Locationally

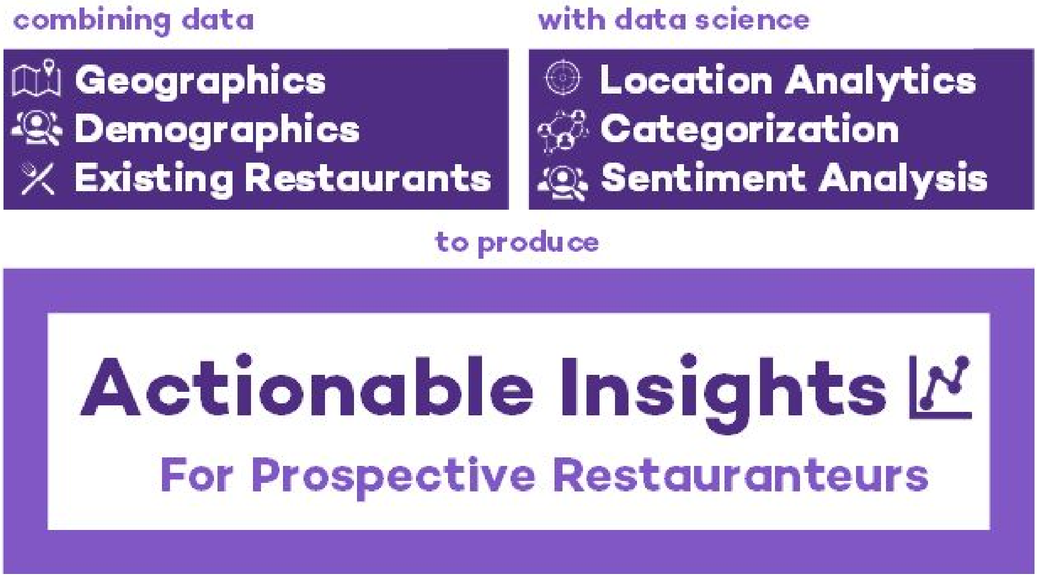
Team 54 | Final Report

2019SU_MSDS_498-DL_SEC55

Latifi, Mahoney, Marass, Scanlon, Vekaria

1. Introduction

Locationally believes that restaurants are a vital and vibrant element of any city. It's in the best interest of the city, the people living in that city, and the restaurant owners to have restaurants located in the most suitable spots. By providing data-driven location recommendations, Locationally is helping prospective restaurateurs improve the odds of their culinary success.



combining data

Geographics
Demographics
Existing Restaurants

with data science

Location Analytics
Categorization
Sentiment Analysis

to produce

Actionable Insights

For Prospective Restauranteurs

Figure 1. How Locationally provides value.

Starting a restaurant can be an intimidating undertaking. In addition to developing a menu, sourcing ingredients, and finding the right employees, prospective business owners need to identify the right location for their restaurant. Finding an area with the right kinds of customers is crucial and can be the difference between a business succeeding or failing. Locationally, LLC (henceforth, Locationally) provides insights for business location planning and specializes in the restaurant industry. Locationally utilizes cutting-edge data science techniques to integrate datasets and analyze them to equip prospective business owners with the same kind of information that large national chains use to make decisions about location selection. Locationally's approach can be used for major metropolitan areas across the United States. With Locationally's applications, restaurateurs can have confidence in their location selection and spend their time focused on finalizing the menu.

2. Problem Statement

The U.S. Chamber of Commerce (https://www.uschamber.com/) has contacted Locationally to apply their analysis tools to generate a tool to identify areas throughout several metro cities that may be a good match for different types of restaurant businesses. Locationally plans to develop web-based dashboard with mobile capabilities that can be utilized and promoted by the U.S. Chamber of Commerce. This project fits into the broader vision of the U.S. Chamber of Commerce to drive a dynamic economy by encouraging prospective restaurateurs. As part of this effort, Locationally plans to target Phoenix, Arizona for its proof of concept.

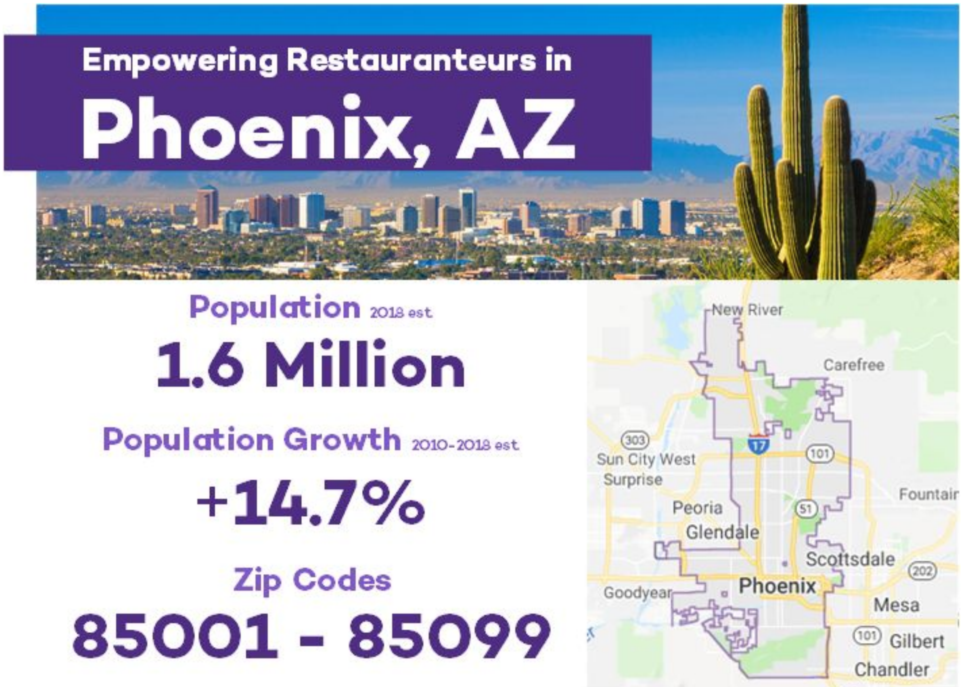


Figure 2. Overview of Phoenix, Arizona[3,4,5].

Phoenix is a major city in the United States and has seen large population growth, creating opportunity for new restaurateurs. Some of the questions we wish to answer for specific neighborhoods are as follows:

- Are there hotspots within a city that are ripe for opening a restaurant business catering to different customer segments defined by prospective restaurateurs?
- Is there a particular type of cuisine that competitors aren’t catering to successfully?
- Are certain types of cuisines rated higher in a specific area than others?

3. Project Goals

1. Leverage text analytics (i.e. sentiment analysis, topic modeling, etc.) on reviews as well as geosocial analytics to create new features for our dataset with basis variables such as demographics and attitudinal data.
2. Determine the number of K clusters to use by performing an initial K-Means
3. Once the number of clusters have been determined, perform segmentation analysis utilizing a variety of clustering techniques such as the following:
    - K-Means Clustering
    - Hierarchical method
    - PAM
    - Model-based approach
4. Based on these findings as well as demographic data gathered from the census, create segment profiles for each geographical location
5. Provide interpretability of the dashboards and recommendations on how to make the dashboards usable to make decisions.
6. Determine which features to include on the app drop downs to help provide new businesses with the best suggestions for locations as well as insights on demographics in those areas.

Our deliverables will include the following:

- Exploratory data analysis of restaurant demographics for each neighborhood
- Combined data set utilizing existing data as well as new features from our text analytics and geosocial analytics
- Cluster analysis results and method selection
- Segment Profiling by neighborhood
- User friendly online dashboard with mobile capabilities to display recommendations based on input criteria with interpretability on how to utilize them to make informed decisions

4. Data

4.1 Data Overview

A major hurdle we initially faced was obtaining the relevant data for our analysis. Eventually, we were able to pull our data from just two sources and five files. First, we gathered a handful of datasets from Kaggle, which consisted of subsets of Yelp's businesses, reviews, and user data. Originally we had found large JSON Yelp files, but upon de-nesting some of the fields in Python, we discovered many of those newly de-nested fields were further nested. Going through the process of de-nesting numerous layers for a plethora of columns seemed a bit too time consuming for the purpose of this analysis, so fortunately we discovered de-nested csv files that were originally organized for a Yelp Dataset Challenge[1]. Second, we pulled Census data from 2010 by zip code[2].

Table 1 provides each dataset name, the total number of records in each entire file, the number of records we used in this analysis after filtering by Phoenix and restaurants only, and the total number of variables.

Table 1. Summary of Datasets.

Dataset	Total Records	Filtered Records	Variables
yelp_business.csv	174,567	3247	13
yelp_business_attributes.csv	152,041	3247	73
yelp_business_hours.csv	174,567	3247	8
yelp_review.csv	5,261,668	323,750	9
Yelp_tips.csv	1098324	72495	5
USA_Census_2010.csv	32,977	41	122

The following is a general description of each dataset:

- **Yelp_business:** This is the main Yelp data file that provides a unique (masked) business id and other business related fields such as name, address, zip code, open yes/no, average number of stars, latitude, longitude, and a categories field that the business uses to describe what service it provides (i.e. “restaurants,” “hair salon,” etc.).
- **Yelp_Business_Attributes**: This file contains a multitude of binary features by business id such as “parking,” “accepts credit cards,” etc.
- **Yelp_Business_Hours:** This file contains fields for each day of the week grouped by business id. Each field is a text item, such as “9:0-18:0.” These need to be parsed to be useful in our analysis.
- **Yelp_Review:** This file contains review level data, organized by review id, user id, and its corresponding business id. The features include the number of stars given, the review date, a text field with the review, and other attributes such as whether or not the review was determined to be “funny” based on other users’ opinions. The text field with the reviews will be used for the text analytics part of the analysis.

- **Yelp_Tip:** This file contains tips data, organized by user id and its corresponding business id. The features include the likes, the review date and a text field with the tips (one or two line reviews). The text field with the tips will be used for the text analytics part of the analysis.
- **USA_Census_2010:** This file is organized by zip code and contains demographic fields related to population, household income, family size, age, ethnicity, gender by age, education levels, utilities, etc.

Filtering the data was not as simple as it seems. We tried filtering on just the city name "Phoenix," but we encountered a couple of issues. First, there were variations in the spelling of Phoenix, such as "Phx", "Phoenix Valley", some misspellings of Phoenix, etc. We also discovered that the city inputted into the dataset was done so at the discretion of the business owner. After plotting the businesses on a map it became clear not all businesses listed as being in "Phoenix" were actually in Phoenix proper. To resolve this, we worked backwards and pulled a list of all the Phoenix zip codes and filtered the data that way.

Next, we needed to remove any businesses that were not restaurants. We did this by grabbing anything with the word "restaurant" in the Category field within Python. From here, we noted that the Category field needed to be broken down further, as it contained a comma-separated text string with all the categories that the owner listed it under. For example, a business might have "restaurants, mexican, casual, tacos" in its Category field, which we would need to parse through to identify what type of restaurant it was. This is addressed later on.

Once we had a filtered down dataset, we tried matching the corresponding zip codes to the Census data to confirm the legitimacy of the business location information. However, we noticed that not every business had a corresponding match. This meant that some of the zip codes were not included in the Census data. We suspected that perhaps some of those unmatched zip codes were perhaps P.O. boxes, so we manually went through all 51 mismatches. This was done by googling the business and address, so those that had listed a P.O. box zip code were updated with their actual location zip codes. Some were also typos, so those were updated as well. The rest of them turned out to be businesses we didn't need to include in our analysis, such as food trucks, catering services, and a handful of businesses that didn't even belong in the restaurant category.

Figure 3 maps out how we connected each data file to form one large dataset:

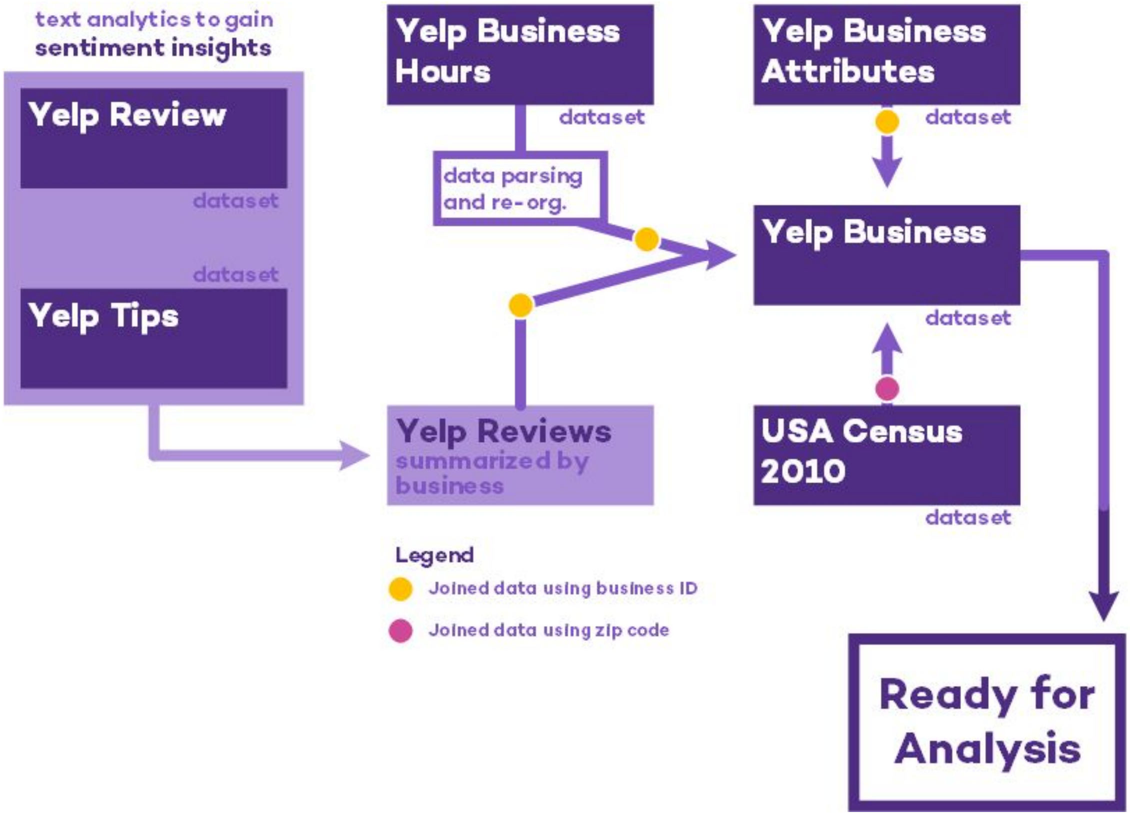


Figure 3. Dataset Integration Map.

The Join Key demonstrates what fields were used to join the datasets together, centering around the “Yelp Business” file. The Yelp Review and Yelp Tip file are highlighted on the right with a mention of text analytics being performed on the dataset. The goal is to create a single line item by business id that represents the overall sentiment for all the associated reviews. Then this revised Review dataset can be joined to the Yelp business file by business id, as discussed further in later sections.

