Restaurants Analysis on Yelp through Visualizations

Jonathan Mendes, Terrance Thomas
National College of Ireland
Dublin, Ireland
X18179584@student.ncirl.ie, X18184928@student.ncirl.ie

***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.**

I. 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. Review 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 segments of job roles and businesses.

A. 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 review the most
- To find number of reviews an average user gives
- To analyze which state has the most restaurants
- Understand how much reviews are useful for restaurants
- Find the number of restaurants in a state that were rated 5 stars by a user

B. Motivation

Analysing the yelp data will help the people from various departments and business with insights which can be used in the future. By analysing the number of restaurants in different cities; the yelp sales and marketing team can target areas 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.

C. Relevance

In today's world the restaurant industry is ever growing, flourishing and is one of the most competitive businesses. Numerous opportunities have been created due to the extensive spread and popularity of online reviews platforms like Yelp in the restaurant industry. Most of these platforms are free and the listing of businesses can be carried out with the help of a few clicks .This makes it ideal for low budget restaurant businesses to grow and flourish with minimal marketing campaign costs. Also, a growing marketing channel can be created by utilizing various social parameters like reviews, ratings, following users, etc. This will spread one's restaurant business to a huge number of users on social networks [1].

D. Elicitation of appropriately formed research question

The principal contributions presented in this work can be summarized as follows:

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

II. Related Work

The project 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 on the cuisine of the restaurant which would be more helpful [2]. Study shows how geographical proximity is important in popularity of a business in different cities [3]. To predict whether a restaurant(business) belongs to positive reviewed class or below it [4].

III. Methodology

To attain the project objective, the KDD approach is followed as mentioned by Fayyad et al. (1996) 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 [5]. 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.

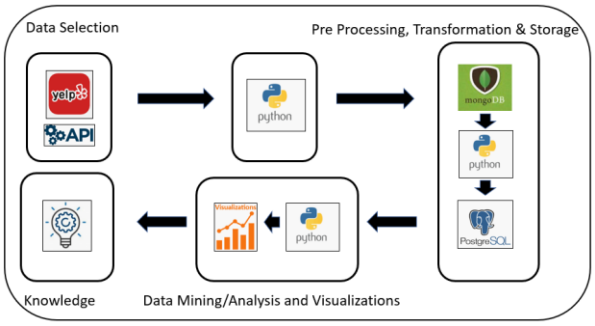

Figure 1: KDD Life Cycle

A. DATASET

The dataset of yelp was taken from Kaggle [6].

- Dataset Justification:

Yelp is a service with has various business 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 desktop and mobile respectively in the first half of 2019. It also stated that its website has 192 million reviews [7]. 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.

- Business Data:

The dataset contains 192,609 rows and 14 variables which comprises of 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.

Business

Columns	Data Type	Description
business_id	string	character unique string business id
name	string	the business's name
address	string	the full address of the business
city	string	the city
state	string	2-character state code, if applicable
postal code	string	the postal code
latitude	float	latitude
longitude	float	longitude
stars	float	star rating, rounded to half-stars
review_count	integer	number of reviews
is_open	integer	0 or 1 for closed or open, respectively
attributes	object	business attributes to values. note: some attribute values might be objects
categories	object	an array of strings of business categories
hours	object	an object of key day to value hours, hours are using a 24hr clock

Figure 2: Busin [3]ess Data Variables

- 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.

Review

Columns	Data Type	Description
review_id	string	character unique review id
user_id	string	character unique user id
business_id	string	character business id
stars	string	star rating
date	string	date formatted YYYY-MM-DD
text	string	the review itself
useful	integer	number of useful votes received
funny	integer	number of funny votes received
cool	integer	number of cool votes received

Figure 3: Review Data Variables

- 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 variables contains list of check-in timings in a timestamp format for a business. Figure 4 displays the check-in variables, datatypes and description.

Checkin

Columns	Data Type	Description
business_id	string	character business id
date	string	string which is a comma-separated list of timestamps for each checkin, each with format YYYY-MM-DD HH:MM:SS

Figure 4: Check-in Data Variables

B. 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 [8]. There are approximately more than 72,000 open source libraries provided by python [9]. 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.

