Abstract:

The report presents a complete analysis of the restaurant business by studying its characteristics/ features and drawing insights to help the industry. We considered the yelp dataset for this analysis due to its size and variation. Firstly, we consider the location as it plays a significant role in this field. Secondly, we analyze reviews (Statements) and ratings (1 to 5 stars) of the restaurants in the dataset on a city and state level; find how popular a restaurant is in a city. Thirdly, we understand the number of restaurants that are reviewed in different cities and states. Finally, analysis of the number of reviews by a user to understand how much reviews usually a user puts (user behavior). We will also throw light on the restaurants with the highest number of check-ins which infer how frequently the restaurants are visited. The reviews given by the users are examined and their relevancy can be used in the improvement of these restaurants.

**Introduction:**

Restaurants has played an important role in business, social, intellectual and artistic life. Besides great food, restaurants are important for meeting friends, relatives; spending some ‘me’ time; office meetings. Public opinion and expert reviews play an important factor for a restaurant’s success or failure. Reviews sites like Zomato, Yelp are platforms where such reviews are shared. The main objective of this report is to use the data of yelp and provide insights that can be useful for different segment of job roles and business.

Statement of the project objectives:

* To find which city has the most restaurants
* To visualize the popular restaurants in a city
* To find which restaurant has the highest number of check-ins
* To discover which restaurants users reviews the most
* To find how much reviews does a user give
* To analyse which state has the most restaurants
* Understand how much reviews are useful for restaurants
* Find how much restaurants are there in a state that were rated 5 stars by a user

Motivation of the problem:

Analysing the yelp data will help the people from various departments and business with insights which can be used for future. By analysing the number of restaurants in different city; the yelp sales and marketing team can target the area with less numbers to register with them. By analysing the popularity of a restaurant; a restaurant franchise can decide where to put up a new outlet. By analysing the user review number; users can be given incentives in order to increase the review count and encourage them to visit the restaurant more often thus helping yelp to get more search traffic

Relevance of chosen topic

The fact that how much importance restaurants plays in part of today’s world for an individual and how reviews and ratings given by a user effects the business and how a new business can use this to open an outlet; we decided to analyse the data of restaurants to get those insights

Elicitation of appropriately formed research question(s):

The principal contributions presented in this work can be summarized as follow

* To analyse restaurants data of yelp
* Analyse reviews, check-ins and stars of restaurants to derive insights to understand user behaviour, restaurant popularity and restaurants registered on yelp

Related Work:

The paper has addressed the problem of setting up a new restaurant. How location can be used to boost profit for the restaurant. Location suggestion could have been done based of the cruise of the restaurant which would be more helpful [1]. Study shows how geographical proximity is important in popularity of a business in different cities [2]. To predict whether a restaurant(business) belongs to positive reviewed class or below it [3].

used in the improvement of these restaurants.

Methodology:

To attain the project objective, the KDD approach is followed as mentioned by Fayyad et al. (1996) [1] which proposes that a distinct number of stages must be followed for the implementation of a project which involves using a dataset to generate knowledge. This can be clearly seen in figure 1. Each stage of the KDD life cycle will be explained in detail in the rest of this section.

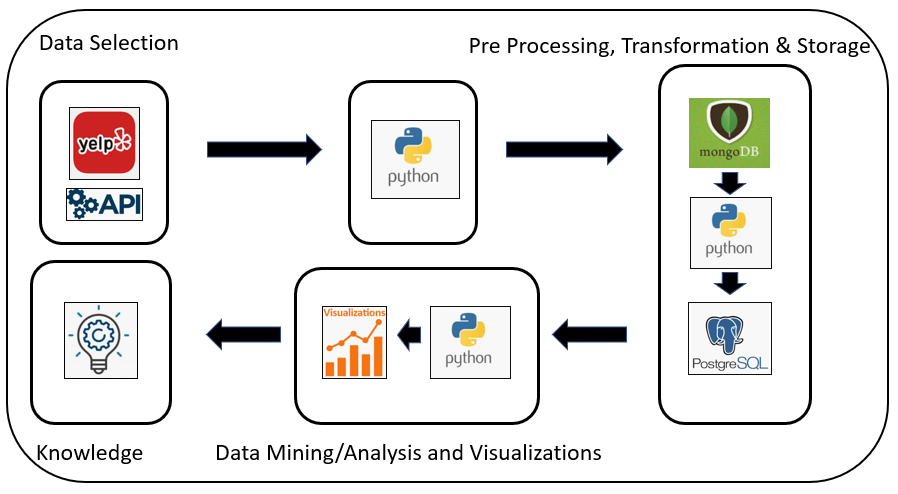


Figure 1: KDD Life Cycle [1]

