is equally important.

Our next visualization, shown to the left, gives a representation of each Tier 2 categories average Score value within the dataset.

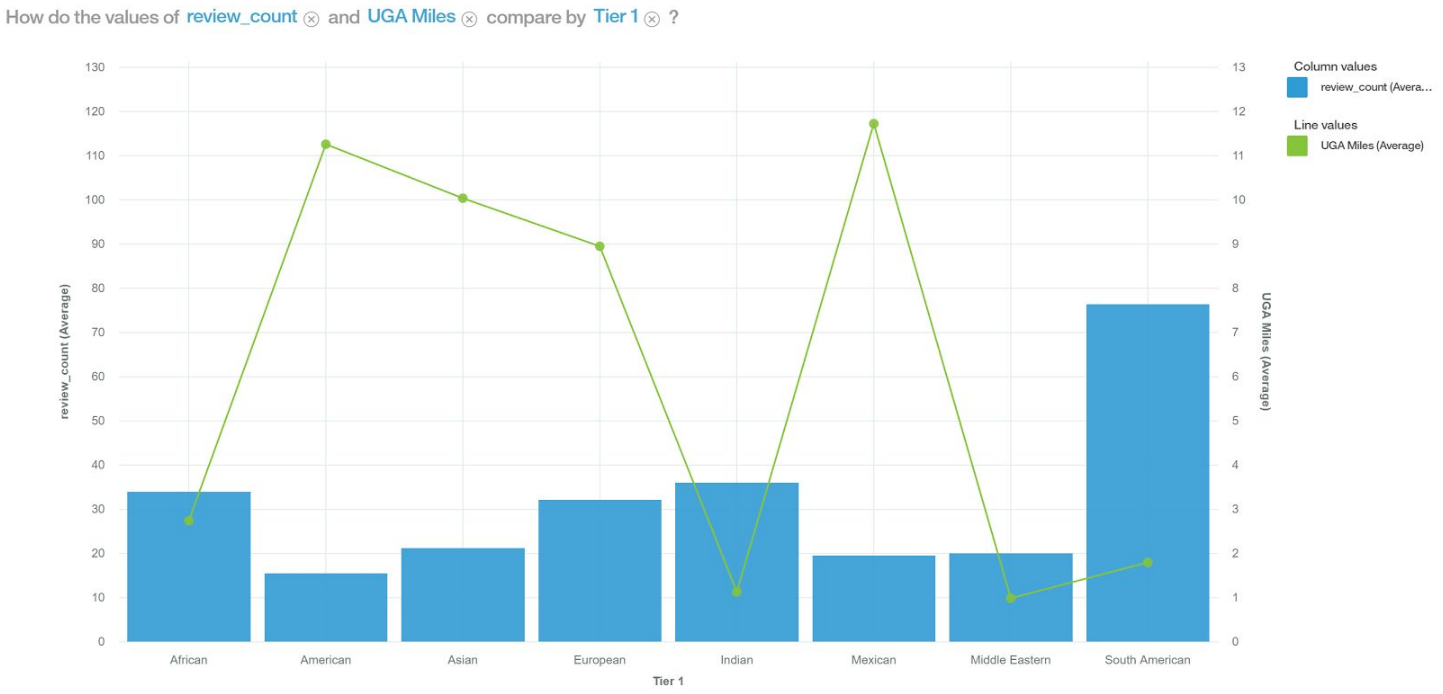


The data chart below shows a sample of the data that was used in the remainder of the visualizations in this report. It expands upon the previous dataset by including each restaurant's name, Yelp rating, number of reviews, and both reverse and regular milage. Each of the columns containing numbers except is were set to use average for aggregation.

