

Project final report towards partial fulfillment of the course MSDS498

Let Us Yelp You:

Leveraging Data from Yelp to Create New Restaurants

MSDS 498, SEC58, WI2019
SPS, Northwestern University, Evanston, IL

Team Member 1: Aamer Khan, (Team Lead, Research/Communication/ Document Preparation)

Team Member 2: Crystal Mosley, (Senior Analyst, Communication/ Document Preparation)

Team Member 3: Laketa Hendrix, (Senior Analyst, Research/ Document Preparation)

Team Member 4: Saranya Vaidyalingam, (Tech Lead, Python Programmer, Research/Document Preparation)

Faculty Advisor: Nethra Sambamoorthi, Ph.D

Team Roles and Responsibilities

Team Member	Role	Responsibility
Aamer Khan	Team Lead, Researcher, Writer	Ensure completion of project, analyze data, validate results, and keep client and advisor communications
Crystal Mosley	Senior Analyst, Communications, Writer	Acts as the intermediary between our team and client. Contributes to the overall document prep and communication.
Laketa Hendrix	Senior Analyst, Researcher, Writer	Collects the data, analyze data, and validate results. Contribute project artifacts.
Saranya Vaidyalingam	Tech Lead, Programmer, Writer	Codes & Creates models, collect data, analyze data, and validate results. Contribute project artifacts.

Index

Team Roles and Responsibilities	2
Abstract	5
Introduction	6
Profile of Let Us Yelp You	6
Research Questions and Key Metrics	8
Research Questions.....	8
Key Metrics	8
Data Requirements and Strategies	10
Metadata and Data Dictionary	10
Issue Tree	15
Data Strategy	16
Data QA Report	16
Variable Reduction.....	22
Statistical Measurements	24
EDA	25
Data retrieval.....	25
Data Integration	25
Data EDA and Transformations	25
Training/Test Data Sets	27
Methodology, Visualizations and Implementation	29
Computational Method	29
Results Visuals	30
Implementation	31
Fitted Models	32
Project Plan	35
Project Management Approach.....	35

Project Deliverables	36
<i>Dashboard (Mockup)</i>	38
Project Requirements.....	40
Project Constraints.....	40
Project Assumptions	41
Project Risks	41
Gantt Chart	42
Cost	42
Resources.....	42
Endnotes	44
APPENDIX	45

Abstract

The purpose of this analysis is to consult our client, Yum! Brands on which characteristics contribute to earning a highly-rated Yelp status for restaurants in Phoenix, AZ. Our client will use our analysis as an input into the decisions they make when opening a new restaurant in Phoenix. We worked with open-source data acquired from the Yelp developer portal site. Four separate tables were evaluated, with two tables being useful for the analysis. The two useful tables were joined and nested fields were parsed to allow more extensive evaluation. After variable reduction was performed to eliminate useless fields and those with a high missing-value percentage, logistic regression was performed. The results of the logistic regression provided which variables were influential in determining whether a restaurant would be highly-rated on Yelp and then with each combination of characteristics, our model determined what percentage chance the restaurant has to be highly-rated. The key metric for a restaurant being highly-rated is *stars*, the average star rating for the business. The analysis culminated in an interactive Restaurant Calculator dashboard.

Introduction

Introduction

Opening a restaurant can often be intimidating; the list of supplies, staff, and work needed to make it successful can often seem endless to the owner. There are apparent necessities such as linens, plates, busboys, servers, menus, chefs, ingredients. However, one thing that isn't so obvious in a successful restaurant is the use of analytics. According to a 2005 study published by H.G. Parsa, et al. in the Cornell Hotel and Restaurant Administration Quarterly, it is approximated that 30 percent of independent restaurants fail within the first year during the period covered by the study¹. As a business, the goal is to reduce your overall risk and increase profit, and that is where Let Us Yelp you can help you lead the way by helping your restaurant steer clear of these points of failures.

Let Us Yelp You is a marketing and analytics firm that utilizes openly sourced datasets such as data from Yelp, to give restaurants the competitive advantage by outsmarting their competition. We understand and break down the needs of every restaurant chain and create a personalized plan that brings their vision to life using openly sourced Yelp data. The use of analytics can be beneficial to restaurant owners because it allows them to gain a deeper understanding of their customers and allows them to cut down on wasteful spending.

Profile of Let Us Yelp You the Organization and Background of the Opportunity

Yum!, a restaurant chain is actively looking for the best neighborhood in Phoenix, AZ to open a new successful restaurant. Our goal at Let Us Yelp You is to help them determine the best attributes that their restaurant should have to compete with the top restaurant in the area. We plan to go to the drawing board and gain an understanding of Yum!'s vision and come up with a personalized plan that helps them achieve their goal. We will use the Yelp dataset to figure out certain features that are critical factors to a successful restaurant in that market. We have a multi-faceted analytical approach which will involve a preliminary exploratory data analysis, visualizations, word clouds, predictive modeling, and natural language processing of reviews. This information is provided and explained to the customer, which will allow them to make an informed decision on achieving their goals for their restaurant.

Research Questions and Key Metrics

Research Questions

Our broad research questions are based around what value the Yelp datasets can provide. As experts of this data, we first use our knowledge to determine which questions the data can answer and if those answers would benefit our client. Some research questions that Yelp datasets can answer:

- What are some common characteristics of top Yelp rated restaurants?
- What are some common characteristics of Yelp users?
- What are the essential concerns of Yelp users?

Our analysis will serve as an extra layer of insights by which to promote a competitive advantage in the marketplace. All the insights as mentioned earlier would be useful information for our client. By understanding the characteristics of which restaurants get top ratings, along with the profile of users and their concerns, our client can make more informed decisions. Additionally, as a result, they may see higher Yelp reviews which may contribute to a positive online brand presence.

Key Metrics

Our primary metric will be Yum! opening a restaurant in Phoenix that garners a 4 to a 5-star rating on Yelp. We will use the various Yelp datasets first to determine which variables will provide us with the best indicators of higher overall star ratings for restaurants in Phoenix. Using this information, we will assess our key measurements. As we would like to provide a layer of insights for decision-

making and competitive edge, we will also take special care to ensure accurate information. The following accuracy measures will be reported: Missing Value Percentage and Model Accuracy (dependent on model type).

There are some variables in the datasets that we are already aware will be important to our analysis. Primarily, the "Stars" field in the Review dataset that ranges from 0 to 5 and will be our key metric. This information will be used as an input measure to categorize restaurants. Restaurants with 4 or more stars will be clustered and analyzed. Within the 4-star category of restaurants, the cluster analysis will yield insights largely measured by term frequencies. For instance, commonly used words for a restaurant cluster characterized by its great food may be "tasty," "best," or "delectable." We may also use measures of positive sentiments or negative sentiments. It is important to note that custom measures can also be made for our client. They will be created based on client interests, delving deeper into specific insights into the request.

Data Requirements and Strategies

Metadata and Data Dictionary

Our primary source of data is the datasets provided by Yelp. There are five different datasets including business, check-in, photo, review, tip, and user. Our primary source will be the review data. This dataset contains the full text of the reviews on Yelp with the unique Yelp user id and business id of the restaurant being reviewed. This dataset also includes the number of stars awarded to the restaurant as well as the amount and type of votes the review received.

Field Name	Data Type	Data Format	Description	Example	Derived?	Direction of impact on response variable (stars)
review_id	string		Unique review id	zdSx_SD6obEhz9VrW9uAWA	N	-
user_id	string		Unique user id	Ha3iJu77CxlRfm-vQRs_8g	N	+
business_id	string		Unique business id	tnhfDv5II8EaGSXZGiuQGg	N	-
stars	integer		Rating	3	N	neutral
date	string	YYYY-MM-DD	Date of review	2018-02-24	N	-
text	string		Review text	Great place to eat	N	-

useful	integer		Number of useful votes	1	N	+
funny	integer		Number of funny votes	0	N	+
cool	integer		Number of cool votes	0	N	+

The business dataset contains information on each restaurant reviewed. There are data on the physical aspects of the restaurant such as an address. There is also data on the category of restaurant, hours open as well as other key feature such as the type of parking available if Wi-Fi is available and more. These data give us a clear picture of the restaurant, its amenities, and general classifications.

