

EAS 503 Project Report

Yuan Hui, Yiqing Zhong, Terrance Christian



December 1, 2018

University at Buffalo

Contents

[Table of Figures: 2](#_Toc531474432)

[Acknowledgement 3](#_Toc531474433)

[Abstract 4](#_Toc531474434)

[1. Introduction 5](#_Toc531474435)

[1.1 Problem Summary 5](#_Toc531474436)

[1.2 Objective of work 5](#_Toc531474437)

[1.3 Data Description 5](#_Toc531474438)

[2. Data Dictionary 6](#_Toc531474439)

[2.1 Relational Schema: 7](#_Toc531474440)

[3. Exploratory Data Analysis 8](#_Toc531474441)

[4. Interactive visualisations in user selected city 12](#_Toc531474442)

[5. Restaurant Filter and Restaurant Recommender 18](#_Toc531474443)

[6. Conclusions and Future Work 20](#_Toc531474444)

[7. References 21](#_Toc531474445)

# Table of Figures:

[Figure 1: Relational Schema 7](#_Toc531474459)

[Figure 2: Top 10 cities with most data 8](#_Toc531474460)

[Figure 3: Top 10 restaurants with most reviews 8](#_Toc531474461)

[Figure 4:Star distribution for hottest restaurants 9](#_Toc531474462)

[Figure 5: Time series plot of reviews for hottest restaurants 9](#_Toc531474463)

[Figure 6: Top 10 active users 10](#_Toc531474464)

[Figure 7: Star distribution for top active users 10](#_Toc531474465)

[Figure 8: Open hours on Sunday 11](#_Toc531474466)

[Figure 9: All restaurants on World map 11](#_Toc531474467)

[Figure 17: GUI for city selection 12](#_Toc531474468)

[Figure 18: Scatter plot for selected city 12](#_Toc531474469)

[Figure 10:Price distribution for restaurants 13](#_Toc531474470)

[Figure 11: Restaurant count for different categories 13](#_Toc531474471)

[Figure 12:Hottest restaurant from each cuisine 14](#_Toc531474472)

[Figure 13: Top 10 restaurants with most 5 star reviews 14](#_Toc531474473)

[Figure 14: Price range VS Restaurant Attire 15](#_Toc531474474)

[Figure 15: Review-star relationship 15](#_Toc531474475)

[Figure 16: 5 star review for restaurant type 16](#_Toc531474476)

[Figure 19: Animation screenshot 1 16](#_Toc531474477)

[Figure 20: Animation screenshot 2 17](#_Toc531474478)

[Figure 21: Filter GUI 18](#_Toc531474479)

[Figure 22: Filter Result 18](#_Toc531474480)

[Figure 23:Recommender GUI 19](#_Toc531474481)

[Figure 24: Recommended restaurants plot 19](#_Toc531474482)

# Acknowledgement

We have put efforts in this project, however, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

We are highly indebted to Dr. Varun Chandola for his guidance and supervision as well as for providing necessary information regarding the project. We take this opportunity to express our sincere gratitude and indebtedness with a sense of appreciation to the people who encouraged and extended their professionally competent suggestion during the whole work and provided useful review, feedback and all time support.

We are thankful to our university, University at Buffalo for providing all required resources and good working environment.

Thanking you,

Yuan Hui

Yiqing Zhong

Terrance Christian

# Abstract

Yelp Restaurant Data Analysis aims at doing exploratory data analysis for the restaurants that are available in the Yelp data provided by kaggle. The key idea is to understand different types of restaurants and the ratings of those restaurant and provide different visualisations that provide meaningful insights about those restaurants.

The project also focuses on providing recommendation to users based on the attributes they select. These recommendations are also plotted on a map to provide better visualisation.

# Introduction

* 1. Problem Summary

The key idea behind this project is to build SQL database for Yelp restaurants data and to implement Exploratory Data Analysis (EDA) by providing various graphical visualizations and analysis based on restaurant characteristics such as ratings, expenses, distance etc. Another objective is to provide filter and recommender tools to users based on their preferred characteristics.

* 1. Objective of work

Following are the main ideas behind the study:

1. Database construction for the restaurants and review database using MySQL;
2. User-friendly graphical visualization of restaurants on Google Maps using Python;
3. EDA on restaurants in interested city and analysis correlations between variables using Numpy, Pandas and Matplotlib
4. Provide filters on restaurant selections based on user preferences using Pandas
5. Study on the selected restaurant characteristics and recommend similar restaurants
   1. Data Description

The dataset used for this project is Yelp dataset provided on kaggle. For the scope of this project, business.json and review.json files are used. The business.json file provides characteristics of various business, of which restaurants are of our interest and the review.json file provides reviews and ratings of the business.