DATASET:

Dataset Justification:

Yelp is a service with has various businesses directories listed and reviews that are sourced by people. It was reported by Yelp that there were 61.8 million and 76.7 million different users that visited it through the desktop and mobile respectively in the first half of 2019. It also stated that its website has 192 million reviews [4]. Due to its huge volume of data its dataset has been considered for analysis.

Three semi-structured datasets from Yelp which has a huge repository of restaurant data is used. The data is in the JSON format and is fetched from Kaggle through an API. The business, review and check-in data of restaurants is taken into consideration for the analysis.

Dataset Description:

1. Business Data

The dataset contains 192,609 rows and 14 variables which contains the information for various businesses represented by their ids. The data includes attributes, location data, open-timings categories for the businesses from which the restaurant businesses are filtered. Figure 2 displays the variables with their corresponding datatypes and description.

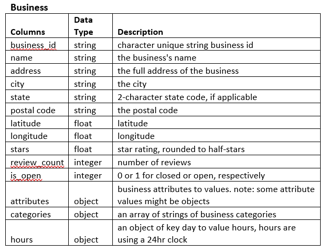


Figure 2: Business Data Variables

1. Review Data

The dataset contains 6,685,900 rows and 9 variables which list the reviews data given. There are user ids which write a review in a text format and rate a business id in the form of stars. Figure 3 displays the user review variables, datatypes and description.

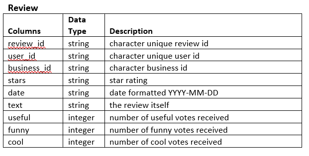


Figure 3: Review Data Variables

1. Check-in Data

The check-in dataset contains 161,950 rows and 2 variables which give the total check-ins made by a user at a business. The data variable contains list of check-in timings in a timestamp format for a business. Figure 4 displays the check-in variables, datatypes and description

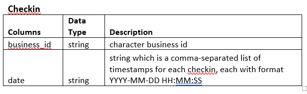


Figure 4: Check-in Data Variables

TECHNOLOGIES USED:

Python is an open sourced language, flexible and has huge, efficient libraries for data manipulation, making it a principal coding language for data visualization, cleaning and processing[5]. There are approximately more than 72,000 open source libraries provided by python[6]. MongoDB classified as a No-SQL database is used because the storage is internal which causes the data to be fetched quicker. In addition, it is schema less and easily scalable. PostgreSQL is used because it is opensource, has multitasking and is a high-performance database. Figure 5 displays the detailed flow of the project.

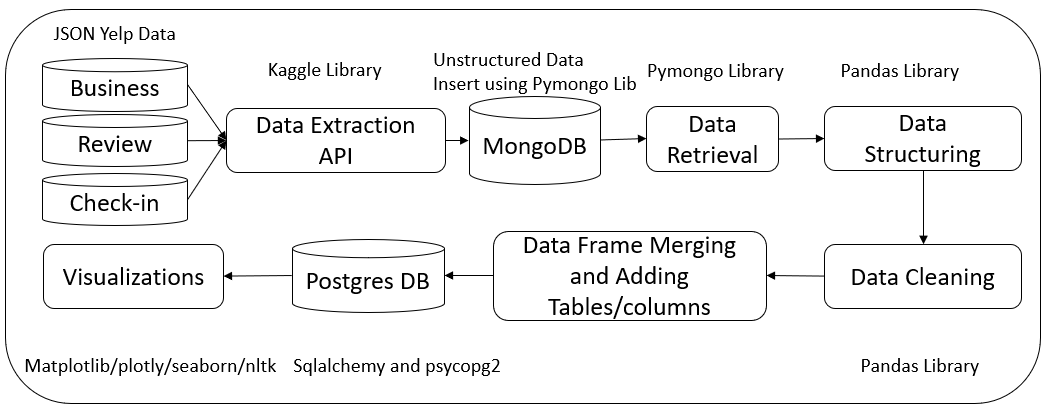


Figure 5 : Detailed Flow Diagram

DATA EXTRACTION:

The three Yelp datasets which are business, reviews and check-in are in the JSON format which are automatically pulled through the API in python using the Kaggle libraries. The zip file is downloaded and is unzipped placing the json files in the local path which is picked up. This data is then put into MongoDB by obtaining a DB connection.