Field Name	Data Type	Data Format	Description	Example	Derived?	Direction of impact on response variable (stars)
business_id	string		Unique business id	tnhfDv5lI8EaGSXZGiuQGg	N	-
name	string		Business name	The Restaurant	N	-
address	string		Business address	253 4 th St	N	-
city	integer		City	Algonquin	N	+
state	string		State	IL	N	-
postal code	string		Postal Code	60056	N	-

latitude	float		Address latitude	37.7817529521	N	-
longitude	float		Address longitude	-122.39612197	N	-
stars	float		Average rating	4.3	N	+
review_count	integer		Number of reviews	526	N	+
is_open	integer	1 or 0	Business open	1	N	+
attributes	object			{"RestaurantsTakeOut": true,},	N	+
categories	array of strings			"categories": ["Italian",],	N	-
hours	object			"hours": { "Monday": "10:00-21:00", "Tuesday": "10:00-21:00",	N	+

The user dataset consists of profile information on the users including the user id, name, total number of reviews posted as well as details on the votes and compliments the user has received.

Field Name	Data Type	Data Format	Description	Example	Derived ?	Direction of impact on response variable (stars)
------------	-----------	-------------	-------------	---------	-----------	--

user_id	string		User id	Ha3iJu77CxlRm-vQRs_8g	N	-
name	string		User name	John Doe	N	+
review_count	integer		Number of reviews	92	N	+
yelping_since	integer	YYYY-MM-DD	Date joined	2013-05-25	N	+
friends	array of strings		User IDs of friends	"friends": ["wqoXYLWmpkEH0YvTmHBsJQ"],	N	+
useful	integer		Useful votes sent by user	6	N	+
funny	integer		Funny votes sent by user	9	N	+
cool	integer		Cool votes sent by user	7	N	+
fans	integer		User's fans	15	N	-
elite	array of integers		Years user was elite	"elite": [2012, 2013],	N	-
average_stars	float		Average of all ratings	4.56	N	+
compliment_hot	object		Hot compliments received by user	1	N	-

compliment_more	array of strings		More compliments received by user	0	N	-
compliment_profile	integer		Profile compliments received by user	9	N	-
compliment_cute	integer		Cute compliments received by user	5	N	-
compliment_list	integer		List compliments received by user	4	N	-
compliment_note	integer		Note compliments received by user	0	N	-
compliment_plain	integer		Plain compliments received by user	1	N	-
compliment_cool	integer		Cool compliments received by user	4	N	+
compliment_funny	integer		Funny compliments received by user	2	N	+
compliment_writer	integer		Writer compliments received by user	12	N	-

compliment_photo	integer		Photo compliments received by user	0	N	-
------------------	---------	--	------------------------------------	---	---	---

The remaining two datasets will likely be used less extensively. The tip dataset has the text tips left by the user for a business along with the compliments that tip has obtained. These tips are generally shorter than the reviews and are often suggestions about the business. The check-in dataset is a listing of the check-ins for each company.

If the scope of our project requires it, we may also use additional datasets. The photo dataset included within the more significant Yelp data contains photos uploaded with captions, the business id, and then the photo categories. While this information is not currently foreseen to be necessary for this project, it can potentially serve as an additional resource. Furthermore, we also have access to restaurant inspection information that may provide another context if needed.

Issue Tree

An issue tree can be utilized to create a data strategy. Defined explicitly, a data strategy is a way of figuring out what data are needed and what you will negotiate to use for the project given the time frame, budget, and resources. However, in the case of our project with our client YUM! we have already collected our data from the beginning. We were provided the precompiled data which was taken directly from Yelp's website as an open source dataset. We are unable to ask Yelp for any additional data if anything is missing. On the bright

side, they have provided us with simple documentation that explains the structure of their dataset. Nonetheless, some challenges may arise from utilizing existing data such as only being able to answer questions pertaining to a narrower scope.

Data Strategy

The methodology to create a solution will consist of a multiphase approach. We begin by performing an exploratory data analysis (EDA) on the Yelp dataset. This examination will help us get a better understanding of each JSON files and its contents. During the EDA we hope to find opportunities to make comparisons between various features or predictor variables and our response variable. As a result, it will drive us further on understanding what influences the ratings of such successful restaurants. We aim to figure out what the typical user, business, review, and check in looks like throughout our sample data. This analysis will be followed up by some form of feature engineering in which we will use domain knowledge of the data to create features that will make machine learning algorithms have higher fidelity for accuracy in our predictive model. Through an iterative process that begins with EDA, we will make use of visualizations and word clouds to help drive our decision on selecting the best predictive model for the intended goal of our client.

Data QA Report

Our primary datasets will be coming from Yelp directly. As such, they are delivered as five individual nested JSON files. We will also receive a data dictionary detailing the specific variables in each file and the count of total

records included. We will use the other datasets if the need arises, but the bulk of our analysis will be conducted on the Review and Business datasets. The Review dataset was stated to contain 5,261,668 records with nine columns and the Business dataset to contain 188,593 records and 15 columns. In our preliminary checks, we found our data matches these parameters. We pulled the top and bottom ten records for both JSON files and verified that it matched the specifics of the data dictionaries provided. (See Appendix on page for the top 10 records) These data encompass Yelp reviews for all North America and include all businesses. For our purposes, we are only interested in restaurant reviews for Phoenix, Arizona. To this end, we have merged our two datasets on the business id. We have also filtered our data to only include records where the state is “AZ,” the city is “Phoenix,” and the categories contains “Restaurants.” This filtering leaves us with 486,825 records and 23 columns of data including two nested variables.

The numerical variables in the merged dataset are coded numerically generally as integers. The latitude, longitude, postal code and average stars for the restaurant are coded as a float. Except for postal code, all the numerical variables are entirely populated and have no missing data.

	star s_x	use ful	fun ny	coo l	is_o pen	latitu de	longit ude	postal_c ode	review_c ount	avg_st ars
--	---------------------	--------------------	-------------------	------------------	---------------------	----------------------	-----------------------	-------------------------	--------------------------	-----------------------

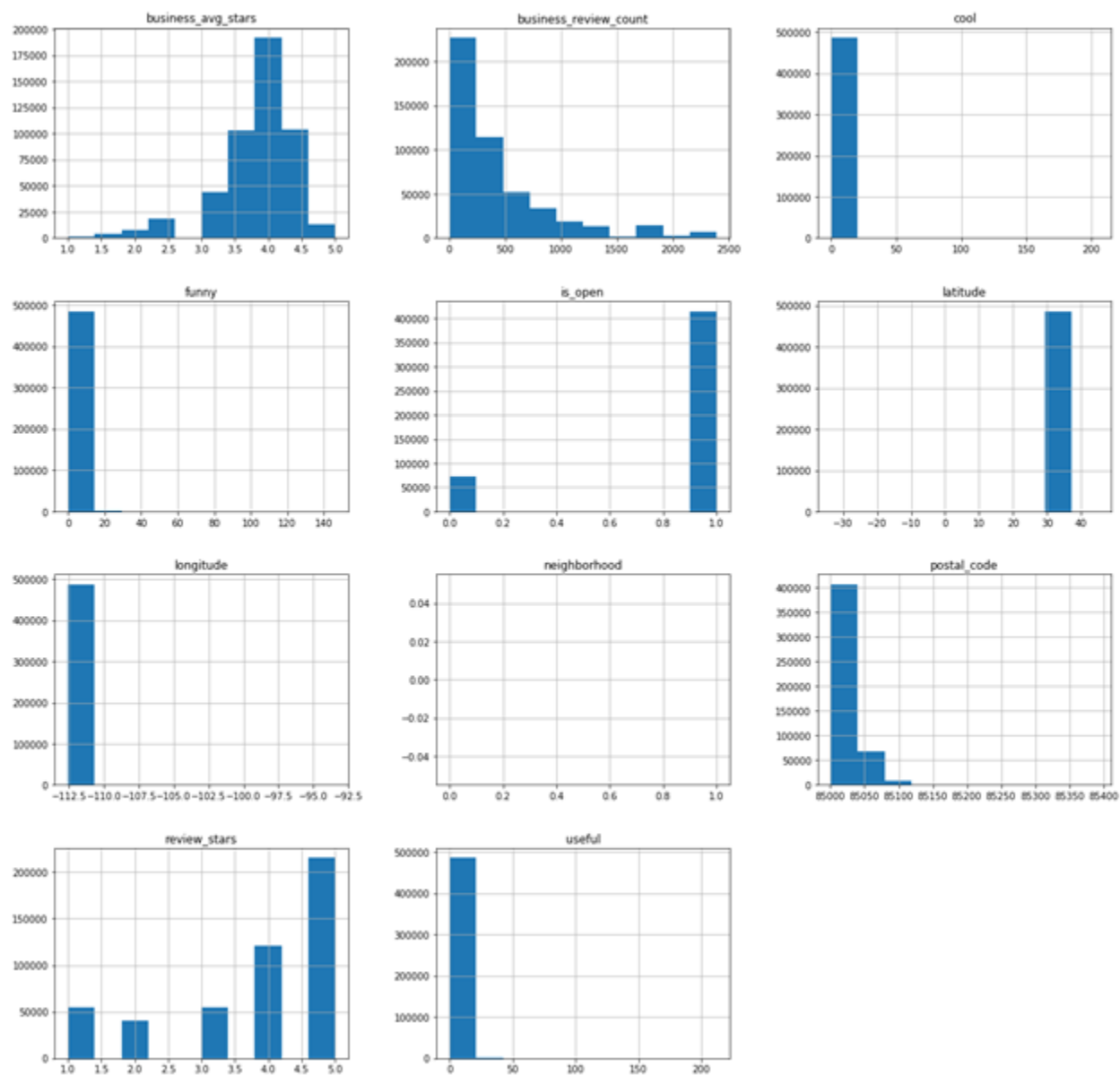
count	486825	486825	486825	486825	486825	486825	486825	486564	486825	486825
mean	3.821833	1.268756	0.514074	0.631644	0.852015	33.50977	-112.05	85024.03	432.754	3.825615
std	1.370398	2.911303	1.876417	2.282446	0.355086	0.614360	0.064938	26.67360	474.4231	0.637123
min	1.000000	0.000000	0.000000	0.000000	0.000000	-33.5086	-112.572	85001	3.000000	1
25%	3.000000	0.000000	0.000000	0.000000	1.000000	33.45885	-112.074	85012.00	110.0000	3.5
50%	4.000000	0.000000	0.000000	0.000000	1.000000	33.50579	-112.063	85018.00	267.0000	4
75%	5.000000	2.000000	0.000000	1	1	33.57084	-112.005	85032	560	4
max	5.000000	212	146	205	1	45.0175	-93.4598	85392	2391	5

