|  |
| --- |
| **Team members:**  Arunkumar Senthilkumar Student number: 100888951  Jayarani Rajesh Student number: 100818850  Rahul Sapparapu Student number: 100890409  2-3-2023 |

|  |
| --- |
| Yelp Inc. |
| Proposal report |
| Capstone DATA2205 |

**Table of Contents**

[About Yelp: 2](#_Toc126152681)

[Project description: 2](#_Toc126152682)

[Objectives: 2](#_Toc126152683)

[Data Source: 2](#_Toc126152684)

[Methodology: 2](#_Toc126152685)

[Tools and Technologies: 3](#_Toc126152686)

[Deliverables: 3](#_Toc126152687)

[Execution: 3](#_Toc126152688)

[About the Dataset: 3](#_Toc126152689)

[Conclusion: 5](#_Toc126152690)

[References: 5](#_Toc126152691)

# **About Yelp:**

**Yelp Inc. is an American company that develops Yelp.com and the Yelp mobile app, which publishes reviews submitted by customers about businesses.** The company also operates Yelp Guest Manager, a service that allows users to book tables. The company is headquartered in San Francisco, California.

**Russel Simmons** and **Jeremy Stoppelman**, former PayPal employees, founded Yelp in **2004**. Following its launch, Yelp raised several rounds of funding and grew in usage. About 4.5 million crowd-sourced reviews were published by the website by 2010, generating $30 million in revenue.

# **Project description:**

The Yelp dataset contains a wealth of information about the restaurant industry, including the type of cuisine, location, ratings, and user reviews. This project aims to use the Yelp dataset to gain insights into the restaurant industry and answer key questions, such as:

* What type of cuisine is the most popular among Yelp users?
* What are the top factors that determine the ratings of a restaurant?
* How does the location of a restaurant impact its popularity among Yelp users?

# **Objectives:**

* To explore the Yelp dataset and gain an understanding of the data.
* To perform data cleaning and preprocessing to prepare the data for analysis.
* To conduct exploratory data analysis to identify patterns and relationships in the data.
* To build predictive models to determine the factors that impact the ratings of a restaurant.
* To visualize the results of the analysis and interpret the findings.

# **Data Source:**

The data for this project will be obtained from the Yelp Open Dataset, which contains information about local businesses and user reviews for millions of businesses in the United States. The dataset includes data on various business attributes such as business type, location, and pricing, as well as information on user ratings and reviews.

Link: <https://www.yelp.com/dataset>

# **Methodology:**

* Data collection: The Yelp dataset will be collected from the official Yelp website.
* Data preprocessing: The collected data will be cleaned and preprocessed to prepare it for analysis. This includes dealing with missing values, removing duplicates, and converting categorical variables into numerical variables.
* Exploratory Data Analysis: The preprocessed data will be used to perform exploratory data analysis using various techniques, including univariate and bivariate analysis, data visualization, and hypothesis testing.
* Predictive Modeling: Machine learning algorithms will be used to build predictive models that determine the factors that impact the ratings of a restaurant.
* Evaluation: The performance of the predictive models will be evaluated using appropriate metrics, such as accuracy, precision, recall, and F1 score.

# **Tools and Technologies:**

* Python programming language
* Libraries such as Pandas, Numpy, Matplotlib, Seaborn, and Scikit-learn.
* Jupyter Notebook for data analysis
* Power BI for visualization
* Microsoft Office suite

# **Deliverables:**

Upon completion the following will be delivered to the client:

1. Exploratory analysis script in python
2. Predictive model
3. Dataset used to train/test the model
4. Visualization for the analysis results
5. Project related documents

# **Execution:**

**Resources:**

The resources readily available for the project are:

* Dataset available in json format
* Tools and software

**Process:**

* Research
* Design
* Development
* Visualization
* Documentation

# **About the Dataset:**

The Yelp Dataset is a comprehensive collection of information about local businesses and user reviews. It includes data on businesses in the United States and is updated regularly to ensure accuracy and relevance. The dataset includes information on a wide range of business attributes, including the type of business, location, and pricing, as well as user ratings and reviews.

