Customer Feedback Analysis

By

Brahma Reddy Gade

Submitted to
The University of Roehampton

In partial fulfilment of the requirements
for the degree of

MASTER OF SCIENCE IN DATA SCIENCE

Abstract

Internet users are growing exponentially for the last one decade and almost every purchase or business depends on the customer feedback. Feedback from the customers helps both the business and customers in making the right decisions for improving the business and purchasing the best product for customers. So, customer feedback plays an important role in today's internet world. In order to analyse the customer feedback manually it is impossible to get the insights from these reviews as we had large volumes of information. Analysing the customer feedback through sentimental analysis and application of machine learning techniques helps to automate the feedback process and provide valuable insights to the business and enhance the customer satisfaction. In this research I had formulated three different aims to analyse the customer feedback. All the project aims and objectives are successfully evaluated and plotted the results. The main goal of my project was to analyse the customer feedback and provide valuable insights to the business and enhance the customer satisfaction. I had done major research on dataset analysis and applied different statistical approaches for analysing the data also I had successfully implemented different machine learning algorithms by finding the best efficient algorithm for customer feedback data. Moreover, I had successfully implemented sentimental analysis for analysing the sentiments of customer review data. Lastly, I had visualized various types of plots to analyse the feedback and view the insights in a clean and understandable format of any individual user. Each and every mentioned aims and objectives of my customer feedback project are evaluated successfully and furnished with results. We can do a lot of research in this customer feedback but due to project time constraint, I had sticked to specific goals of the project in order to complete within the time frame.

Declaration

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.

Enter your name here
Brahma Reddy Gade
Date: 06/09/2023

Signed (apply signature below)

Brahmareddy

Acknowledgements

Firstly, I would like to express my heartfelt gratitude for the unwavering support and encouragement to the people who have provided throughout my academic journey and in the completion of my Customer Feedback Analysis project at **University of Roehampton**.

I am very grateful to my project supervisor **'Mohammad F Khan'** for his guidance, feedback and encouragement throughout the project. His support has been invaluable in my pursuit of knowledge and personal growth towards the success of my project.

Table of Contents

1. Introduction

With the increase of online platforms & social media, businesses receive vast amount of customer feedback data in the form of surveys, emails and social media posts. Customer feedback consists of wide range of topics & issues, our project aims to categorize different feedback into relevant topic and translate to valuable insights to the company. We are going to analyse the customer mood towards a particular restaurant either it may be positive, negative or neutral. Our project main goal was to analyse the customer feedback and provide useful insights to the company. The reviews are taken into consideration by that respective company towards their product price, brand, features so that it can enhance more positive feedback from customers by rectifying the problems.

Research Question or Problem statement:

As internet users are growing exponentially many users are preferring to purchase on online and to know the review of business or company. Opinions reviewed by other customers plays an important role for the successful running of their businesses. As many people prefer to write online reviews it is difficult to track every customer need because as some reviews are too long, ambiguous and short which makes it more complex for analysing the data. Almost all the organizations and stakeholders who runs the business will be affected by this problem. Even customers also affected by this problem because of unaddressed issues, lack of product improvements etc. Basically, these problems occur in almost all the businesses and companies in many ways. They are: Businesses will completely degrade their performance because as it is crucial to understand customer needs and preferences. The product development team gets valuable insights from customers to make various improvement of the product without customer feedback it suffers a major loss in terms of sales and potentially losing out to its competitors. Marketing and sales team of the company will degrade their performance without customer feedback. Future customers of the company will also be impacted without proper customer feedback for a business as it reduces brand values of the company or product. According to recent study shows that about 70% of users read online review and make purchase. About 90% of consumers read online reviews before visiting a business or restaurant etc. People are trusting online reviews as much as personal recommendations. Online reviews have become a perfect validator for the users to know about the business. So, Customer Feedback Analysis plays an important role for analysing the data and present it in a visualized format for the business.

Aims:

1. **Understand Customer Sentiment and Monitor Brand Reputation:** One of the main aims of the project is to identify the customer needs and preferences thereby evaluating the sentiment towards the product. Based on the customer feedback organizations respond to customer concerns, issues effectively, so that it can proactively manage their brand image.
2. **Apply different Machine Learning models to Customer Feedback Analysis:** Other aim of the project is to find an efficient machine learning model and achieve better accuracy. Through this we can employ different machine learning models on the same data and finalize efficient model.
3. **Improve Customer Experience and Enhance Decision Making through Visualization:** Enhance the customer feedback by understanding the customer challenges, issues etc. Companies can identify where necessary improvements are needed for enhancing the customer experience. By analysing the customer feedback visualizations, organizations can improve product design, customer service strategies and resource allocation.

Objectives:

1. **Data Collection:** On of the main objective of the project was to collect efficient dataset so that it can be analysed effectively to get useful insights.
2. **Data Pre-processing:** In order to maintain quality of the dataset we should effectively pre-process the dataset such as Data Cleaning, Data Transformation and Organizing the data.
3. **Model Development:** Another important objective is to find different Natural Language Processing techniques, so that it helps in processing human language, enabling sentimental analysis and text categorization.
4. **Model Training:** Here we will employ different machine learning techniques to the processed data. This shows a comparative case study between different machine learning models to perform sentiment analysis on the customer feedback and finding the best efficient model for the given data.
5. **Visualization:** After employing Natural Language Processing and Machine Learning techniques we will provide valuable insights to the organizations by visualizing the data in meaningful and understandable format.
6. **Performance Evaluation:** The final objective is to evaluate the performance of the model by using various performance evaluation metrics and validation techniques.

Legal, Social, Ethical and Professional Considerations:

The execution of this project involves various Legal, Social, Ethical and Professional Considerations that requires careful attention for protection of the rights.

Data Usage Rights, Data Privacy are the legal considerations which are to be handled in this project. Ensure that we have the legal rights to access and analyze the Yelp business dataset. Always review the terms of usage and licensing agreements provided by the dataset source. We had followed all the legal rights and licensing agreements in order to protect the Data Usage Rights. Data Privacy means to protect user privacy such as their names, contact details by Anonymizing the data. Here the sensitive information such as names, business id's data are completely anonymized throughout the project.

Result Analysis of the businesses could impact their reputation and success. However, biases could impact businesses and communities unfairly. As we are not publishing our findings online, we no need to consider the social impacts of our project.

User Data and Privacy are the important Ethical considerations need to be taken care for our project. As most of the user content data is anonymized, we are able to handle the ethical consideration of our project. Always define the purpose, scope, methodology of our project by clearly documenting the process, tools used and decisions made for the analysis process.

Always adhere to professional attributes such as ensuring the quality of the data, provide credit to external sources of used tools or libraries. Approach the analysis in a systematic, scientific and unbiased manner and present the findings truthfully.

The above-mentioned points are the important legal, social, ethical and professional considerations which are to be considered in our project. All the considerations are handled effectively by following the systematic procedures.

Background:

In today's digital age, online customer reviews have become a significant source of information for consumers seeking insights into various products, services, and businesses. Platforms like Yelp provide users with the ability to share their experiences and opinions about local businesses, contributing to a vast repository of user-generated content [6]. Netflix always asks new users to rate their interest on watched movies so it can recommend new movies to users. If I have collected the business feedback over recent years, 80% of the reviews was completely online. Even after having a reasonable amount of customer feedback to analyze but after going through the feedback you don’t have any takeaways or insights to share with the team. Even 40% of the marketer’s insights are not actionable. Coming to vast amount of feedback organizations receive everyday we can’t get the useful insights directly from the customer feedback data. Recent studies indicate that 70% of the users read and consult online review before making a purchase. 90% of the consumers read online reviews before visiting a business or purchasing a product. [3]These research shows how people are addicted towards online review for purchasing a product, investing in a business. Feedback can shed light on employee behaviour and performance. Businesses can use this information to improve training and customer interactions, which, in turn, can positively impact customer experiences. Monitoring customer feedback can reveal emerging trends and shifts in customer preferences, allowing businesses to anticipate and prepare for future demands. Innovative ideas often stem from customer suggestions and feedback. Analysing customer feedback can inspire businesses to develop new features, services, or offerings that stand out in the market[6]. Sentiment analysis helps businesses gain a clear understanding of the sentiment expressed in customer feedback. It classifies feedback as positive, negative, or neutral, allowing for a quick summary of overall sentiment. Sentiment analysis can aggregate feedback from various sources, including surveys, social media, reviews, and customer support interactions. This holistic view of sentiment helps in comprehensive decision-making. Sentiment analysis can be applied in real-time, allowing businesses to monitor customer sentiment as it evolves. Immediate responses to negative sentiment can prevent issues from escalating. In summary, customer feedback analysis empowers businesses to align their offerings with customer expectations, leading to improved customer satisfaction, loyalty, and competitiveness. It transforms subjective opinions into actionable insights that guide decision-making and strategy development across all aspects of a business. The above points show that why Customer Feedback Analysis is an important topic in today’s digital world. So that’s why I had chosen Customer Feedback Analysis as my dissertation topic to make more research on analysing the customer feedback.

Report overview:

This report tends to explore the various trends of the customer feedback, sentimental analysis of customer feedback and application of machine learning algorithms. Coming to view the upcoming sections of report I had categorized into different sections.

Literature Review section describes a comprehensive study of existing research techniques for analysing the customer feedback. Both Literature Review and Technology Review of various researches are tabulated in a tabular format to describe the prominence of each research paper.

Coming to methodology section it describes about the systematic approach and techniques that I used to carry out my research or analysis. Data Collection, Applying Statistical Analysis, Data Pre-processing, Application of Machine Learning algorithms, Sentimental Analysis, Visualization are the different techniques used for my project.

The Implementation section describes about how the experimental details are carried out and it involves executing the tasks and activities outlined in my project methodology and plan. It describes each and every methodology technique in detail.

The results section describes the findings of the project. It is the crucial part of the project where as it provides evidence to support my conclusions and helps readers to understand the significance of my work. It shows about complete evidence to visually represent the findings of my work.

Conclusion section is a critical section where it summarizes the key findings, discuss their implications, and offer insights or recommendations based on my project work. It describes about what we had achieved through this project. It represents about future enhancements towards the existing research.

All the references related to this report are furnished in reference section. Project related work that supports the research is defined in the appendix section.

2. Literature - Technology Review

Literature Review:

The authors of [1] proposed a comparative case of different machine learning models for analysing the sentiment on amazon customer product reviews. In this paper [1] after pre-processing they used Exploratory data analysis to showcase various visualizations of the data. Secondly, they used different machine learning models to show a comparative case study on the same data. Out of different machine learning models [Naive Bayes, Random Forest Classifier, Stochastic Gradient Descent, BERT (Bidirectional Encoder Representations from Transformers)]. BERT model produces more efficiency when compared with other models. Coming to pros and cons of this research are making Exploratory data analysis with different visualizations, they also performed a comparative case study of various machine learning models on the same data, they also performed tuning the parameters of previously executed models and tested it using CountVectorizer and TF-IDF Tokenizer. This research paper was closely related to my project goals, as this model helps to utilize various machine learning models and make a comparative case study of various different machine learning techniques on the same data, it also helps me to analyse the data and visually represent the data in various formats of visualization. This research was specifically applied on Electronics category of Amazon dataset and it can't be analysed on other categories of the dataset as different categories has different schema data. This research was very close to my project goal for making a comparison of different machine learning models. It also helps me to visualize Exploratory analysis on the data. This paper avail me to learn tune the parameters of executed models and also tested it using CountVectorizer and TF-IDF Tokenizer for improving the accuracy of the models.

Towards analysis for sentiment detection of customer reviews was published in [2]. This paper proposed the detection of sentiment analysis using voice components and deep learning on amazon dataset. Firstly, through pre-processing they cleaned the data and later by using various word embedding techniques [BERT, ELMo, GloVe, FASTTEXT] for creating a bag of word vectors. After completion of word embedding, they used different deep learning models [CNN, BLSTM, Multichannel CNN, RMDL] to carry out the sentiment detection. Among all of the models, the Multi-channel CNN model with Fast Text classifier produces better accuracy when compared with the other models. Regarding pros and cons of this research are it uses different Deep Learning techniques to analyse the data, moreover this research uses different word embedding techniques to embed various kinds of data. This paper avails me to learn word embedding techniques, applying different deep learning models using NLP techniques and making a comparative case study of the models. This research proves that among all the combinations of the various parts of speech, a verb, an adverb and an adjective is the most effective combination. It helps me to learn various embedding techniques to create a bag of word vectors of different voice component data.