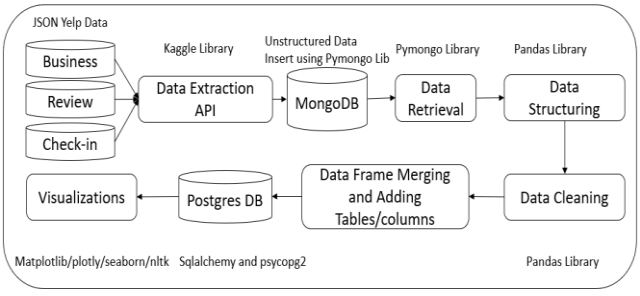

Figure 5: Detailed Flow Diagram

C. 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.

D. 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 cleaned 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 through the lambda function for insertion into the database.

E. FEATURE ENGINEERING

For the purpose of data visualization new columns 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 to be used for visualizations through charts and graphs.

F. DATA STORAGE

The data is stored in MongoDB and PostgreSQL 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 chuck size 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.

G. 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 [9]. 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.

IV. RESULTS

Using the data after cleaning process and merging the required data the following visualizations were created to answer the research question.

- States with most 5-star restaurants

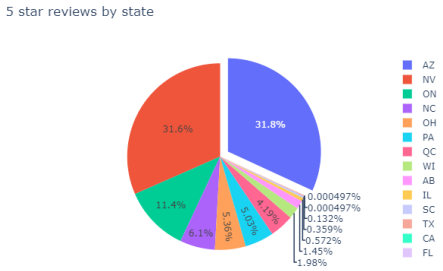

Figure 6: Pull out chart of states with most 5- star restaurants

The chart in Figure 6 gives details of 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 used for a franchise which has always maintained good reviews to target the states with less 5-star restaurants to open new outlets.

- Reviews by relevancy

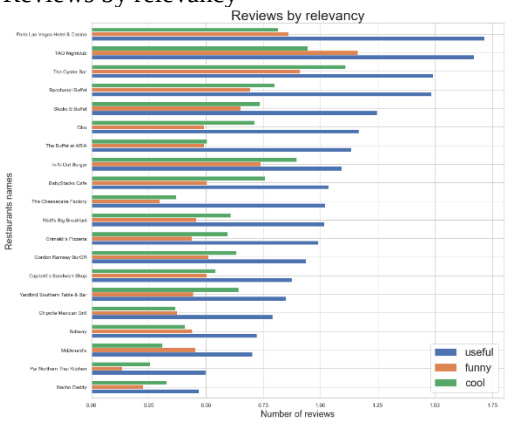

Figure 7: Horizontal graph of review relevancy

Figure 7 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 is of useful.

- Number of reviews given by average user

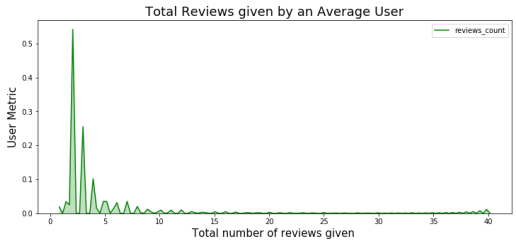

Figure 8: Reviews given by user

Figure 8 is used to understand the user behaviour. 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 reviews given by an average user 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

Figure 9: Donut Chart of restaurant rating distribution

This is one of the most important insights. Figure 9 shows the distribution of ratings of restaurants in entire dataset. As per Figure 9, 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

Figure 10: Line chart of Restaurant number in different cities

Figure 10 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

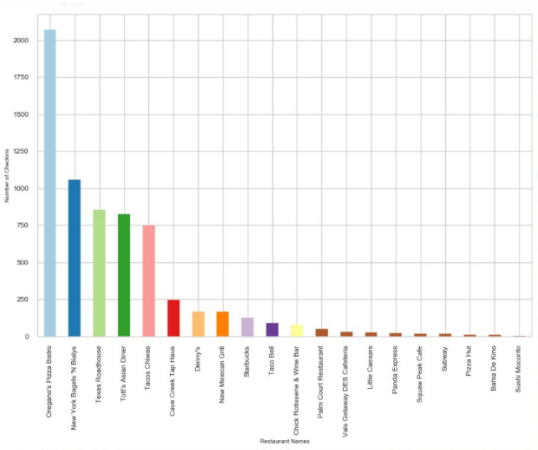

Figure 11: Histogram of restaurants by number of check-ins

Figure 11 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 have few numbers of check-ins. This data can be used by Restaurants to encourage user to check-in in yelp and review them.