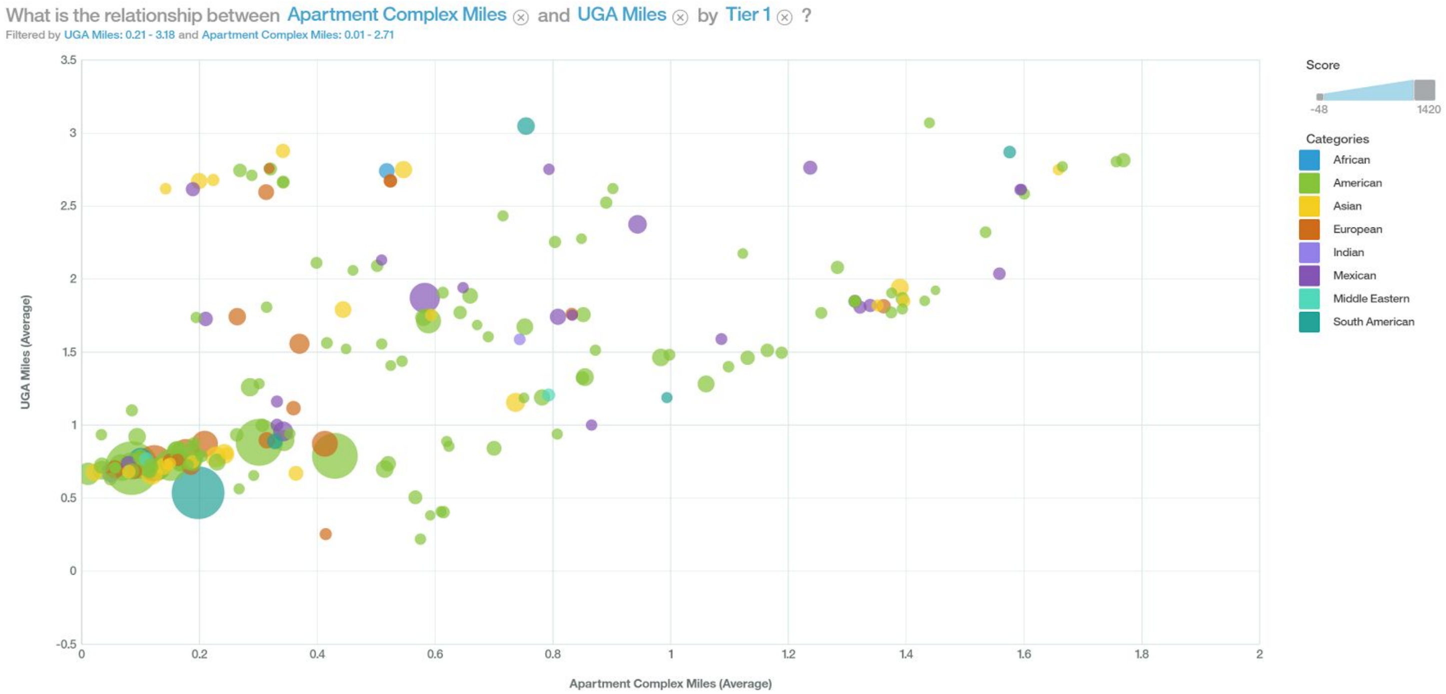


| id | name | Score | review_count | rating | Categories | Tier 1 | Tier 2 | Reverse UGA ... | Reverse Shop... | Reverse Apart... | UGA Miles | Shopping Cen... | Apartment Co... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 385 | Outback Stea... | -57.0 | 19 | 2.0 | Steakhouse | American | Steakhouse | 8.82156614710... | 7.4176937599... | 7.2968203900... | 21.5417118695... | 19.87925774 | 20.10264425 |
| 14 | Applebee's | -48.0 | 16 | 2.0 | Bar | American | Bar | 28.502661453... | 27.020556617 | 26.016972201 | 1.9234542744... | 0.346366633 | 1.449630249 |
| 477 | Sonny's BBQ | -45.0 | 15 | 2.0 | Barbeque | American | Barbeque | 7.0988333444... | 6.362049389... | 6.5442871999... | 21.240603214... | 19.56337514 | 19.79123686 |
| 17 | Applebee's | -39.0 | 13 | 2.0 | Bar | American | Bar | 6.8894491664... | 8.4623923499... | 7.3644377699... | 23.536666561... | 18.9045309 | 20.10216468 |
| 504 | Taco Bell | -38.5 | 11 | 1.5 | Fast Food | American | Fast Food | 25.627840266... | 25.767452416 | 26.054969552 | 2.71159629187... | 0.157972114 | 0.280554508 |
| 419 | Pizza Hut | -33.0 | 11 | 2.0 | Italian | American | Italian | 8.4656841818... | 11.39752212 | 10.58965229 | 19.873752376... | 14.82883981 | 16.0465193 |

