RestViz Analytics a Visualization Tool

Ashutosh Bhadke
ASU ID -1011095675
abhadke@asu.edu

Laveena Bachani
ASU ID -1213246721
lbachani@asu.edu

Mayuri Kambli
ASU ID -1212890144
mkambli@asu.edu

Shikhar Sharma
ASU ID -1213091904
sshar133@asu.edu

ABSTRACT

Restviz Analytics is a tool, that provides insights into the restaurant business of an area. We developed this tool to help investors open a restaurant and also to help restaurant managers to improve their business. We performed sentiment analysis on restaurant reviews along with rating which helps to determine the overall quality of the restaurant. Finally, we implemented various interactive visualization charts that give the user the ability to compare the good restaurants with average and poor, and also analyze particular restaurant.

KEYWORDS

Data Visualization, Yelp, Chord Chart, Sentiment Analysis,Area Chart, Info Graphics, Map, Word Cloud

1 INTRODUCTION

Restaurant Industry is competitive and tricky industry, which is not only limited to the quality of food but various other features such as parking, Wi-fi, restaurant location etc. Restaurant owners have to keep up with the customer demands, and should continuously update their restaurants according to the trends. Otherwise, they could fall behind. To analyze their business we need a lot of data. Fortunately, there is an abundant data that is available for the public to analyze and make some inferences that will positively affect businesses performance. In this project, we are using Yelp Dataset. We are limited to the data of Arizona State. In this paper, we design a tool to analyze restaurant business on the basis of important factors that might be helpful in the success of a restaurant. This tool would be helpful for new restaurant investors and restaurant managers to guide and improve their business.

2 MOTIVATION

Food and drinks sales of the restaurant industry in the United States has reached 782 billion in 2016. The statistics show that it is a highly profitable industry. But that is only true if every step made in correct direction because statistics also shows that 23%of restaurants fail in their first year only. Today, the restaurant industry is not comprised of only food as a lot of people see it not only a place to eat but also for a hangout. To open a successful restaurant, not only we need to know the popular cuisine in an area but also what other feature do people demand. According to Nerd wallet [8], the demand for a restaurant is dependent on the population growth and density and also median annual income and income growth. To open a restaurant investor has to take a lot of things in the mind. As it is an expensive business and failure can lead to the high loss. So, our motivation for this project is to build a tool that can help investors and restaurant manager to leverage the datasets available This tool would be helpful to know the demographics, public preferences and other features that can affect the industry.

3 RELATED WORK

Yelp dataset has been used for various research purposes and a lot of work has been done in the area of using the dataset to improve the quality of restaurants. James Huang et. al 2013 used Latent Dirichlet allocation (LDA) algorithm on the reviews of the restaurants and found that areas of interest for specific restaurants. Overall, it turned out that users care most about service, and subsequently value, take out, and decor [1]. Michail Alifierakis in his article about predicting restaurant closure in using Yelp, shown various statistics that affect the success of a restaurant. He also showed the price range for food is dependent on the demographics of an area[2]. Aileen Wang et. al. have studied on the weight of the features that contribute most to the success of the restaurant. They have applied the chi-squared test and stochastic gradient descent to identify the features that hold the most weight. It found out that ambiance, ages allowed, parking and music have the high weight in the success of the restaurants[3]. Sindhu Hegde et. al. have analyzed the Yelp academic dataset to figure out the most crowded day in a restaurant to help to improve the business of a restaurant. They have used kd tree for this purpose[7]. Mengqi Yu et. al. have tried predicting the ratings of a restaurant based on its statistics and reviews[5]. In our project we have taken their ideas and results to develop a tool that can help restaurant manager to improve their business.

4 VISUALIZATION DESIGN

First of all, we take the input as cities and categories from the user that they are interested in.Next according to selected cities map is opened with the markers which indicate whether the restaurants exist in chosen cities and categories. We have shown three colors of markers i.e. green, yellow and red, which show if the restaurants fall under good, average or poor labels. We have generated these labels using sentiment analysis and the rating. Using this kind of division, user can compare what facilities good restaurants provide and improve on them and decide on the price range for his restaurant by comparison with these restaurants.