More specifically: the business.json file has 15 features including business\_id(Primary key), name(String),neighborhood(String), address(String), city(String),  state(String), postal code(String), latitude(Float), longitude(Float), stars(Factor), review\_count(Int), is\_open(Factor), attributes(Key-value), categories(Categorical) and hours(Key-value);

The review.json file has 9 features including review\_id(Primary key), user\_id(Foreign key), business\_id(Foreign key), stars(Factor), date(DateTime), text(String), useful(Int), funny(Int) and cool(Int).

1. Data Dictionary

Table: Yelp1

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Constraint |
| Business\_id | Varchar(25) | Not Null (PK) |
| Name | Varchar(25) | Default Null |
| Categories | Varchar(50) | Default Null |
| Latitude | Float | Default Null |
| Longitude | Float | Default Null |
| Address | Varchar(100) | Default Null |
| City | Varchar(30) | Default Null |
| State | Varchar(20) | Default Null |
| Review\_count | Small int(6) | Default Null |
| GoodForKids | Varchar(7) | Default Null |
| NoiseLevel | Varchar(12) | Default Null |
| OutdoorSeating | Varchar(7) | Default Null |
| RestaurantAttire | Varchar(7) | Default Null |
| RestaurantsDelivery | Varchar(7) | Default Null |
| RestaurantsGoodForGroups | Varchar(7) | Default Null |
| RestaurantsPriceRange2 | Small int(6) | Default Null |
| RestaurantsReservations | Varchar(7) | Default Null |
| RestaurantsTakeOut | Varchar(7) | Default Null |
| Alcohol | Varchar(10) | Default Null |
| Caters | Varchar(7) | Default Null |
| DogsAllowed | Varchar(7) | Default Null |
| DriveThru | Varchar(7) | Default Null |
| RestaurantsTableService | Varchar(7) | Default Null |
| WheelChairAccesible | Varchar(7) | Default Null |
| WiFi | Varchar(7) | Default Null |
| CoatCheck | Varchar(7) | Default Null |
| GoodForDancing | Varchar(7) | Default Null |
| HappyHour | Varchar(7) | Default Null |
| Open24Hours | Varchar(7) | Default Null |

Table: Yelp2

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Constraint |
| Business\_id | Varchar(25) | Not Null (FK) |
| Review\_id | Varchar(25) | Not Null (PK) |
| User\_id | Varchar(25) | Not Null |
| Date | Date | Default Null |
| Stars | Small int(6) | Default Null |
| Useful | Small int(6) | Default Null |
| Cool | Small int(6) | Default Null |
| Funny | Small int(6) | Default Null |

Table: yelp\_h

|  |  |  |
| --- | --- | --- |
| Column Name | Datatype | Constraint |
| Business\_id | Varchar(25) | Not Null (PK) |
| Monday | Varchar(15) | Default Null |
| Tuesday | Varchar(15) | Default Null |
| Wednesday | Varchar(15) | Default Null |
| Thursday | Varchar(15) | Default Null |
| Friday | Varchar(15) | Default Null |
| Saturday | Varchar(15) | Default Null |
| Sunday | Varchar(15) | Default Null |

## 2.1 Relational Schema:

The following image is the relational schema of our database.



Figure 1: Relational Schema

# Exploratory Data Analysis

The data provided was from locations around the globe, so we identified the cities for which maximum number of data was available. The below chart shows top 10 cities that have the most data.

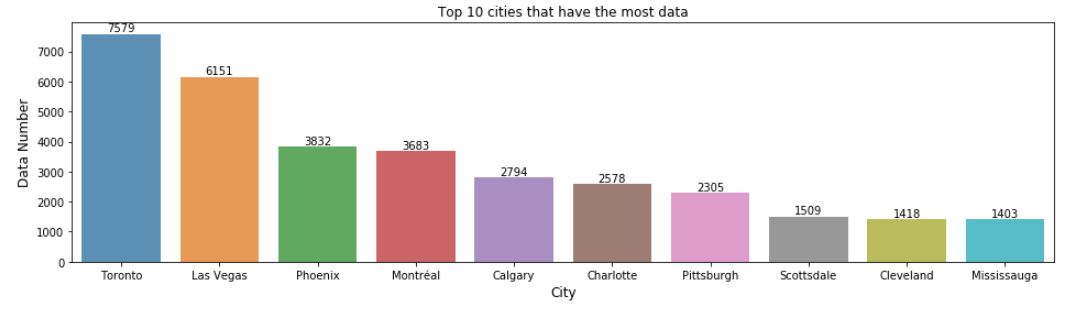


Figure 2: Top 10 cities with most data

The x-axis show the names of cities and y-axis has the number of records available.