DATA CLEANING:

The data fetched from MongoDB is structured into a DataFrame by using Pandas Library. The three datasets are converted into three different DataFrames and are cleaning separately. The unwanted columns are removed from the three DataFrames and the rows with NAs and Nulls are discarded for accurate visualizations. Columns with dictionaries are converted into string formats through lambda for insertion into the database.

FEATURE ENGINEERING:

For the purpose of data visualization new columns are and tables/DataFrames are created and merged from the existing data in python. A ‘total number of check-ins’ column is added in the check-in table to visualize restaurants with maximum check-ins. The business and check-in data are merged by using an inner join on the business id. Tables for states with maximum stars and review counts are created for visualization charts and graphs.

DATA STORAGE:

The data is stored in MongoDB and Postgres databases at different stages. The unstructured data in the JSON format fetched from the datasets are stored in MongoDB first. The database connection is established with the help of the pymongo library. A new database is created through python and individual collections are created for the three datasets.

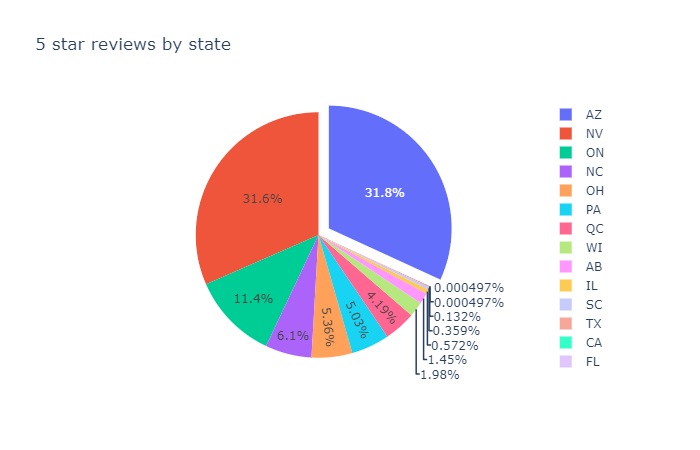
After structuring, cleaning and transforming the data in python with pandas it is loaded into PostgreSQL. A connection is created through the psycopg2 library by giving the connection parameters like username, password, host ip, port and database. After achieving the database connection, a database is created in PostgreSQL by executing a create database query. The tables are created and loaded in the database using Sqlalchemy. The chucksize attribute is used to push rows in batches due to the large size of the data. The tables are then picked up from PostgreSQL through sqlalchemy by establishing a new connection and the data is used for visualizations.

DATA VISUALIZATION:

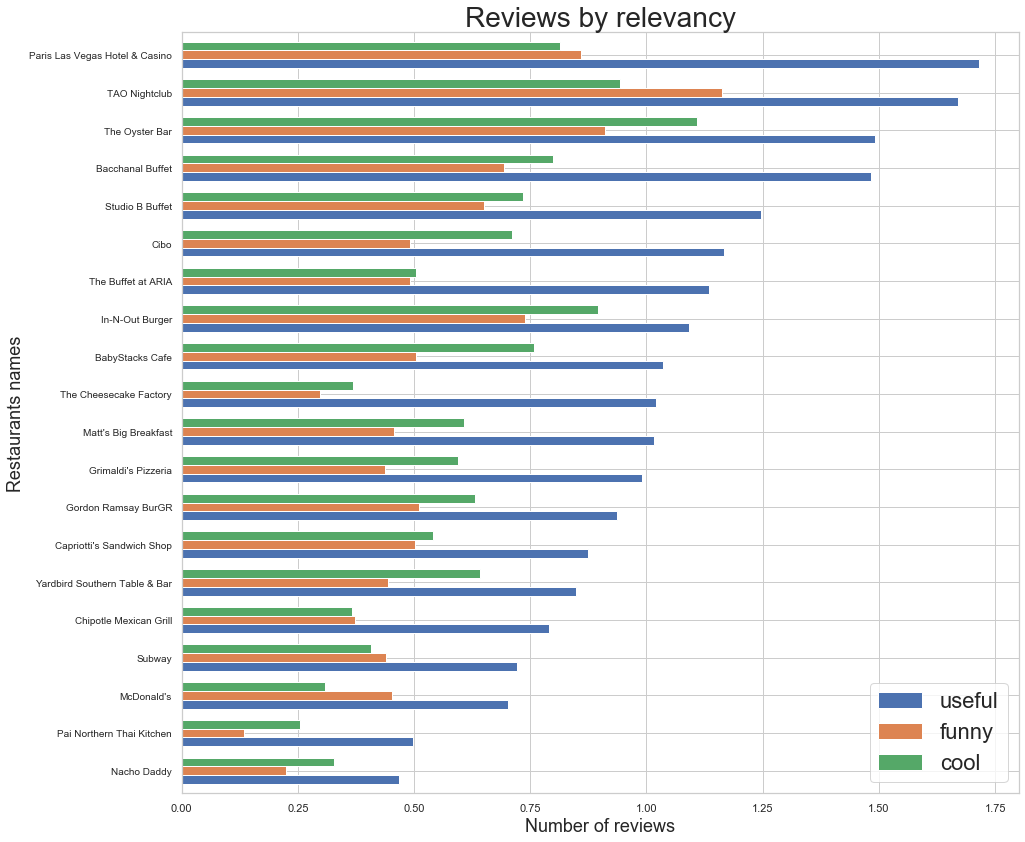
Plotly, seaborn and Matplotlib libraries are used to visualize the data. Most standard plots are supported by Plotly. The plots seen in MATHLAB can be implemented in python using Matplotlib. A variety of customizations can be implemented using this library [7]. In the project the Matplotlib library is used for visualizing restaurants with maximum check-ins in a bar graph. The distribution of star review ratings by restaurants with the help of a pie chart and the states with maximum 5 stars using a pulled-out chart are executed using the Plotly library. To visualize the cities with the maximum number of restaurants the seaborn library is used. Matplotlib is used for getting the heatmap of the cities with the most restaurants based on popularity. Las Vegas and Toronto have been considered for it. Matplotlib is used further to display the review breakdown of most reviewed restaurants.

Results:

* States with most 5-star restaurants

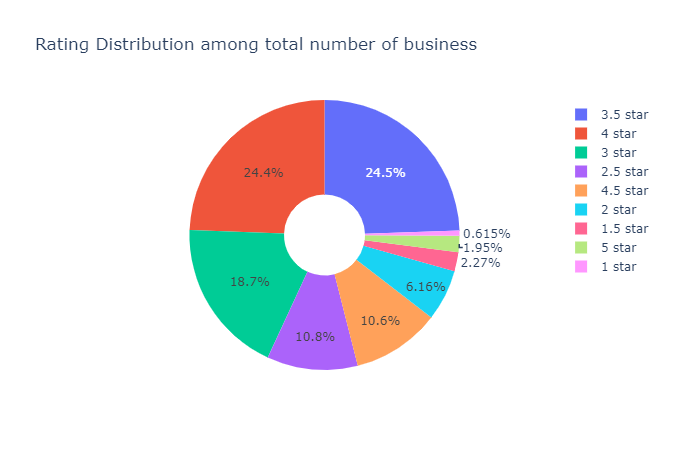
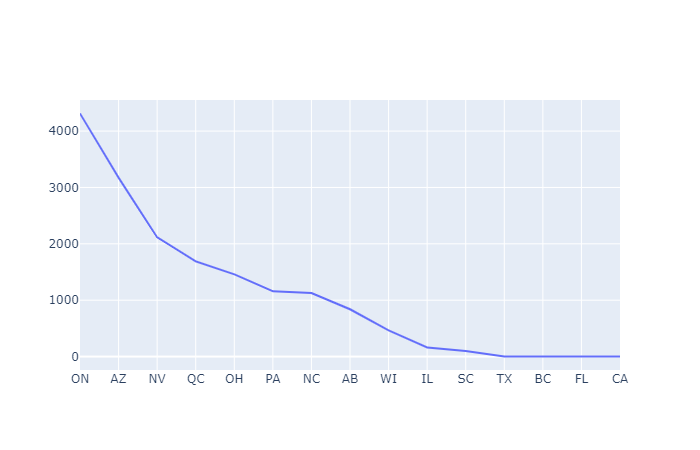
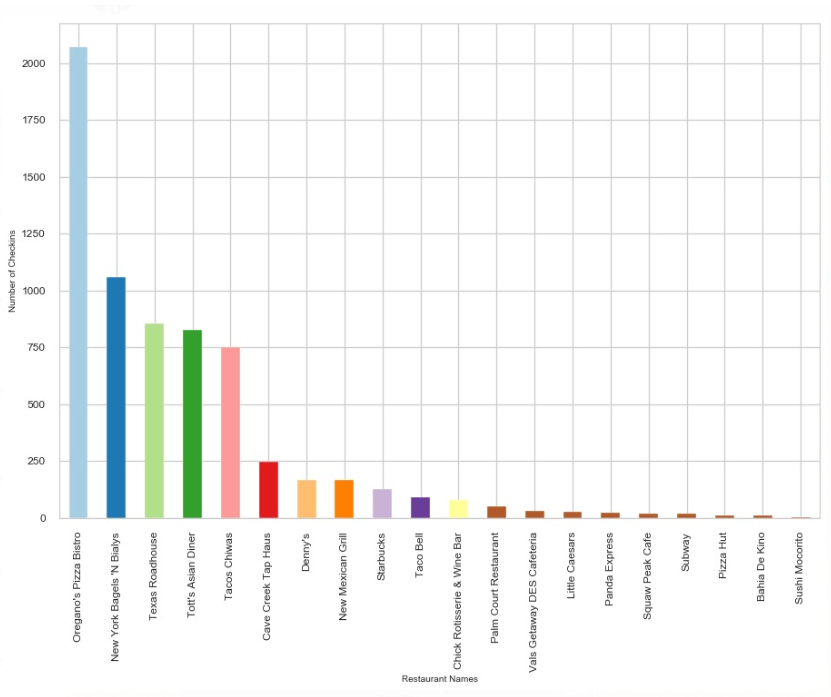
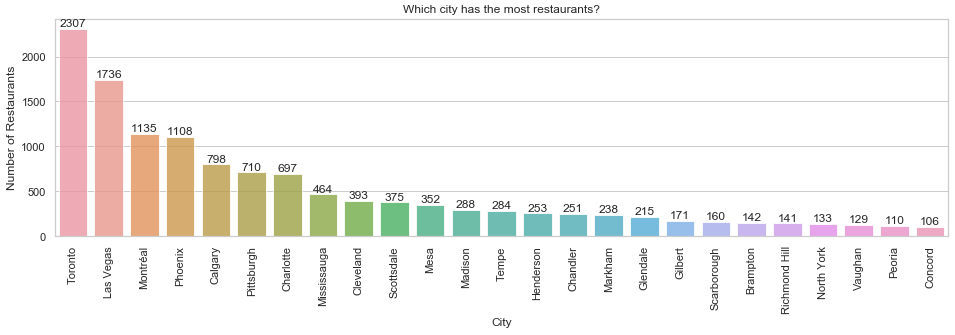
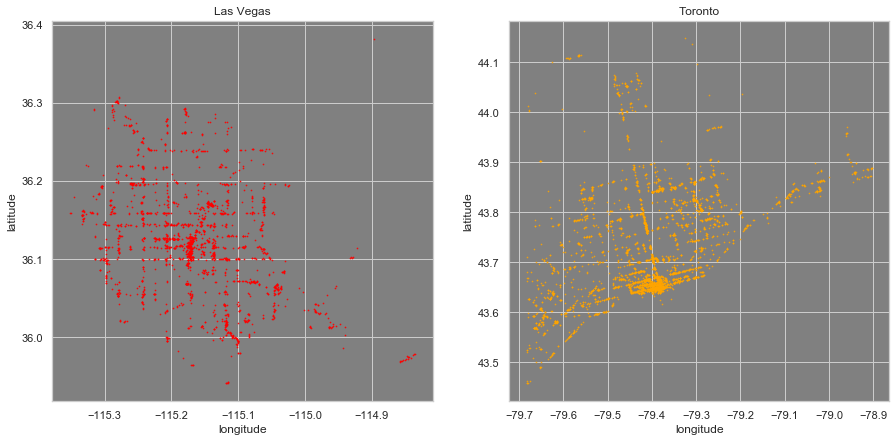