After merging the data, some further cleanup was needed. True/False values were converted to binary 1/0 values. The categories field (discussed in further detail later) was converted from one categorical variable to multiple indicator variables by each category. For example, the variable “Pizza_dum” was created, indicating whether or not the restaurant was listed by the owner under the “Pizza” category. Missing values were examined, and almost all of the features from the attributes file that were merged into the dataset, such as “BusinessAcceptsCreditCards,” had missing data for well over 50% of the businesses (some closer to 99%). We decided to remove these variables, with the exception of the 5 parking variables: BusinessParking_garage, BusinessParking_street, BusinessParking_validated, BusinessParking_lot, Businessparking_valet. A new “Has_Parking” variable was created, indicating whether or not the business offered any parking based on any of those 5 parking variables, which were subsequently removed.

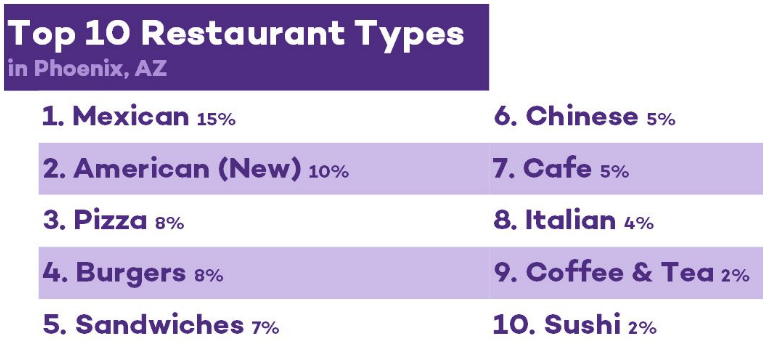
4.2 Data Transformation Process

The Yelp dataset contains many different categories for each restaurant. A primary category for each restaurant is useful in performing analyses. In the dataset, a restaurant could be tagged as being an “American (New)”, “Burger”, and “Fast Food”. For the purposes of data analysis, it is useful to understand which categorization is the most useful in representing the type of restaurant.

Locationally has developed a proprietary classification algorithm that uses hierarchical relationships among categories based on real-world experience and analytic research. These unique relationships allow for rapid categorization of restaurants. Python code was used to reduce and identify the primary restaurant categorization. The general approach utilized for performing this task was to parse the restaurant categorization labels into individual entries for each restaurant and search for increasingly specific restaurant categories. Broad categories (e.g. “Comfort Food”, “Bars”, “Breakfast & Brunch”, etc.) were identified first and specific categories (e.g. “Seafood”, “Gluten-Free”, “Donuts”) were then identified and used to replace previously identified broad categories. Returning to the example in the preceding paragraph, that restaurant falls under the broad category of “Fast Food”, the next more specific category is “American (New)”, followed by the most specific category of “Burger”. As a result, the restaurant in this example would be classified as primarily being a “Burger” place.

These relationships, which all filter from broad categories to increasingly specific categories, cover different food types (e.g. “American (Traditional)”, “Steakhouses”, etc.) and different ethnic cuisines (e.g. “Modern European”, “French”, etc.). Locationally also applies text analytics to parse restaurant names as a means of cross-validating the restaurant category. For example, a restaurant that has gone to the effort of placing the word “pizza” in their name is likely a pizza restaurant and should be classified as such. Other tasks handled by the classification algorithm include combining rare restaurant categories with broader ones to better classify a restaurant (e.g. “Bavarian” restaurants, a very rare classification, can be classified as “German” restaurants). Finally, the algorithm also filters out food businesses that do not have a physical public space open to the public (e.g. “Caterers”, “Personal Chefs”, “Food Trucks”, etc.). There were no missing category labels in the Yelp dataset that had to be addressed. Table 2 and Figure 4 shows the restaurant frequencies for the top-ten categories present. The top 35 restaurant types are presented in Figure 5. Finally, Figure 6 shows the review counts against the categories. It is observed that Mexican cuisines have the highest number of ratings followed by the Italian cuisine.

Table 2. Top 10 restaurant types/categories in Phoenix, AZ.



Top 10 Restaurant Types
in Phoenix, AZ

1. Mexican 15%	**6. Chinese** 5%
2. American (New) 10%	**7. Cafe** 5%
3. Pizza 8%	**8. Italian** 4%
4. Burgers 8%	**9. Coffee & Tea** 2%
5. Sandwiches 7%	**10. Sushi** 2%

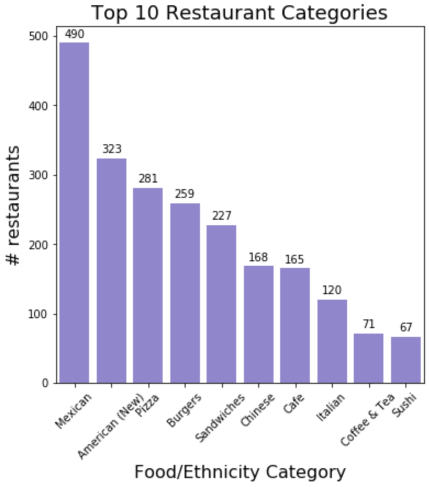


Figure 4. Top 10 restaurant types/categories in Phoenix, AZ.

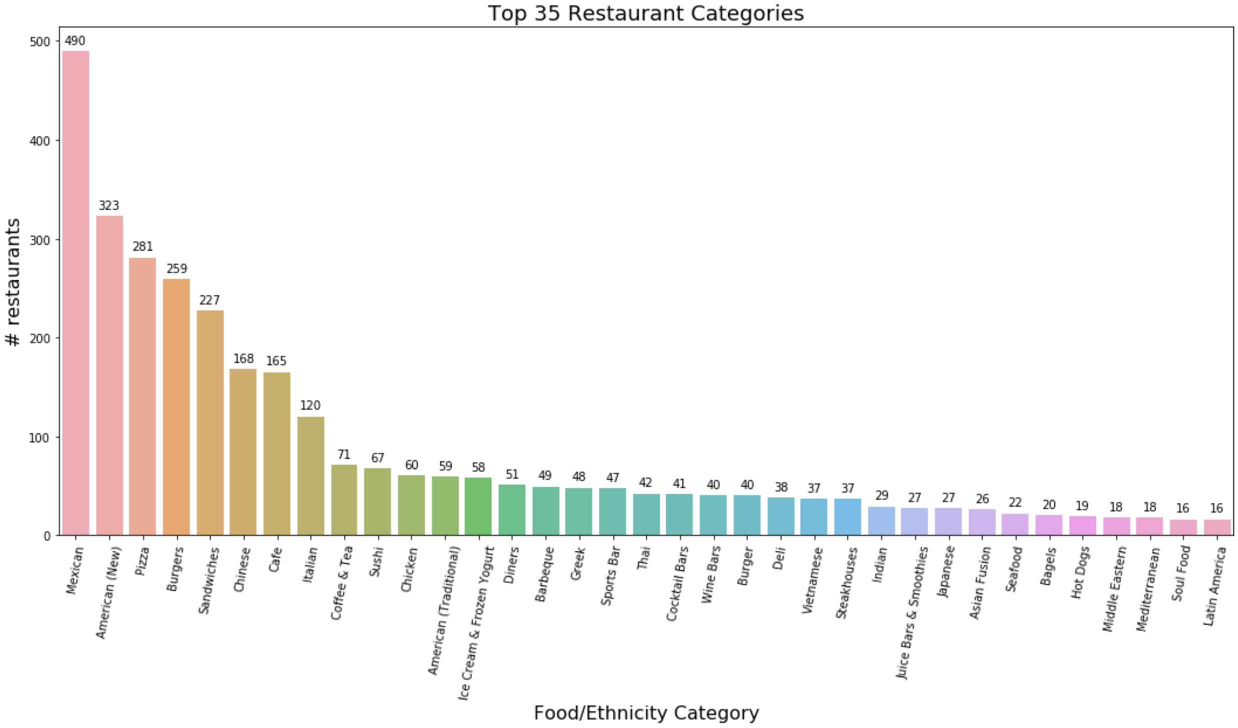


Figure 5. Top 35 Restaurant Categories.

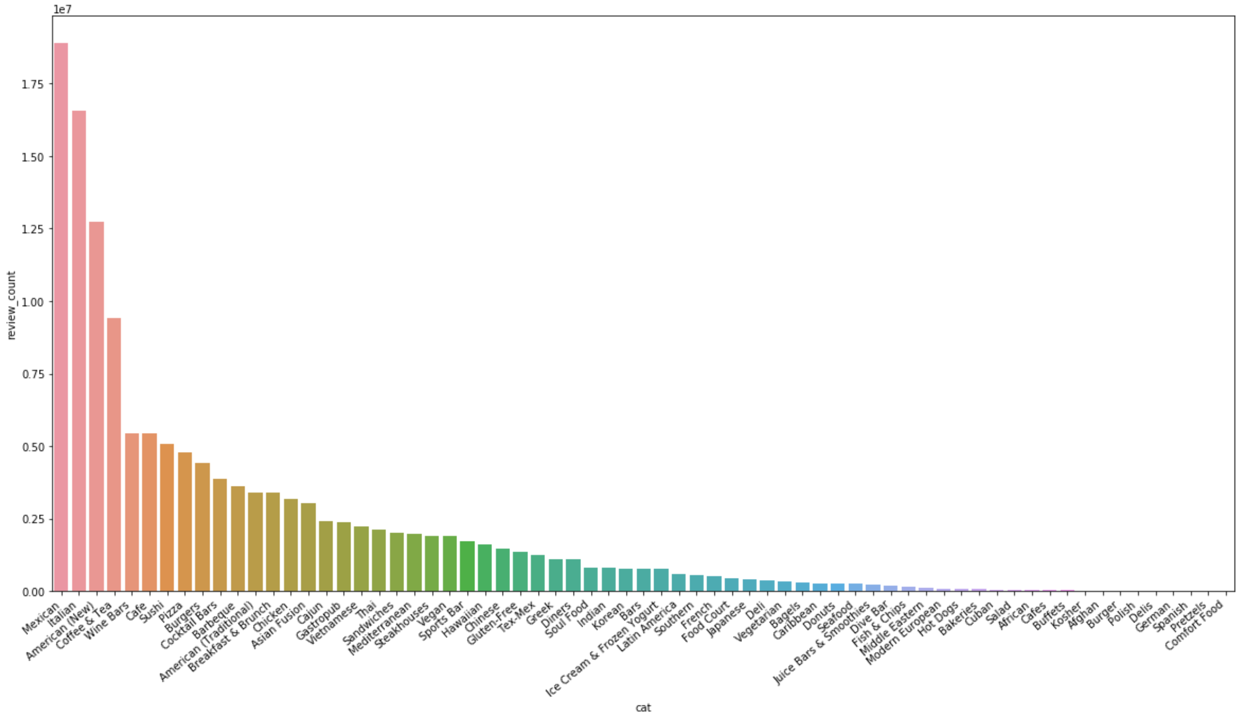
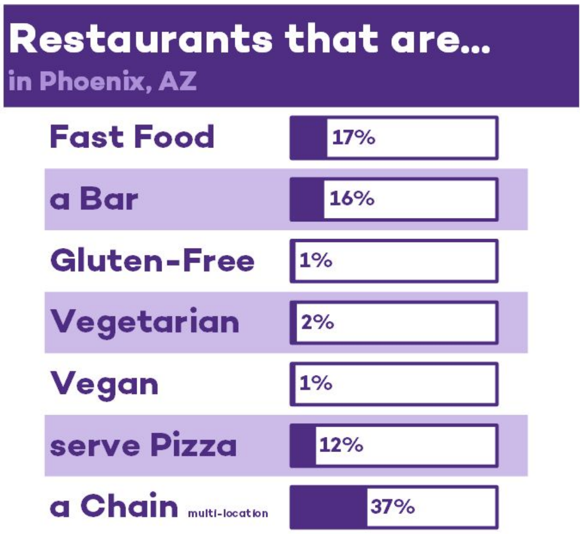


Figure 6. Top reviewed Categories.

In addition to primary-restaurant-type categorization, each restaurant is assessed to set indicator variables to cover aspects of the restaurant that may not be reflected in the primary categorization. These indicator variables include if a restaurant is considered fast food (e.g. a restaurant could be considered a burger restaurant that is fast food), is considered a bar, or has a gluten-free menu, a vegetarian menu, a vegan menu, or has pizza (e.g. some American restaurants also offer pizza as a secondary menu item). These very broad indicator variables allow for additional insights into common restaurant offerings. Table 3 shows the summary of counts for each of the secondary indicator variables produced from the dataset; these variables are also presented in Figures 7 through 13.

Table 3. Summary of indicator variables.



Restaurants that are...
in Phoenix, AZ

Fast Food	17%
a Bar	16%
Gluten-Free	1%
Vegetarian	2%
Vegan	1%
serve Pizza	12%
a Chain multi-location	37%

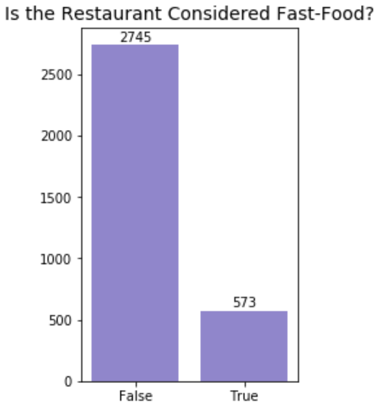


Figure 7. Fast-Food Categorization.

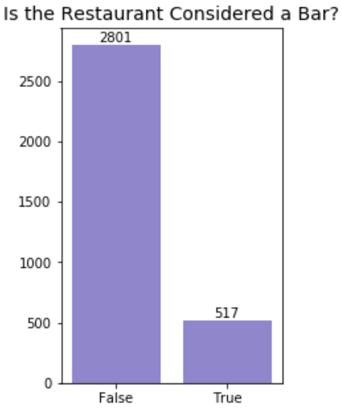


Figure 8. Bar Categorization.

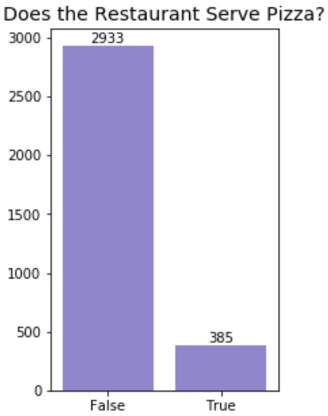


Figure 9. Pizza Categorization.

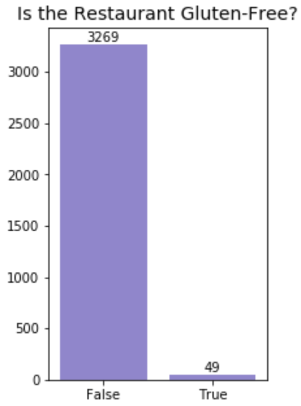


Figure 10. Gluten-Free Categorization.