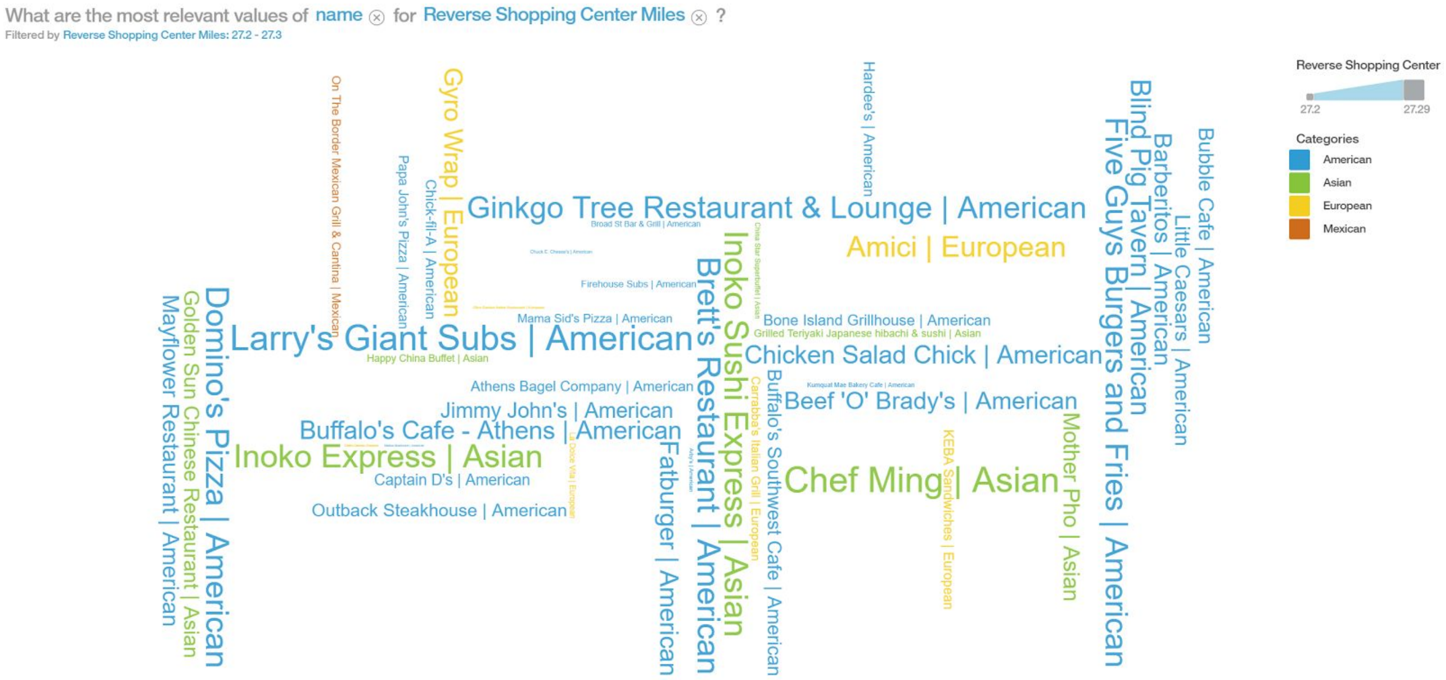
Our first visualization utilizing the expanded data set is shown below. The blue columns represent the average number of reviews per Tier 1 category. The green line shows each categories average distance from the UGA arch. From this graph it can be seen that food categories that are less well known among the general population, such as African, Indian and South American restaurants garner the most average review counts while they are the closest on average to the UGA campus.