Only the neighborhood, hours, attributes, address and postal code variables have missing values. The neighborhood variable is completely missing and will be dropped from further analysis. The remaining variables are missing values within our threshold of 50% and will be used as needed. If more of the data could be filled in, it will lead to better outcomes from using the data. However, since we are using open source data, we are not in the position to bring about changes to the inputs by communicating with the IT department. As it stands, we will have to decide to impute or delete variables with missing values above our threshold.

	Missing Values	% of Total Values
neighborhood	486825	100.0
hours	28889	5.9
attributes	6189	1.3
address	4672	1.0
postal_code	261	0.1

To check for consistency of our Yelp dataset, we were able to cross check a few random restaurants in Phoenix to see if those establishments and their reviews matched what we had in our dataset. From our initial check, everything seems to be consistent.

To build a model, we must take into consideration the distribution of our data points. We have created distribution plots for each of our numerical variables. Analyzing these will help us determine if we need to transform our variable to facilitate better analysis or remove the variable entirely.



We also un-nested and created frequency tables for our one of our significant categorical variables, categories. We have 416 unique categories in our dataset. We have listed the top 10 categories with their frequency for our merged dataset. As part of our continued data exploration, we will also un-nest the attributes variable and use frequency to help analyze our data.

<i>Categories</i>	<i>Frequency</i>
<i>Restaurants</i>	485974
<i>Nightlife</i>	124906
<i>Bars</i>	121723
<i>Food</i>	108349
<i>American (New)</i>	103844
<i>Breakfast & Brunch</i>	93881
<i>American (Traditional)</i>	82348
<i>Mexican</i>	80715
<i>Sandwiches</i>	75193

Since our goal is to help Yum! International open a new 5-star Yelp-reviewed restaurant in the Phoenix area, we are going to focus on Phoenix restaurants that are in the same category as Yum! International chain restaurants. We can view our sample as Phoenix restaurants that fit the relevant Yum! International categories (i.e., Fast Food, Chicken Shop, Chicken Wings, Taco, Tex-Mex, Pizza, Italian, Mexican, Food Delivery Services) and the total population is Phoenix restaurants of any category. Our sample will be the target data for our predictive modeling. We can note that the composition of reviewers for all restaurants and determine if weighting our sample using an auxiliary variable to

compensate for the lack of representation is necessary. However, dates of reviews are more appropriate for weighting to make sure we have an even representation across time.

Variable Reduction

The idea of variable reduction involves the process of choosing the most relevant features for the data. We will remove features to help the model generalize better for new data, and that creates a more interpretable model. By using the test-train method, variable selection will be crucial to our machine learning model and its ability to learn from the data we provide to it and subtracting features so that we are left with the features that are the most important. Being the subject matter experts, we do an initial review of the content provided by each variable and immediately determined columns that will not be useful to our model: *user_id*, *business_id*, *address*, *city*, *state*, *hours*, *longitude*, *latitude*, *name*, *neighborhood*, and *postal_code*.

To start, we merged our Review dataset and our Business dataset by using '*Business ID*' as the unique key. This new set forms our working dataset by which we will start variable reduction. After our initial EDA, we may wish to append additional variables from our 'Users' and 'Tips' datasets, if needed to increase model accuracy. With the current merge, we are left with 23 columns. However, 'attributes' and 'categories' are nested columns. We will unpack the nested columns which will result in approximately 91 variables.

We will remove columns with a high percentage of missing values. Any variables with more than 50% missing values, will result in removal of that variable from the dataset. Also, we will remove outliers, using the fencing method. We will plot the distribution with a box plot to identify the upper and lower quartiles. Once we understand the distribution of each of our variables, we can determine whether the variable is normally distributed and will be useful to our model. If the variable distribution does not appear normal, our first action will be to transform the variable. If transformation does not result in a normal distribution, we can remove the feature from the model.

From there, we plan to explore some statistical methods for feature selection: Pearson Correlation Coefficients (PCC), Clustering, and Principal Component Analysis (PCA). For example, we can use the values of PCC to quantify the relationship between the variables. PCC is scored on a range of -1 to 1. If a relationship has a value of -1, it has a direct negative relationship. If a relationship has a value of 1, it has a direct positive relationship. Predictors that are strongly correlated with each other are known as being collinear. By considering the idea of collinearity between variables, it only helps the machine learning model general and be more interpretable. We explore collinearity by using a method called the Variance Inflation Factor (VIF). If we find predictor with collinearity, we will remove one of those variables.

Statistical Measurements

By utilizing the correlation pair matrix (refer to Appendix) to measure the shape of the distributions for skewness, and to verify the impact on the direction of our response variable, stars, we were able to come up with initial best guesses. The Review data set contains 44% positive correlations, 0.22% as neutral, and 50% as having some impact on stars. The Business data set includes 43% positive correlations, and 57% as being potentially impactful to stars. The Users data set contains 36% positive correlations, and 64% as possibly affecting stars.

EDA

Data retrieval

Yelp has provided public datasets, as mentioned previously, that our team is using to identify critical characteristics that Yum! will need to open a new successful restaurant in Phoenix and keep a competitive drive. All data are in the form of JSON files ranging from 0.05GB to 4.6GB.

Data Integration

During Phase 1, we will not be integrating the data with any source systems as we are only analyzing the data to identify critical needs for the future restaurant to be successful in one location. The next phase will utilize Yelp's Data Ingestion API which provides a means for partners to programmatically perform updates on many businesses asynchronously. Partners can deliver updates on various attributes (Yelp). This API will enable Yum!'s dashboard to continually be fed updates on restaurant data attributes within Phoenix to stay abreast of any triggers they should pay attention to for their restaurant to have a competitive advantage.

Data EDA and Transformations

It is essential to have data preprocessed and cleansed before the modeling, evaluation and deployment steps of any data analysis project. To effectively

select predictive algorithms, we need to understand the makeup of the data; therefore, we will explore each dataset for the following:

- Missing data – Any variables that contain missing data will be reviewed and handled via the below:
- Dropping variable or attributes – if the variable or attribute is missing 50% or more observations, it will be cut.
- Most frequent – we will take the most frequent data point of the variable and place this value into the missing observation with the notion that this could potentially bring noticeable variance into our model.
- Outliers – If more than 2% of a variable's data are outliers, the team will assess the variable in more detail to determine whether it can be used or discarded.