For all the reviews we had, we identified the top restaurants for which maximum number of reviews are available. The following chart shows the top 10 restaurants on y-axis and the number of reviews available on the x-axis.

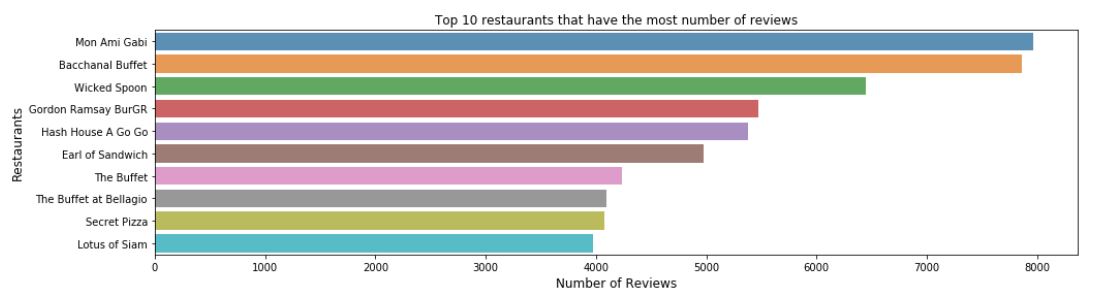


Figure 3: Top 10 restaurants with most reviews

For the restaurants with most reviews, we then find the number of stars that they received. The number of star components for the top restaurants is shown in the pie chart below.

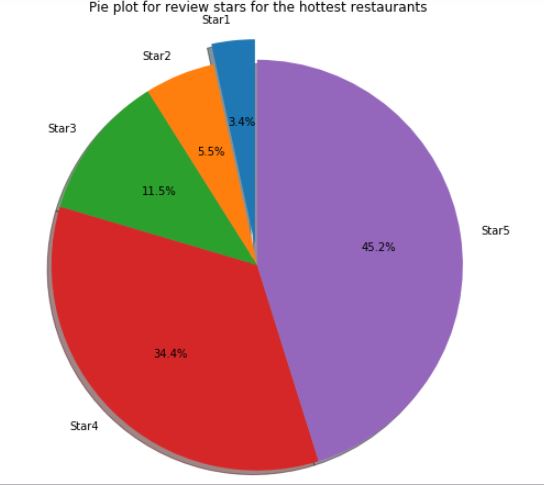


Figure 4:Star distribution for hottest restaurants

We then made a time-series plot for all the reviews that were made for the top restaurants from 2005 to 2018. The below plot shows the result.

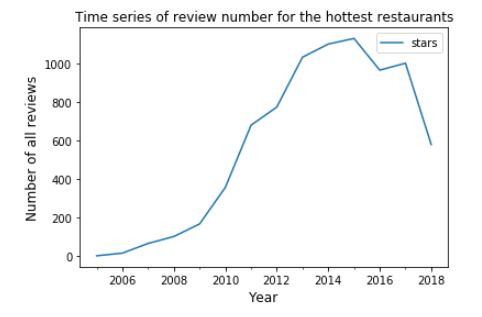


Figure 5: Time series plot of reviews for hottest restaurants

The above plot shows the number of reviews for top restaurants on a time series.

The most important contributors to the data were the users who were giving the reviews for the restaurants. We identified the users that are highly engaged in reviewing the restaurants and plotted their star-giving trend on a pie plot.