In the next paper authors proposed a machine learning model for analysing the sentiment on amazon customer product reviews. The authors of [3] have proposed a Machine Learning model to analyse the customer feedback analysis. In this paper they had extracted data using 'Webharvy' tool for reviews from amazon sites. In the next step they had used Pre-processing techniques to remove any stop words etc. Here they segregated into two categories like Product related and Support related review comments. By using Support Vector Machine classifier, they classified the features and represented them in a statistical report. With respect to pros and cons of this research are, they extracted the data using an extraction tool from amazon website without going for traditional approach like downloading the dataset, they used a machine learning algorithm to classify the product and support related review comments. It attains me to learn different pre-processing techniques which is helpful to my problem, moreover it assists to get the data in different formats from various websites using web extraction tools. This research helps me to provide proper insights to the company by analysing the data in a clean and attractive format.

The research paper [4] aims at running sentimental analysis on product reviews. Here they proposed three phases for analysing the data. Firstly, they used Data Filtration technique to gather the data from different sources and later combining them into a single source and later cleaning the data into usable format. In the Training phase they extract the text paragraph from dataset, later they extract words corresponding to adjective, adverb, verb. Then all positive words are labelled as 'pos' words and negative words are labelled as 'neg'. In the Final step they proposed an algorithm to perform sentiment analysis. Here the combination of adjective, adverb and verb are turned out to be the best combination among various parts of speech on test data. The pros and cons of this research are analysing the reviews using parts of speech words and this approach can be used for text related feedback and this approach is not used for other than text related reviews. This research throws a different approach for analysing the data using parts of speech approach. This research was near to my problem as this involves applying machine learning algorithms, which relates to goals of my problem. This paper helps me to apply machine learning models for the customer feedback data and produces accuracy results.

Coming to paper [5] proposed a technique to summarize and analyse the product reviews using text mining and Neuro Linguistic programming approach. In this approach the input is taken through a web interface from various clients. The reviews data is stored in a dedicated database which in turn to be processed for reviews. They proposed a classifier set of NLP predicates which is compared with the processed text and the results will be displayed. Here their main objective is to provide a platform for the customer to provide their feedback directly so that it can be analysed individually and this is mostly useful for Beta-version related products because we can analyse it for a smaller number of customers. This approach is not useful for processing the customer reviews of large-scale companies like amazon, eBay etc, because we can't process each and every customer feedback individually. This literature was helpful for my problem as this relates to review the feedback of customers using a Neuro Linguistic approach. This research technique helps me to know about processing feedback by developing a web interface, handling the individual customer feedback and also classifying the data by comparing with classifier set of NLP predicates.

Coming to entire section of literature review, each of the paper reviewed was relevant to my problem statement. Most of the techniques and methodology used for analysing the customer feedback data of various research papers are related to my project aims and objectives. Almost all the research papers follow similar aims and objectives of my project. All the research papers assist me to achieve my project aims and objectives through the techniques and methodology followed in those researches. Each of the research paper has its own advantages and disadvantages and I will use these pros to achieve the goals of my project.

S. No	Research Paper	Published Year	Technologies Used	Relevant to my Project
1.	Sentiment Analysis on Amazon product reviews	2022	Naive Bayes, Random Forest Classifier, Stochastic Gradient Descent, BERT Models Fine tuning techniques	This research was more closely related to my project as it involves application of various machine learning models.

2.	Towards improving e-commerce customer review analysis for sentiment detection.	2022	Word Embedding Techniques [BERT, ELMo, GloVe, FASTTEXT], Deep Learning models [CNN, BLSTM, Multichannel CNN, RMDL]	This paper helps to learn new techniques like Word Embedding and Deep Learning models.
3.	Customer Feedback Analysis Using Machine Learning.	2019	WebHarvy Extraction Tool, Support Vector Machine Classifier	This research also related to my project as it involves employing machine learning technique
4.	Sentimental Analysis of product reviews.	2019	Data Filtration, Used Parts of Speech approach	This research assists me to learn a new approach for analysing the data.
5.	Neuro Lingusitic Programming Approach.	2019	Web Interface, NLP predicates	This research helps to process customer feedback individually.

Table 1: Comparison of different literature researches

Technology Review:

Coming to technology review of customer feedback analysis project we had wide range of options to analyse the feedback and visually represent it.

We can use either Python or R programming language to interpret the feedback and represent it visually. Each of the programming language has its own advantages and disadvantages based on their abundant libraries and packages.

We can use different Natural Language Processing libraries to analyse customer review text for various functionalities for tokenization, named entity recognition, speech tagging etc.

We had various sentiment analysis API's like Google Cloud Natural Language API, IBM Watson, or Microsoft Azure Text Analytics to quickly determine the sentiment of customer reviews (positive, negative, neutral) without building the model from scratch.

Train different text classification models using machine learning algorithms like Support Vector Machines, Naive Bayes to categorize them into classes or topics.

Use prebuilt Aspect-Based Sentiment Analysis Tools to extract the sentiment towards specific aspects or features of the reviews. Employ Named Entity Recognition (NER) models[21] to identify and classify named entities such as product name, location and people from customer review text.

Employ different data visualization libraries such as Seaborn, Matplotlib or Plotly to create customer feedback analysis using word clouds, sentiment heatmaps and bar charts.

We can reduce lengthy customer reviews by using text summarization algorithms such as TextRank or BERT-based models.

If we want to analyse large volumes of customer data with scalable and cost-effective solutions we use cloud-based NLP services like Amazon Comprehend, Google Cloud Natural Language, or Azure Text Analytics.

In order to analyse customer reviews and feedback from various social media platforms we use social media monitoring tools and platforms like Hootsuite, Brandwatch, or Sprout Social.

If we want to extract specific theme or topic from the customer reviews, we can use topic modeling algorithms like Latent Dirichlet Allocation (LDA) or Non-Negative Matrix Factorization (NMF) to extract the theme or topics in the reviews.

We had different technology options to analyse the customer feedback but it is necessary to know which techniques are essential for our project data. We need to choose these techniques wisely among the available options which should produce comprehensive insights to the companies.

Coming to my Customer Feedback Analysis project we need to employ some of these techniques in order to produce better and efficient visualizations. Coming to programming language we are going to use Python in our project as it is general purpose, high level, interpreted programming language. Python has different and useful libraries when compared with R programming language. Moreover, python had large community and for data analysis python was the best programming language whereas R was specifically designed for mathematical and statistical problems. So, we are going to use Python programming language and we had different python libraries for visualizations. For tokenizing the words of customer reviews, I am going to use Natural Language Processing (NLP) libraries. As I am going to build a model from the scratch, I am not going to utilize Sentiment Analysis API's. I am not going to utilize any text summarization techniques as we can process text feedback data directly, I am not going to process any data from the social media platforms so it is not necessary to employ social media platform tools, as our data can be managed directly I am not going to utilize any cloud based NLP services for processing large amounts of data. As I need to analyse text classification, I am going to different text classification models for classifying the data. Support Vector Machines, Random Forest Classifier, Stochastic Gradient Descent, Navie Bayes are the different machine learning models I am going to employ in my project. I am not going to extract aspect based sentiment analysis as I am not focusing on any specific aspect to extract the sentiment. Coming to data visualization I am going to employ different libraries of python like Matplotlib, Seaborn or Plotly to visually represent the analysis of our data.

S. No	Technology	Technique Using in Project[Yes/No]	Explanation
1.	Natural Language Libraries	Yes	As I need to tokenize the data I need to use the NLP libraries
2.	Sentiment Analysis API's	No	As I am building from the scratch, there is no need of using Built-in API's
3.	Text Classification Models	Yes	It is necessary to classify the data so I need to use text classification models.
4.	Aspect-Based Sentiment Analysis Tools	No	Here I am not to target specific features or aspects of the dataset.
5.	Data Visualization Libraries	Yes	Need to visually represent the data by data visualization libraries.
6.	Social Media Monitoring Platforms	No	No need to use social media monitoring platforms as I am not using any social media platforms data.
7.	Text Summarization	No	As the reviews from the customers are not too lengthy no need to use text summarization techniques.
8.	Cloud-Based NLP Services	No	Not going to utilize Cloud-Based NLP services as I am not going to process large amount of data.

Table 2: Comparison of different Technology researches

3. Methodology

For Analysing the customer feedback, we are going to employ different machine learning algorithms and find the best machine learning model for this Yelp business dataset. We are going to represent the entire dataset visually and statistically to understand its main characteristics and features through Exploratory Data Analysis, this EDA technique is employed and cited to this paper [1]. By using Feature Engineering, I need to create new feature that might provide more meaningful relationship to the data. By using different pre-processing techniques, I am going process the data for better analysis of the data and this pre-processing technique is cited to this paper [2]. For employing different machine learning algorithms, I had referenced to these papers [1][4]. The reason for choosing specific machine learning algorithms are, these algorithms proven a case study of performing better in case customer review data and also, we are employing some other machine learning techniques to find out best machine learning model. The reason for choosing specific pre-processing techniques are these methods are mostly used for different customer analysis datasets. Here I am going to visually represent the data by analysing with various visualizations in a clean and understandable format.

Data Collection:

Data collection is the process of gathering and measuring information on variables of interest in a systematic and organized way. It is a fundamental step in the research and analysis of data for various purposes, such as scientific research, business decision-making, policy development, and more[8][11]. In order to analyse customer feedback, we need a business data. I had used Yelp Business dataset for customer feedback analysis.

Statistical Analysis:

Statistical analysis is a process of collecting, cleaning, summarizing, analysing, interpreting, and presenting data to uncover patterns, trends, relationships, and insights within the data. In this step we are going to analyse the dataset completely as it contains three different datasets. In this section we will apply Statistical Analysis and find the mathematical results of the dataset. We are going visualize the dataset using various plots (box plots, bar graphs) to know the complete analysis of the dataset.

Data Pre-processing:

Data pre-processing is a crucial step in the data analysis pipeline, where raw data is cleaned, transformed, and prepared for further analysis. Proper data pre-processing ensures that the data is of high quality, free from errors or inconsistencies, and in a format that can be effectively used for modelling, visualization, or other analytical tasks[17]. Next, we will try to remove any unwanted columns in the dataset. We will try to add any new columns which will be used for performing better data analysis[14]. We are going to use different NLP techniques in order to remove punctuations, stop words and escape characters.

We will go through different pre-processing techniques such as:

- Removing Punctuation
- Removing Stop words
- Lemmatization

Applying Machine Learning Techniques:

Machine Learning is the process of enabling computers to learn and make predictions or decisions based on data, without being explicitly programmed. Here we will employ different machine learning algorithms to the pre-processed data and find out the best accuracy among the trained models. The different machine learning algorithms we are going to apply are Multinomial Naive Bayes, Logistic Regression, Gradient Boosting Classifier[16] etc.

Sentimental Analysis:

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine and extract the sentiment or emotional tone expressed in text data[17]. In our project we will use sentiment analysis to analyse the text review as positive, negative or neutral.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is a critical step in the data analysis process. It involves examining and summarizing the main characteristics of a dataset to gain insights and inform further analysis. EDA helps you understand your data, identify patterns, and detect anomalies. It involves presenting data, information, or concepts in a graphical or visual format to enhance understanding, engage the audience, and convey complex ideas more effectively. Here, I am going to analyse the processed data and provide valuable insights to the company.

Project Management:

A project management tool is used to help teams and individuals to plan, organize, track their projects effectively and efficiently. I had used Trello as my project management tool to perform my customer feedback analysis tasks. The reason why I had chosen Trello as my project management tool was because it was highly visual tool that uses boards, lists and cards to help and manage teams tasks and projects effectively. It is well suitable for small, medium and teams. It follows agile workflows as my project follows agile methodology. For further details regarding URL refer to Appendix B.

For Managing and Maintaining the code I had used GitHub as my storage for code repository. As GitHub is a Free open source repository with a wide adoption of vast community developers. It is user friendly and can easily integrate with third party services. For further details regarding URL refer to Appendix C.