Visualizing the data is also a key component of an EDA as it will allow us to spot any patterns or connections to explore further. We will use the following visualization methods:

- Univariate visualization to summarize and potentially find any patterns.
- Bivariate visualization for assessing the relationship between each variable and the target variable – stars.
- Multivariate visualizations to understand interactions between different fields within the data.

After an initial analysis of the data, we will focus on a few more in-depth data manipulations.

- Dimensionality reduction to understand the areas that account for the most variance between observations and allow for the processing of a reduced volume of data.
- Clustering of similar observations into differentiated groupings. This way, patterns of behavior can be more easily identified.

Features within the data may have differing values/ranges; therefore, a variable transformation may be necessary before it is ready for model analysis. Here are a few variable transformation methods to consider:

- Normalization – Scaling data to handle skewness and aggregation of attributes.
 - Min-Max: Moves the values towards the mean of the column
 - Z score: If the variable has outliers, this method will be used so we don't lose the variables impact
- Skewness – Measures the asymmetry of the probability distribution of a real-valued random variable about its mean. We will utilize the square root, cube root or log method.

Training/Test Data Sets

We will split the data into three different sets:

- Training (70%)– This set will be used to fit the parameters of each classifier.
- Test (30%) – This set will be used to assess the performance of a fully quantified classifier.

Methodology, Visualizations and Implementation

Computational Method

Our target variable is Stars. Stars are the average rating for a restaurant based on the reviews it receives. Since we are giving consultation advice to help our client open a restaurant that would be highly rated, we are going to gather the data of Phoenix restaurants and use them as inputs (feature variables) in determining which factors are hugely influential in achieving high review ratings. We have set our goal rating for the new restaurant to be 4.5 or higher, and this is our definition for when we refer to a restaurant as being highly rated. Based on the distribution of ratings, this would put the new restaurant in the top 15% of highly reviewed restaurants in the Phoenix area. There are several characteristics that we want to measure to determine whether they will have a significant contribution to ratings. Ultimately, we want to end up with a list of features that we can say positively correlate to a restaurant that is highly rated. The list will note each characteristic as being either positively or negatively correlated with our target.

The modeling technique we will use is Logistic Regression. This modeling technique is commonly used when the target variable is a binary class or nominal variable. In our case, each restaurant will be classified as highly rated or not highly rated. Logistic regression will return probabilities on whether, given

specific inputs, the restaurant is likely to be classified as highly rated. We will also use text analysis to get the most common words used in reviews for highly rated restaurants.

Results Visuals

After performing our logistic regression, we check our performance with a ROC Chart and a confusion matrix. With these, we can visualize and describe the performance of our classification problem. We also have created a Gantt chart to guide our project along.

For our clients, we will provide a simple chart explaining the result of the model we implemented. This chart will give clear direction to our client advising them which variables are important in achieving a highly rated status.

Variable	% Influential	Recommendation
Variable a	30%	recommend
Variable b	21%	recommend
Variable c	6%	Do not recommend

Additionally, we provide a list of inconsequential variables to help our clients focus on more important decisions with their new restaurant.

Variable
Bike Parking
TV
Outdoor Seating

Implementation

We are contracted to guide Yum! as they explore their options for building a new restaurant in Phoenix, Arizona. We will use the results from our analysis to learn what variables impact a restaurant becoming highly rated on Yelp which is receiving a 4 or 5-star rating. We will create an interactive dashboard that allows our clients to see the viability score, projected potential customer population and the level of expected demand for the restaurant as they consider different options. They will primarily be able to choose a cuisine type and, based on that choice, the relevant variables that our analysis has deemed signifiers of higher ratings will populate. The clients will then be able to choose options within those variables. As they change their options, the key metrics will update. The viability score is the probability that their new restaurant is highly rated. Probabilities of 80% or higher will be green, 60-79% will be yellow, and all others will be red. The

projected customer population is the number of people in the population that fit the customer profile for a restaurant with the chosen variables. The projected level of demand measures the demand for a restaurant with the selected attributes in the area. The dashboard will also show the density of similar restaurants in the proposed restaurant location. For example, the client could choose to offer valet parking or not and be able to see the subsequent changes in the probability of becoming a highly rated restaurant. They can then make an executive decision based on this information. With this dashboard, we hope to provide a complete picture of the options for opening a restaurant in Phoenix.

Fitted Models

To analyze the data, we began with the Business dataset that was filtered to only include the information for all restaurants in Phoenix, AZ. After conducting some initial descriptive statistics, any variable that was missing more than 50% of data was removed. Variables that had True or False were transformed the data, so the data read as 1 or 0 respectively. Variables that contained categorical data were turned into separate columns. For example, the WiFi column included information that was either 'Yes,' 'No' or 'Free.' These features became three separate columns titled WiFi_no, WiFi_Yes and WiFi_Free with 1s or 0s as indicators. For variables that still contained missing data, an imputation using the most frequent data point was conducted. The target variable was set as restaurants with a 4.5 or higher rating having a 1 and 0 for all others. We began

with a linear regression using all remaining variables and further regressions with variables added or removed with the goal of improving our adjusted R-squared value. While we were able to develop this score to some degree, as the nature of our data was not quite linear in life, we chose to use logistic regression. We used the RFE feature selection from *sklearn* in Python to reduce our variables before running the regression. Our data were also split into test and training data with 30% and 70% respectively split. We were able to reach 67% accuracy on our test dataset.

Fitted Model for Linear Regression (OLS):

TARGET = Restaurants with 4.5-5 Star Ratings

$$\begin{aligned} \text{TARGET} = & 0.0285 \times \text{OutdoorSeating} + 0.0439 \times \text{RestaurantsReservations} + \\ & 0.1116 \times \text{Caters} - 0.0172 \times \text{NoiseLevel_average} - 0.0503 \times \text{NoiseLevel_loud} - \\ & 0.1115 \times \text{NoiseLevel_very_loud} - 0.0737 \times \text{RestaurantsAttire_casual} + 0.1565 \times \\ & \text{Street_Parking_ Fals} + 0.2656 \times \text{Street_Parking_ True} - 0.1084 \times \\ & \text{Alcohol_full_bar} + 0.3183 \times \text{intimate_ True} + 0.1093 \times \text{hipster_ True} + 0.1104 \times \\ & \text{classy_ True} + 0.1394 \times \text{upscale_ True} \end{aligned}$$

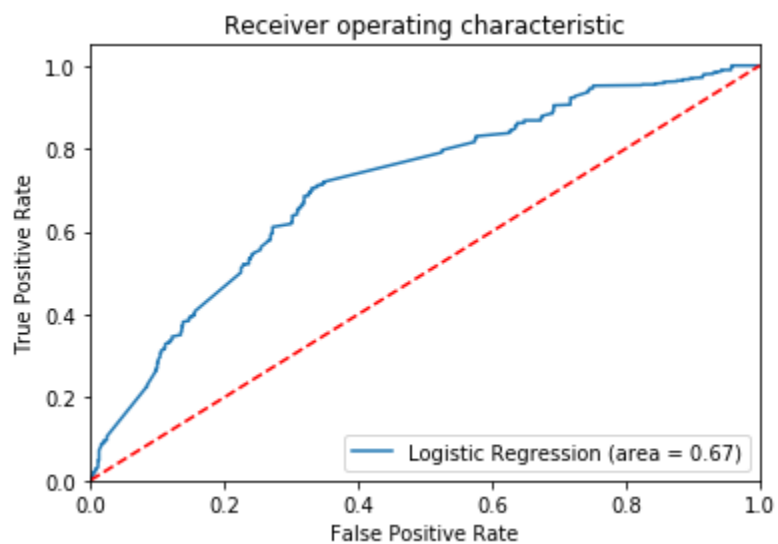
Fitted Model for Logistic Regression (Logit):

TARGET = Restaurants with 4.5-5 Star Ratings

$$\begin{aligned} \text{TARGET} = & [\text{logit} (-0.8085 \times \text{BusinessAcceptsCreditCards} + 0.4376 \times \\ & \text{RestaurantsReservations} + 0.7519 \times \text{Caters} + 1.0093 \times \text{WheelchairAccessible} - \\ & 0.4266 \times \text{NoiseLevel_loud} - 2.6864 \times \text{NoiseLevel_very_loud} - 1.1308 \times \\ & \text{RestaurantsAttire_casual} - 1.2093 \times \text{RestaurantsAttire_dressy} + 0.8316 \times \\ & \text{Street_Parking_ True} + 0.8743 \times \text{Valet_ True} - 0.4051 \times \text{Alcohol_beer_and_wine} \\ & - 1.1175 \times \text{Alcohol_full_bar} + 1.8494 \times \text{intimate_ True} + 0.7620 \times \text{hipster_ True} + \\ & 2.0950 \times \text{touristy_ True} + 0.5984 \times \text{casual_ True} - 0.8698 \times \text{breakfast_ True} + \\ & 0.6753 \times \text{brunch_ True})] \end{aligned}$$

Accuracy of logistic regression classifier on test set: 0.67 or %67.