- Number of Restaurants by City

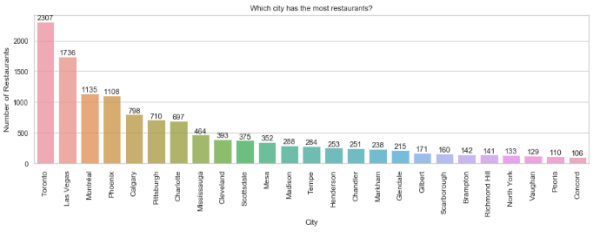

Figure 12: Histogram of Restaurants in different cities

Figure 12 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 businesses for setting up new restaurants and by yelp sales and marketing team to target cities with fewer number or restaurants.

- Popularity of Restaurants

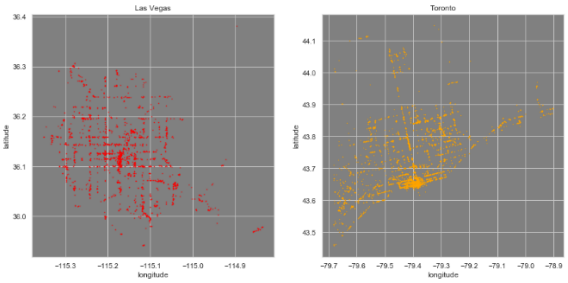

Figure 13: Heatmap of Restaurant popularity

Figure 13 is 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. Opening a restaurant in the more popular area as this would help their business as those are the areas with the high review count.

V. CONCLUSIONS AND FUTURE WORK

In this project, we analyse the data for restaurants registered in yelp. We found that how frequently a user reviews; the number of restaurants 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 the 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 review 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, easily accessible restaurant, etc. affect the restaurant performance.

VI. CHALLENGES

- The massive data size was one of the biggest challenges which we resolved using the chunk size attribute while inserting the data into the database. The data was inserted in batches to decrease the RAM usage and prevent memory errors
- The code had to be optimised cause the automation was crashing due to the low machine config. It worked smoothly on a 16GB RAM configuration
- Fetching the datasets through API from website. A significant amount of research and time was spent on it
- Merging the dataset code together and implementing the views created in PostgreSQL in python using pandas for automation was a challenge which was overcome by brainstorming and research learning

VII. REFERENCES

[1] V. L. K. J. S. a. K. P. S. M. Prithivirajan, "Analysis of star ratings in consumer reviews: A case study of Yelp," in *2015 IEEE International Conference on Big Data (Big Data)*, Santa Clara, CA, USA, 2015.

[2] S. S. a. S. S. S. Hegde, "Restaurant setup business analysis using yelp dataset," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, India, 2017.

[3] S. S. a. B. M. A. K. Bhowmick, "Effect of Information Propagation on Business Popularity: A Case Study on Yelp," in *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, Daejeon, South Korea, 2017.

[4] D. Kaing, "Yelp business rating classification using hybrid ensemble," in *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, London, UK, 2017.

[5] G. P.-S. a. P. S. Usama Fayyad, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 3, p. 18, 1996.

[6] I. Yelp, "Kaggle," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/yelp-dataset/yelp-dataset. [Accessed 14 October 2019].

[7] "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Yelp. [Accessed 20 November 2019].

[8] C. O. DONNABHAIN, "Irish Times," 2 Janurary 2019. [Online]. Available: https://irishtechnews.ie/how-python-is-used-in-data-science/. [Accessed 10 December 2019].

[9] A. Ohri, Python for R Users: A Data Science Approach, Wiley, 2017.