Project Report Delivery Schedule Note: Reorder the sections in the order that you plan to complete them.	Deadline Date
Abstract	4th August 2023
Declaration	23rd August 2023
Acknowledgements	23rd August 2023
Introduction	21st July 2023
Literature - Technology Review	21st July 2023
Methodology	21st July 2023
Implementation and Results • Evaluation • Related Work	25th August 2023

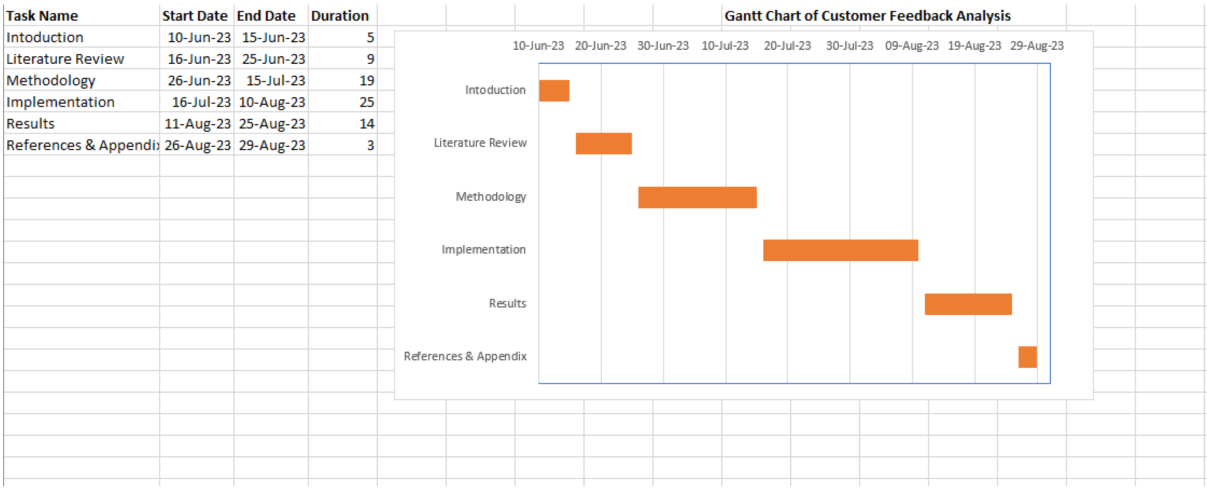
Conclusion • Reflection • Future Work	27th August 2023
References	29th August 2023
Appendices	29th August 2023

Table 3: Project Report Delivery Schedule

Artefact Delivery Schedule Note: Reorder the activities in the order that you plan to complete them.	**Deadline Date**
Artefact Planning and Resourcing	21st July 2023
Artefact Design	21st July 2023
Artefact Procurement Activities (e.g., data collection, source framework etc.)	21st July 2023
Artefact Development, Deployment, Implementation	30th August 2023
Artefact Evaluation and Testing	2nd September 2023
Artefact Presentation and Demonstration	8th September 2023
Artefact Screencast	8th September 2023

Table 4: Artefact Delivery Schedule

The project duration and Gantt chart are displayed below in the below figure.



Task Name	Start Date	End Date	Duration
Intoduction	10-Jun-23	15-Jun-23	5
Literature Review	16-Jun-23	25-Jun-23	9
Methodology	26-Jun-23	15-Jul-23	19
Implementation	16-Jul-23	10-Aug-23	25
Results	11-Aug-23	25-Aug-23	14
References & Appendi	26-Aug-23	29-Aug-23	3

Figure 1: Project Gantt Chart

4. Implementation

Project implementation is the crucial phase in a project's life cycle where the planned activities, tasks, and deliverables are put into action to achieve the project's objectives and goals. Customer Feedback Analysis has successfully demonstrated the effectiveness in analysing the customers feedback of Yelp business dataset. The key steps and considerations required for project are described in Methodology section are going to be explained in a very detailed manner in this section. This section furnishes the complete information about how the project is executed in-order to achieve its aims and objectives.

A system design diagram, often referred to as an architectural diagram is a visual representation of the high-level structure and components of a computer system[12]. These diagrams help communicate how different parts of the system interact and work together. Flowcharts are used to represent processes or workflows in a step-by-step manner, it uses rectangles, diamonds and arrows for controlling the data. In my project I had used flow chart to describe the system architectural diagram of the customer feedback analysis project[13]. The below diagram depicts the flowchart of Customer Feedback Analysis.

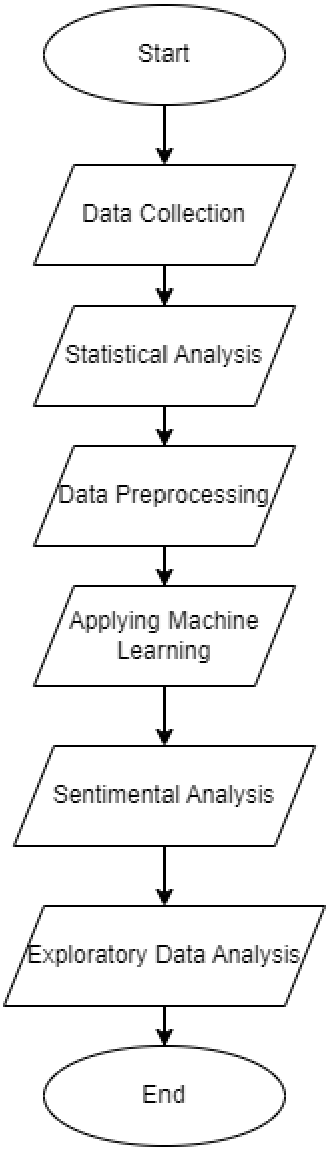
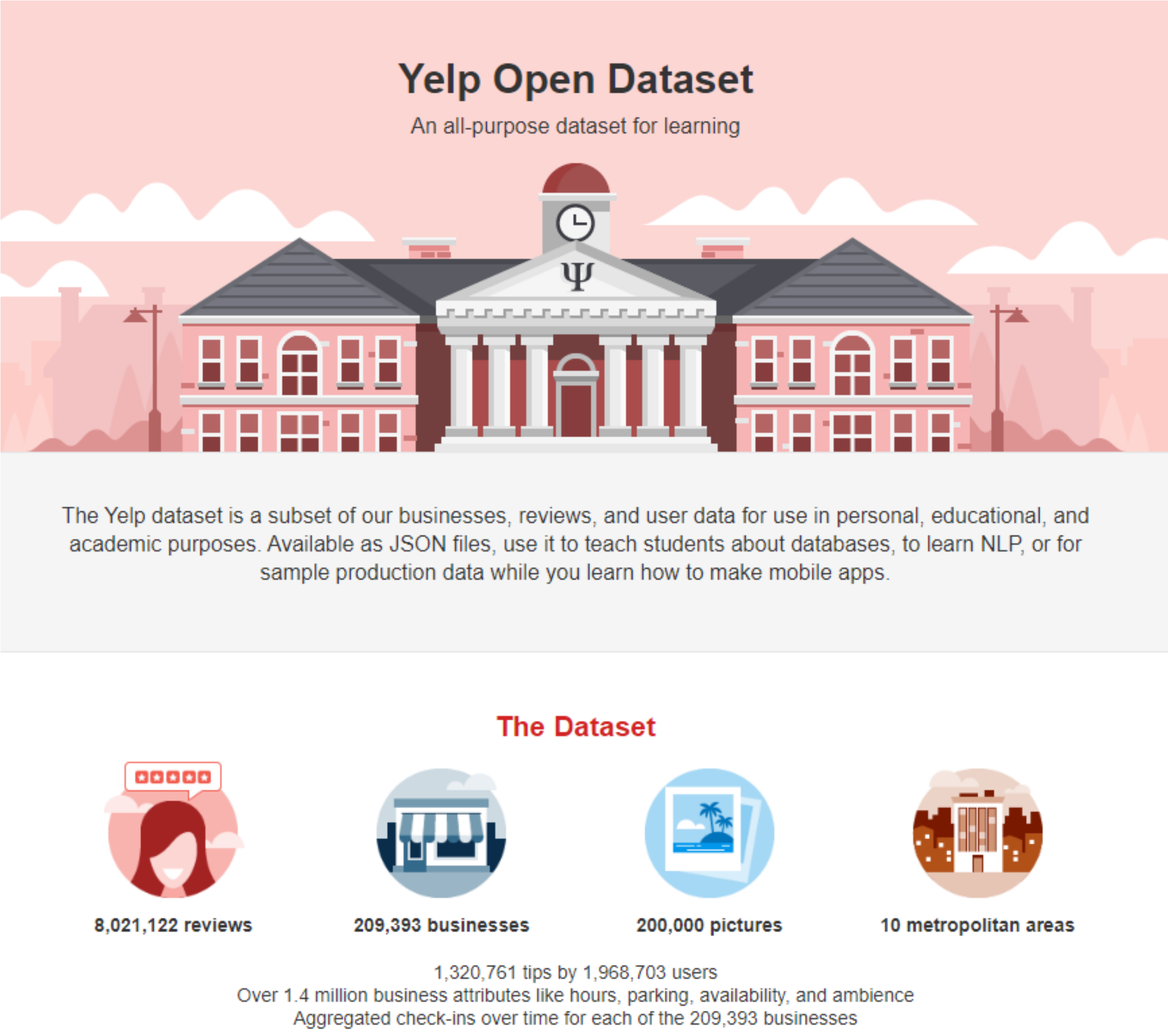


Figure 2: Project Flow Chart

Dataset Collection:

In order to analyse customer feedback, we need a business data. For this I had researched various types of dataset and I had finalized Yelp dataset for customer feedback analysis. The Yelp dataset is an open dataset which is widely used in the field of natural language processing and machine learning. The Yelp dataset consists of businesses, reviews, user datasets which are available in json format. The business dataset consists of data containing location data, attributes, and categories, the review dataset consists of reviews written by the user to a particular business_id, the user dataset consists of users data associated with each of the user. Some of the other datasets are also available like tip, checkin, photo: checkin dataset is used for business checkin, tip dataset is used for tips from the customers, photo dataset conatins photo data including the caption and classification. In our project we are going merge these three different datasets into one dataset which can be used for analysing the customer feedback. The Yelp dataset is readily available from yelp website by providing valid details and accepting the terms and conditions of Yelp Business data we can download the dataset. The dataset can be readily available through this URL https://www.yelp.com/dataset . The below figure shows the yelp business dataset.



Yelp Open Dataset

An all-purpose dataset for learning

The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

The Dataset

8,021,122 reviews

209,393 businesses

200,000 pictures

10 metropolitan areas

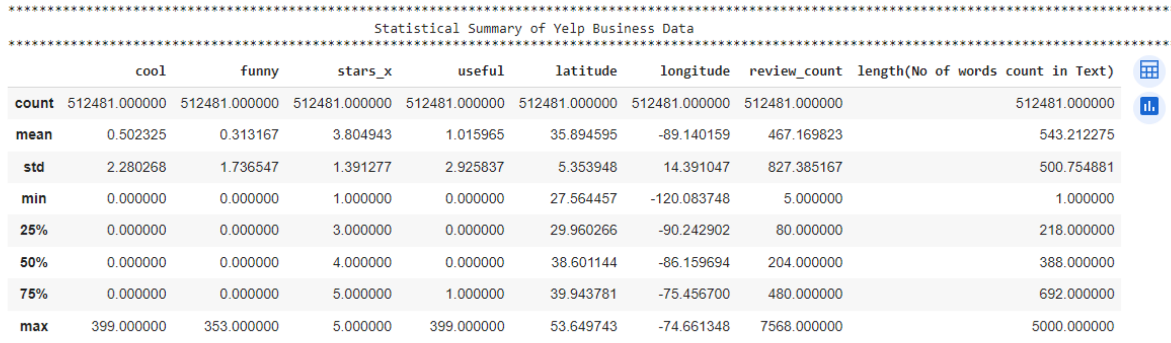
1,320,761 tips by 1,968,703 users
Over 1.4 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 209,393 businesses

Figure 3: Yelp Business Data

Statistical Analysis:

Statistical analysis is the process of using statistical techniques, methods, and tools to analyze and interpret data. It involves examining data sets to identify patterns, relationships, trends, and insights, which can be used for decision-making, hypothesis testing, and drawing conclusions about a population or a sample. Statistical analysis is widely used in various fields, including science, economics, social sciences, healthcare, business, and more[15]. For making complete dataset analysis I had created one more extra column called Length (No of word counts in Review Text) to visually represent other column data with new column data.

As I merged different yelp business data into single dataframe called review_business_data_merged and it contains 24 columns of different datatypes. I had done dataset analysis and statistical analysis on some of the numerical columns of the dataset. The numerical columns are cool, funny, stars_x, useful, latitude, longitude, review_count, length(No of words count in Text) etc. I had tried to find statistical measures of these numerical columns. The below picture depicts the statistical measures of those columns.