	<i>precision</i>	<i>recall</i>	<i>f1-score</i>	<i>support</i>
0	0.68	0.68	0.68	679
1	0.67	0.66	0.67	658



Project Plan

Project Management Approach

With this project's limited time frame and the need to see frequent movement, we will utilize the Agile project methodology. Agile is a method of developing software solutions that focus on high-quality working software frequently and consistently while minimizing project overhead and increasing business value. There are eight benefits to this approach that Yum! felt were satisfactory and necessary to manage this project (SegueTech):

- Stakeholder engagement – There's a high degree of collaboration between the project and client teams.
- Transparency – Clients can be involved as much or as little as they choose.
- Predictable delivery – Sprint time is between 2-4 weeks, which means deliverable dates are always known.
- Predictable cost – With the sprints being a fixed duration, the cost is already known and limited to the amount of work that can be performed by the team.
- Allows for change – There is an opportunity to refine and re-prioritize the overall deliverables for each sprint, continually.


- Focuses on business value - By allowing the client to determine the priority of features, the team understands what's most important to the client's business and can deliver the features that provide the most business value.
- Focuses on users - By focusing features on the needs of real users, each feature incrementally delivers value, not just an IT component.
- Improves quality - By producing frequent builds and conducting testing and reviews during each iteration, quality is improved by finding and fixing defects quickly and identifying expectation mismatches early.

Project Deliverables

Let Us Yelp You will provide a Phase I recommendation report and video presentation, along with a data visualization tool packaged for the company, Yum!, by March 10, 2019. The recommendation report will include an executive summary of findings, current rating assessment with location demand benchmarks and areas of restaurant feature improvements, as well as a data dictionary. The visualization tool will include at least one interactive dashboard with an overview of the findings and all dashboard data. The dashboards will consist of visual plots, maps, and charts with an explanation of what each graphic is used for and how to interrelate with each. Yum! will be able to request more consulting hours with the team to cover more research areas, if necessary,

or to get a better understanding of how to use and update the interactive dashboards.

Dashboard (Mockup)



Yum! Restaurant Group Phoenix Area Restaurant Calculator

Viability Score	Projected Population of Type of Restaurant	Demand of type of Restaurant
GREEN % 80 or higher Probability of being a highly rated restaurant	Approximately 1200	HIGH

Choose Your Cuisine

Mexican

Chinese

Indian

American

Average Restaurant Stars

3.5

Number of Restaurant

457

Price Range

Type of Parking

WiFi

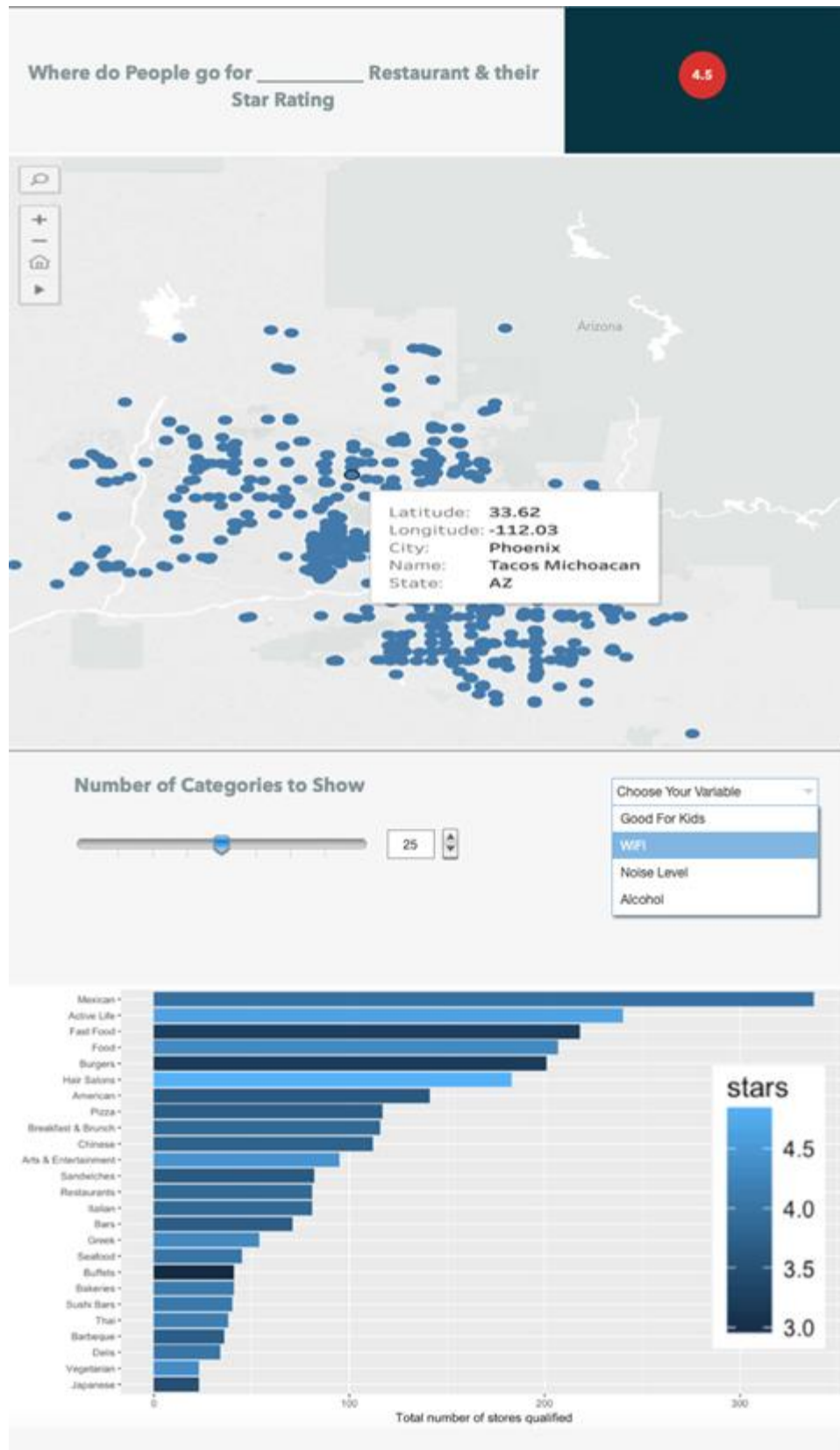
Noise Level

Serves Alcohol

Ambiance

Dogs Allowed

Happy Hour



Project Requirements

This project has been approved to meet a business need for expanding the Yum! Restaurant group to a new location, Phoenix, AZ. To achieve this business need, we need specific requirements to be met which will act as a catalyst for the successful execution of this project. The recommendation results must match or exceed a possible Yelp rating of 4 stars. The restaurant features needing improvement or needing to exist for the new location to be competitive, must meet or exceed the model performance of 70%. Requirements may be added as required, but the project sponsor will need approval before moving forward; however, Phase I must be complete by March 10, 2019.

Project Constraints

Several constraints have been identified and are imperative that considerations be made for these constraints throughout the project lifecycle. All Stakeholders must remain mindful of these constraints as they must be carefully planned to prevent any adverse impacts on the project's schedule, cost, or scope. The major constraint for this project is time. The team will need timely responses and access to the client and data, time to analyze the data/build models/analyze performance and results, all while meeting each milestone deadline. The deliverables are due to the client by March 3, 2019. The data sets necessary for this project must be readily available for use during the entire duration of this project to ensure consistency of metrics and reliability of data sources.