Some of the key attributes included in the Yelp Dataset are:

1. Business Information: This includes the name of the business, address, city, state, postal code, latitude and longitude, and phone number.
2. Business Attributes: This includes information on the type of business, such as restaurants, shopping, or entertainment, as well as details on the business's attributes, such as whether it has a parking lot, wheelchair accessibility, and more.
3. User Ratings: This includes the overall rating for the business, as well as individual ratings for various aspects of the business, such as the quality of service, atmosphere, and food.
4. User Reviews: This includes written reviews from users, providing valuable insights into the customer experience and the business's strengths and weaknesses.
5. Categories: This includes a list of categories associated with the business, such as cuisine type, entertainment type, or shopping type.
6. Check-in Data: This includes information on the number of check-ins to the business, as well as the days and times that the business is busiest.
7. Tips: This includes written tips from users, providing additional insights into the customer experience.

Overall, the Yelp Dataset provides a wealth of information that can be used to explore the relationship between businesses and users. Whether a data analytics project is focused on understanding the impact of business location on user ratings or identifying the key drivers of high ratings, the Yelp Dataset provides the data necessary to support these analyses.

Here are some **potential questions** that could be explored using the Yelp dataset:

1. What is the distribution of restaurant ratings on Yelp?
2. Which type of cuisine is the most popular among Yelp users?
3. What are the factors that contribute to a high rating for a restaurant?
4. Does the location of a restaurant impact its popularity among Yelp users?
5. How does the price range of a restaurant relate to its rating?
6. Are there any trends in the type of cuisine or restaurant attributes that are popular in different regions?
7. Can we predict the rating of a restaurant based on its attributes and user reviews?
8. How does the number of reviews for a restaurant impact its rating?
9. Are there any patterns in the times and days when restaurants receive the most positive reviews?
10. Can we identify the most common reasons for negative reviews for a particular type of cuisine or restaurant attribute?

Here are some **High-level questions** that can be answered using the Yelp dataset:

1. Popularity of Cuisines: What type of cuisine is most popular among Yelp users? How has the popularity of different cuisines changed over time?
2. Restaurant Ratings: What factors contribute the most to the ratings of a restaurant? How do ratings vary by city and state?
3. Location Analysis: How does the location of a restaurant impact its popularity among Yelp users? How does the population density in a particular area affect the number of restaurants and their ratings?
4. User Demographics: Who are the typical users of Yelp, and how do they differ in their dining preferences and ratings?
5. Sentiment Analysis: What are the most common sentiments expressed in Yelp reviews, and how do these sentiments vary across different restaurants and cuisines?
6. Time Series Analysis: How do the ratings of a restaurant change over time? What are the seasonal trends in the restaurant industry?
7. Recommendation Systems: How can a recommendation system be developed using the Yelp dataset to suggest restaurants to users based on their dining preferences and previous ratings?
8. Predictive Modeling: Can predictive models be built to predict the ratings of a restaurant based on its attributes and user reviews?
9. Clustering: Can the restaurants be grouped into clusters based on their attributes and ratings, and what do these clusters reveal about the restaurant industry?
10. Network Analysis: Can the relationships between restaurants and users be analyzed using network analysis techniques to identify influencers and popular restaurants?

# **Conclusion:**

Our team would be elated to work on a project that provides insights into the restaurant industry and answers key questions about why restaurants are rated as they are. As a result, restaurant owners and customers will receive valuable information that will assist them in making informed decisions in the future.

# **References:**

Yelp. (2023, January 14). In *Wikipedia*. <https://en.wikipedia.org/wiki/Yelp>

Yelp,Inc. (2022,February 04). *Yelp dataset.*  <https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?resource=download>

Yelp,Inc. (2023). *About Yelp.* <https://www.yelp.com/about>