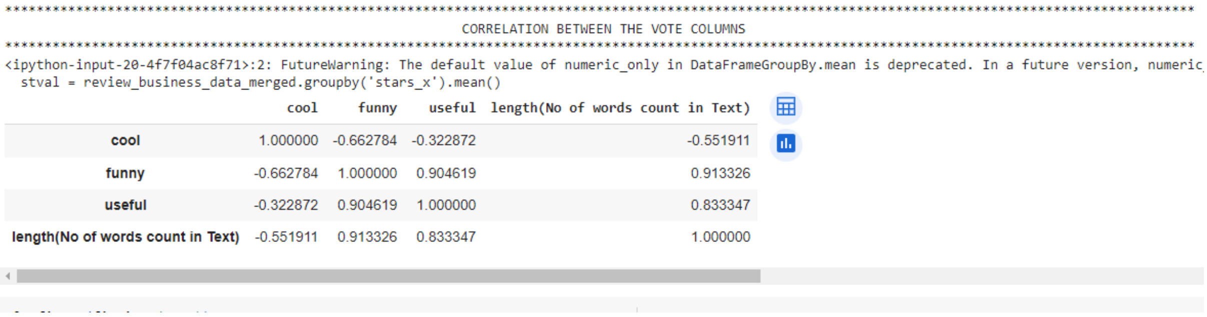


```
****************************************************************************************************************
                                     Statistical Summary of Yelp Business Data
****************************************************************************************************************
```

	cool	funny	stars_x	useful	latitude	longitude	review_count	length(No of words count in Text)
count	512481.000000	512481.000000	512481.000000	512481.000000	512481.000000	512481.000000	512481.000000	512481.000000
mean	0.502325	0.313167	3.804943	1.015965	35.894595	-89.140159	467.169823	543.212275
std	2.280268	1.736547	1.391277	2.925837	5.353948	14.391047	827.385167	500.754881
min	0.000000	0.000000	1.000000	0.000000	27.564457	-120.083748	5.000000	1.000000
25%	0.000000	0.000000	3.000000	0.000000	29.960266	-90.242902	80.000000	218.000000
50%	0.000000	0.000000	4.000000	0.000000	38.601144	-86.159694	204.000000	388.000000
75%	0.000000	0.000000	5.000000	1.000000	39.943781	-75.456700	480.000000	692.000000
max	399.000000	353.000000	5.000000	399.000000	53.649743	-74.661348	7568.000000	5000.000000

Figure 4: Statistical summary of Yelp Dataset

Correlation is a statistical measure that quantifies the degree to which two or more variables are related or associated with each other. It indicates whether and how changes in one variable are linked to changes in another. Correlation does not imply causation, meaning that a correlation between two variables does not necessarily mean that one causes the other. In my research I had made correlation analysis on cool, funny, useful and length (No of words count in Text) columns of yelp merged business dataset. It depicts negative correlation between funny and useful columns and rest of the columns had positive correlation. The below figure shows the correlation of different columns.



```
****************************************************************************************************************
                                           CORRELATION BETWEEN THE VOTE COLUMNS
****************************************************************************************************************
<ipython-input-20-4f7f04ac8f71>:2: FutureWarning: The default value of numeric_only in DataFrameGroupBy.mean is deprecated. In a future version, numeric
  stval = review_business_data_merged.groupby('stars_x').mean()
```

	cool	funny	useful	length(No of words count in Text)
cool	1.000000	-0.662784	-0.322872	-0.551911
funny	-0.662784	1.000000	0.904619	0.913326
useful	-0.322872	0.904619	1.000000	0.833347
length(No of words count in Text)	-0.551911	0.913326	0.833347	1.000000

Figure 5: Correlation of vote columns

A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution and summary statistics of a dataset. It provides a visual way to understand the central tendency, spread, and skewness of the data, as well as identify potential outliers. Box plots are particularly useful for comparing the distribution of multiple datasets or variables. I had used box plot to find out the statistical analysis of many columns of the dataset. For instance, I had used average stars and length of the text columns to visually represent spread of the data between these two columns. The below figure shows the box plot information of stars and length of the text information.

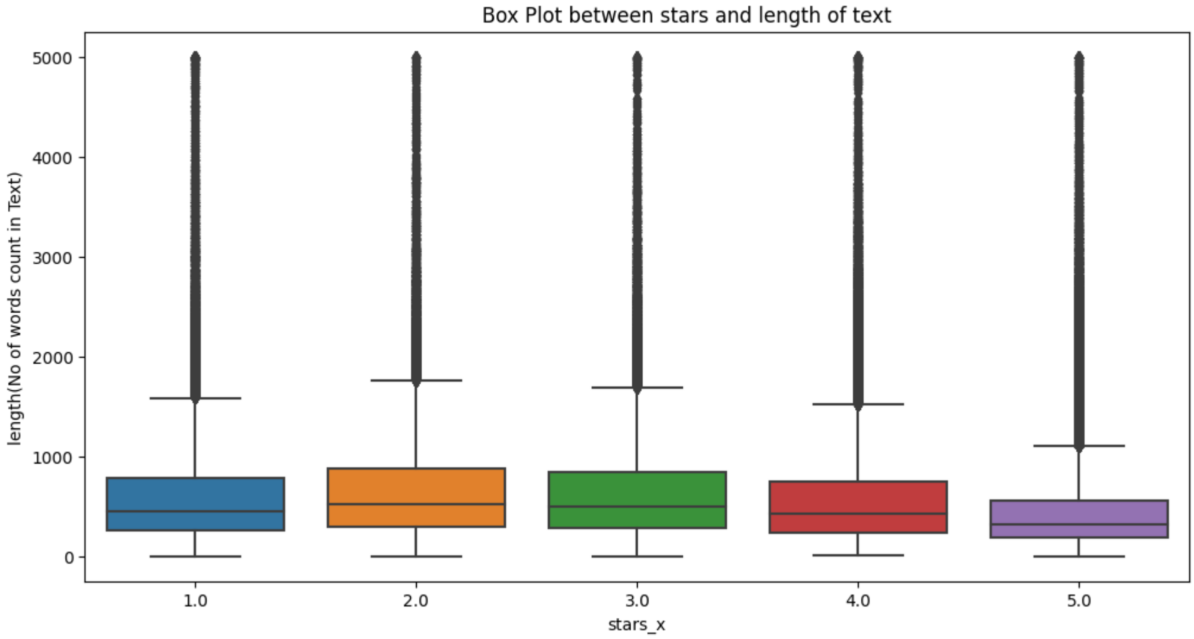


Figure 6: Box plot of reviews data

Data distribution refers to the way data is spread or distributed across different values in a dataset. Understanding the distribution of data is a fundamental aspect of data analysis, as it provides insights into the characteristics of the data, including central tendency, spread, and potential patterns. The below figure displays the normal distribution of stars information.

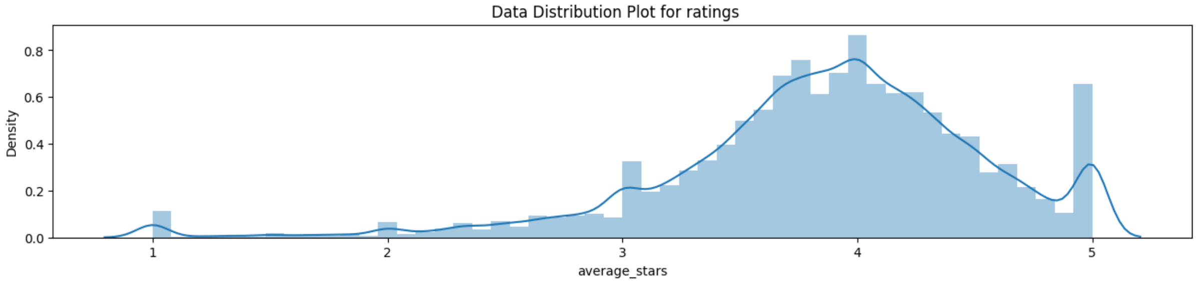
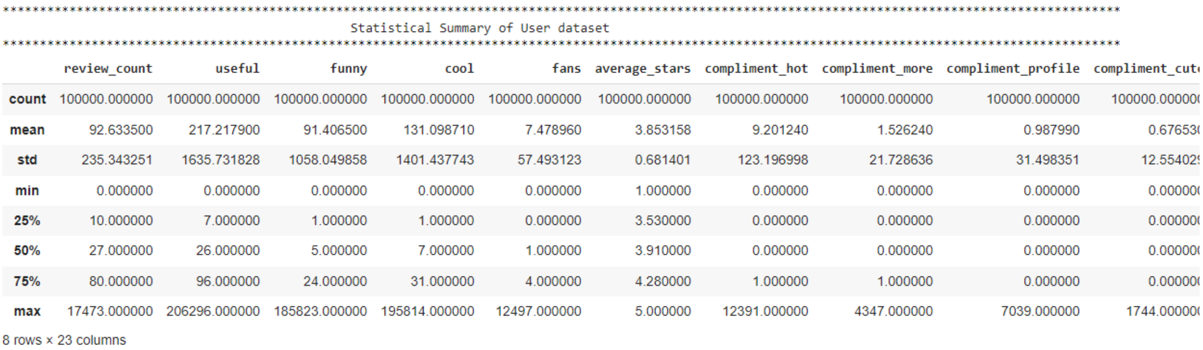


Figure 7: Data distribution of Ratings data

User Dataset was analysed statistically in different ways to measure its dispersion of data, calculating statistics summary, normal distribution, calculating covariance etc. The statistical summary of user dataset is depicted in below figure.



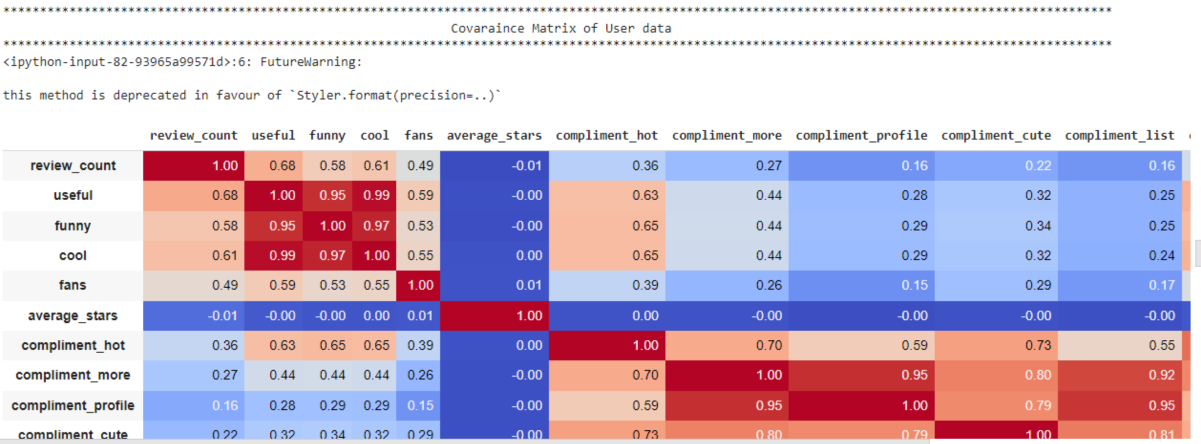
```
**********************************************************************************************************************
                                      Statistical Summary of User dataset
**********************************************************************************************************************
```

	review_count	useful	funny	cool	fans	average_stars	compliment_hot	compliment_more	compliment_profile	compliment_cut
count	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.00000
mean	92.633500	217.217900	91.406500	131.098710	7.478960	3.853158	9.201240	1.526240	0.987990	0.67653
std	235.343251	1635.731828	1058.049858	1401.437743	57.493123	0.681401	123.196998	21.728636	31.498351	12.55402
min	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.00000
25%	10.000000	7.000000	1.000000	1.000000	0.000000	3.530000	0.000000	0.000000	0.000000	0.00000
50%	27.000000	26.000000	5.000000	7.000000	1.000000	3.910000	0.000000	0.000000	0.000000	0.00000
75%	80.000000	96.000000	24.000000	31.000000	4.000000	4.280000	1.000000	1.000000	0.000000	0.00000
max	17473.000000	206296.000000	185823.000000	195814.000000	12497.000000	5.000000	12391.000000	4347.000000	7039.000000	1744.00000

8 rows × 23 columns

Figure 8: Statistical summary of User Dataset

Covariance is a statistical measure that quantifies the degree to which two random variables change together. It indicates whether there is a linear relationship between the variables and whether they tend to move in the same direction or in opposite directions. In other words, covariance measures how changes in one variable are associated with changes in another variable. Here we employed the covariance of user business dataframe and trained dataframe of user business dataset. The user dataframe covariance is depicted in below figure.



```
**********************************************************************************************************************
                                         Covaraince Matrix of User data
**********************************************************************************************************************
<ipython-input-82-93965a99571d>:6: FutureWarning:

this method is deprecated in favour of `Styler.format(precision=..)`
```

	review_count	useful	funny	cool	fans	average_stars	compliment_hot	compliment_more	compliment_profile	compliment_cute	compliment_list
review_count	1.00	0.68	0.58	0.61	0.49	-0.01	0.36	0.27	0.16	0.22	0.16
useful	0.68	1.00	0.95	0.99	0.59	-0.00	0.63	0.44	0.28	0.32	0.25
funny	0.58	0.95	1.00	0.97	0.53	-0.00	0.65	0.44	0.29	0.34	0.25
cool	0.61	0.99	0.97	1.00	0.55	0.00	0.65	0.44	0.29	0.32	0.24
fans	0.49	0.59	0.53	0.55	1.00	0.01	0.39	0.26	0.15	0.29	0.17
average_stars	-0.01	-0.00	-0.00	0.00	0.01	1.00	0.00	-0.00	-0.00	-0.00	-0.00
compliment_hot	0.36	0.63	0.65	0.65	0.39	0.00	1.00	0.70	0.59	0.73	0.55
compliment_more	0.27	0.44	0.44	0.44	0.26	-0.00	0.70	1.00	0.95	0.80	0.92
compliment_profile	0.16	0.28	0.29	0.29	0.15	-0.00	0.59	0.95	1.00	0.79	0.95
compliment_cute	0.22	0.32	0.34	0.32	0.29	-0.00	0.73	0.80	0.79	1.00	0.81

Figure 9: Covariance matrix of User data

Principal Component Analysis (PCA) is a dimensionality reduction technique and a powerful tool for data analysis and feature engineering. PCA is used to transform a high-dimensional dataset into a lower-dimensional form while preserving as much of the original variability as possible. It achieves this by creating new variables, called principal components, that are linear combinations of the original variables. PCA is widely used in various fields, including statistics, machine learning, and data science, for tasks such as data visualization, noise reduction, and feature selection. Here we analysed Principal Component Analysis (PCA) on various combinations of different columns.