Project Assumptions

Assumptions have been identified for this project. All Stakeholders must be mindful of these assumptions as they introduce some level of risk to the project until they are confirmed to be true. During the project planning cycle, every effort must be made to identify and mitigate any risk associated with these assumptions. We assume the data is accurate to reflect the current state of reviews posed upon the Yum! Restaurant group by consumers. We believe higher Yelp reviews will affect the bottom line, and that a Yelp review of 5 stars constitutes a successful restaurant.

Project Risks

Risks have been identified as part of the project analysis. The Project Manager will determine and employ the necessary risk mitigation/avoidance strategies as appropriate to minimize the likelihood of these risks:

- Poor data: We rely on the public data sets for Yelp. These data sets have not been cleaned; therefore, we are unsure of how useful the data will be until we have gone through thorough EDA (exploratory data analysis).
- Inaccurate estimates for each team member: We have allotted each resource with a specific number of hours; however, until we dive deeper into the data, we will not know whether these estimates are accurate enough.

Gantt Chart

Activity	Duration	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9
Complete project analysis and SDaIP - V1	40									
Data cleansing and EDA	40									
Complete SDaIP - V2	40									
Predictive modeling, Training and Test	40									
Analyze results, Complete SDaIP - V3	40									
Video presentation	120									
Complete final project report	40									
Duration										
Total Hours	360									

Cost

Let Us Yelp You has negotiated a fixed price per hour, for each team member at a rate of \$75.00. The team anticipates allocating 8-10 hours per person, per week, across nine weeks for the duration of this project. This estimate leaves a total budget of \$21,600 - \$27,000 for Phase I. Previous project experience suggests the need for contingency hours of at least 15% of the estimated hours allotted for the project; therefore, we have approval for an extra 54 hours, which leaves \$4,050 on reserve. The Project Manager will be responsible for managing and reporting on the project's cost throughout the duration of the project.

Resources

As mentioned in the cost description, the team will allocate 8-10 hours per person, per week, and will have four resources available. Each resource will have a hybrid role, one being a data analyst while the other will coincide with each team member's previous work experience. The team will utilize both Python and R, which are free open-source software. Microsoft Office is already available

to each team member and will not incur a cost. All meetings will be held online via a communication tool, Bluejeans, at no charge. We do not anticipate any face-to-face meetings; however, should this be required, the individual team members will absorb this cost.

Endnotes

1. View. "Why Restaurants Fail - H. G. Parsa, John T. Self, David Njite, Tiffany King, 2005." Journals.sagepub.com. n.d. Web. 20 Jan. 2019.

<<https://journals.sagepub.com/doi/abs/10.1177/0010880405275598>>

2. "8 Benefits of Agile Software Development." Seguetech.com. Web. 1 Feb. 2019

<<https://www.seguetech.com/8-benefits-of-agile-software-development/>>

APPENDIX

Figure 1

Review Dataset (top 10 records)

review_id	user_id	business_id	stars	date	text	useful	funny	cool
x7mDlID B3jE iPG PHO mDz yw	msQe 1u7 Z_X uqj Go qhB 0J5 g	iCQ piavj jPzJ 5_3 gPD 5Eb g	2	2 / 2 5 / 2 0 1 1	The pizza was okay. Not the best I've had. I prefer Biaggio's on Flamingo / Fort Apache. The chef there can make a MUCH better NY style pizza. The pizzeria @ Cosmo was over priced for the quality and lack of personality in the food. Biaggio's is a much better pick if youre going for italian - family owned, home made recipes, people that actually CARE if you like their food. You dont get that at a pizzeria in a casino. I dont care what you say...	0	0	0
dDI8 zu1v WPd KGih Jrw Qbp w	msQe 1u7 Z_X uqj Go qhB 0J5 g	pomGBq fbxc qPv 14c3 XH- ZQ	5	1 1 / 1 3 / 2 0 1 2	I love this place! My fiance And I go here atleast once a week. The portions are huge! Food is amazing. I love their carne asada. They have great lunch specials... Leticia is super nice and cares about what you think of her restaurant. You have to try their cheese enchiladas too the sauce is different And amazing!!!	0	0	0
LZp 4UX 5zK 3e- c5Z GSe o3k A	msQe 1u7 Z_X uqj Go qhB 0J5 g	jtQA RsP 6P- Lbky jbO1 qNG g	1	1 0 / 2 3 / 2 0 1 4	Terrible. Dry corn bread. Rib tips were all fat and mushy and had no flavor. If you want bbq in this neighborhood go to john mulls roadkill grill. Trust me.	3	1	1
Er4 NB WC mC D4n M8_ p1G Rdo w	msQe 1u7 Z_X uqj Go qhB 0J5 g	elqb BhBf EIM NSrj FqW 3no w	2	2 / 2 5 / 2 0 1 1	Back in 2005-2007 this place was my FAVORITE thai place EVER. I'd go here ALLLLL the time. I never had any complaints. Once they started to get more known and got busy, their service started to suck and their portion sizes got cut in half. I have a huge problem with paying MORE for way less food. The last time I went there I had the Pork Pad se Ew and it tasted good, but I finished my plate and was still hungry. I used to know the manager here and she would greet me with a "Hello Melissa, nice to see you again, diet coke & pad thai or pad se ew?" Now a days, I know she still knows me but she disregards my presence. Also, I had asked her what was up with the new portion sizes and she had no answer for me. Great food but not worth the money. I havent been back in over a year because I refuse to pay \$10-15 for dinner and still be hungry after. Sorry PinKaow, you are not what you used to be!!	2	0	0

jsDu 6QE JHb wP2 Blo m1P LCA	ms Qe 1u7 Z_X uqj Go qhB 0J5 g	Ums 3ga P2q M3 W1X cA5r 6Ss Q	5	9 / 5 / 2 0 1 4	Delicious healthy food. The steak is amazing. Fish and pork are awesome too. Service is above and beyond. Not a bad thing to say about this place. Worth every penny!	0	0	0
pfav A0hr 3nyq O61 oupj- IA	ms Qe 1u7 Z_X uqj Go qhB 0J5 g	vgfc TvK 81o D4r5 0NM jU2A g	1	2 / 2 5 / 2 0 1 1	This place sucks. The customer service is horrible. They dont serve food unless you order a pizza from a neighboring restaurant. Who does that? They dont control their crowd. Many times I've gone I've seen fights. The bartenders suck - I've almost got in a fight with one because she was a complete bitch. Refused to serve me a drink because she was "busy" celebrating her friends birthday BEHIND THE BAR. This place is ridiculous. I will NEVER go there again.. EVER.	2	0	0
brok Eno 2n7s 4vrw mm Udr9 w	ms Qe 1u7 Z_X uqj Go qhB 0J5 g	Axe QEz 3- s9_1 Tylo - G7U Qw	5	1 0 / 1 0 / 2 0 1 1	If you like Thai food, you have to try the original thai bbq. Their pad se ew is to DIE for. Their thai egg rolls are delicious. Basil beef will not let you down (its not on the menu anymore, you have to ask for it!) \n\nYes, the building is not as fancy as some other places. Yes, I've batted a fly off my plate more than once. Yes, I do NOT go to the bathroom their because I dont even WANT to know what it looks like... \n\nBUT... the thai food is the best in town. The service rocks. And you can get a \$25 gift cert. on Restaurant.com for \$2. Can you beat that? I think NOT.\n\nThis is the only place my husband and I go for anniversaries, date nights, birthdays.. anything!! I recommend it to everyone I know. If you KNOW good thai food, go here.	1	0	0
kUZ WB VZv hWu C8T WUg 5AY yA	ms Qe 1u7 Z_X uqj Go qhB 0J5 g	zdE 82Pi D6w quvj YLy hOJ NA	5	4 / 1 8 / 2 0 1 2	AMAZING!!!!\n\nI was referred here by a friend and at first thought "Korean and Mexican?? WEIRD... i dont know"... but my husband and I were in the area and thought why not, lets try it... after my friend had nothing but great things to say about this place. We tried the chimichanga and it was ABSOLUTELY AMAZING.... We actually did 1/2 chicken 1/2 beef (have you EVER been to a place that will make a burrito half and half? i havent!) The meat has fantastic flavor. I'm not a beef girl but i'd order a whole plate of their beef - its that good. The chimichanga was HUGE enough for both of us to share it (and my husband also had a beef taco which also was delicious) and we were FULL and it was \$11 total with drinks. YOU cant beat that.\n\nThe service was great. The server was so nice and attentive- you could tell she was the owner. \n\nThe only thing I'd change is their chips they serve. They were burnt, too thick and kind of bland but their salsa is great. And lets face it, i'm fat so i ate two baskets anyway so they cant be that bad. lol\n\nAll in all, food was delicious, A LOT of food for cheap, great prices, great service great flavor. Hell even my burps taste delicious. Bahaha!	0	1	0
wcqt 0III8 8LE cm1 9lxF FyA	ms Qe 1u7 Z_X uqj Go qhB 0J5 g	EAW h1O mG6 t6p3 nRa ZO W_A A	4	2 / 2 5 / 2 0 1 1	Ribs = amazing\n2 hour wait time= not so amazing, but understandable. \n\nThis place would get 5 stars if they expanded their BBQ restaurant. Their ribs are AMAZING. You get SO much food for the price and it tastes sooo good. Plus, the two hour wait isnt always a bad thing because it gives you an excuse to drink and gamble while you wait!	0	0	0
LWU tzqN thM M3v pWZ IFBI Pw	ms Qe 1u7 Z_X uqj Go qhB 0J5 g	atVh 8viq Tj- sqD J35t AYV g	2	1 1 / 9 / 2 0 1 2	Food is pretty good, not gonna lie. BUT you have to make sacrifices if you choose to eat there. It literally takes an hour to an hour and a half to deliver food. Seriously. EVERY SINGLE TIME. Doesnt matter if we order at 8am, 10am or 1pm. Never fails, they take F-O-R-E-V-E-R. If you dont get what you ordered or you are upset by them delivering your breakfast around LUNCH time, be ready to have the owner talk down to you and be a total bitch to you for i dont know, just wanting what you pay for?! \n\nIts over priced. But its decently tasteful food. Takes forever. Owners a witch. And i'm pretty sure that they continuing forget to pack my extra ranch just to piss me off. \n\nEnd Rant. \n\nPS- I've never gone in there to eat because i frankly, i'd rather tip the nice delivery driver then the ignorant imbeciles that work in the dining area. \n\nPPS- My hot chocolate today was cold. They should call it Cold Chocolate. Or start caring if their hot chocolate is hot. One of the two would be great!	1	2	1