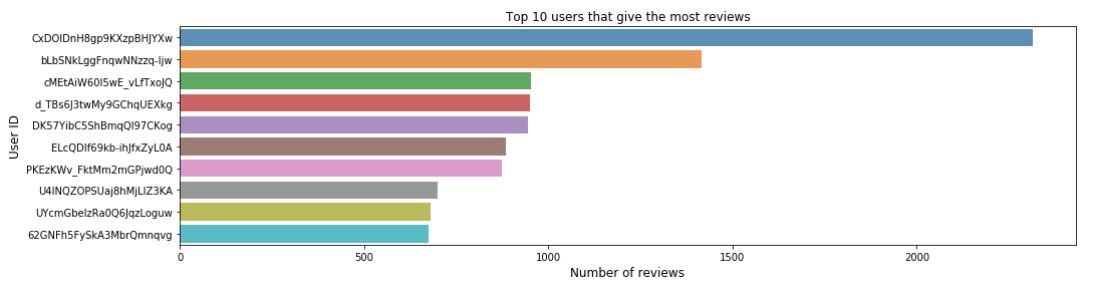


Figure 6: Top 10 active users

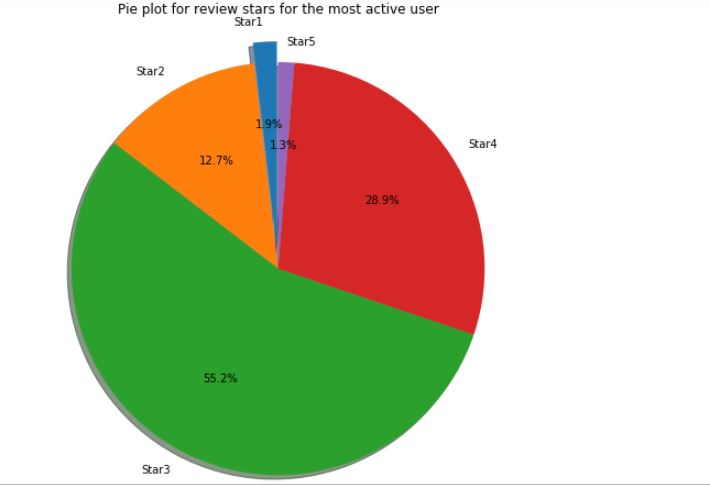


Figure 7: Star distribution for top active users

From the plot, it is observed that the top active reviewers give an average of 3 stars to restaurants. The majority of reviews are 3 stars and 4 stars by the top reviewers.

Apart from this, we have made different plots such as number of restaurants with opening time on Sunday as well as plotted all the restaurants on the map of North America.

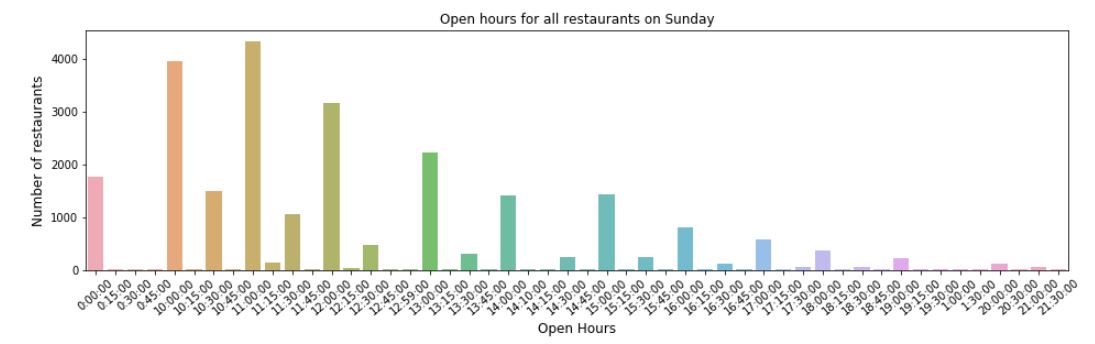


Figure 8: Open hours on Sunday

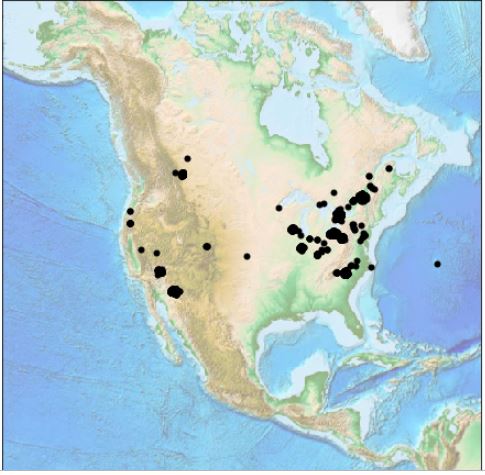


Figure 9: All restaurants on World map

1. Interactive visualisations in user selected city

The notebook provides a GUI to enter a city from a list of predetermined cities and a plot showing all restaurants in that city is displayed. For example, consider the city Phoenix is Arizona.

The GUI for entering location and the resultant plot is as follows.

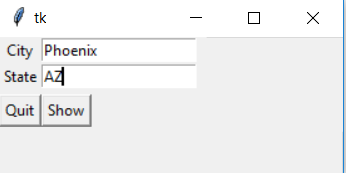


Figure 17: GUI for city selection

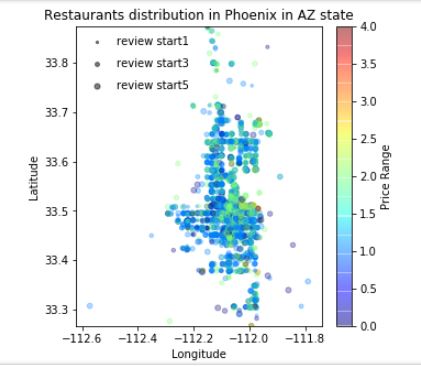


Figure 18: Scatter plot for selected city

In this plot, restaurants are plotted based on their latitude-longitude and the marker colour shows the price range for that respective restaurant. Also, the marker size shows the average star-rating of that respective restaurant.