I had analysed various distributed curves to display the feedback of various cluster objects of Compliments feedback and Active Feedback. I had made a correlation between popularity feedback and average stars and it displays the information between number of clusters between these two dataframes. The below figure depicts the information of popularity feedback and average stars.

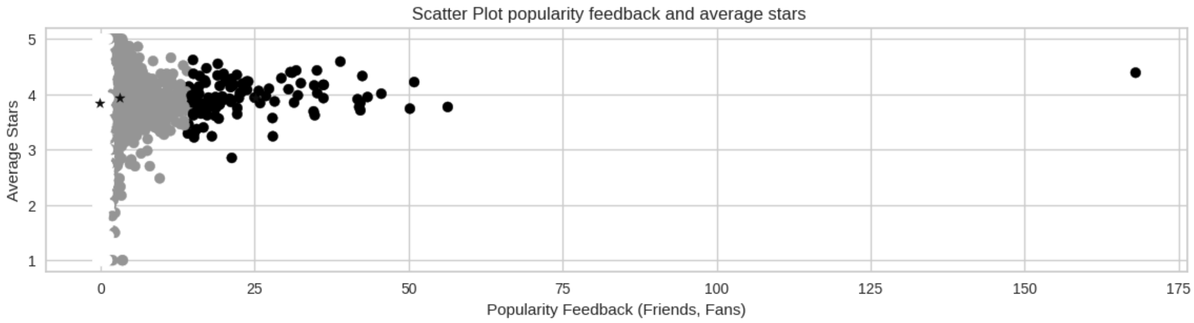


Figure 10: Scatter Plot of Popularity Feedback and Ratings

Data Pre-processing:

Data pre-processing is a crucial step in the data analysis and machine learning pipeline. It involves cleaning, transforming, and organizing raw data into a format suitable for analysis or model training. Proper data pre-processing is essential because the quality of the input data significantly impacts the results and performance of analytical or machine learning models. In data pre-processing step we had created one new column called length(No of words count in Text) and also I had removed some of the extra columns that are not useful for analysing the customer feedback such as Unnamed: 0, attributes, is_open, serves_food, hours etc.

In Yelp business dataset we had different forms of datasets such as business, user, reviews datasets, these datasets are mainly used for analysing the customer feedback. All the three datasets are combined together and formed a review_busisness_data_merged dataset. Most of the part of the analysis is done through this dataset.

Removing Punctuation: Removing punctuation from text data is a common preprocessing step in natural language processing (NLP) and text analysis tasks. Punctuation marks, such as periods, commas, exclamation points, and question marks, often do not carry much semantic meaning on their own and can be safely removed to simplify text data. Some of the models like Bert, GPT3 are unable to select on this punctuation rather than helping them predict the right class, it often throws the model. Here I had defined a function to remove the punctuation of given text data.

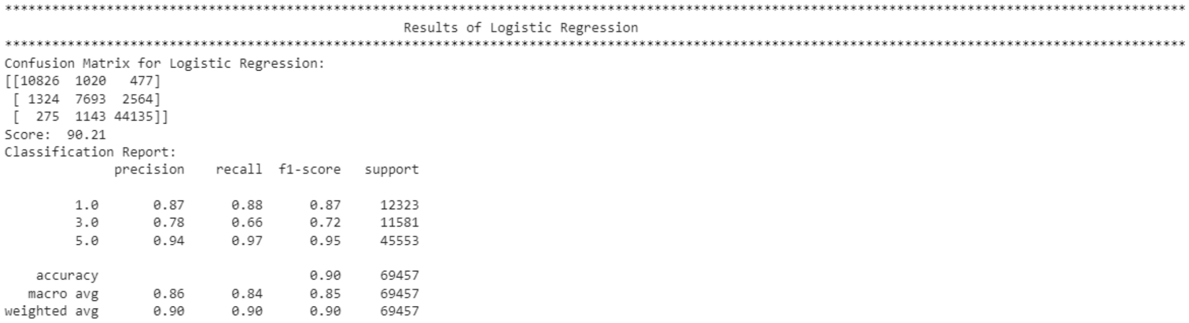
Removing Stopwords: Removing stopwords is another essential preprocessing step in natural language processing (NLP). Stopwords are common words in a language (e.g., "the," "and," "in") that are often removed from text data because they typically do not carry significant meaning and can introduce noise into text analysis. In my project I had used stopwords function to add the list of English stopwords so that it will remove the stopwords of given text data.

Lemmatization: Lemmatization is a text normalization technique used in natural language processing (NLP) to reduce words to their base or root forms, which are also known as lemmas. The goal of lemmatization is to transform words in a way that they make sense and are valid words in a language. This can improve the accuracy and interpretability of text analysis tasks. Lemmatization works better with different words as it provides best lemmatized words when compared with stemming. So, in my project I had used lemmatization technique for processing the words.

Applying Machine Learning Algorithms:

Machine learning (ML) is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to learn from and make predictions or decisions based on data. ML systems are designed to automatically improve their performance on a specific task through experience and data, without being explicitly programmed. In my research I had employed six different machine learning algorithms and made a comparison of these algorithms and found the best efficient algorithm for customer feedback data. The different machine learning algorithms used for analysing customer feedback are Logistic Regression, Multinomial Navie Bayes, Gradient Boosting Classifier[19][20]. Among the three algorithms Logistic Regression algorithm performs better with an accuracy of about 90% when compared with other two algorithms.

Logistic regression is a statistical model and classification algorithm used for binary and multi-class classification problems. Despite its name, logistic regression is primarily used for classification tasks, not regression tasks. It's a simple yet effective algorithm for estimating the probability that an instance belongs to a particular class. In this section I am going to present the Logistic Regression results of analysing the customer feedback data of Yelp business data. The below figure depicts the information of logistic regression on customer feedback.



```
*********************************************************************************************************************************************************
                                                   Results of Logistic Regression
*********************************************************************************************************************************************************
Confusion Matrix for Logistic Regression:
[[10826  1020   477]
 [ 1324  7693  2564]
 [  275  1143 44135]]
Score:  90.21
Classification Report:
              precision    recall  f1-score   support

         1.0       0.87      0.88      0.87     12323
         3.0       0.78      0.66      0.72     11581
         5.0       0.94      0.97      0.95     45553

    accuracy                           0.90     69457
   macro avg       0.86      0.84      0.85     69457
weighted avg       0.90      0.90      0.90     69457
```

Figure 11: Logistic Regression results

Sentimental Analysis:

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique that involves determining and extracting subjective information from text data to understand the sentiment or emotional tone expressed within the text. Sentiment analysis is commonly used to automatically classify text into categories such as positive, negative, or neutral, or to assign a sentiment score indicating the intensity of sentiment. In my project I had categorized positive and negative based on trending_keywords function. I had used pie chart to display the sentiments of the feedback of respective company. Moreover, I had furnished the recent reviews of the feedback in a table based on their year. Below figure displays the information about distribution of review sentiments.

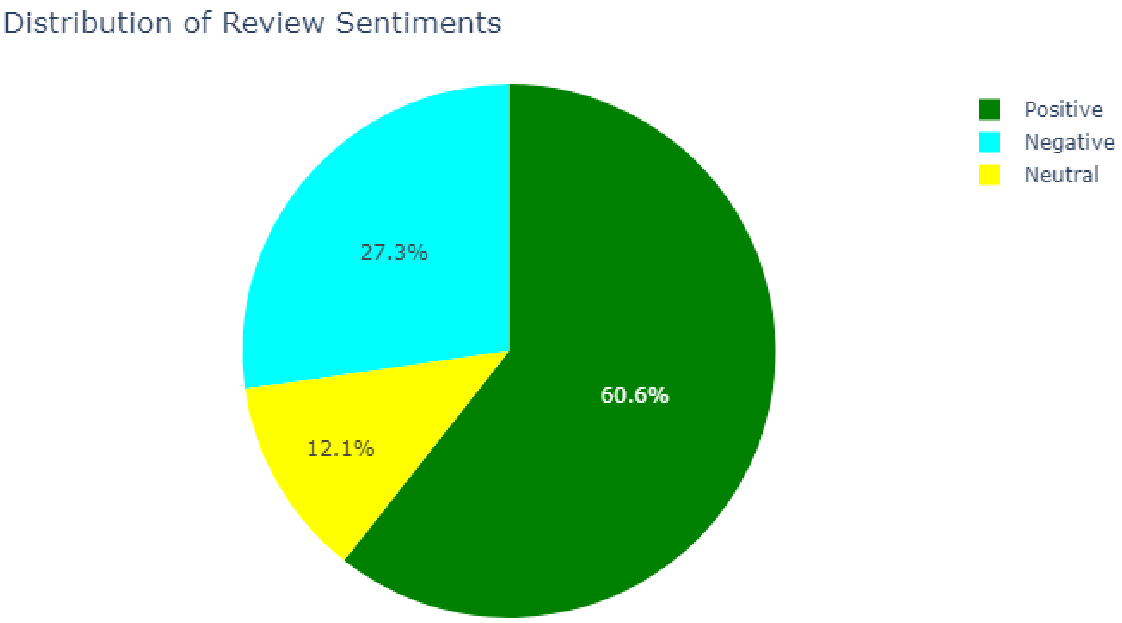
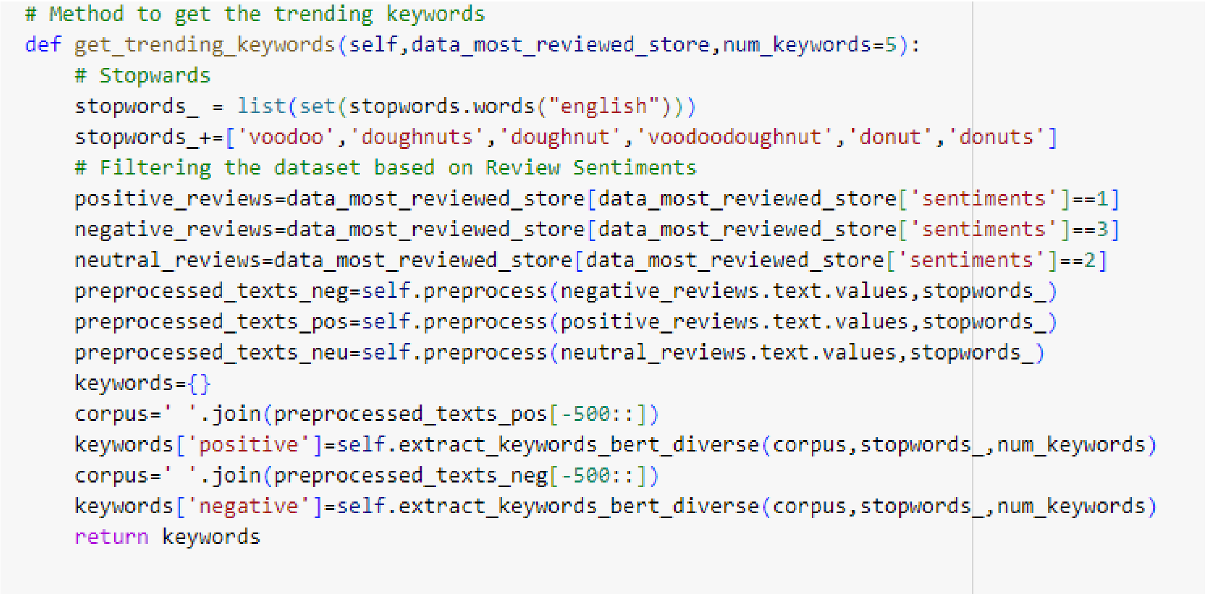