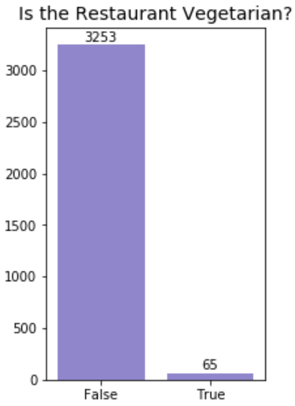


Figure 11. Vegetarian Categorization.

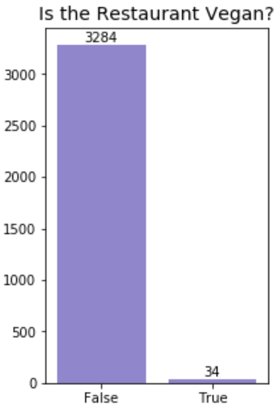


Figure 12. Vegan Categorization.

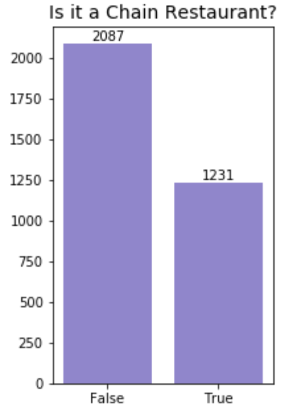
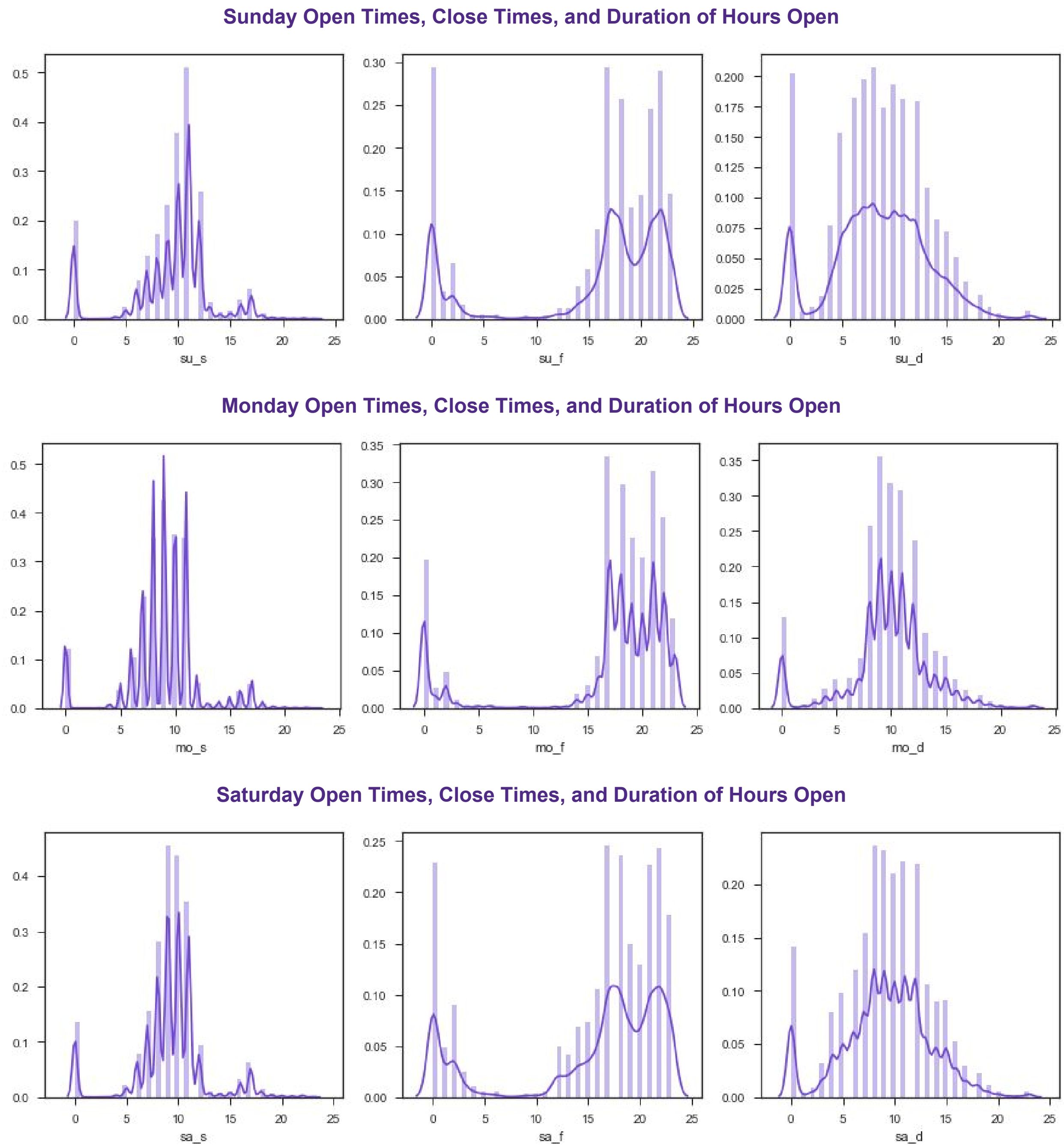


Figure 13. Chain Categorization.

The information in the Yelp business hours data set consisted of daily hours for each business. So each day was represented as a column and the values were a range of hours, e.g. 11:00-8:00. If a business was not open on a given day, the hours were null for that respective column. To make this information more useful, the hours were parsed into daily open times and daily closing times. So, each day was separated into two columns, one for opening hours and one for closing hours. Additionally, three binary columns were created that represent whether a business is open during the weekdays, Fridays, and/or weekends. Weekdays are considered as Monday through Thursday and weekends are Saturday and Sunday. Exploratory analysis of the weekday and weekend classification revealed that there was no statistical significance in including Fridays as part of the weekday versus weekend, therefore, it was left as its own category.

Similarly, binary columns were created representing whether a restaurant was open for breakfast, lunch, dinner, and/or late-night. The categorization of hours was a combination of subjective logic derived by Locationally's expertise and an assessment of the duration of hours a business is open, business open times and business close times. Figure 14 displays distributions of Sunday, Monday, and Saturday opening times, closing times, and duration of hours open. It is clear that there is greater deviation of duration on the weekends (Saturday and Sunday) compared to the weekdays. Especially note the spike at zero for both the Sunday and Saturday closing time plots, as well as the plots showing duration of hours open. Having a close time of zero indicates the business closes at midnight (12:00 a.m. or 0:00 UTC), whereas having a duration of zero indicates that the business was open for zero hours.


Sunday Open Times, Close Times, and Duration of Hours Open

Monday Open Times, Close Times, and Duration of Hours Open

Saturday Open Times, Close Times, and Duration of Hours Open

Figure 14. Distributions of Business Open Times, Close Times, and Duration of Hours Open.

The Yelp Review and Yelp Tip data were first merged together. Both the Review and Tip tables had User ID, text, business IDs data. First, a subset of the datasets was created for both the datasets individually based on the Business IDs in the Business dataset. Then the Tip dataset was appended to the Reviews dataset inorder to merge the User ID and Text columns for the merged dataset. This combined dataset was then merged on to the Business dataset.

Some of the columns were renamed in the combined dataset such as stars column from the Business dataset was renamed to Business_Star_Rating and stars column in Review dataset was renamed to User_Star_Rating. The name column from the business dataset was renamed to Business_Name. Columns such as cool, funny, likes and useful were dropped as they are not being used for the text analytics piece.

5. Exploratory Data Analysis

5.1 Exploratory Data Analysis

During the exploratory analysis (EDA) phase, it was quickly discovered that a majority of data was missing in the Yelp business attributes file. Over 75% of the data were missing in most of the columns, as displayed in Figure 15. There were only seven columns that had less than 75% of data missing. Therefore, due to the high volume of missing data and the increasing width of the modeling data frame, columns that were missing more than 99.0% of data were removed from the business attributes data set. Additionally, columns that were irrelevant to restaurants (e.g. ‘HairSpecializesIn_coloring’) were also removed from the data set.

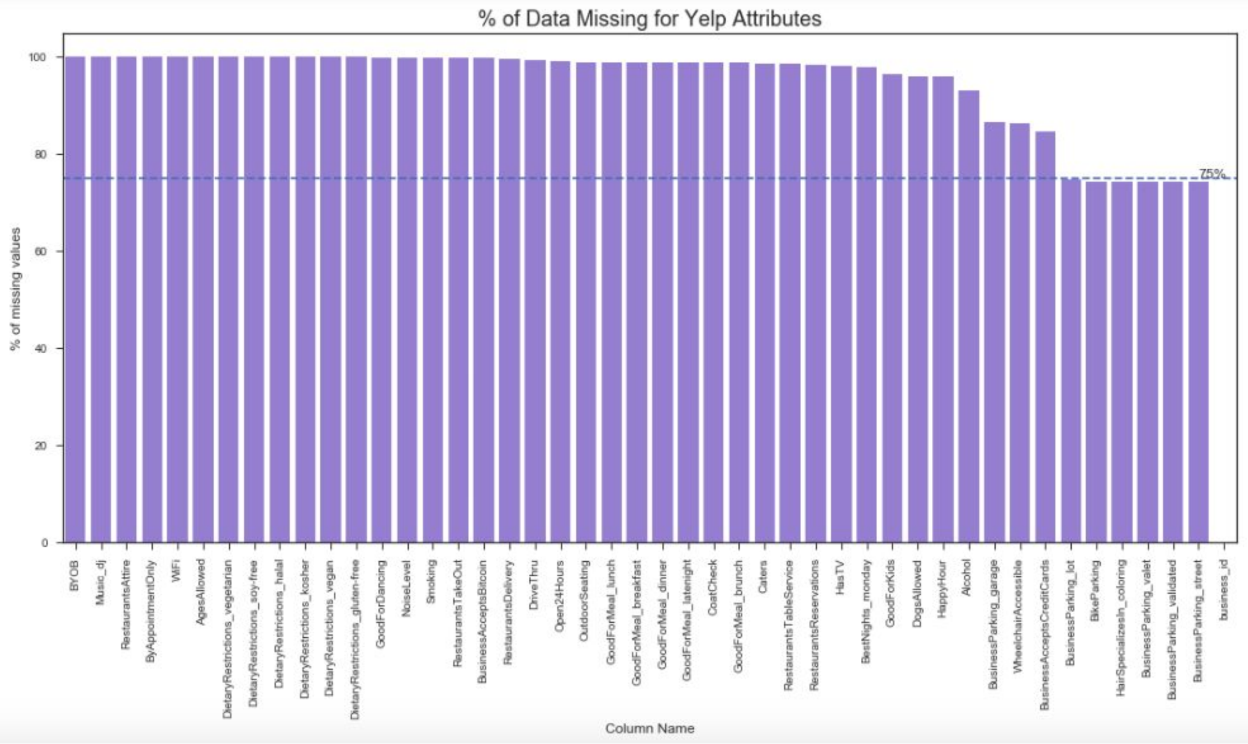


Figure 15. Percent of Missing Data per Business Attribute Column.

To better understand customer ratings patterns and temporal restaurant reviews, Locationally analyzed monthly and yearly review trends as well as spatial trends. Figure 16 displays the yearly number of Yelp reviews and the monthly number of Yelp reviews across all available years. Phoenix has been a steady growing city[1] over the last 10 years, so yearly growth in number of reviews for Phoenix restaurants is expected. However, it is interesting that less reviews are posted during February and December, as shown in Figure 16. There also appears to be a decline in the number of Yelp reviews posted towards the end of each year. In the view of daily Yelp reviews, there is defined seasonality that also displays the observed pattern of fewer reviews during the February and December months, as well as the decline in posted reviews at the end of each year. It is important for prospective business owners to be aware of these types of trends, especially if fewer reviews associates to less business.

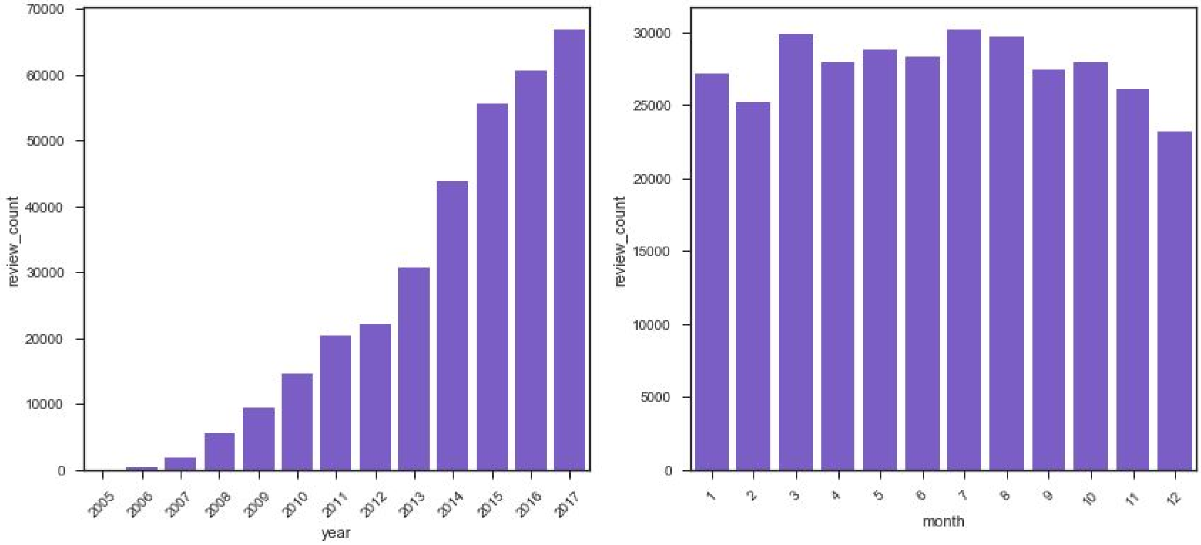


Figure 16. Yearly and Monthly Number of Yelp Reviews for Phoenix Restaurants.