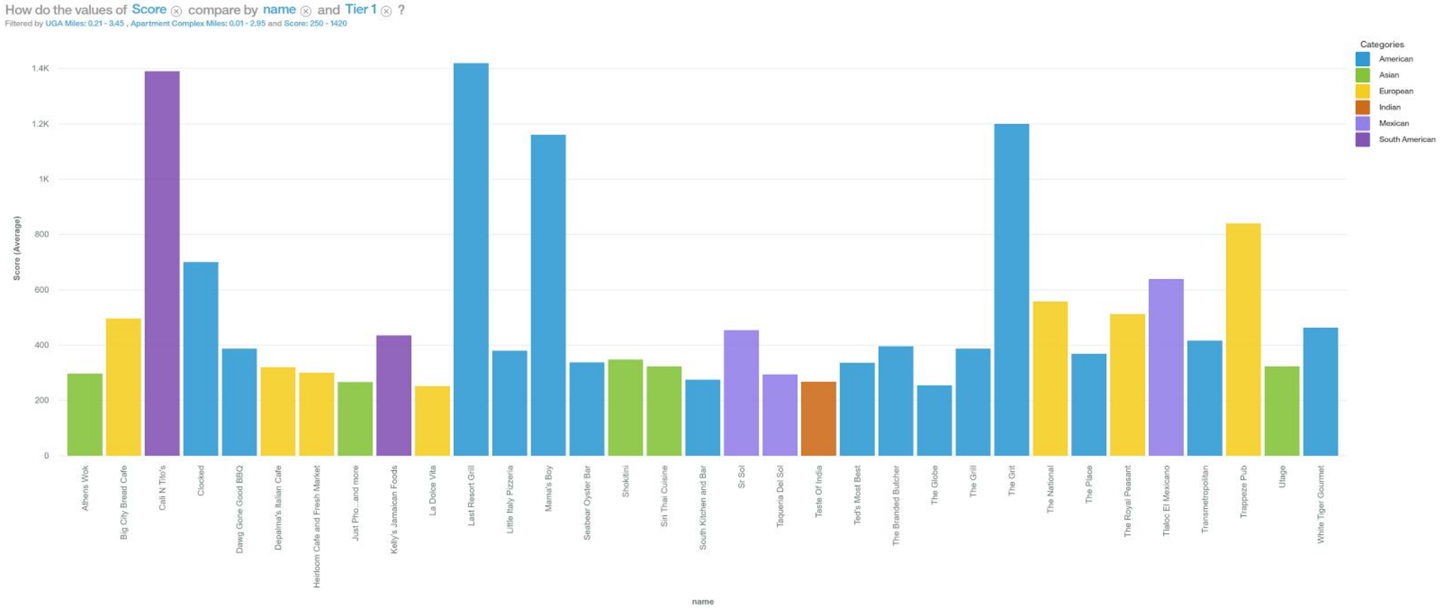
Identifying the relationship between how how many restaurants exist close to both the Apartment complexes within Athens and also UGA is shown in the illustration below. This graph shows the value of Score (rating times review count) as the size of each bubble. In addition the color of each bubble is determined by its corresponding restaurants  Tier 1 category.  This illustrates that the majority of the restaurants with the largest scores are clustered close to both UGA and to the various apartment complexes.



The word cloud shown below consists of the names of restaurants located in shopping centers (within 0.1 miles). Only four categories are represented, with the American and Asian categories dominating.



Our last visualization shows the restaurants with the top scores in Athens, colored by their Tier 1 Category.



# Conclusion

Yelp data shows that ‘unique’ and ‘ethnic’ restaurants should be located closer to campus and apartment buildings. This is be illustrated by each of our key visualizations and is highly pertinent  information for potential restaurant owners in Athens.

While not everyone uses Yelp to find restaurants, many consumers still do. A positive score on Yelp, especially with many reviews backing this score speaks well of a restaurant, particularly if that restaurant’s cuisine is categorized as ‘ethnic’ or ‘atypical’ to regular cuisine in America. Watson found that the two major predictors of score were location and restaurant category. Restaurant owners in Athens will be able to use this insight to make strategic decisions about where to locate and how to market their restaurant based on their location. Our Yelp data shows that many restaurants in Athens adhere to our findings (ex. ethnic cuisines like Middle Eastern and South American are located on Broad St. while chain restaurants are located farther away from downtown Athens).

In conclusion, we were able to help restaurant owners identify ideal locations key drivers of success. We were pleased with the outcome of our project and excited to share our results.