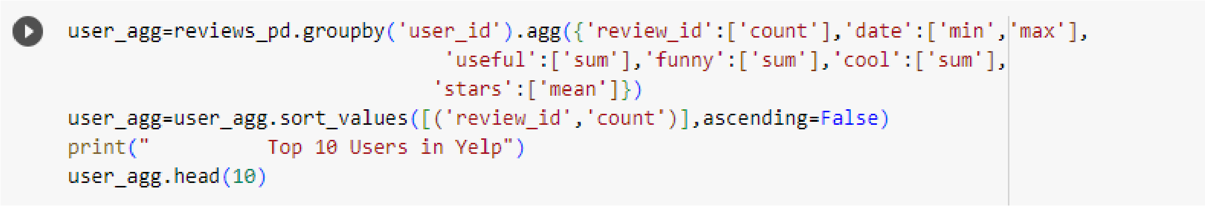
Figure 12: Pie Chart of sentiment reviews

The code snippet for getting the trending keywords is shown in below figure.



```
# Method to get the trending keywords
def get_trending_keywords(self,data_most_reviewed_store,num_keywords=5):
    # Stopwards
    stopwords_ = list(set(stopwords.words("english")))
    stopwords_+=['voodoo','doughnuts','doughnut','voodoodoughnut','donut','donuts']
    # Filtering the dataset based on Review Sentiments
    positive_reviews=data_most_reviewed_store[data_most_reviewed_store['sentiments']==1]
    negative_reviews=data_most_reviewed_store[data_most_reviewed_store['sentiments']==3]
    neutral_reviews=data_most_reviewed_store[data_most_reviewed_store['sentiments']==2]
    preprocessed_texts_neg=self.preprocess(negative_reviews.text.values,stopwords_)
    preprocessed_texts_pos=self.preprocess(positive_reviews.text.values,stopwords_)
    preprocessed_texts_neu=self.preprocess(neutral_reviews.text.values,stopwords_)
    keywords={}
    corpus=' '.join(preprocessed_texts_pos[-500::])
    keywords['positive']=self.extract_keywords_bert_diverse(corpus,stopwords_,num_keywords)
    corpus=' '.join(preprocessed_texts_neg[-500::])
    keywords['negative']=self.extract_keywords_bert_diverse(corpus,stopwords_,num_keywords)
    return keywords
```

Figure 13: Code snippet for trending keywords

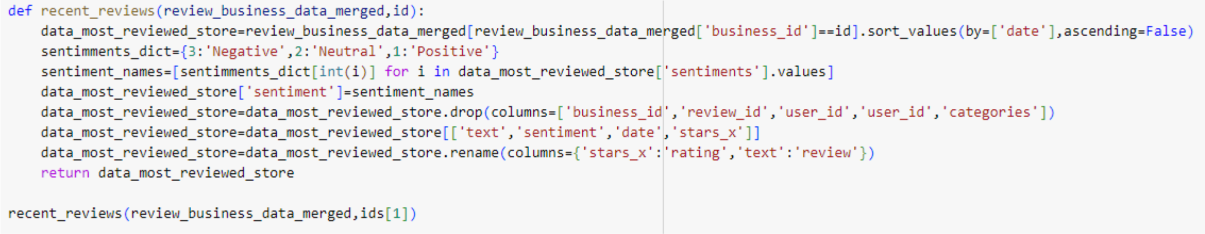


```
user_agg=reviews_pd.groupby('user_id').agg({'review_id':['count'],'date':['min','max'],
                                            'useful':['sum'],'funny':['sum'],'cool':['sum'],
                                           'stars':['mean']})
user_agg=user_agg.sort_values([('review_id','count')],ascending=False)
print("          Top 10 Users in Yelp")
user_agg.head(10)
```

Figure 14:Code snippet for Top performing users

The above code snippet is used to display the top users based on their ratings.

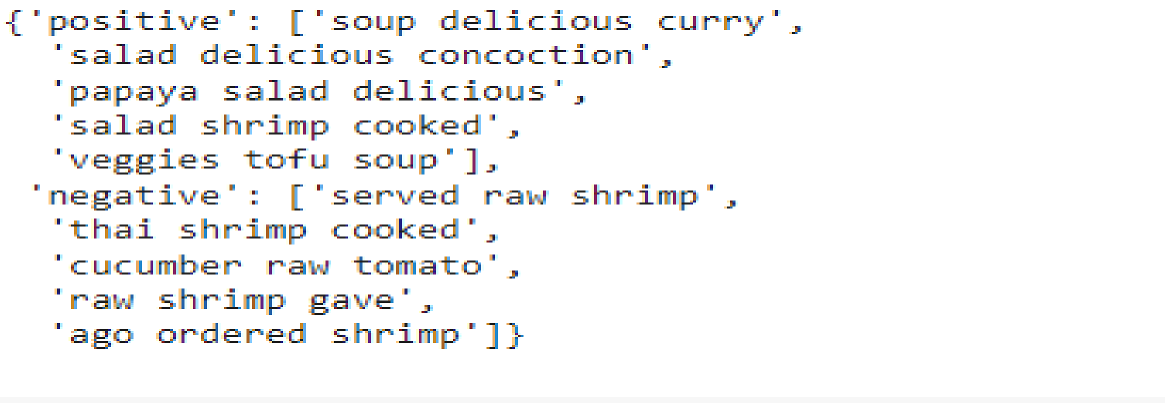
The below code is used for getting the most recent_reviews of a particular business_id.



```
def recent_reviews(review_business_data_merged,id):
    data_most_reviewed_store=review_business_data_merged[review_business_data_merged['business_id']==id].sort_values(by=['date'],ascending=False)
    sentimments_dict={3:'Negative',2:'Neutral',1:'Positive'}
    sentiment_names=[sentimments_dict[int(i)] for i in data_most_reviewed_store['sentiments'].values]
    data_most_reviewed_store['sentiment']=sentiment_names
    data_most_reviewed_store=data_most_reviewed_store.drop(columns=['business_id','review_id','user_id','user_id','categories'])
    data_most_reviewed_store=data_most_reviewed_store[['text','sentiment','date','stars_x']]
    data_most_reviewed_store=data_most_reviewed_store.rename(columns={'stars_x':'rating','text':'review'})
    return data_most_reviewed_store

recent_reviews(review_business_data_merged,ids[1])
```

Figure 15: Code snippet for most recent reviews

Sentence Transformers are a type of deep learning model specifically designed for encoding and transforming sentences or text paragraphs into fixed-length vector representations. These vector representations are highly useful for various natural language processing (NLP) tasks, such as semantic similarity analysis, text classification, clustering, and information retrieval. Sentence Transformers leverage transfer learning techniques and pre-trained language models to generate meaningful sentence embeddings. The sentence transformers is very new topic for me to understand after learning, it was very useful for my text analysis. After getting the trending keywords from the function the keywords are passed to function where it was categorized into positive and negative keywords. Below figure is categorized into positive and negative keywords based on their trending keywords from the text.



```
{'positive': ['soup delicious curry',
   'salad delicious concoction',
   'papaya salad delicious',
   'salad shrimp cooked',
   'veggies tofu soup'],
  'negative': ['served raw shrimp',
   'thai shrimp cooked',
   'cucumber raw tomato',
   'raw shrimp gave',
   'ago ordered shrimp']}
```

Figure 16: Categorization of Positive and Negative keywords

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is an approach in data analysis and statistics that focuses on summarizing and visualizing a dataset to gain insights and understanding of its main characteristics. The primary goal of EDA is to help analysts, data scientists, and researchers get a better grasp of the data, detect patterns, identify outliers, and generate hypotheses for further analysis. I had visualized various analysis to represent the data in a meaningful and understandable format. I had used bar plots to plot the top categories and top cities of best performing business. Below figure shows the information about top performing categories of yelp business dataset.

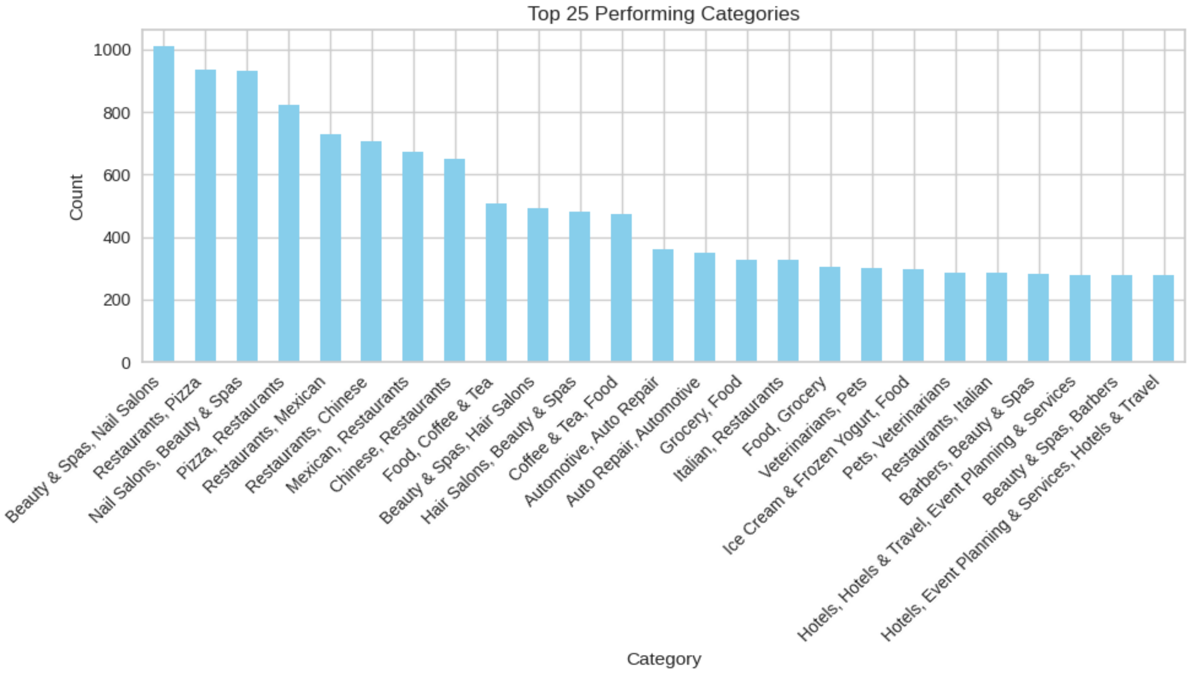


Figure 17: Top performing business

A word cloud, also known as a tag cloud or text cloud, is a visual representation of text data in which words are displayed in varying sizes and colors. The size of each word corresponds to its frequency or importance within the given text. I had used word clouds to display the important words of the text data, positive words of business and negative words of business. Below figure shows the top trending words of a particular business.

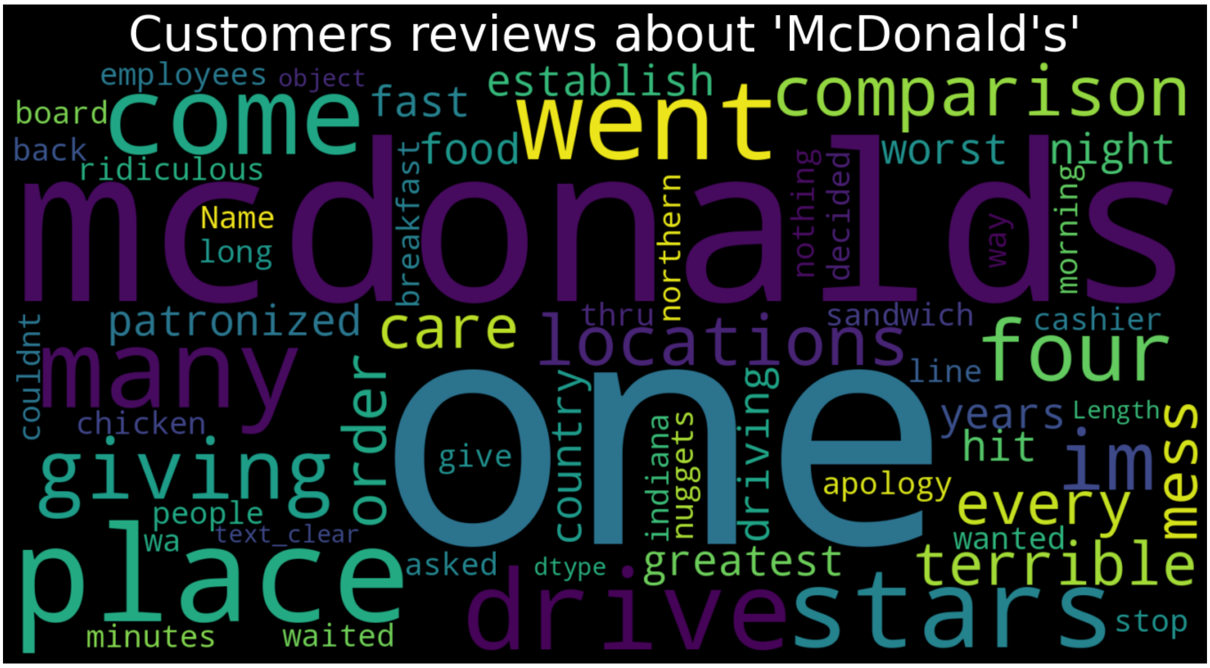


Figure 18: Word Cloud of McDonald’s review

I had displayed the review ratings in a rotational bar graph. Its an automated bar graph where I can display a rating count based on review rating count. Below figure depicts the information about distribution of review ratings.