[1] https://www.abc15.com/news/region-phoenix-metro/central-phoenix/phoenix-leads-us-in-population-growth-new-census-data-shows

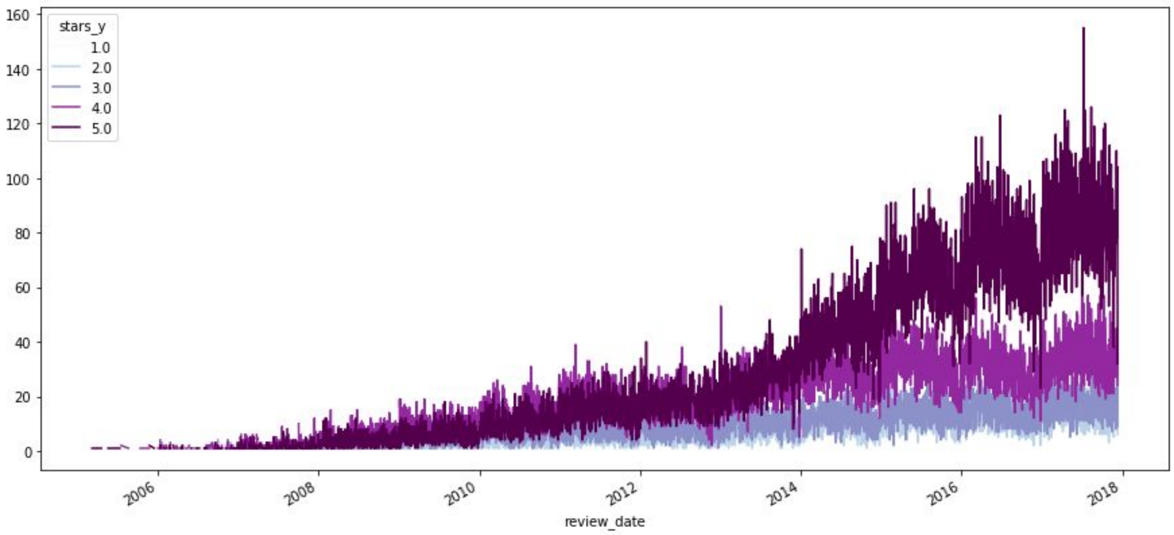


Figure 17. Daily Number of Yelp Reviews for Phoenix Restaurants.

In addition to temporal trends, Locationally supplies their clients with geo-spatial information to help guide decisions focused on selecting the best restaurant location. Therefore, the zip codes within Phoenix were analyzed to understand if there are areas that are more dense with restaurants, areas that are underserved for certain food types, areas that receive better ratings, areas that are catered to certain demographics, etc. Figure 18 contains four charts that look at the top 25 zip codes with the highest number of restaurants, highest number of reviews, highest average star rating, and highest availability of food categories. A notable observation from these charts is that for the zip codes with the most restaurants listed, those same zip codes don’t necessarily have the highest number of reviews. A logical assumption prior to this analysis would be that areas with more restaurants would probably have more reviews (assuming more restaurants equates to more opportunities to leave/submit a review), however, we see this is not necessarily the case within Phoenix. This could indicate that other factors may affect whether or not a review is submitted for a given business.

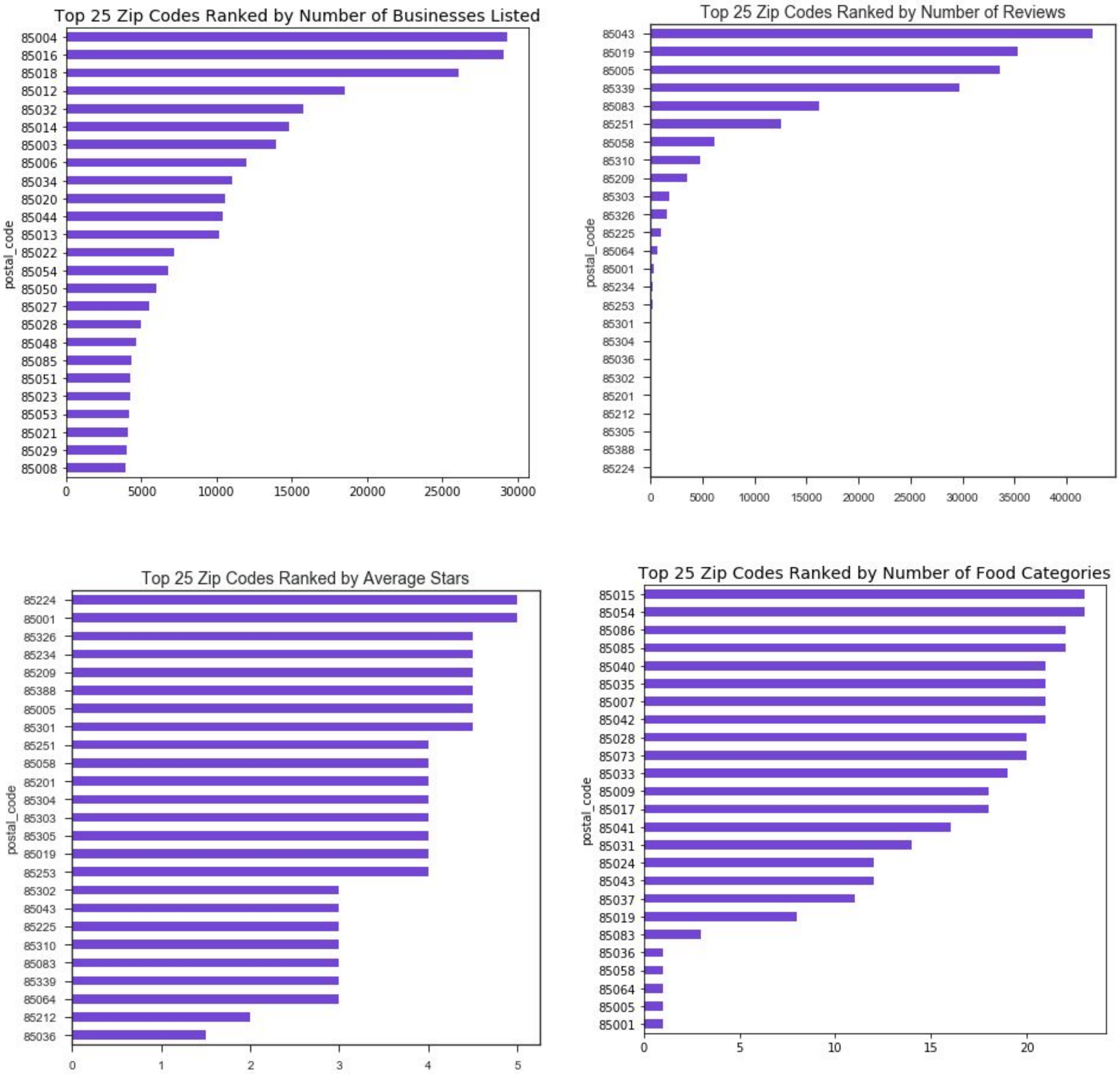


Figure 18. Top 25 Zip Codes by Various Rankings.

Similarly, we see that having a greater number of restaurants within a zip code does not guarantee greater cuisine diversity within that same zip code. That is, zip codes with more restaurants do not necessarily have a greater variety of cuisines. We also looked at locations that had higher ratings (more stars) on average given the normalized review count was greater than a set threshold. In other words, ratings were normalized to avoid skewed average results for restaurants within zip codes that did not receive a significantly large enough number of reviews. As with the other rankings, zip codes with more restaurants or more reviews do not necessarily have higher overall star ratings.

Additionally, we evaluated the locations and distributions of how the restaurants are clustered across the Phoenix area. First we assumed there would be a large segment of restaurants in the city center itself, however, our findings showed that there are several larger clusters of restaurants just outside the city center, as well. See Figure 19.

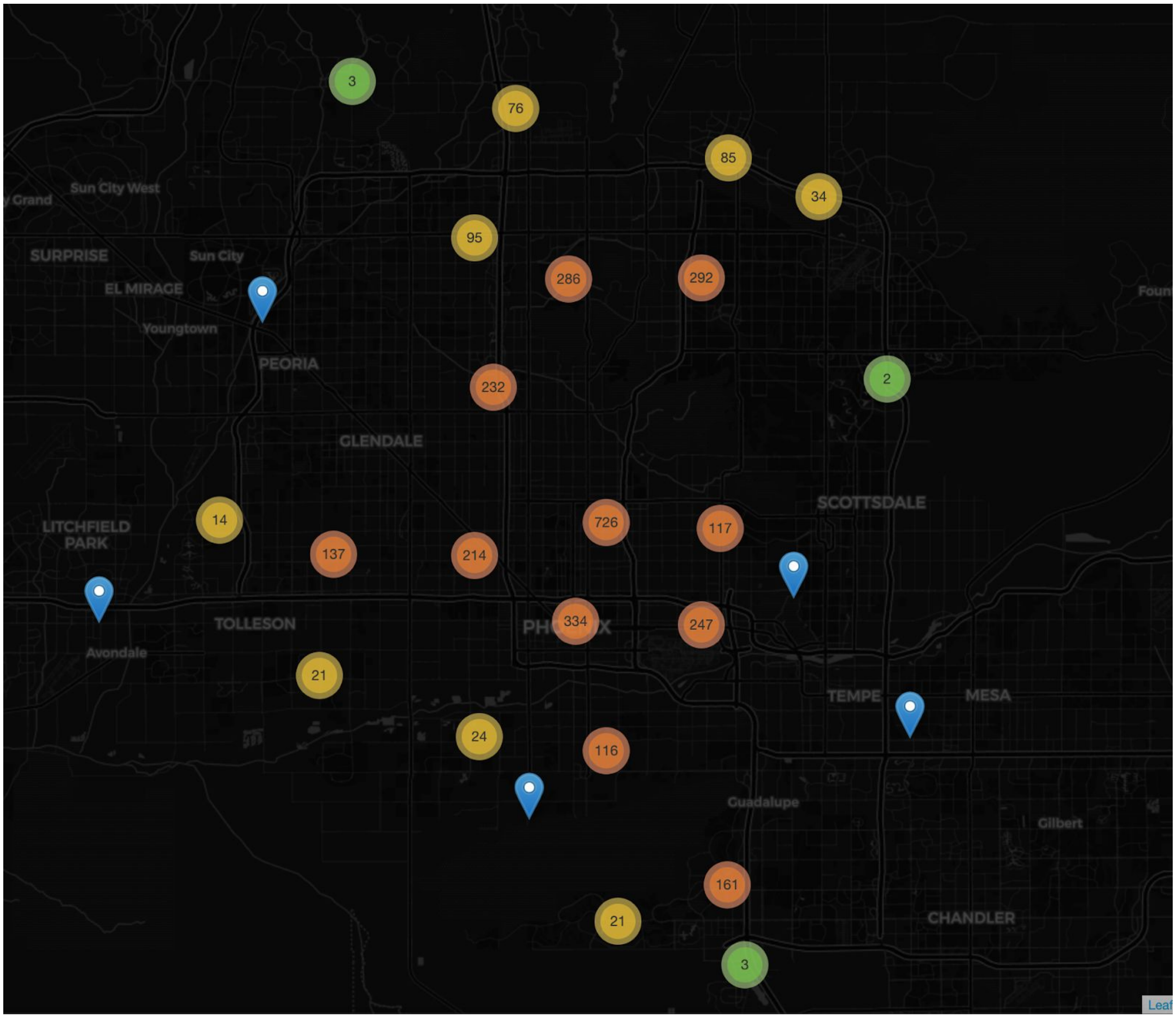


Figure 19. Restaurant Distribution by Location.

Next, we wanted to see where the best restaurants (by star rating) in Phoenix were located. Figure 20 illustrates a heat map that has the highest rated restaurants in the greater Phoenix area. The majority of the highest rated restaurants are in the city center, but there is also a pocket of several highly rated restaurants located north of the city which corroborates with our previous findings regarding clusters of restaurants north of the city center.

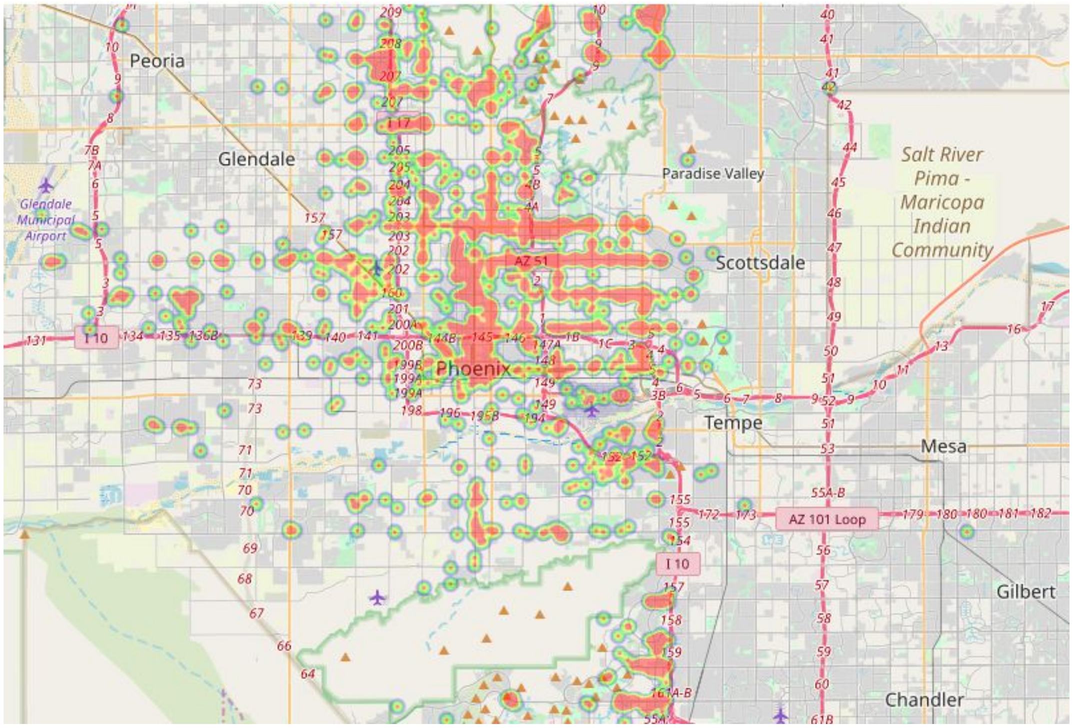


Figure 20. Phoenix Heat Map Based on Average Star Rating per 100 Reviews.

Lastly, we need to understand the difference in star ratings between the business data set and the reviews data set. Both of these data sets contained a column for the star rating, but the distribution of stars were drastically different, as seen in Figure 21. In the business data set, each restaurant has a star rating that is an overall weighted aggregate of the individual ratings received through reviews. The left chart that shows the distribution of these overall ratings at a restaurant level, whereas the chart on the right shows the distribution of stars per review. These stars represent the individual customer ratings at a review level. These are discrete compared to the business stars because customers only have the option to select 1,2,3,4, or 5 stars and not half stars within the Yelp mobile application.