The following pie chart shows the distribution of restaurant prices in user selected city.

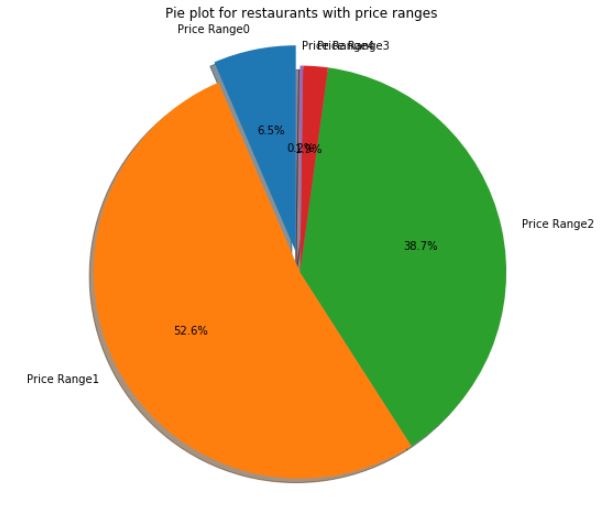


Figure :Price distribution for restaurants

Number of restaurants for different categories:

There are number of restaurants that offer different cuisines. These cuisines are categorized in 11 classes and the number of restaurants for each category in a given city are provided in the plot below.

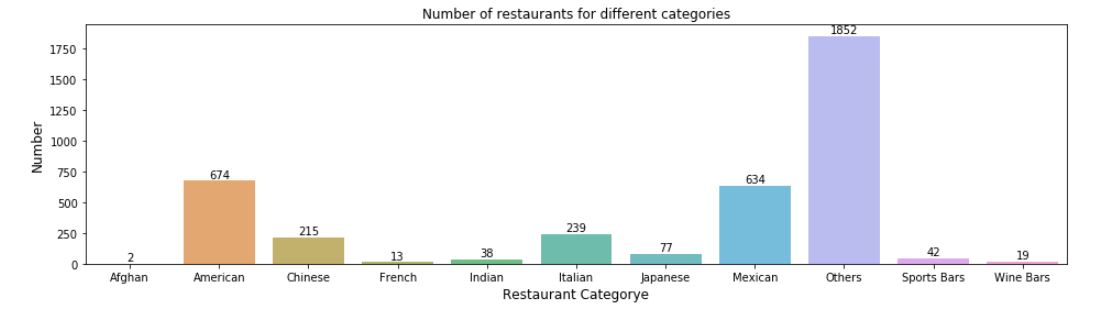


Figure : Restaurant count for different categories

Top restaurant for each category:

As there are different categories of restaurants based on the cuisine offered, we identified the top restaurant for each category for the user specified city and plotted it on the same bar graph for number of restaurants in each category. The plot is shown below.

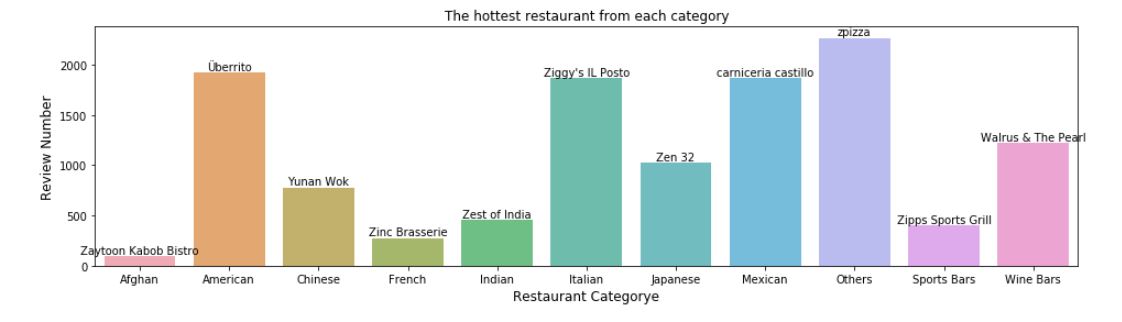


Figure :Hottest restaurant from each cuisine

Top 10 restaurants with most 5 star ratings:

Often the search is limited to the top restaurants, therefore we identified top 10 restaurants that has most number of 5 star ratings and plotted them on a bar graph with number of 5 star ratings on the y-axis. The plot is shown below.