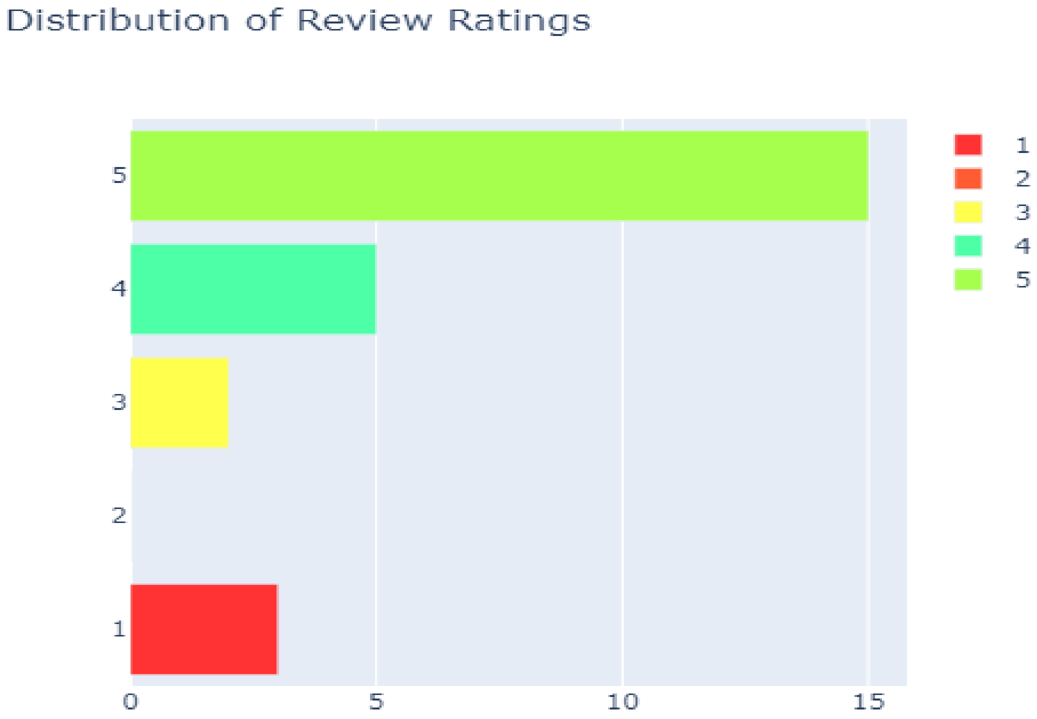


Figure 19: Automated plot for review ratings

Among the entire research on customer feedback analysis, I had faced some difficulties while doing the analysis. Some of the major problems that I had encountered and how I approached to solve them are going to be discussed here.

The Yelp business datasets are available in json formats. I cannot use json formats directly for analysis. For this I tried to use spark in python. PySpark is the official Python library for Apache Spark, an open-source, distributed computing framework designed for big data processing and analytics. PySpark allows Python developers to harness the capabilities of Spark, enabling distributed data processing, machine learning, graph processing, and more while using the Python programming language. It took some time to learn about spark as I don’t have any idea to use it. The size of Yelp business data is very large as it’s very difficult to make analysis directly by downloading data in colab or in visualstudio environment. It took lot of time to download and process the data. I had decided to upload the dataset in google drive so then we can process the data whenever I want. It reduces so much of time to process the data without downloading. As the Yelp business dataset has different datasets, I need to merge the datasets, as most of the individual datasets have same column names it creates duplicate columns. So, I tried to delete the duplicate columns and merged the dataset individually. It creates some chaos while merging the datasets as column data to get merged.

I tried to implement rain plot in my research and it takes me sometime to learn about it and also for analysing which data needs to be used for perfect rain plot. While working on the user dataset it is difficult to understand about the friend_count columns. I tried to do Principal Component Analysis (PCA)[13][22] on new column data with other data but it doesn’t provide best results so I did Principal Component Analysis (PCA) with the existing columns data. I had learnt a new thing to plot the data based on number of clusters with average stars and based on the feedback data. It took some time to research about this topic plotting-based number of clusters. For analysing the textual data, it took lot of time to research about topics Natural Language Processing, Stopwords, WordClouds, Sentence_Transformers etc. I don’t have any knowledge about some topics so it took time to know about it and make practical implementation on those new topics.

Getting trending keywords from the data took lot of time for coding these things as it needed lots of time for referencing the functions. Application of Machine Learning to the text data takes lots of time to train the data as it contains a greater number of records, the data is not able to process through colab environment directly. So, I had used parallel processing technique to train the data parallelly without taking lots of time to process the data. Moreover, I had used vectorizer parallelly to train both the train data, test data at a time.

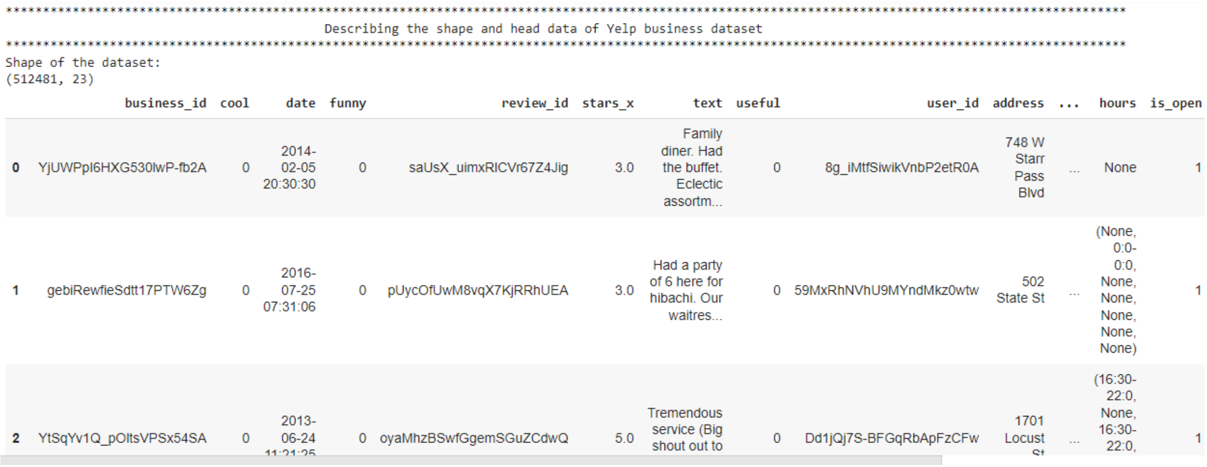
5. Results

In this section, I had presented the outcomes of my research, which aimed to investigate the impact of analysing the customer feedback. The findings offer valuable insights to yelp business organizations and upcoming researchers. My research results can be used as reference for other researchers as I had included various broader contents for analysing the customer feedback. In this research I had lightened on various new topics for analysing the dataset this may include both dataset analysis and statistical analysis, categorization of positive and negative reviews, getting most recent reviews and application of various machine learning algorithms on the same data. Each of the above-mentioned topics are evaluated and the results will be discussed in the below sections.

Dataset Analysis Results:

Dataset analysis, also known as data analysis or data exploration, refers to the process of examining, cleaning, summarizing, and interpreting a collection of data, often referred to as a dataset. The primary goal of dataset analysis is to extract meaningful insights and knowledge from the data, which can be used for decision-making, problem-solving, research, or other purposes. In my project, I had done an extensive analysis on the dataset[9][10]. Below are the findings of yelp business dataset analysis.

This picture shows the information about the shape and top columns data of yelp business dataset. As yelp dataset is large, we had around 5 thousand rows of data and 24 columns of data when dataset was merged.

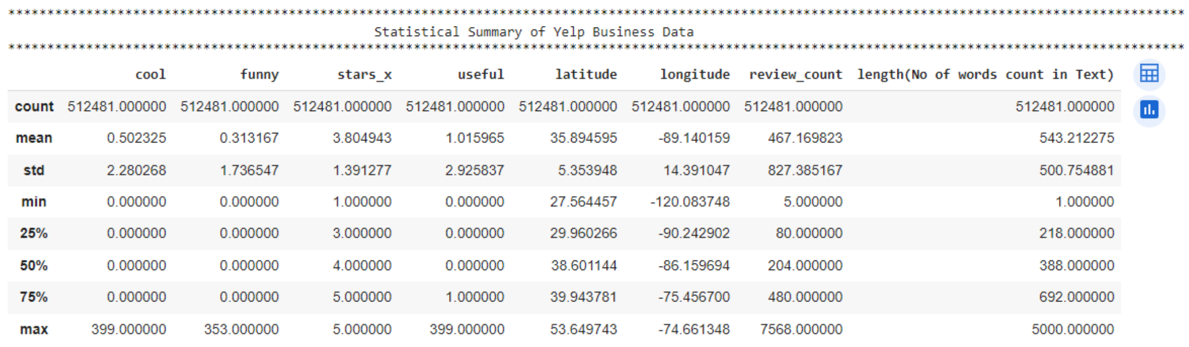


```
****************************************************************************************************************
                                   Describing the shape and head data of Yelp business dataset
****************************************************************************************************************
Shape of the dataset:
(512481, 23)
```

	business_id	cool	date	funny	review_id	stars_x	text	useful	user_id	address	...	hours	is_open
0	YjUWPpI6HXG530lwP-fb2A	0	2014-02-05 20:30:30	0	saUsX_uimxRICVr67Z4Jig	3.0	Family diner. Had the buffet. Eclectic assortm...	0	8g_iMtfSiwikVnbP2etR0A	748 W Starr Pass Blvd	...	None	1
1	gebiRewfieSdtt17PTW6Zg	0	2016-07-25 07:31:06	0	pUycOfUwM8vqX7KjRRhUEA	3.0	Had a party of 6 here for hibachi. Our waitres...	0	59MxRhNVhU9MYndMkz0wtw	502 State St	...	(None, 0:0-0:0, None, None, None, None, None)	1
2	YtSqYv1Q_pOltsVPSx54SA	0	2013-06-24 11:21:25	0	oyaMhzBSwfGgemSGuZCdwQ	5.0	Tremendous service (Big shout out to	0	Dd1jQj7S-BFGqRbApFzCFw	1701 Locust St	...	(16:30-22:0, None, 16:30-22:0,	1

Figure 20: Shape of Yelp dataset

The below picture depicts the statistics summary of yelp business merged dataset. Almost most of the data are in positive expect longitude column which can be avoided as it is not useful for my analysis.

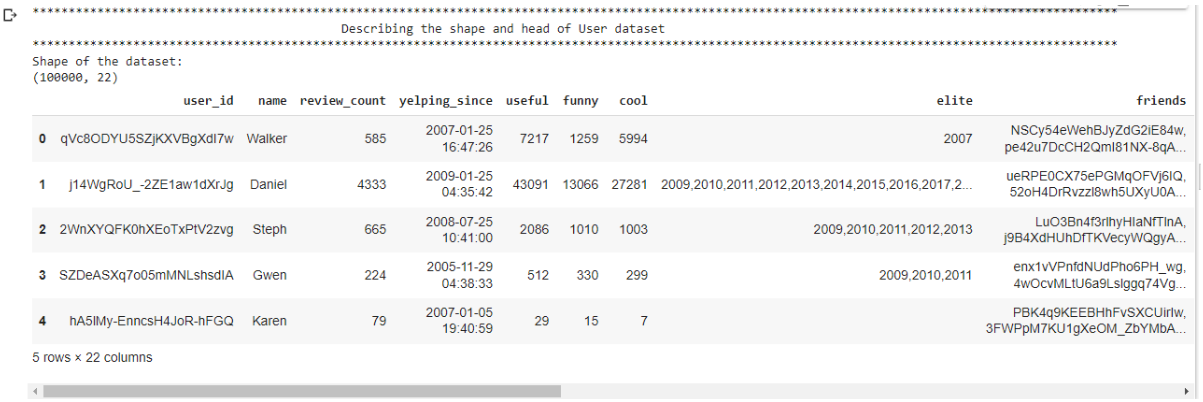


```
****************************************************************************************************************
                                   Statistical Summary of Yelp Business Data
****************************************************************************************************************
```

	cool	funny	stars_x	useful	latitude	longitude	review_count	length(No of words count in Text)
count	512481.000000	512481.000000	512481.000000	512481.000000	512481.000000	512481.000000	512481.000000	512481.000000
mean	0.502325	0.313167	3.804943	1.015965	35.894595	-89.140159	467.169823	543.212275
std	2.280268	1.736547	1.391277	2.925837	5.353948	14.391047	827.385167	500.754881
min	0.000000	0.000000	1.000000	0.000000	27.564457	-120.083748	5.000000	1.000000
25%	0.000000	0.000000	3.000000	0.000000	29.960266	-90.242902	80.000000	218.000000
50%	0.000000	0.000000	4.000000	0.000000	38.601144	-86.159694	204.000000	388.000000
75%	0.000000	0.000000	5.000000	1.000000	39.943781	-75.456700	480.000000	692.000000
max	399.000000	353.000000	5.000000	399.000000	53.649743	-74.661348	7568.000000	5000.000000

Figure 21: Statistical summary of Yelp dataset

User Dataset shape and top column data values are shown in the below figure. The shape of user dataset is 1 thousand and it contains 22 columns data.