We are showing two types of visualization, first is aggregate results which lets the user compare the price range, demographics and the food the restaurants offer in good, average and poor restaurants. We have used multi level area chart and chord chart for the

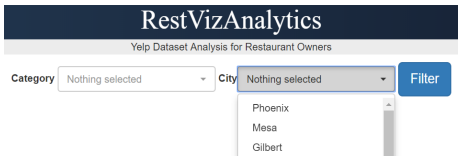

Figure 1: user input

purpose.
The other type of visualizations lets the user explore the individual restaurants. We are showing word cloud that that is the result of the text mining on the reviews that lets the user explore what the visitors most excited about, what are things that they like or dislike. We are also showing the infographics to tell the details of the resturants in a more interactive manner. We are also drawing a bubble chart that shows at what time and what day of the weak there is more crowd in the restaurants.
With all these visualization we are giving the restaurant owner and opener a tool to explore at various levels their business and its requirements.

4.1 Dataset

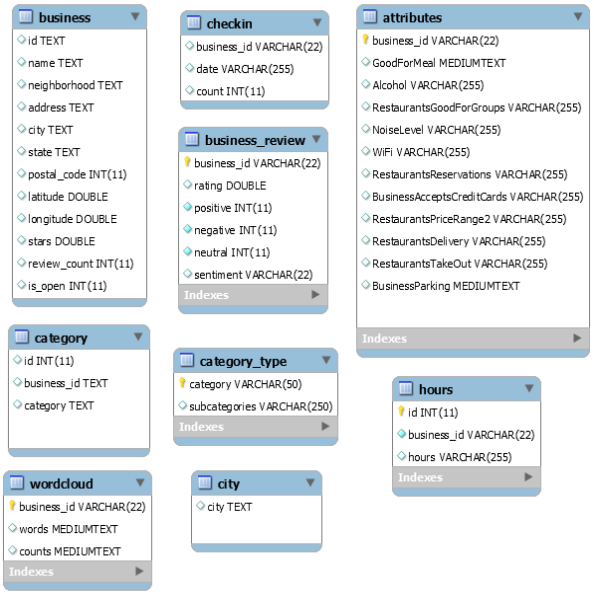

Figure 2: dataset

We have chosen Yelp Dataset of this purpose. It is a huge dataset so we decided to focus on the restaurants data available for arizona state. There are 11054 restaurants in the dataset for Arizona.Based on the popular food categories in Arizona, we filtered the yelp data for 17 categories: American, Bakery, Barbecue, Bars, Breakfast and Brunch, Burgers, Cafeteria, Chinese, Delis/Diners, Indian, Italian, Japanese, Mediterranean, Mexican, Seafood, Steakhouse, Thai. We have denormalized the yelp data and formed 9 relations as follows:
1. Business: Business represents the information belonging to a particular restaurant.
2. Business Review: Business Review represents rating parameters for e.g. sentiment score.
3. Attributes: Attributes relation contains various facilities a restaurants provides.
4. Category: This relation stores all the 18 categories.
5. Wordcloud: This table contains preprocessed words from the reviews and their respective counts. This table has been maintained for faster access to data while loading the word cloud.
6. Category Type: Every category has subcategories. This relation stores subcategories for each main category.
Hours: This stores the open hours of all restaurants.
7. City: This stored all cities belonging to Arizona from yelp data.
8. Check-in: This records the count of check-ins for restaurants at a particular time of day.

4.2 Technologies used

We have used the following technologies in building the software.
1. D3.js - It provided the library for building interactive visualizations using javascript. We have used D3 version 4. The word cloud, scatter plot, chord diagram has been plotted using D3.
2. Google Map - This is an API for rendering maps for spatial data in terms of latitude and longitude.
3. Highcharts - This library provides some easy to use interactive visualizations to represent complex data.
4. Spring Boot - The backend of the web application is developed using Java Spring Boot framework.
5 .User Interface - The user interface for the application has been designed using HTML 5, CSS 3, Bootstrap, JavaScript.