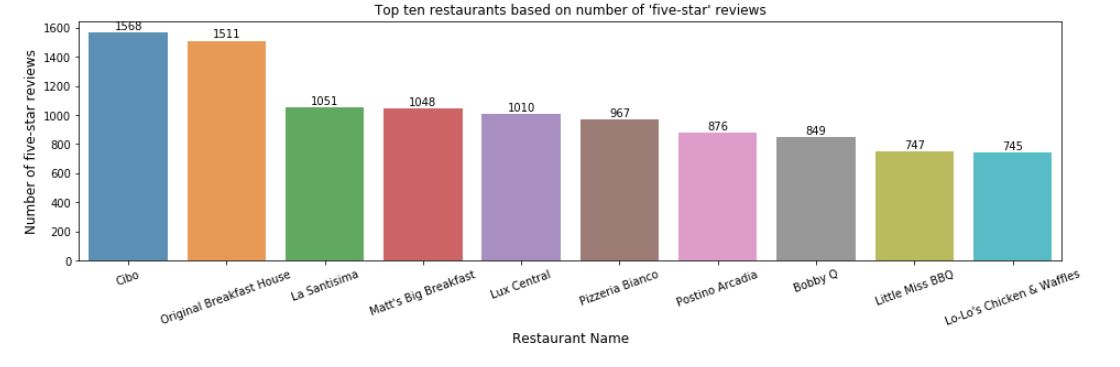


Figure : Top 10 restaurants with most 5 star reviews

Price range VS cloth required:

The below box plot shows the type of attire suited for a restaurant with given price range range.

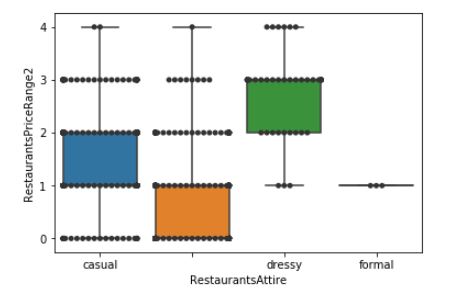


Figure : Price range VS Restaurant Attire

Number of reviews VS Star rate:

The below plot shows the trend between the number of reviews and the star rating.

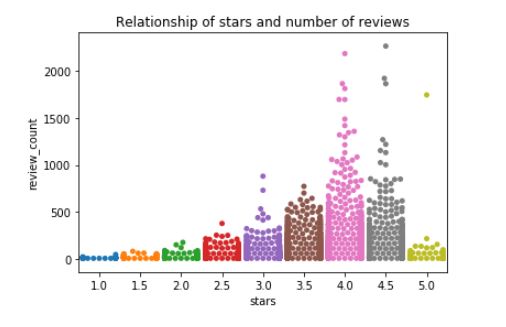


Figure : Review-star relationship

Number of 5-star review VS restaurant type:

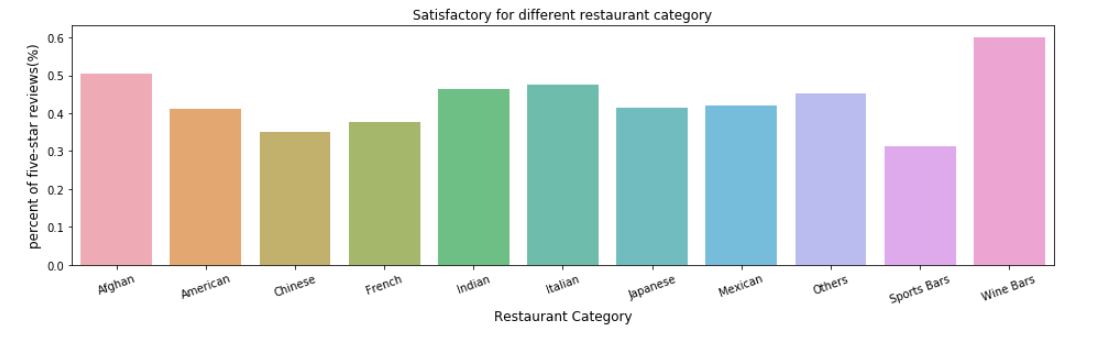


Figure : 5 star review for restaurant type

The above plot shows the percentage of 5-star rating for each category of cuisine.

Apart from the above plots, the restaurants are plotted as an animation with different star rating restaurants plotted together. User can zoom in/zoom out, scroll through the animation and get restaurants on the map. The following images shows the animation.