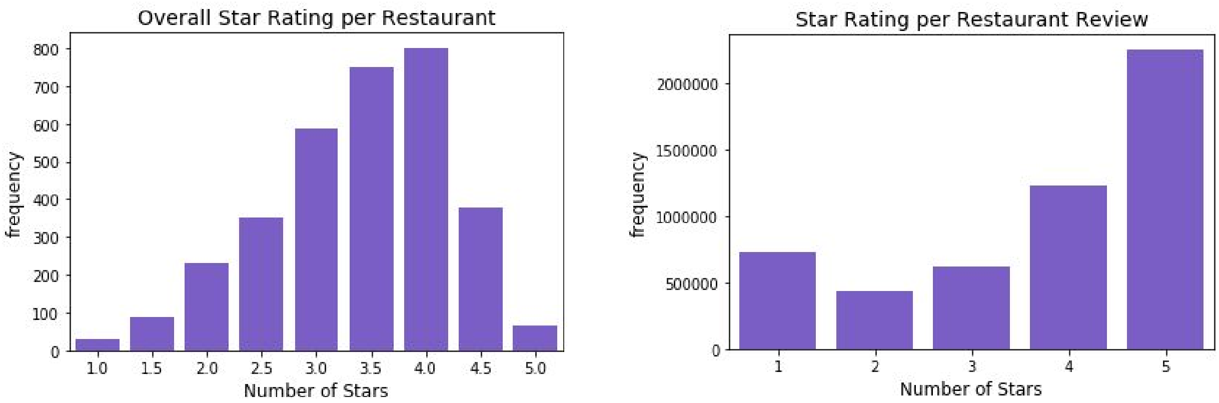
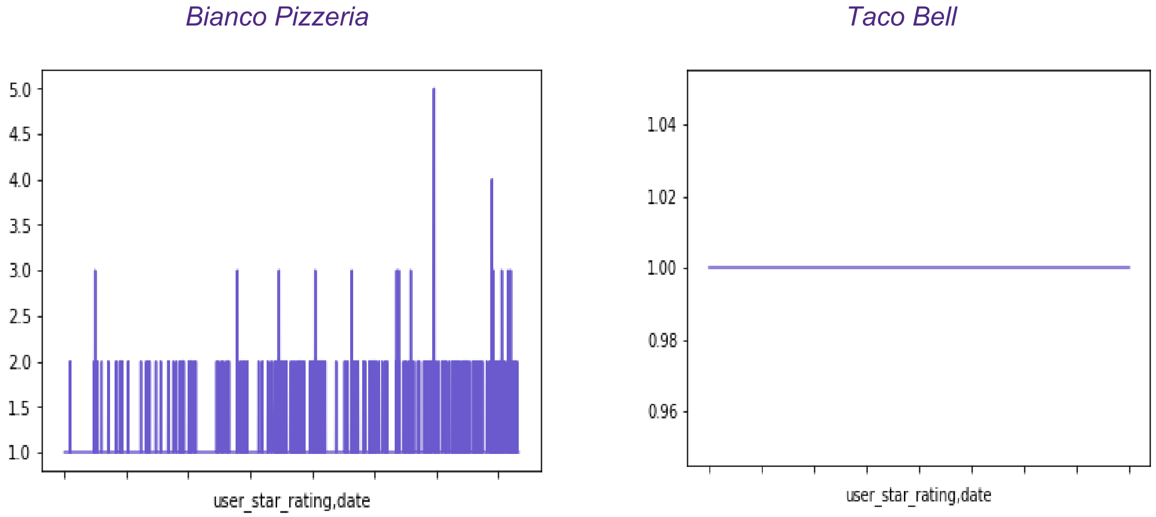


Figure 21. Star Ratings Distributions.

We also reviewed the restaurants to identify restaurants that had the maximum reviews and those that had the minimum reviews. An Italian Pizza restaurant Bianco Pizzeria had the maximum review count and a Taco Bell had the lowest review count. Figure 22 shows the user star rating for Bianco over the entire range of dates in the data. As can be observed the restaurant has consistently received 2 star ratings but in more recent months, the ratings have also gone up to a 4. Figure 22 displays the review distribution for the lowest number of ratings for a Taco bell over the entire period.


Bianco Pizzeria

Taco Bell

Figure 22. Star Ratings Distributions over time for maximum and minimum review count restaurants.

Most of the clustering techniques we will be using are heuristic unsupervised learning methods, which means we aren't making assumptions such as normality, etc. So, since we aren't making those assumptions, we don't need to validate them. Variable correlation is also something that doesn't need to be "fixed" with Clustering analysis; however, it is still worth exploring to gain more insights on the data. A correlation plot was performed on all the numeric variables below in Figure 23.

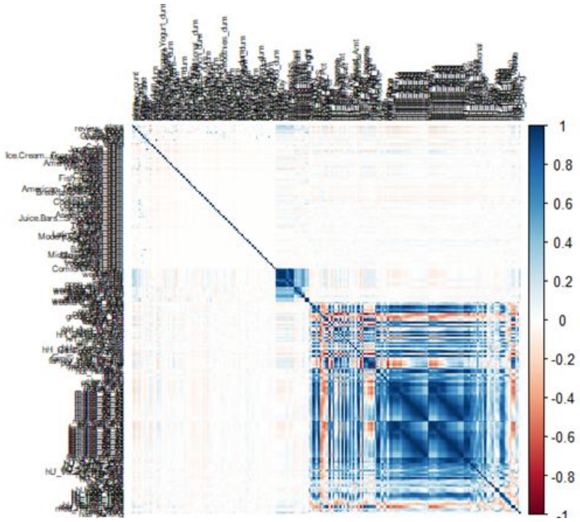


Figure 23. Original Correlation Plot.

Along the diagonals, we are seeing stronger correlation with the dark blue encoding versus the weaker correlation seen with the light areas. The bottom right hand corner, which is primarily made up of Census data, also displays some strong correlation. Some amount of negative correlation is apparent as well, shown by the red/pink areas, so we examined this further by identifying the culprits. These included Census variables such as family size, poverty percent, food stamps, etc. that made logical sense. A revised correlation plot was run, and a decent amount of the negative correlation seems to be removed. See Figure 24 below.

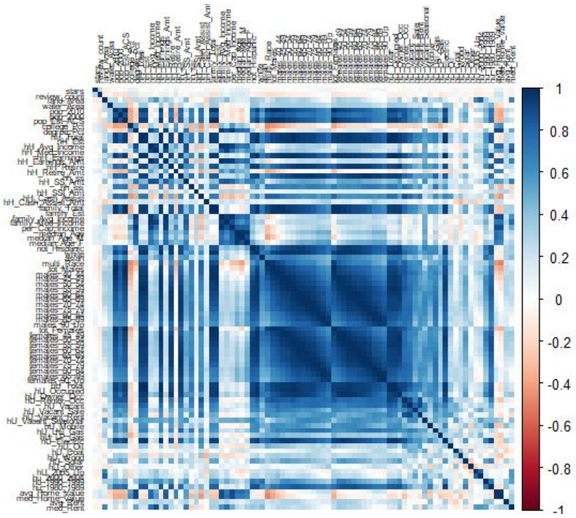
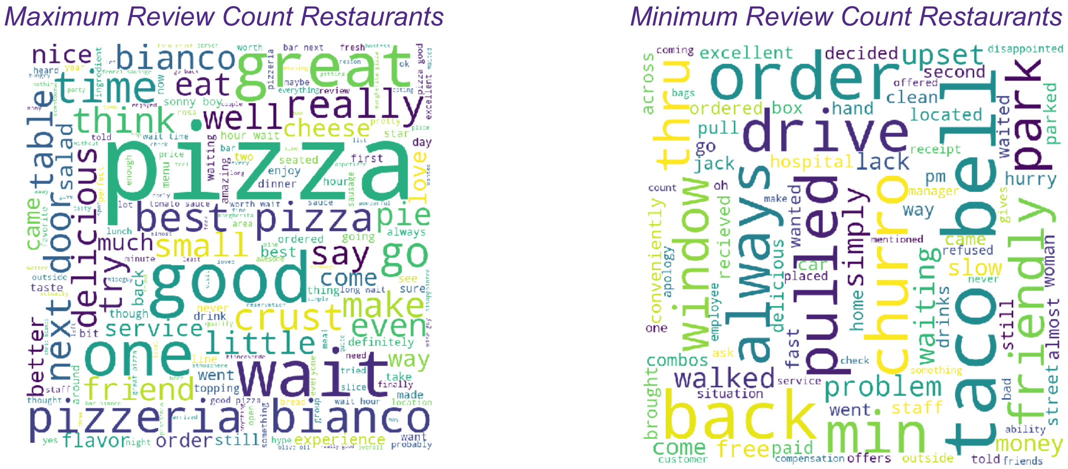


Figure 24. RevisedCorrelation Plot.

5.2 Word Cloud

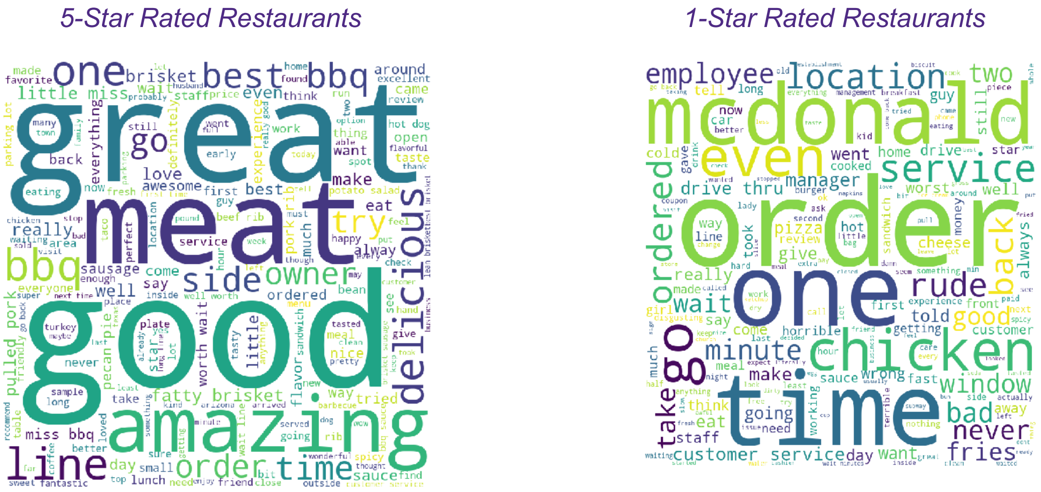
As part of the initial text analytics EDA, word clouds were generated for the restaurants with highest review counts and lowest number of review counts. Figure 25 displays the values for the top and least review counts respectively. As can be seen the most talked about words are pizza, best pizza, great, Bianco (the name of the restaurant), wait, crust, delicious, salads etc for the Bianco restaurant with the top review counts indicating the restaurant has the best and delicious pizza with a good crust. They also offer salads but probably have a wait time. While for the Taco bell with the least review counts, words like taco bell, churro, park, friendly, pulled, drive thru, etc. are the most talked words indicating the Taco Bell is a drive-through (thru) with friendly staff and a good churro.


Maximum Review Count Restaurants

Minimum Review Count Restaurants

Figure 25. Word Clouds for maximum and minimum review count restaurants.

Word Clouds were also generated for all the businesses that had 5 star reviews and also those that had 1 star reviews. Figure 26 displays the word cloud for these 2 respective ratings. It can be seen that the restaurants with 5 star ratings offer great meat, best bbq, delicious, amazing food, briskets., sausage, turkey etc. They also have lines for seating in their restaurants. While the restaurants with one star seem be fast food chains such as McDonalds, their service can be rude and offer limited items such as chicken or fries. These one star restaurants are associated with words like bad, never, wait, long etc,


5-Star Rated Restaurants

1-Star Rated Restaurants

Figure 26. Word Clouds for 5 star and 1 star business ratings restaurants.

6. Cluster Modeling

6.1 Feature Reduction

We decided to leave out the binary variables for our cluster modeling, leaving only continuous variables. However, we still had quite a lot of remaining variables given the Census data. Although clustering doesn't require feature reduction, it is a good way to look for natural clusters in the data, especially in our case given the couple hundred features in our dataset. Principal Components Analysis was performed on a subset of the data consisting of all numeric variables. The second principal component accounted for ~95% of the variance, which is usually good. Figure 27 below shows the first two principal components plotted:

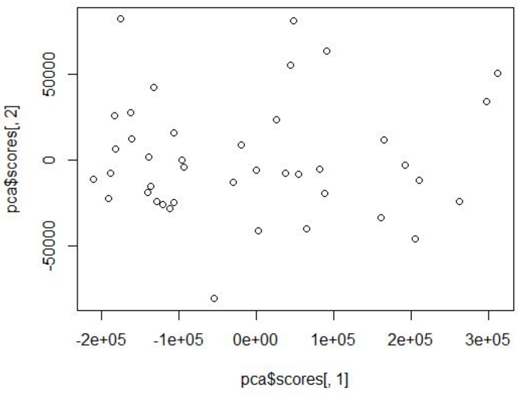


Figure 27. First Two Principal Components.

Here we are seeing some minor natural clusters as well as a few outliers. We proceeded on with the cluster modeling.

6.2 K-Means Clustering

The first step we took was deciding how many clusters we wanted. To do this, we examined a scree plot. See Figure 28 below.

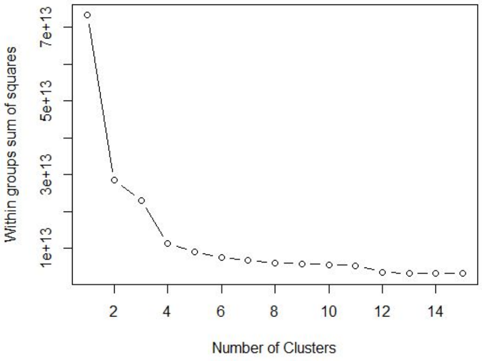


Figure 28. Scree Plot.

The scree plot demonstrates a dramatic drop from 1 cluster to 2 clusters, and it shows a decent “elbow” somewhere between 4 - 6 clusters. Based on this, we decided to try out 5 clusters to start.

Next, we ran K-Means using 5 clusters. See Figure 29.

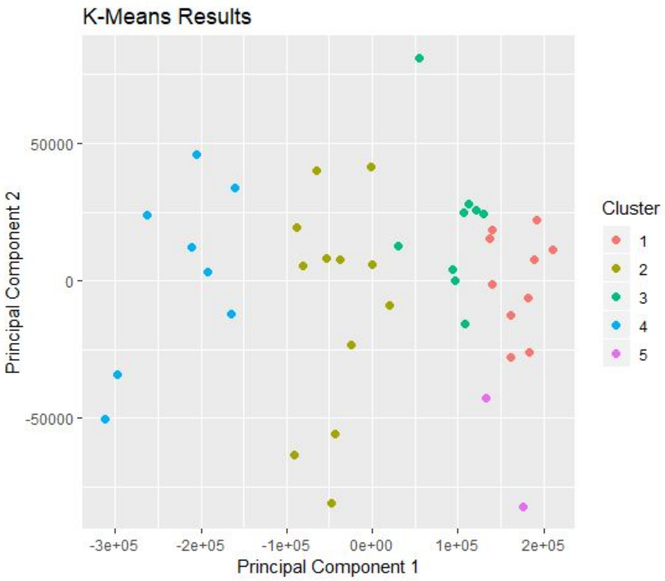


Figure 29. K-Means using 5 Clusters

There doesn't appear to be much overlap between the clusters, so that's good. But we're not seeing many patches (points within clusters aren't that close). We are also seeing a few outliers, specifically in Clusters 2, 3 and 4. The r-squared value obtained was 0.85.