4.3 Map

The yelp data contains tons of restaurants, and interpreting those restaurants just through text visualizations or graphs is not sufficient to capture the crux of the data. The spatial analysis using map, distributes the restaurants according to the sentiments, thus helping the user to know the distribution of the restaurants locally and such visual representation helps user to grasp the information faster and retain it for longer time. There are several map APIs, which we could use, but Google map stands out in them with its easy to use functionalities which are customizable and its capability to show the complex data makes it extensible.

4.4 Aggregate Comparison

4.4.1 Chord Chart. The chord chart is used to show the interconnection between entities. It is ideal to show what is common between two entities and also how much common is between two entities.Nodes are arranged along a circle, with the relationships between points connected to each other either through the use of arcs or Bezier curves. Values are assigned to each connection, which is represented proportionally by the size of each arc. Color can be used to group the data into different categories, which aids in making comparisons and distinguishing groups. [1] Issues with the chord chart is when there is a high number of entities, it is difficult to read the figure clearly and also one can not know the precise values of the interdependency. We are using chord chart to

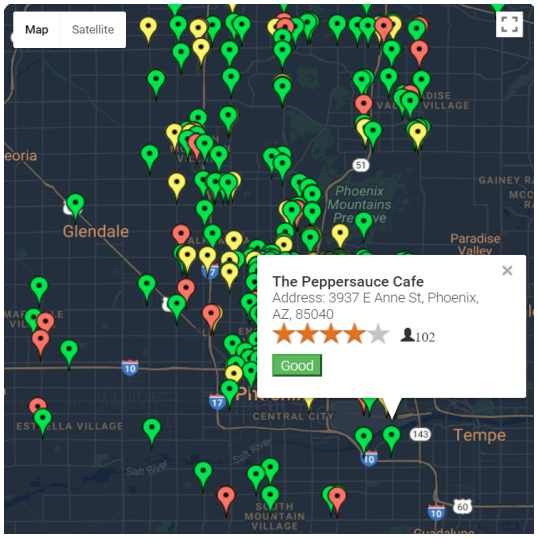

Figure 3: Map showing the filtered restaurants

show the interconnection of good, average and poor restaurants with the facilities they provide.

4.4.2 Area Chart. Area chart connects continuous or discrete data with a straight line. When there are multiple trends to show we can use leveled area chart. It is most useful when you want to show the part-to-whole relationship. Such as with our visualization, we are price range offered by restaurants in a city. The aggregated shows the number of total restaurants in the price range and in parts we have shown the division of the cities. We are giving the flexibility to select and deselect the cities. Also, we have given four buttons i.e good, poor, average and aggregate. On clicking, it will show the data for the restaurant that falls into the category.

4.5 Individual Restaurant Analysis

4.5.1 Scatter Plot. Scatter plot is used to show the data on the 2-D point graph. The advantage of scatter plot is that we can show 4-D data in 2-D. With third dimension representation through the point size and another one dimension through the point color. The scatter plot helps you find potential relationships between values, and to find outliers in data sets. Disadvantages of scatter plot is we can not show the precise value of third and second dimension as the human eye can only compare relative size and change in color up to some extent. So we can only show on the second and third dimension something that simplifies identification. We have used scatter plot to represent the check-in information about one particular restaurant. We are showing weak-days on the x-axis, and day hours on the y-axis. The number of check-ins is shown by the bubble size. As the user would be only interested in figuring the hours when visitors visit the most, it may not need to know the exact value of check-ins. In this case, scatter is appropriate to show the check-in information.

4.5.2 Word Cloud. The word cloud in the group of words shaped in a cloud. A word cloud is used to summarize a large text and give important words in it. The important words can be the words that uniquely appear or the words that most frequently appear. It does not show the stop words, the words that are common to every document. We have used the word cloud to show the summary of the reviews of the restaurant. It shows what uniquely categories the restaurant in the perspective of its visitors.