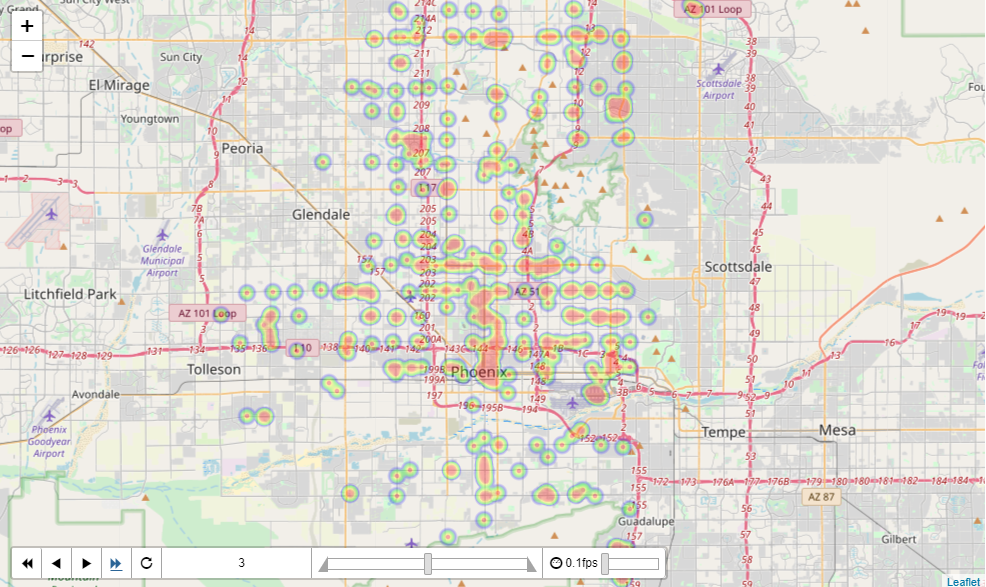


Figure 19: Animation screenshot 1

The above image shows the restaurants in Phoenix, AZ that have a 3-star rating. Similarly, below image shows all the restaurants in Phoenix, AZ that have a 4-star rating.

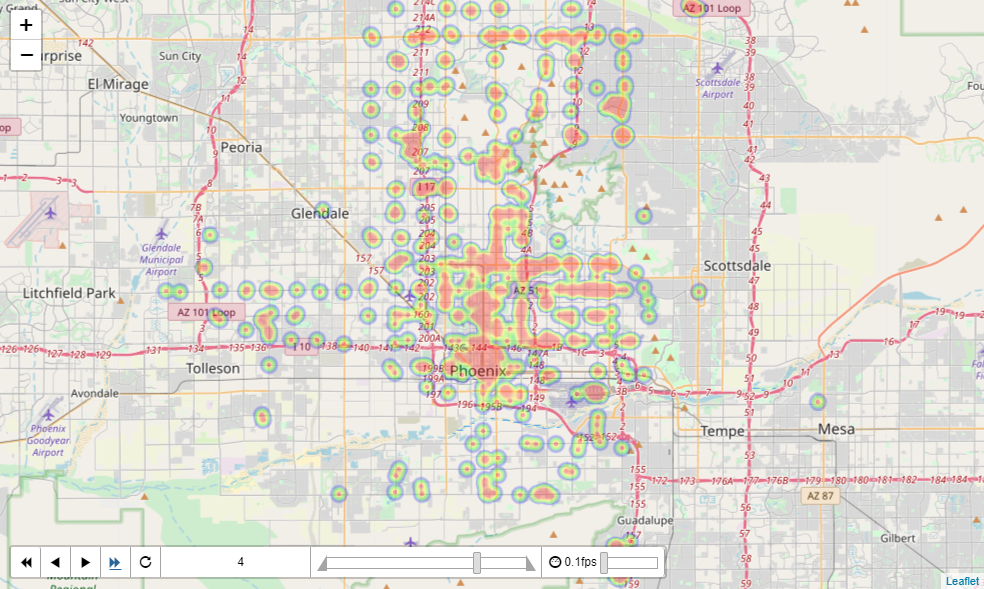


Figure 20: Animation screenshot 2

1. Restaurant Filter and Restaurant Recommender

Apart from the restaurants details on an interactive map, the users are allowed to filter the restaurants based on certain attributes. These filtered restaurants are again displayed on a map. Both the GUI for filter and the resultant plot are provided below.