To check the validity of the clusters, we then created a Silhouette plot. The average silhouette width was 0.59, which is pretty decent. Clusters 4 and 5 are the strongest with respective indexes 0.71 and 0.70.. Cluster 3 is the worst with a Silhouette index of 0.33, but that still isn't bad. See Figure 30 below.

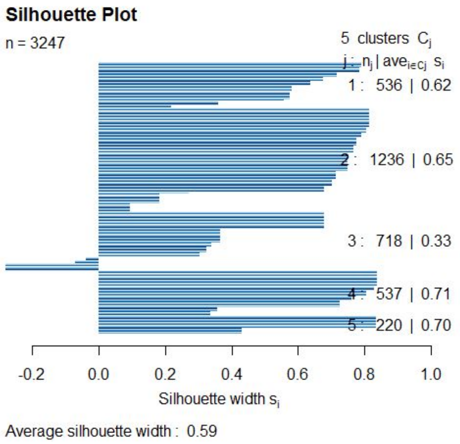


Figure 30. Silhouette Plot.

6.3 Hierarchical Clustering

Since we have a small dataset (less than 10,000 observations), we can perform hierarchical clustering. We tried four methods using 5 clusters: Ward D2, single, average, and complete. Silhouette plots were created for each method, and the one with the highest silhouette index was the Ward D2 method with an index of 0.51. The r-squared value for this model is 0.87. See Figure 31 below.

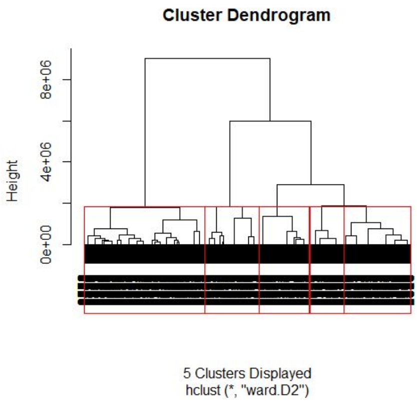


Figure 31. Hierarchical Clustering with Ward D2 Method

The r-squared for this Hierarchical model is slightly higher than the K-Means r-squared of 0.85, but the K-Means cluster method had a higher Silhouette width of 0.59.

6.4 PAM Clustering

Next, a PAM clustering method was tried because it's less sensitive to outliers compared to K-Means. The total average Silhouette width achieved was 0.49. The silhouette width for Cluster 5 was good at 0.71 while Clusters 1 and 4 had the lowest silhouette widths, both at 0.34. See Figure 32 below.

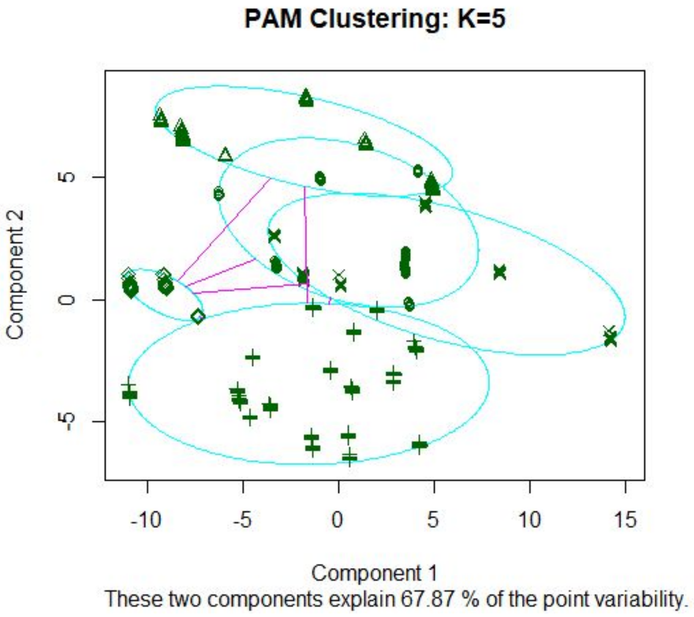


Figure 32. PAM Clustering Method with 5 Clusters

6.5 Model-Based Clustering Method

Given that our data is based on only continuous variables, we could make the normality assumption and attempt model-based clustering. First, a generic fit was tried without specifying the number of clusters. This resulted in a model with only 1 cluster and a BIC of 5156037, which obviously we won't use since it allocated all the data to one cluster. Next, a fit with 5 clusters was attempted, and a BIC of -5077299 resulted. Additionally, a large portion of the observations were put into Cluster 1: 2,477 out of 3,247. This also isn't ideal. When calculating the total average Silhouette width for the fit with 5 clusters, it resulted in -0.35, which isn't good at all.

6.6 Model Comparison

Using the Correct Rand function, all the pairs between all 3,247 observations are compared in the K-Means and the PAM models to see how many pairs end up in the same cluster in both methods. Here we see a 35.79% match, which means that each model clustered 35.79% of the observations the same. When comparing the K-Means model to the hierarchical model, we only get an 43.90% match, and comparing the hierarchical to the PAM models gives us a high 89.28% match.

We examine the Average Silhouette Width for each model in Table 4 below:

Table 4. Clustering Model Comparison.

Model	Average Silhouette Width
K-Means	0.59
Hierarchical with Ward D2	0.51
PAM	0.49
Model-Based Method	-0.35

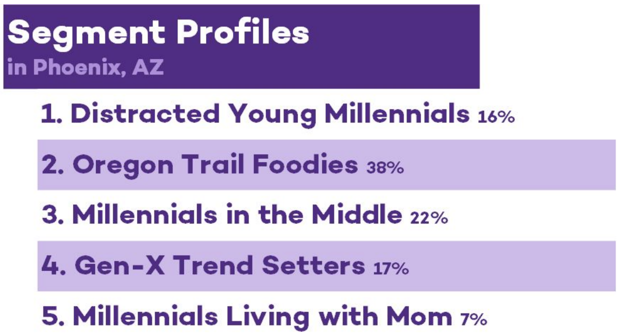
The winning model could be chosen solely on the highest Average Silhouette Width, which in this case is the K-Means model. However, it was interesting that the Hierarchical and PAM methods had such a high % match in clustering results, and given the fact that their average silhouette widths aren't far behind the K-Means model, one could justify choosing one of them. For this purpose, we will base our clustering results on the K-Means model.

6.7 Segment Profiling

Table 5 contains a summary of the proposed segment profiling based on the K-Means results.

Table 5. Summary of segment profile variables by cluster.

Cluster	% Total	Avg. Reviews Per Business	Avg. Population Per Zip	Median Income	Median Age (years)	Degree (%)	Median Home Value
1	16%	36	41,445	$34,674	27.8	9.7	$121,800
2	38%	114	25,392	$43,681	36.0	31.8	$248,014
3	22%	60	49,980	$47,054	32.4	21.4	$172,215
4	17%	127	25,785	$69,131	39.3	46.6	$371,152
5	7%	105	11,381	$16,669	30.9	10.8	$131,492



Segment Profiles
in Phoenix, AZ

1. Distracted Young Millennials 16%
2. Oregon Trail Foodies 38%
3. Millennials in the Middle 22%
4. Gen-X Trend Setters 17%
5. Millennials Living with Mom 7%

Figure 33. Segment Profiles.

6.7.1 **Cluster 1: Distracted Young Millennials** *(approximately 16% of the population):* These consumers are the youngest in the mid-to-late 20's range. They don't have the highest education level and only make an average salary, which is expected given their age. They are constantly distracted by things like Instagram and Snapchat and can't be bothered with things that require much focus like writing yelp reviews. They're more prone to eating at chain restaurants and tend to gravitate towards fast food and pizza.

6.7.2 **Cluster 2: Oregon Trail Foodies** *(approximately 38% of the population):* These consumers are in the mid-30's range with income levels. They are a unique sub-generation in between Gen-Xers and Millennials because they learned to adapt to technology at a young age but remember life without it. This college-educated group enjoys trying new restaurants and reviewing them yelp on a regular basis. They tend to give the highest number of stars for reviews along with the Gen-X trend setters.

6.7.3 **Cluster 3: Millennials in the Middle** *(approximately 22% of the population):* These consumers are mostly millennials in the early 30's with average income, a moderate amount of education and living in highly populated areas. They also enjoy chain restaurants and fast food like the Distracted Young Millennials. This group yelps a little but not nearly as much as the older generations.

6.7.4 **Cluster 4: Gen-X Trend Setters** *(approximately 17% of the population):* Out of all five clusters, this group has the highest average age range in the late 30's to early 40's (which isn't old by any means). They have the highest median income and the highest education levels of all the groups. They enjoy frequenting all sorts of restaurants and like to show off their knowledge to all their friends by their detailed yelp reviews. They also gravitate more towards restaurants with parking.

6.7.5 **Cluster 5: Millennials Living with Mom** (approximately 7% of the population): This group of millennials average around 30 years of age. They have the lowest income of all the groups as well as lower education levels. They generally can't afford to live on their own and often live with their parents! Because of this, they have plenty of time and enjoy yelping almost as much as the Oregon Trail Foodies.

7. Prediction Model

7.1 General Purpose

Given the fact that only the continuous variables were used to perform the cluster analysis, we decided to make use of the binary variables by creating a prediction model. This model predicts the star rating of a given business based on variables such as zip code (which were subsequently converted to binary variables), the category binary variables (such as Mexican yes/no), as well as the variables related to hours open (weekday yes/no). We did not include the Review Counts because that would be a target leak. The reason we are predicting star ratings and provide future restaurants with a sense of how their restaurant might perform given a few inputs. These inputs are initially the restaurant category (type of food the restaurant would serve) and zip code of interest, but may be expanded in the future.

7.2 Feature Importance

Random Forest variable importance plots were performed. The top 20 variables are shown below in Figure 34.

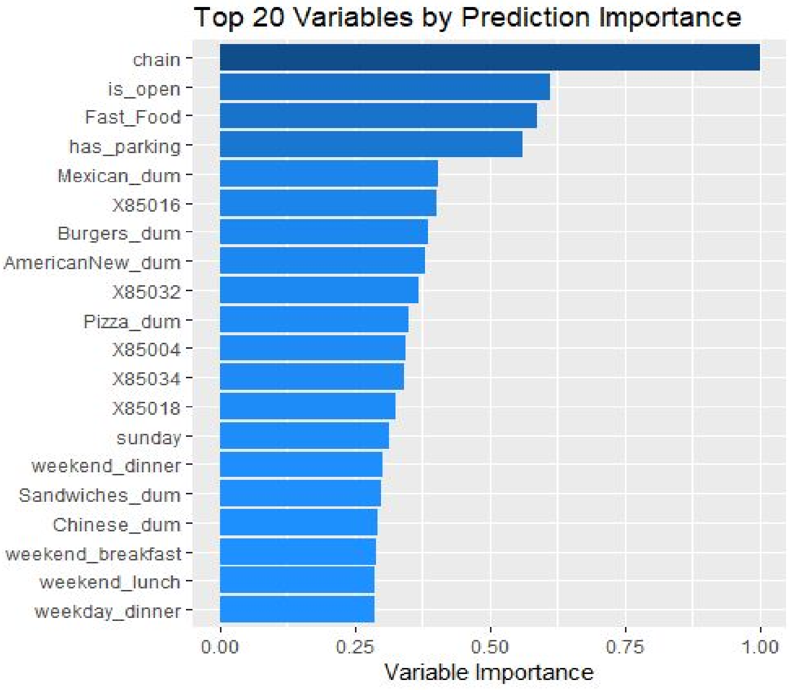


Figure 34. Variable Importance Plot

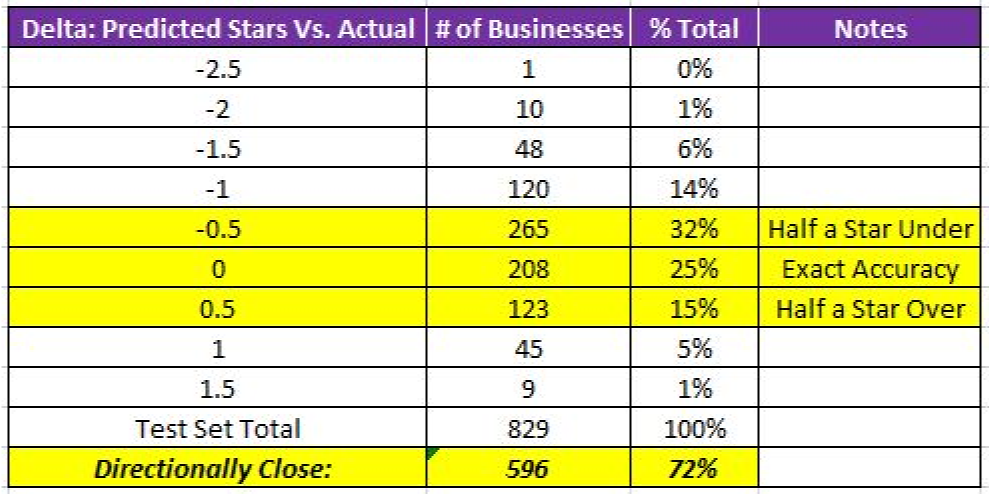
From this, it’s clear that whether or not a restaurant is a chain or not is the most important predictor of stars, followed by the restaurant being open or not. Fast food is also important, which makes sense assuming those types of restaurants have lower stars. Whether or not a restaurant has parking was also deemed important, which could be due to the fact that we are examining a place like Phoenix and not a major city such as New York or Chicago where Ubers and public transit are more present. We also see that Mexican is the most important restaurant category, which makes sense given that we saw during the EDA process Mexican restaurants had the highest volume of businesses in our dataset.

7.3 Modeling and Scoring Approach

We then split the data into train and test datasets (75/25 split respectively). Since we were trying to predict an ordinal categorical response variable, we first attempted various classification models. These models included Random Forest, Support Vector Machines, and Naive Bayes. However, we were getting extremely low AUC and accuracy results, so we decided to try a simple regression model. We chose this because it is very explainable, and in this case we also wanted to keep all of the variables because we would be using them as potential inputs in our dashboard (i.e. we want to keep all the restaurant categories).

Traditionally one would evaluate a linear regression model using measures such as adjusted r squared and/or root mean squared error. However in this case, we decided that we were okay with providing a more directional prediction that was within the range of 0.5 stars above or below actual star rating. Additionally, we needed to round the predicted results down to the 0.5 level. Below in Table 6 is our approach based on our predictions on the test set.

Table 6. Prediction Accuracy on Test Set



Delta: Predicted Stars Vs. Actual	# of Businesses	% Total	Notes
-2.5	1	0%	
-2	10	1%	
-1.5	48	6%	
-1	120	14%	
-0.5	265	32%	Half a Star Under
0	208	25%	Exact Accuracy
0.5	123	15%	Half a Star Over
1	45	5%	
1.5	9	1%	
Test Set Total	829	100%	
Directionally Close:	*596*	*72%*	

7.3 Model Application

The resulting linear regression model was then used to generate predictions for the star rating a given type of restaurant would be expected to have based on its location (represented in this model by a selected zip code in Phoenix, AZ). A data file was generated that contained an entry for each of the identified restaurant types as if they were in each of the Phoenix zip codes. This data file was then processed and scored by the linear regression model to obtain predicted restaurant star ratings. This information is presented on a map in the resulting dashboard and allows a prospective restauranteur to visually look for locations in a given area. For example, if someone is interested in opening a Japanese restaurant, they can look at a map of predicted star ratings for a Japanese restaurant in each of the zip codes in Phoenix. A zip code with a high predicted rating indicates that an area may be a good location. This kind of visualization allows users to quickly review the predictive results of the linear regression model. (For the purposes of this predictive modeling, it was assumed that the restaurateur would provide parking, not be a fast food restaurant, not be a chain restaurant, and be open each day of the week, but only serve dinner.)