Figure 4: Word cloud of the selected restaurant reviews

4.5.3 Infographic. Infographic is the information represented with graphics. It is an alternative to the text, which people never read. Infographic is an interactive way of telling the information. People are visual, we only remember 20% of what we read. Whereas, we process visual information much faster. Brain process the information through converting it to the visuals in one or another way.So to show the details about particular restaurant we used infographics. As these details are not statistics to plot on a graph and to show in form a text would be uninteresting.[6]

5 METHODOLOGY

We have first taken performed semantic analysis on the reviews. With the results of the semantic analysis and ratings of the user, we have tagged the restaurants as good, average and poor. We have seen that reviews are a good measure to figure out the success of a restaurant [5]. Then we have shown collective results according to our segregation of the restaurants. We have also seen that some features affect the restaurant success more than others [2]. Finally, we are giving the user the functionality analyses the statistics of a particular restaurant with the word cloud, infographic and scatter plot. The number of people can be helpful to know when a restaurant should serve special menu and so special arrangements [7]. Scatter plot shows the most crowded hours of a restaurant. Details of the charts are described below.

5.1 Sentiment Analysis

Business owners may want to see the facilities provided by restaurant based on their popularity. We used reviews of a restaurant along with average rating to get sentiment score of a particular restaurant for that we used the textblob library to get the polarity of each review of a business and aggregated the sentiment and classified the restaurants into good, average and poor categories. We used NLTK library - TextBlob for this purpose. The score given by text blob is divided into three ranges - if greater than 0.2 then positive, if less than -0.1 the negative, otherwise neutral. If a score is positive by the rating is less than 2.5, then the restaurant is considered as neutral. If a score is neutral and rating is less than 2.5, then it is considered as negative. This is how we have done the combined scoring for restaurants. Because we utilized both sentiment score and ratings, we were able to accurately determine the restaurant's ranking instead of just relying on the yelp's restaurant rating.

5.2 Map

To perform the spatial analysis, we used google maps and plotted the latitude and longitude of each restaurant on the google map. We used roadmap type of google map, with the bound restricted to Arizona. Then we supplied the latitude and longitudes of the restaurants retrieved from the filter query for cities and categories and plotted markers based on the sentiment type assigned to the restaurants i.e. Good, Average, Poor. The green markers represent good restaurants, yellow markers represent average restaurants and red marker represent poor restaurants. On hover, the details of the restaurant, such as address, stars, number of reviews are displayed. When the user clicks on a restaurant, the visualizations for that individual restaurants are displayed.

5.3 Aggregate Comparison

For the given set of cities and categories, we have aggregated the data and analyzed it city wise and restaurant's attributes wise. The city wise distribution is represented using multi-level area chart. The attribute wise distribution is represented using chord diagram. This chart helps to look the restaurants in individual rating - Good, Average and Bad.

5.3.1 Chord Chart. The chord chart represents the overall attribute distribution of the restaurants for selected categories and cities. The outer arc is divided into 18 attributes that include restaurant attributes and 3 parameters that rate restaurants(Good, Average, Poor). The green, yellow and red colors are used to represent these ratings and other colors represent the attributes.The ribbons represent the portion of attributes that fall under these three ratings. Each arc represents a node and the chords show the connectivity among these nodes. This visualization helps the owner to understand why some restaurants stand good, some average and some poor in the domain of the facilities that they provide to the customer. We selected 18 nodes to draw the chord diagram. The attributes are:

1. Good for Meal - Breakfast, Lunch, Brunch, Dinner, Dessert, Late night
2. Alcohol - This represents whether a restaurant provides alcohol or not, the restaurant might have full bar or it just provides beer and wine.
3. Loud - If the noise level is loud or very loud then that restaurant is represented as loud.
4. Wi-fi - This tells if the restaurant provides Wi-fi (free or paid).
5. Business Parking - If the restaurant provides one of the kinds(Garage, Street, Validated, Lot and Valet) of parkings.
6. Credit card - This attribute represents whether the restaurants accepts the credit card.
7. Delivery - This attribute represents if the restaurant provides delivery.
8. Take Out - This attribute represents if the restaurant provides take out.
9. Good for Groups - This attribute states if the restaurants if good for groups.
10. Reservations - This attribute shows if the restaurants provides reservations in advance.