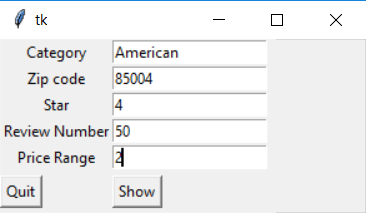


Figure 21: Filter GUI

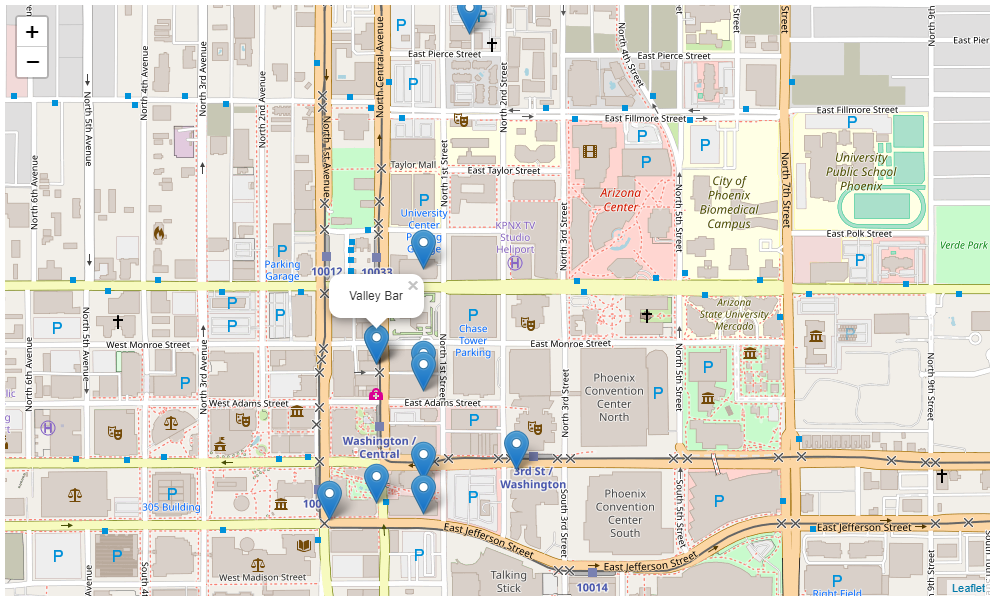


Figure 22: Filter Result

One of the main feature of our project was to provide a recommendation system for the users. The user is provided with a graphical interface where he/she can enter the attributes of restaurants as they like and the system will provide them with a list of restaurants that are similar to the attributes they provided.

These restaurants are plotted on a map to provide better visualisation. Both the GUI and the resultant plot for recommended restaurants is provided below.

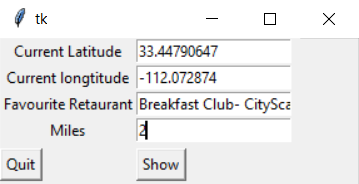


Figure 23:Recommender GUI

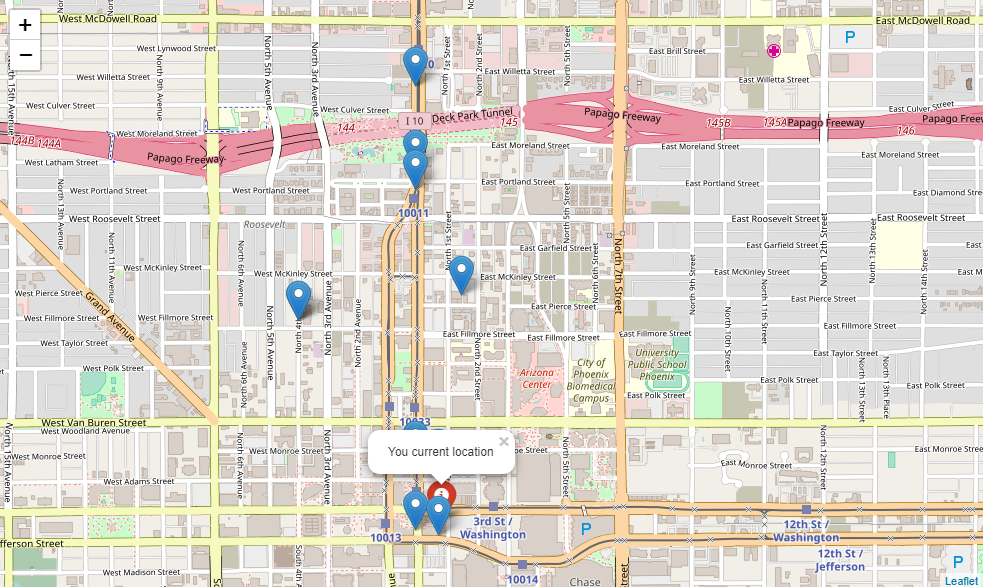


Figure 24: Recommended restaurants plot

1. Conclusions and Future Work

The project included the analysis of the Yelp data set. Restaurants business was explored and different insights were drawn from the data like which restaurants have highest ratings, which restaurant has the highest rating for a particular cuisine, who are the top users that provide highest number of ratings etc.

The project covered visualisations that show different plots of restaurants with different star ratings on a selected city map.

A GUI to provide filter and recommendation feature for different types of restaurants and for different cities were provided.

In future, we would like to extend our project for different type of business like hotels.

1. References

[1]. Chao Shi, Sam O’Mullane, Sean Kickham, Reza Rad, Andrew Rubino,” *A Hybrid Recommender with Yelp Challenge data -- Part 1*”, NYC Data Science Academy (blog), June 27, 2017

[2]. https://nycdatascience.com/blog/student-works/yelp-recommender-part-1/