The above pull out chart gives details the distribution of 5-star restaurants among the different states. As seen Arizona has the Highest number of restaurants with 5-star ratings followed by Nevada and Ontario. This graph is particularly helpful to understand that the reviewers in other states are not completely satisfied with the restaurants and this can be also be used for a franchise which has always maintained good reviews to target the states with less 5-star restaurants to open a new outlet.

* Reviews by relevancy  
    
  The above bar graph gives a brief about the restaurants with the breakdown of the reviews. The reviews are classified as useful (good or bad), funny and cool. Of the 3 type of reviews the restaurants use the useful reviews as they are a proper feedback of the food and service and can be used to enhance in any area they are lagging. The Graph also shows that of all the 3 type of reviews; the maximum number of reviews are useful.
* Number of reviews given by a user



This graph is used to understand the user behaviour. For this graph we have kept the count of reviews to 40 i.e. no user who gave more than 40 reviews have been considered. It is seen that the average number of reviews that a user give is less than 5. This can be used by the yelp team to promote users to give more reviews as more the number of reviews the more the search traffic visits their page.

* Rating distribution  
    
  This is one of the most important insight. This pie chart shows the distribution of ratings of restaurants of the entire dataset. As per most of the restaurants are rated 3.5 or 4 stars (48.9%). Only 0.615% restaurants have the least rating i.e. 1 star and only 1.95% restaurants are rated 5 stars.
* Restaurants by State  
    
  The above chart shows the number of restaurants in different states. Ontario has the highest number of restaurants followed by Arizona. States like Texas, Florida and California have very few restaurants registered. This data is helpful for the sales team and marketing team of yelp to target these states and get more restaurants.
* Restaurants with highest number of check-ins  
    
  The above histogram shows the number of check-ins in the state of Arizona (state which has the maximum 5-star restaurants). Oregano’s Pizza Bistro has the highest number of check-ins as per the dataset. Famous Restaurant chains such as Subway and Pizza hut has few numbers of check-ins. This data can be used by Restaurants to encourage user to check-in in yelp and review them.
* Restaurants by City  
    
  The histogram shows the number of restaurants in a city. Toronto has the highest number of restaurants followed by Las Vegas. Such data can be used by business for setting up new restaurants and by yelp sales and marketing team to target cities with fewer number or restaurants.
* Popularity of Restaurants  
    
  This is a heatmap of restaurants based by their popularity. The popularity was calculated by considering the Review count \* Number of stars. The heatmap is of Toronto and Las Vegas which are the top 2 cities by number of restaurants. This chart can be used for business before setting up a new restaurant by opening a restaurant in the more popular area as this would help their business as those are the areas with the high review count areas.

Conclusions and Future Work

In this project, we analyse the data for restaurants of yelp. We found that how much usually user reviews; how much restaurants are there in a city and state; how popular the restaurant is and how much of the user reviews are useful. Finally, which restaurants user check-ins the most in a city and how the data can be useful for Yelp sales and marketing team as well as restaurant owners.

For the future, more analysis can be done on a user to find how does he/she reviews restaurants by a specific cuisine and if the reviews of user have a pattern. Also, how opening of a new restaurant in the same locality has affected the restaurant reviews and popularity. Finally, how factors such as amenities, easy of finding restaurant etc affect the restaurant performance.

References:

1. Restaurant Setup Business Analysis Using Yelp Dataset
2. <https://ieeexplore.ieee.org/document/7962431>
3. <https://ieeexplore.ieee.org/document/8169897>

[1] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, AI Magazine 17(3): 37.

[2] Recommender system for food in a restaurant based on Natural Language Processing and Machine Learning MSc Research Project Data Analytics Kedar Ratnaparkhi x17111013