Figure 2

Business Dataset (top 10 records)

ad dr es s	attr ibu tes	busines s_id	cat ego ries	cit y	h o u r s	is_ op en	lat itu de	lon git ud e	n a m e	neigh borh ood	post al_c ode	revie w_co unt	s t a r s	s t a t e
131 4 44 Ave nue NE	{BikeParking: False, BusinessAcceptsCreditCards: True, BusinessParking: {garage: False, street: True, validated: False, lot: False, valet: False}, GoodForkIds: True, HasTV: True, NoiseLevel: average, OutdoorSeating: False, RestaurantAttire: casual, RestaurantDelivery: False, RestaurantGoodForGroups: True, RestaurantPriceRange2: 2, RestaurantReservations: True, RestaurantTakeOut: True}	Apn5Q_b6Nz61Tq4XzPd f9A	Tours, Breweries, Pizzas, Restaurants, Food, Hotels & Travel	Calgary	{Monday: 8:30-17:0, Tuesday: 11:0-21:0, Wednesday: 11:0-21:0, Thursday: 11:0-21:0, Friday: 11:0-21:0, Saturday: 11:0-21:0}	1	51.09181	-114.032	MinhasMicroBrewery		T2E6L6	24	4	AB
NaN	{Alcohol: none, BikeParking: False, BusinessAcceptsCreditCards: True, BusinessParking: {garage: False, street: True, validated: False, lot: True, valet: False}, Caters: True, DogsAllowed: True, DriveThru: False, GoodForkIds: True, GoodFormeal: dessert: False, latenight: False, lunch: False, dinner: False, breakfast: False, brunch: False}, HasTV: False, OutdoorSeating: False, RestaurantAttire: casual, RestaurantDelivery: False, RestaurantGoodForGroups: True, RestaurantPriceRange2: 2, RestaurantReservations: True, RestaurantTakeOut: True}	AjEblBw6ZFfIn7ePHha9P A	Chicken Wings, Burgers, Caterers, Street Vendors, Barbeque, Food Trucks, Food, Restaurants, Event Planning & Services	Henderson	{Friday: 17:0-23:0, Saturday: 17:0-23:0, Sunday: 17:0-23:0}	0	35.96073	-114.94	CK'SBBQ & Catering		89002	3	4.5	NV

	ating': True', 'Restaurant sAttire': 'casual', 'Restaurant sDelivery': 'False', 'Restaurant sGoodForG roups': True', 'Restaurant sPriceRang e2': '2', 'Restaurant sReservatio ns': 'False', 'Restaurant sTableServi ce': 'False', 'Restaurant sTakeOut': True', 'Wheelchair Accessible': True', 'WiFi': 'no'												
133 5 rue Bea ubie n E	{Alcohol: 'beer_and_ wine', 'Ambience': 'romantic': False, 'intimate': False, 'classy': False, 'hipster': False, 'touristy': False, 'trendy': False, 'upscale': False, 'casual': False}, 'BikeParkin g': True', 'BusinessAc ceptsCredit Cards': 'False', 'BusinessP arking': 'garage': False, 'street': False, 'validated': False, 'tot': False, 'valet': False}, 'Caters': 'False', 'GoodForKi ds': True', 'GoodForM eal': 'dessert': False, 'latenight': False, 'lunch': False, 'dinner': False, 'breakfast': False, 'brunch': False}, 'HasTV': True', 'NoiseLevel': 'average', 'OutdoorSe ating': 'False', 'Restaurant sAttire': 'casual', 'Restaurant sDelivery': 'False', 'Restaurant sGoodForG roups': True', 'Restaurant sPriceRang e2': '2', 'Restaurant sReservatio ns': True', 'Restaurant sTableServi ce': True', 'Restaurant sTakeOut':	Breakfast & Brunch, Restaurants, French, Sandwiches, Cafes	Mont rÃfÃ ©al	{Monday: '10:0- 22:0', Tuesday: '10:0- 22:0', Wednesday: '10:0- 22:0', Thursday: '10:0- 22:0', Friday: '10:0- 22:0', Saturday: '10:0- 22:0', Sunday: '10:0- 22:0'}	0	45.5 405	- 73.59 93	La Ba stri ngu e	Rosemo nt-La Petite- Patrie	H2G 1K7	5	4	Q C

	{False, 'WiFi': True}													
211 W Monroe St	NaN	bFzdJJ3wp3PZssNEsyU23g	Insurance, Financial Services	Phoenix		1	33.45	-112.077	Geico Insurance		85003	8	1.5	AZ
2005 Alyth Plac e SE	{BusinessAcceptsCreditCards': True}	8USyCYqpScwiNEb58Bt6CA	Home & Garden, Nurseries & Gardening, Shopping, Local Services, Automotive, Electronics Repair	Calgary	{Monday: '8:0-17:0', Tuesday: '8:0-17:0', Wednesday: '8:0-17:0', Thursday: '8:0-17:0', Friday: '8:0-17:0'}	1	51.03559	-114.027	Action Engine		T2H0N5	4	2	AB
20235 N Cave Creek Rd, Ste 1115	{BikeParking': True, 'BusinessAcceptsCreditCards': True, 'BusinessParking': {'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}, 'Caters': False, 'OutdoorSeating': True, 'RestaurantPriceRange2': '1', 'RestaurantsTakeOut': True, 'WheelchairAccessible': True, 'WiFi': True}	45bWSZtniwPRiqlivpS8Og	Coffee & Tea, Food	Phoenix	{Monday: '5:30-20:0', Tuesday: '5:30-20:0', Wednesday: '5:30-20:0', Thursday: '5:30-20:0', Friday: '5:30-21:0', Saturday: '5:30-21:0', Sunday: '6:30-19:0'}	1	33.67138	-112.03	The Coffee Bean & Tea Leaf		85024	63	4	AZ