8. Text Analytics

8.1 Sentiment Analysis

To start off sentiment analysis, the very first lexicon we have used is the Afinn lexicon. Afinn has been recognized as a simple but popular lexicon. The lexicon is recommended to be used with word scores between -5 to 5. Considering the amount of text data involved in our Yelp reviews and tips dataset, we look to analyze if the Afinn lexicon is a good fit to classify our data into positive and negative sentiments. To start off, we created subsets of data from the merged business, review and tips data (other datasets were excluded). The subsets were created based on the business star rating of the businesses. Therefore, we created 5 separate datasets for each rating. We began our analysis with the 3 star dataset. We chose this dataset because it had about 34461 records, second highest in the group and was faster to evaluate. 4 star rating had the highest with 152081 records and 1 star rating had the lowest.

Afinn scoring was performed on the 3 star dataset and the sentiment score and sentiment category lists were created. Sentiment category consisted of positive, negative and neutral records based on scores > 0, < 0 and 0. A word count for each of the text records in the dataset was also calculated. The sentiment score, sentiment category and word count were added to a new dataset with the business ID, category and text information. Since the text records were going to be larger, to get the appropriate scores for these text records, an adjusted sentiment score was calculated by dividing the original sentiment score by the word count of the text. Adjusted sentiment category was then created with the new scores. Below are some results of the Afinn analysis based on the above steps -

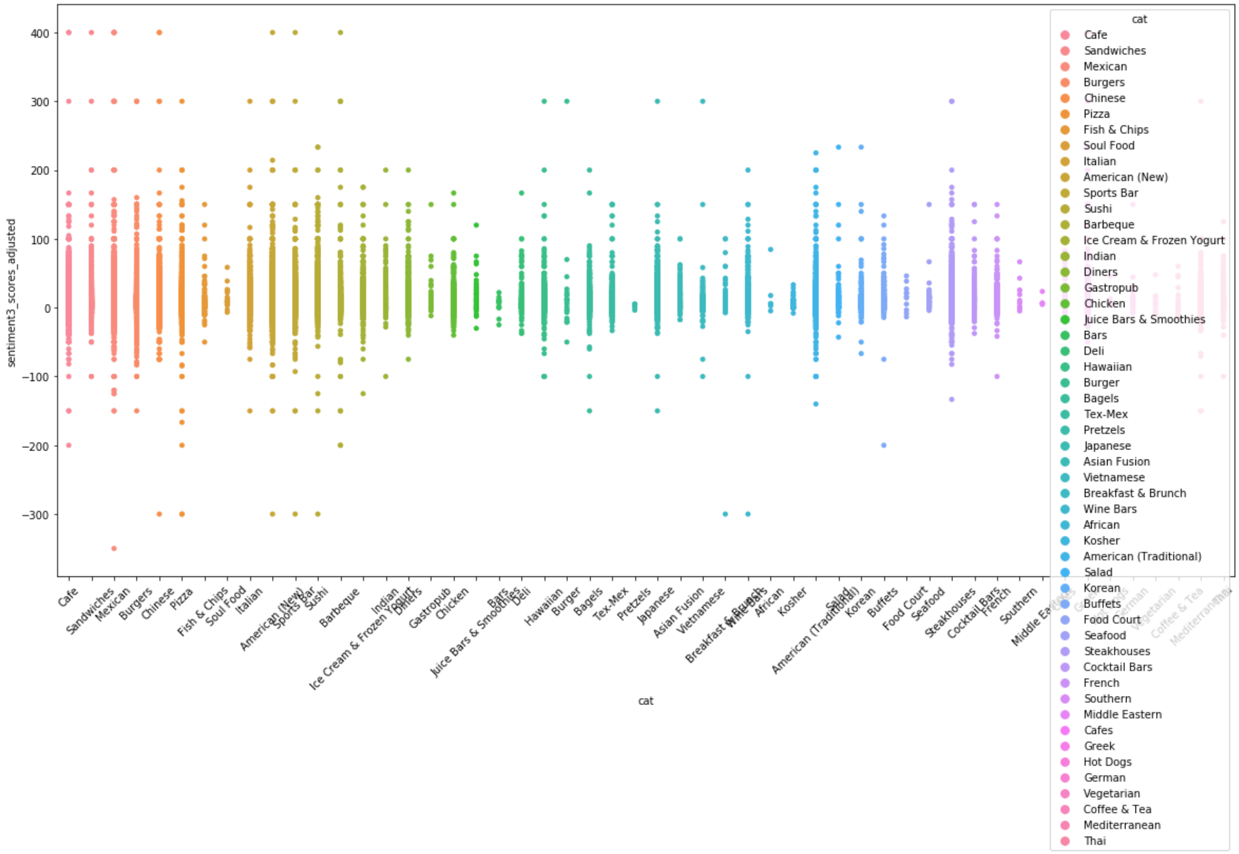


Figure 35. Sentiment Score by Categories.

The strip plot of Sentiment Score by Categories, Figure 35, displays the distribution of sentiments score across Restaurant categories. As can be seen on an average the sentiment scores range between -100 and 200 on an average. Cafe, Sandwiches, Mexican, Chinese, Italian, American New, Sports Bar and Barbecues are the categories that tend to have much higher positive sentiment scores. Mexican, Pizza, American(New) Sports Bar and Sushi categories seem to have the lowest negative sentiment scores. It can be seen from the above that categories like Mexican, American(New), Sports Bar tend to have a much higher polarity.

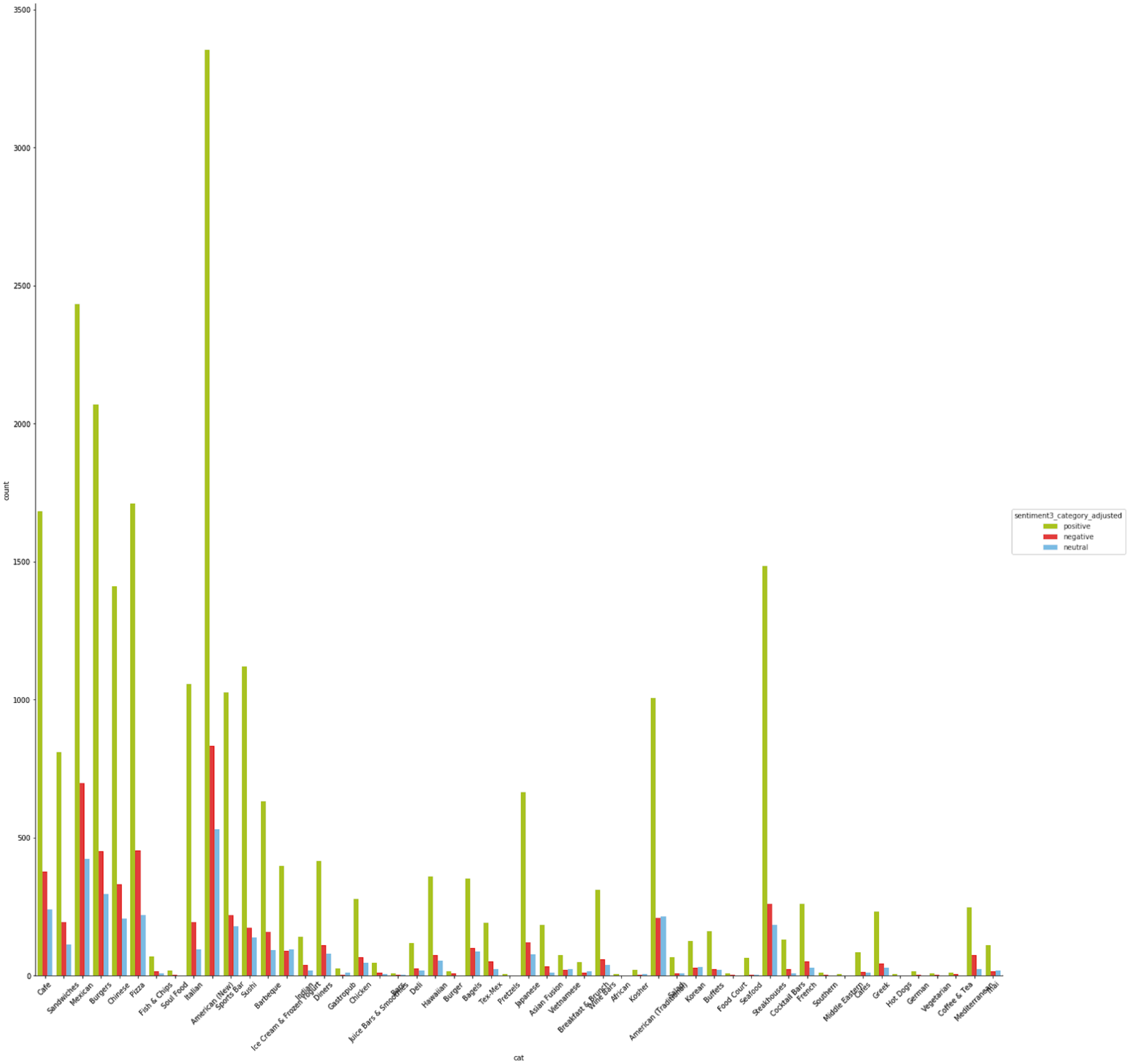


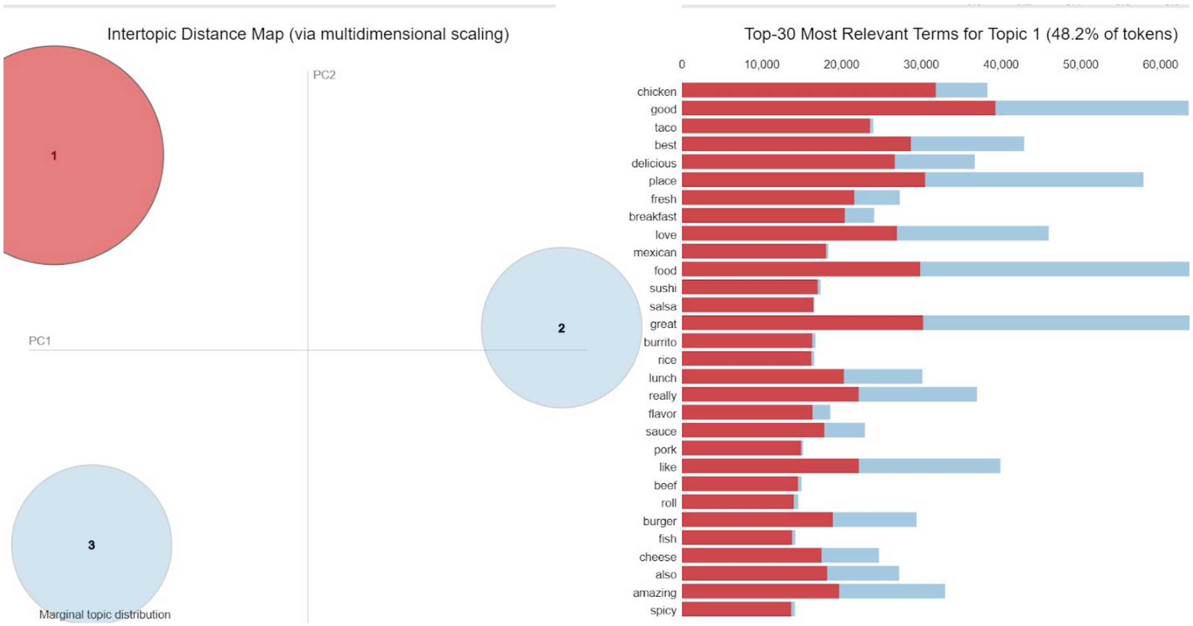
Figure 36. Sentiment categories by Restaurant Categories.

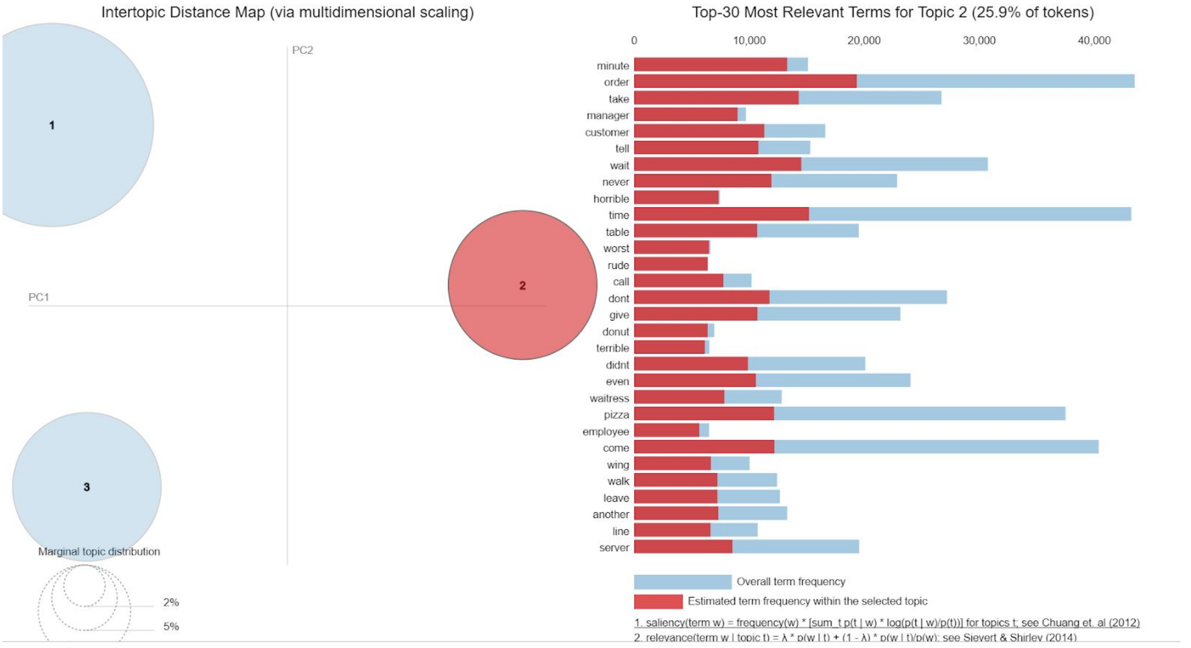
Inferences that can be made from the Sentiment categories by Restaurant Categories factor plot are a little different. Italian category has the highest number of positive reviews/tips followed by Mexican, Burgers and Pizza. American New category seems to have the highest number of negative reviews.