For the selected restaurant we get these attributes. These attributes are then represented as nodes in the form of the matrix. These matrix values indicate the edge strengths of the chord diagram. The edges are directional, from restaurant attributes to ratings good/bad/poor and vice versa.

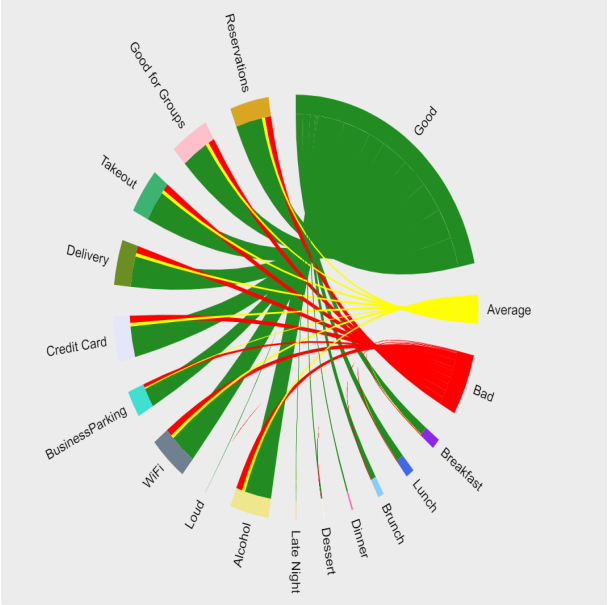

Figure 5: chord chard showing the facilities provided by good, average and poor restaurants

5.3.2 Area Chart. The area chart represents the distribution of restaurants over the prices ranges for selected categories and cities. We aggregated all the restaurants in prices ranges 1, 2, 3 and 4 and plotted across the x and y-axis where the x-axis represents the price ranges and y-axis represent the number of restaurants. The different lines in the area chart represent cities. The distribution can be seen in all restaurants or good/average/poor restaurants separately. This tells the owner how the rated restaurants are distributed in terms of price i.e. whether they are on the cheaper side or costly side, which city contains more contains more costly restaurants etc. The owner

can focus on few cities and compare them using the selection tool. On hover, the chart displays the number of restaurants in particular price range.

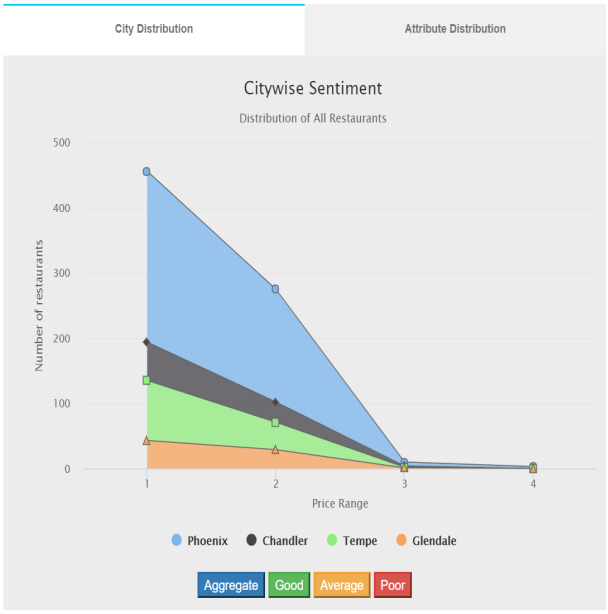

Figure 6: area chart showing the restaurants that fall in a price range for good,average and poor category for selected cities.