631 Bloor St W	{BusinessParking: 'True', 'garage': False, 'street': False, 'validated': False, 'lot': False, 'valet': False}, {RestaurantPriceRange: '1'}	9A2quhZLy Wk0akUetBd 8hQ	Food, Bakeries	Toro nto		0	43.6 643 8	- 79.41 44	Bn c Ca ke Ho use	Koreatow n	M6G 1K8	7	4	O N
341 7 Derr y Roa d E, Unit 103	{Alcohol: 'none', 'BusinessAcceptsCreditCards': 'True', 'BusinessParking: 'True', 'garage': False, 'street': False, 'validated': False, 'lot': False, 'valet': False}, {GoodForkIds: 'True', 'OutdoorSeating': 'False', 'RestaurantAttire': 'casual', 'RestaurantGoodForGroups': 'True', 'RestaurantPriceRange: '2', 'RestaurantTableService': 'False', 'RestaurantTakeOut': 'True'}	6OuOZAok8i KONMS_T3 EzXg	Restaur ants, Thai	Mississau ga		1	43.7 129 5	- 79.63 28	Th ai On e On	Ridgewo od	L4T 1A8	7	2	O N
144 0 N. Dys art Ave	{Alcohol: 'none', 'Ambience': 'romantic', 'False', 'intimate': False, 'classy': False, 'hipster': False, 'diver': False, 'touristy': False, 'trendy': False, 'upscale': False, 'casual': True}, {BikeParking: 'True', 'BusinessAcceptsCreditCards': 'True', 'BusinessParking: 'True', 'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}, {Caters: 'False', 'GoodForkIds: 'True', 'GoodFormeal: 'False', 'dessert': False, 'latenight': True, 'lunch': False, 'dinner': False, 'breakfast': False, 'brunch': False}, {HasTV: 'False', 'NoiseLevel': 'average', 'OutdoorSeating': 'False', 'Restaurant	8- NRKkPY1Ui FXW20WXXi Xg	Mexican Restaur ants	Avo ndal e	{Monday: '0:0-0:0', 'Tuesday: '0:0-0:0', 'Wednesday: '0:0-0:0', 'Thursday: '0:0-0:0', 'Friday: '0:0-0:0', 'Saturday: '0:0-0:0', 'Sunday: '0:0-0:0'}	1	33.4 481 1	- 112.3 41	Fili ber to's Me xic an Fo od		85323	40	2. 5	A Z

	{sAttire: 'casual', 'Restaurant sDelivery': 'False', 'Restaurant sGoodForGroups': 'True', 'Restaurant sPriceRange2': '1', 'Restaurant sReservations': 'False', 'Restaurant sTableService': 'False', 'Restaurant sTakeOut': 'True', 'Wheelchair Accessible': 'True', 'WiFi': 'no'}													
209 Oakl and Ave	{BikeParking: 'True', 'BusinessAcceptsCreditCards': 'True', 'BusinessParking': 'garage', 'street': 'False', 'validated': 'False', 'lot': 'False', 'valet': 'False', 'Restaurant sPriceRange2': '2'}	UTm5QZThPQIT35mkAcGOjg	Flowers & Gifts, Gift Shops, Shopping	Pittsburgh	{Monday: '9:0-18:0', Tuesday: '9:0-18:0', Wednesday: '9:0-18:0', Thursday: '9:0-18:0', Friday: '9:0-17:0', Saturday: '10:0-17:0'}	1	40.44142	-79.9565	Maggie & Stella's Gifts	Oakland	15213	3	3.5	PA

Figure 3

Boxplots of the numerical variables in the merged dataset

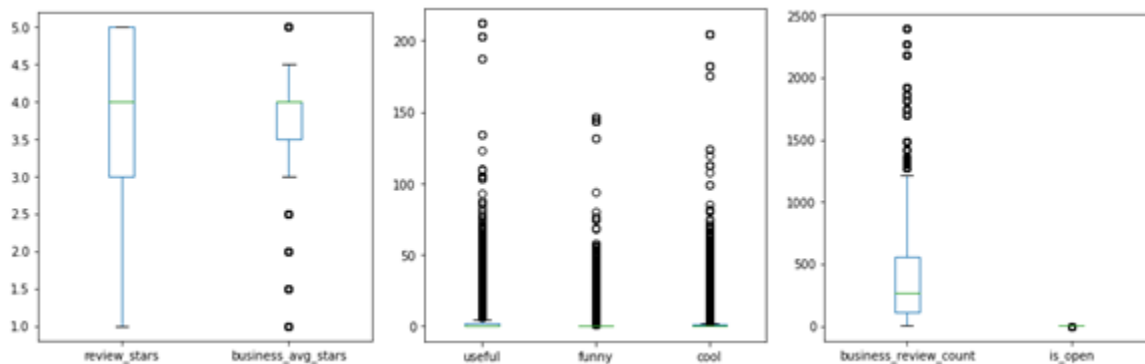


Figure 4

Correlation plot of the numerical variables in the merged dataset

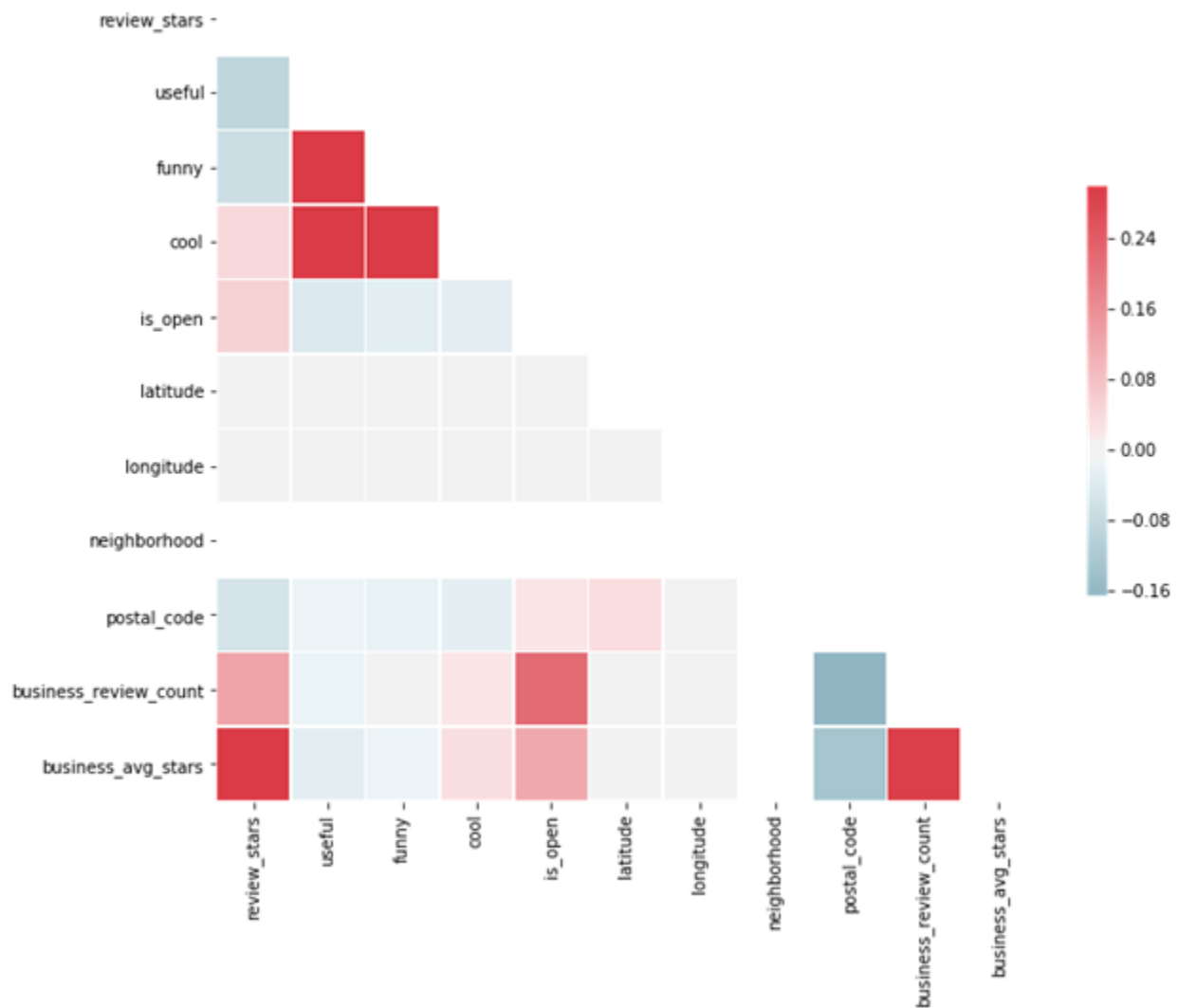


Figure 5

Pair plot of the numerical variables in the merged dataset