Since we identified that the Afinn model scores were too high. We decided to utilize Vader lexicon to do further analysis. Vader analysis identified the positive, negative and neutral sentiments in each of the reviews along with a compounded vader score. The compounded vader score was then utilized to redefine the user rating labels. If the vader compound score was less than -0.5, it was a negative rating, between -0.5 and 0.5 was a neutral rating and greater than 0.5 was a positive rating. The 0.5 value classification have been defined as default classification values by the writers of the vader lexicon. The neutral ratings were then dropped off the entire dataset to help identify positive and negative sentiments through supervised classification techniques. The revised vader user ratings were defined as the labels for these classification techniques. Logistic regression classification, random forest classifier, support vector machines and xgboost were used as classification models. In the first round of analysis, the prediction accuracies were identified to be extremely high with the random forest classifier being the highest at 97%. However, this was before the updated user star ratings were used for the vader analysis. With the revised star ratings, the logistic regression prediction accuracies went down to 87% which were still relatively good. The prediction accuracies for other models couldn't be recalculated due to compute capability issues. The sentiment analysis modeling had the capability to provide revised business ratings which could help decide how sentiments of people affect business. However, due to lower processing power, these conclusions are not available as part of these findings.

8.2 Topic Modeling

Topic Modeling was also implemented on the entire data. The text corpus was first scrubbed to create tokens, remove punctuation and lowercase all text and then lemmatized to convert the text data of the reviews and tips. Bag of words models and Terms Frequency-Inverse Data Frequency (TF-IDF) vectorization was run on these text pieces. Finally, Latent Dirichlet Analysis(LDA) and Latent Semantic Analysis(LSA) Models were run on this data to identify the top 3 prominent topics in the data and also the 30 most prominent words in these different topics. Below are the topic distributions with their words -


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 1 (48.2% of tokens)
0
10,000
20,000
30,000
40,000
50,000
60,000
chicken
good
taco
best
delicious
place
fresh
breakfast
love
mexican
food
sushi
salsa
great
burrito
rice
lunch
really
flavor
sauce
pork
like
beef
roll
burger
fish
cheese
also
amazing
spicy


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
Top-30 Most Relevant Terms for Topic 2 (25.9% of tokens)
0
10,000
20,000
30,000
40,000
minute
order
take
manager
customer
tell
wait
never
horrible
time
table
worst
rude
call
dont
give
donut
terrible
didnt
even
waitress
pizza
employee
come
wing
walk
leave
another
line
server
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

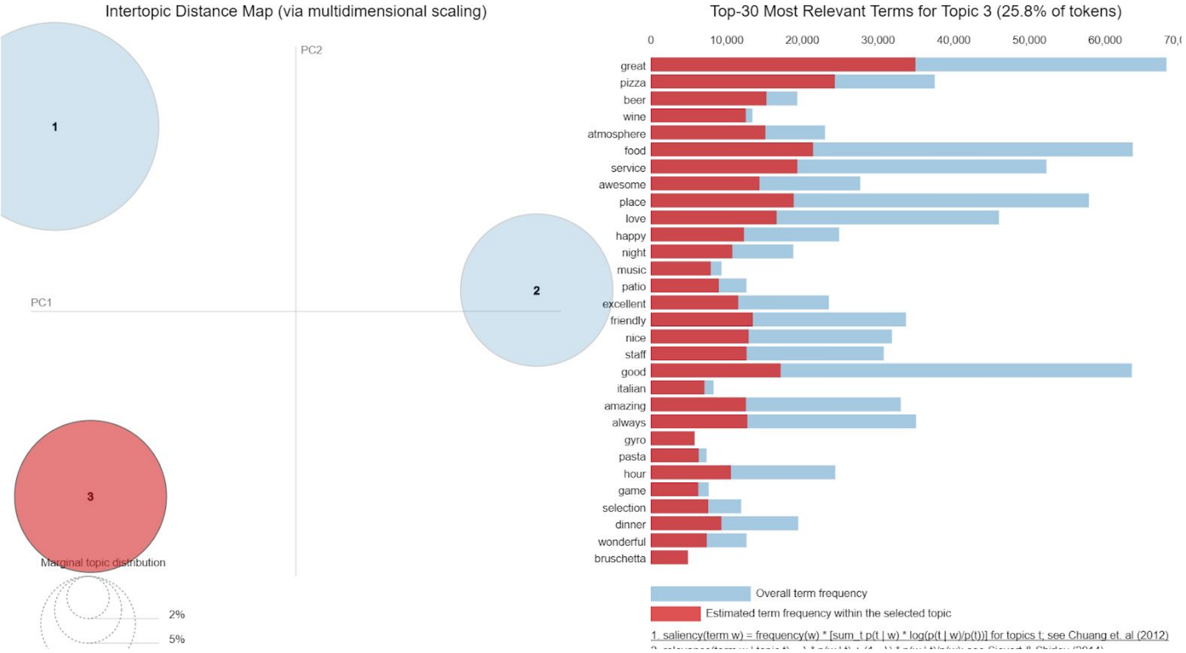


Figure 37. Top 3 Topics with the most relevant terms

As can be seen from the topics above, the 1st most relevant topic talks mostly about some of the prominent cuisines like Mexican, Sushi, Burger and also good experiences indicating this topic is mostly found in positive reviews. The second topic seems to be associated with negative experiences at restaurants and mostly negative reviews. Finally the third topic seems to be associated with positive ambience experiences and also is associated with Pizza, Beer and Bars. All three topics give interesting insights into what cuisines and experiences are most liked by them and also what brings out a negative experience for customers. When an individual is looking to make decisions, information from these topics can help them decide what would make the experience of their customers a positive one.

9. Conclusions

Locationally believes that at the end of this proof of concept we have created a holistic picture that shows the customer segments that are relevant in the Phoenix area, the optimal location to open a potential restaurant, and current customers are saying about the existing restaurants. We believe that a combination of all three will allow potential restaurateurs and investors to make more informed decisions about their ventures.

The biggest hurdle was getting the right data and transforming it into a working dataset ready for analysis. The Yelp dataset was compressed using a nested JSON format. To parse and flatten this data out took considerable time. However, this allowed for the utilization of additional attributes. Next, joining that data to external sources such as the U.S. Census data increased the data drastically by enhancing the number of features. All of these data manipulation and munging tasks required work from all team members. This ensured the clustering and predictive models would have more than enough features to be fed in.

In both modeling methodologies, it was found that less is more. The simpler we kept the model, the better they performed. For the clustering model, a simple K-Means outperformed all other models with the highest silhouette width of 0.59. When using the binary variables to predict optimal locations of a restaurant, a simple linear regression model was the best performing able to predict with a 72% accuracy the customer rating score +/- 0.5 rating.

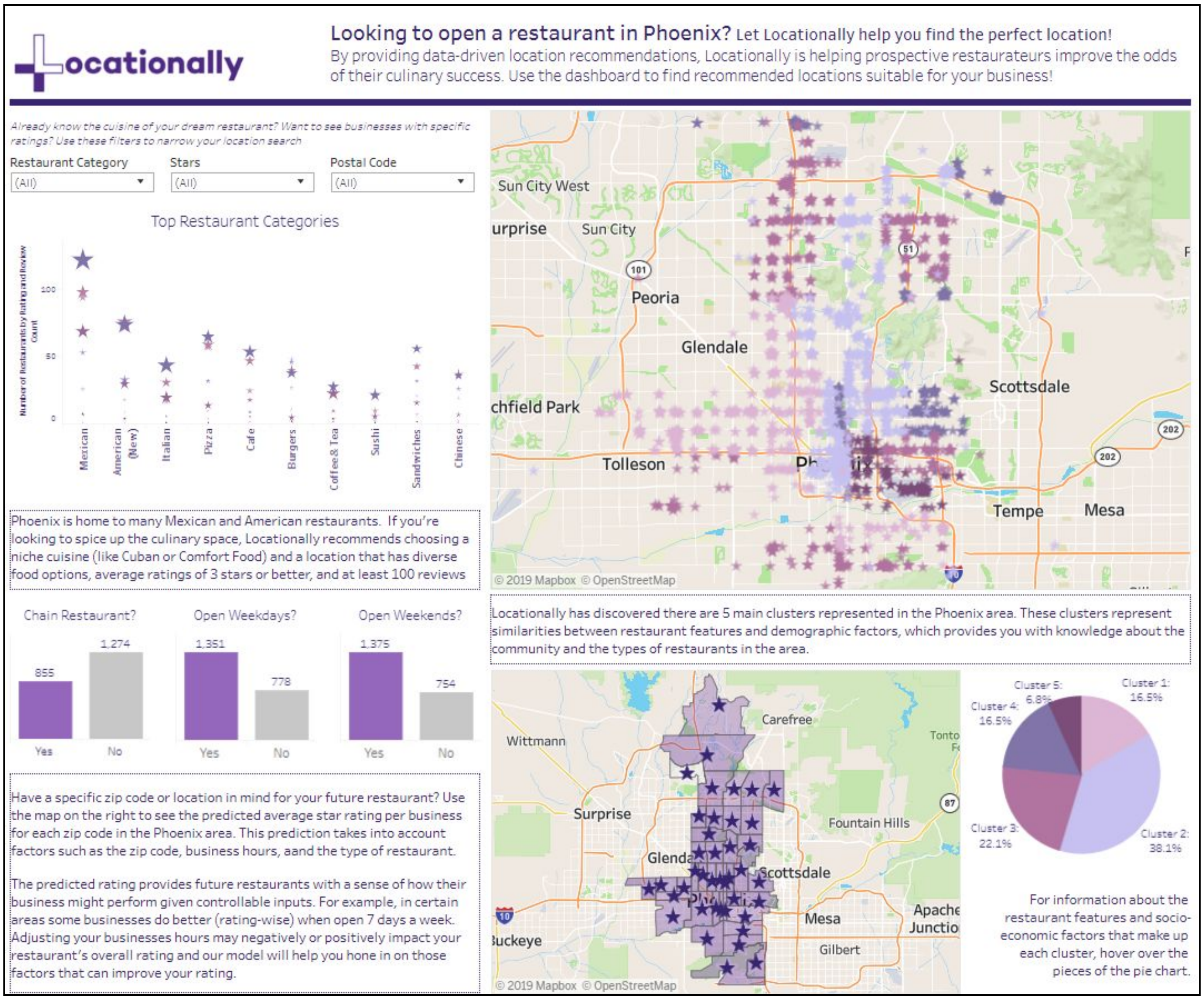
Within the customer sentiment portion, there were several business insights gathered. It was found that there are certain types of restaurants that generally have higher sentiment scores. Cafe's, sandwich shops, or American bistros tend to have higher scores. This makes sense as those are not very complex restaurant categories and generally serve standard menus. Mexican, pizza, and Italian had more polarizing opinions. In the food industry, specific food types or cuisines will have higher critics.

In the future, Locationally would like to invest in more data research or hire a research firm to conduct a study or survey or find additional data sources that can be used to enhance our models. Having more features for this type of assessment will only enable smarter training sets and better model predictions. Furthermore, Locationally will deploy its technologies to other cities in the United States and use broader national datasets to better inform its models.

10. Dashboard and Mobile User Interface

The insights Locationally has identified are accessible via a Tableau dashboard, which is publicly accessible via laptops, desktops, ad via mobile devices. The dashboard contains a map view of the zip codes that comprise the city of Phoenix, AZ and allow a prospective restauranteur to visualize the location of different types of existing restaurants, demographic information (via the identified demographic clusters), and predicted scores for different types of restaurants in different zip codes. This dashboard provides a holistic view of considerations one must take into account when evaluating the location and type of restaurant.

Link to dashboard.



Looking to open a restaurant in Phoenix? Let Locationally help you find the perfect location!
By providing data-driven location recommendations, Locationally is helping prospective restaurateurs improve the odds of their culinary success. Use the dashboard to find recommended locations suitable for your business!

Already know the cuisine of your dream restaurant? Want to see businesses with specific ratings? Use these filters to narrow your location search

Restaurant Category	Stars	Postal Code
(All) ▼	(All) ▼	(All) ▼

Phoenix is home to many Mexican and American restaurants. If you're looking to spice up the culinary space, Locationally recommends choosing a niche cuisine (like Cuban or Comfort Food) and a location that has diverse food options, average ratings of 3 stars or better, and at least 100 reviews

Locationally has discovered there are 5 main clusters represented in the Phoenix area. These clusters represent similarities between restaurant features and demographic factors, which provides you with knowledge about the community and the types of restaurants in the area.

Have a specific zip code or location in mind for your future restaurant? Use the map on the right to see the predicted average star rating per business for each zip code in the Phoenix area. This prediction takes into account factors such as the zip code, business hours, aand the type of restaurant.

The predicted rating provides future restaurants with a sense of how their business might perform given controllable inputs. For example, in certain areas some businesses do better (rating-wise) when open 7 days a week. Adjusting your businesses hours may negatively or positively impact your restaurant's overall rating and our model will help you hone in on those factors that can improve your rating.

For information about the restaurant features and socio-economic factors that make up each cluster, hover over the pieces of the pie chart.

Figure 38. Dashboard.

References

1. Yelp, Inc. (05 February 2019). *Yelp Dataset.* Retrieved from https://www.kaggle.com/yelp-dataset/yelp-dataset.
2. United States Census Bureau. (2010). *USA 2019 Census Dataset.* Retrieved from https://www.census.gov/data.html
3. VRBO. (2019). *Phoenix.* Retrieved from https://www.vrbo.com/vacation-rentals/usa/arizona/phoenix-area/phoenix.
4. Google Maps. (2019). *Phoenix, Arizona.* Retrieved from https://www.google.com/search?q=google+maps+phoenix&rlz=1C1GGRV_enUS796US796&oq=google+maps+phoenix&aqs=chrome..69i57j0l2j69i60l3.4377j0j7&sourceid=chrome&ie=UTF-8.
5. United States Census Bureau. (2018). *QuickFacts* | *Phoenix, Arizona.* Retrieved from https://www.census.gov/quickfacts/phoenixcityarizona.
6. Wu, et al. *A New Semantic Approach on Yelp Review-star Rating Classification.* Retrieved from https://pdfs.semanticscholar.org/3ff1/5b9956aaaf501c0a47d040b09626797984df.pdf
7. Yu, et al. (September 2017). *Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews.* Retrieved from https://arxiv.org/ftp/arxiv/papers/1709/1709.08698.pdf.

Appendices

- **Appendix A,** Python Data Preparation Script V5 Code
    - Github link
- **Appendix B,** Python Mapping Script Code
    - Github link
- **Appendix C,** Python Analysis and Feature Engineering
    - Github link
- **Appendix D,** R Clustering Analysis Code
    - Github link
- **Appendix E,** R Regression Code
    - Github link
- **Appendix F,** Python Sentiment Analysis Code
    - Github link
- **Appendix G,** Python Topic Modeling Code
    - Github link