5.4 Individual Restaurant Analytics

On clicking the pushpins on the map, it shows the analytics of individual restaurants. We are showing three charts, Word chart on the restaurant reviews, Scatter plot for most visiting hours and infographic for specific details about a restaurant. Details about each of them are explained below.

5.4.1 Word Cloud. The yelp data contains lots of textual data in the form of reviews. A restaurant owner cannot go through the all reviews to come up with a business decision as reviews are just one perspective of yelp data. This textual data needs to be represented in such form so that the owner gets significant information through a glance. Thus we decided to generate a word cloud for all the reviews of the restaurant and retrieved top 50 words. We considered all the reviews of the restaurant, removed the punctuation details from them. We ignored the stop words and performed filtering using porter stemming. The resulted words were counted for their frequencies and after sorting in descending order, we chose top 50 ranked words from the results.
The word cloud represents the frequently used words by the users in the reviews. We used d3.layout.cloud.js to render the word cloud. The word cloud contains words that have most significance in the user reviews. These words are animated using transition and rotation. The word cloud helps business owner to know what users think about that restaurant, what attracts them most about the restaurant, the right day to visit the restaurant, the popular recipes, good for which meal, about the atmosphere, is it reasonable or costly from user's point, how is the staff/service, customer moods.

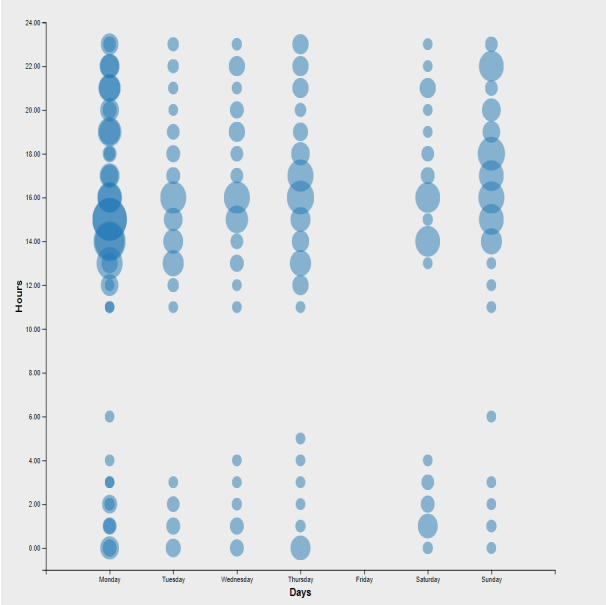

Figure 7: scatter plot for the most visiting hours for a restaurant

5.4.2 Scatter Plot. The scatter plot represents the check-in pattern for the customers in a particular restaurant. It shows the busy hours of the restaurants day wise. It helps to find the trend of visitors, the timings they prefer. The restaurant owners can use this information to figure out, how well a restaurant is doing, how many visitors they are getting on hour basis. This helps in taking decisions about the resources and their efficient utilization in the busy hours. We are showing weak-days on the x-axis, and day hours on the y-axis. The number of check-ins is shown by the bubble size.

5.4.3 Infographics. We have shown the details of the particular restaurant with infographics. We have created SVG images for various details such as a full bar or not, credit card acceptance etc. Finally, we have highlighted those SVG images if the restaurant provides it otherwise it is shown in the background. If a restaurant provides the facility it is shown in green color and otherwise shown in red color. Below are the details of the features that the infographic represent :
1. Good for Meal - Breakfast, Lunch, Brunch, Dinner, Dessert, Late night
2. Alcohol - This represents whether a restaurant provides alcohol or not, the restaurant might have a full bar or it just provides beer and wine.
3. Loud - There are four levels of noise from very loud to quite.
4. Wi-fi - This tells if the restaurant provides Wi-fi (free or paid).
Price Range: there is four price range from 1 to 4 where 4 shows how expensive the dataset is.
5. Business Parking - If the restaurant provides one of the kinds(Garage, Street, Validated, Lot and Valet) of parking. 6. Credit card - This

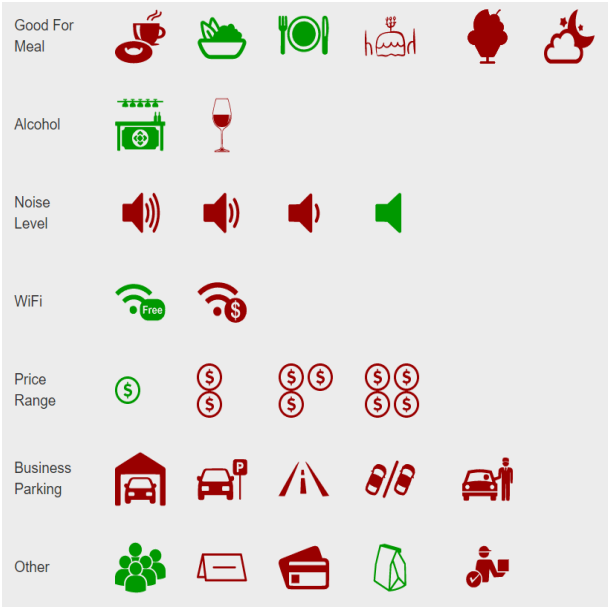

Figure 8: Infographic for the specific details of a restaurant.

attribute represents whether the restaurants accept the credit card.
7. Delivery - This attribute represents if the restaurant provides delivery.
8. Take Out - This attribute represents if the restaurant provides take out.
9. Good for Groups - This attribute states if the restaurants if good for groups.
10. Reservations - This attribute shows if the restaurants provide reservations in advance.

5.5 Evaluation Plan

The evaluation of Restviz Analytics is based on the user query and how they get benefitted from the tool. This system gives the user to explore the demographics and restaurant preferences of an area. However, we have observed the consistency in the plots, that shows that user can get benefit from the system.. Our findings include :

- For a given area, visiting hours are consistent. Mostly crowd like to go to restaurant in evening from 6 pm to 8 pm, with the exception of the restaurant near the business area where most frequent visiting hours were from 12 pm to 2 pm.
- The facilities provided by the good restaurant differ from the average and poor by a significant margin, shows that these facilities affect the success of a restaurant.
- The word cloud for the good restaurant shows positive sentiment words.
- The price range for the good, poor and average restaurant differs from significantly from city to city.

For further evaluation, we can ask the user how much they find the system useful.

6 DISCUSSION AND FUTURE WORK

The application can be extended to cover all the restaurants available in yelp dataset across the world so that it can reach out to a large number of restaurant owners who are handling restaurants over the globe. The category distribution can be done spatially to know in which region what categories are popular so that the owner can decide what kind of food categories he should provide in this restaurant. The data can be compared category-wise with respect to cities and sentiment scores. The time series analysis can be done for reviews, to get how the restaurants performed over a period of time. We can also analyze which attributes have most weight on restaurant success for a given city and show analysis of only those attributes[3]. We can also improve our sentiment analysis using the algorithms such as LDA [4].

REFERENCES

[1] James Huang, Stephanie Rogers, Eunkwang Joo *Improving Restaurants by Extracting Subtopics from Yelp Reviews.* Spring 2013.
[2] Michail Alifierakis *Using Yelp Data to Predict Restaurant Closure* Towardsdatascience
[3] Aileen Wang, William Zeng, Jessica Zhang predicting New Restaurant Success and Rating with Yelp Stanford University, December 16, 2016
[4] D. Blei, A. Ng, and M. Jordan.Latent Dirichlet Allocation. Journal of Machine Learning Research ,3:9931022, January 2003.
[5] Mengqi Yu, Meng Xue, and Wenjia Ouyang.Restaurants Review Star Prediction for Yelp Dataset UCSD
[6] Caitlin Jordan Design an infographic from scratch CANVA
[7] Sindhu Hegde et. al.Restaurant Setup Business Analysis Using Yelp Dataset IEEE
[8] Jonathan ToddBest Cities to Start a Restaurant Nerd Wallet, 2015