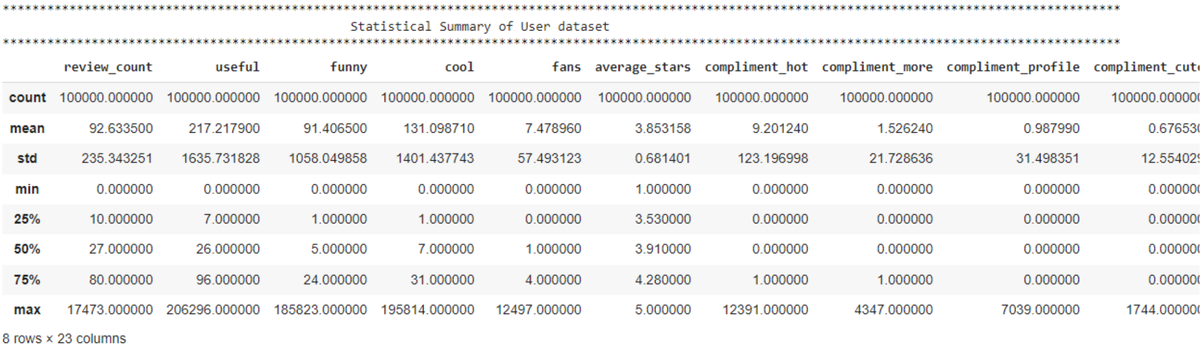
```
****************************************************************************************************
                              Describing the shape and head of User dataset
****************************************************************************************************
Shape of the dataset:
(100000, 22)
```

	user_id	name	review_count	yelping_since	useful	funny	cool	elite	friends
0	qVc8ODYU5SZjKXVBgXdI7w	Walker	585	2007-01-25 16:47:26	7217	1259	5994	2007	NSCy54eWehBJyZdG2iE84w, pe42u7DcCH2QmI81NX-8qA...
1	j14WgRoU_-2ZE1aw1dXrJg	Daniel	4333	2009-01-25 04:35:42	43091	13066	27281	2009,2010,2011,2012,2013,2014,2015,2016,2017,2...	ueRPE0CX75ePGMqOFVj6IQ, 52oH4DrRvzzl8wh5UXyU0A...
2	2WnXYQFK0hXEoTxPtV2zvg	Steph	665	2008-07-25 10:41:00	2086	1010	1003	2009,2010,2011,2012,2013	LuO3Bn4f3rlhyHIaNfTlnA, j9B4XdHUhDfTKVecyWQgyA...
3	SZDeASXq7o05mMNLshsdIA	Gwen	224	2005-11-29 04:38:33	512	330	299	2009,2010,2011	enx1vVPnfdNUdPho6PH_wg, 4wOcvMLtU6a9Lslggq74Vg...
4	hA5lMy-EnncsH4JoR-hFGQ	Karen	79	2007-01-05 19:40:59	29	15	7		PBK4q9KEEBHhFvSXCUirlw, 3FWPpM7KU1gXeOM_ZbYMbA...

5 rows × 22 columns

Figure 22: Shape of user dataset

The below figure depicts the information about the statistics summary of the user's dataset. Almost every column of the dataset contains positive values which denotes a positive correlation among the data.



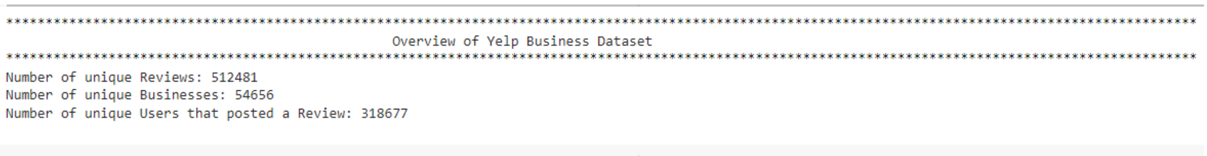
```
****************************************************************************************************
                              Statistical Summary of User dataset
****************************************************************************************************
```

	review_count	useful	funny	cool	fans	average_stars	compliment_hot	compliment_more	compliment_profile	compliment_cut
count	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.00000
mean	92.633500	217.217900	91.406500	131.098710	7.478960	3.853158	9.201240	1.526240	0.987990	0.67653
std	235.343251	1635.731828	1058.049858	1401.437743	57.493123	0.681401	123.196998	21.728636	31.498351	12.55402
min	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.00000
25%	10.000000	7.000000	1.000000	1.000000	0.000000	3.530000	0.000000	0.000000	0.000000	0.00000
50%	27.000000	26.000000	5.000000	7.000000	1.000000	3.910000	0.000000	0.000000	0.000000	0.00000
75%	80.000000	96.000000	24.000000	31.000000	4.000000	4.280000	1.000000	1.000000	0.000000	0.00000
max	17473.000000	206296.000000	185823.000000	195814.000000	12497.000000	5.000000	12391.000000	4347.000000	7039.000000	1744.00000

8 rows × 23 columns

Figure 23: Statistical summary of user dataset

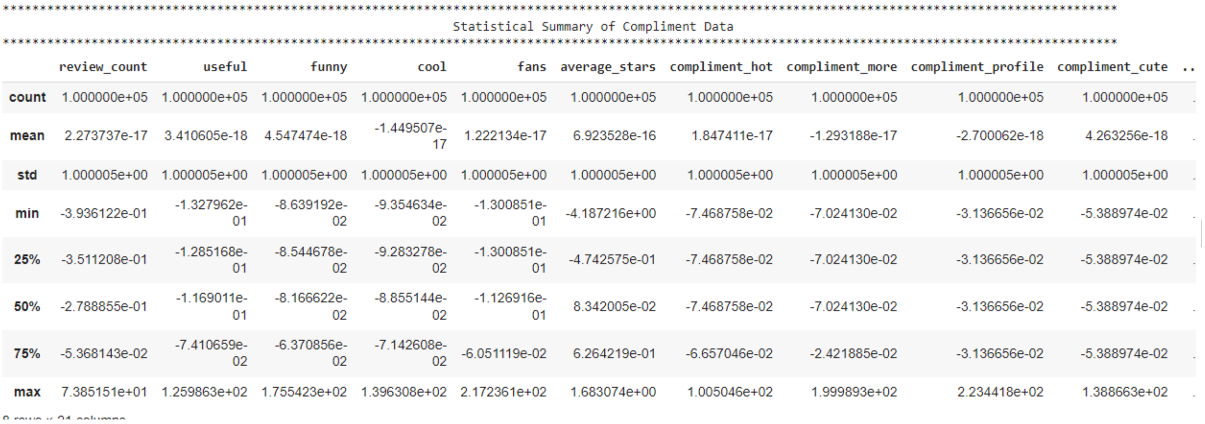
Below picture shows the information about the dataset.



```
****************************************************************************************************
                              Overview of Yelp Business Dataset
****************************************************************************************************
Number of unique Reviews: 512481
Number of unique Businesses: 54656
Number of unique Users that posted a Review: 318677
```

Figure 24: Overview of yelp business dataset

Below figure displays the information about statistical information of compliments data.



```
*****************************************************************************************************************************
                                        Statistical Summary of Compliment Data
*****************************************************************************************************************************
```

	review_count	useful	funny	cool	fans	average_stars	compliment_hot	compliment_more	compliment_profile	compliment_cute	..
count	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	1.000000e+05	.
mean	2.273737e-17	3.410605e-18	4.547474e-18	-1.449507e-17	1.222134e-17	6.923528e-16	1.847411e-17	-1.293188e-17	-2.700062e-18	4.263256e-18	.
std	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	1.000005e+00	.
min	-3.936122e-01	-1.327962e-01	-8.639192e-02	-9.354634e-02	-1.300851e-01	-4.187216e+00	-7.468758e-02	-7.024130e-02	-3.136656e-02	-5.388974e-02	.
25%	-3.511208e-01	-1.285168e-01	-8.544678e-02	-9.283278e-02	-1.300851e-01	-4.742575e-01	-7.468758e-02	-7.024130e-02	-3.136656e-02	-5.388974e-02	.
50%	-2.788855e-01	-1.169011e-01	-8.166622e-02	-8.855144e-02	-1.126916e-01	8.342005e-02	-7.468758e-02	-7.024130e-02	-3.136656e-02	-5.388974e-02	.
75%	-5.368143e-02	-7.410659e-02	-6.370856e-02	-7.142608e-02	-6.051119e-02	6.264219e-01	-6.657046e-02	-2.421885e-02	-3.136656e-02	-5.388974e-02	.
max	7.385151e+01	1.259863e+02	1.755423e+02	1.396308e+02	2.172361e+02	1.683074e+00	1.005046e+02	1.999893e+02	2.234418e+02	1.388663e+02	.

8 rows × 21 columns

Figure 25: Statistical summary of compliments data

Statistical Analysis Results:

Statistical analysis is a process of collecting, cleaning, exploring, and interpreting data to uncover patterns, relationships, and insights. It involves using statistical techniques to summarize and draw meaningful conclusions from data. In my I had applied various statistical techniques to draw more conclusions from the data. I used box plot to find the outliers in the dataset, raincloud plot for extensive analysis, T-Test, P-Test, Covariance, Correlation, Principal Component Analysis, Chi-square test, Confidence Interval of ratings between various columns of the dataset, applying normal distribution on the dataset, probability distribution of the ratings data. I am going to showcase the results of above-mentioned topics.

I had used boxplots to find out the outliers of funny, useful, cool columns with the review ratings of the users. Each of the column is plotted individually in the below figures.

The below figure represents the box plot between funny and length of the text column data.

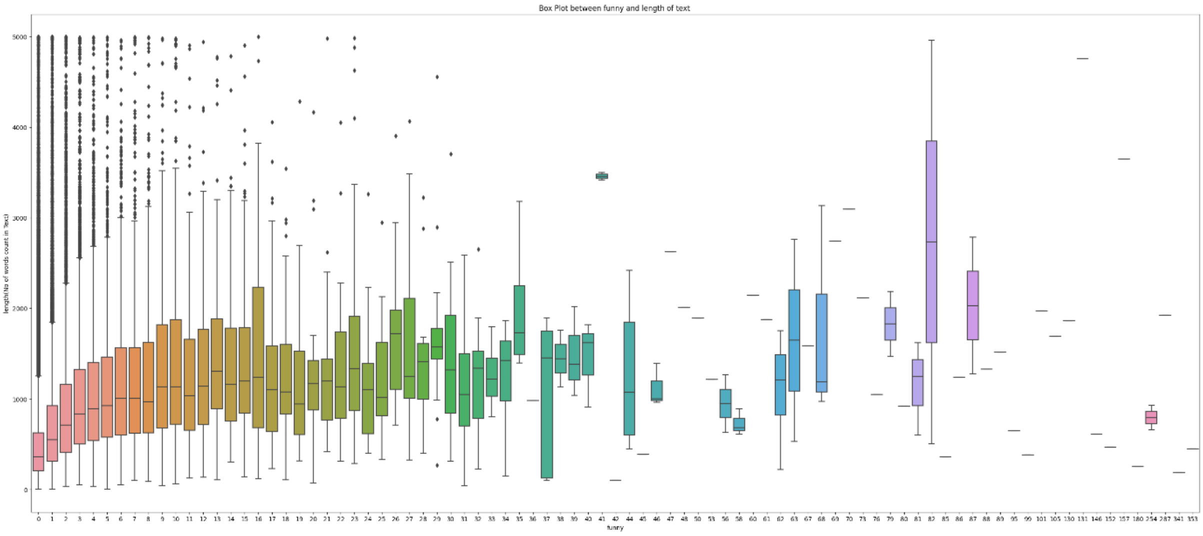


Figure 26: Box plot of funny data

Below figure depicts the box plot information of cool and length of text column data.

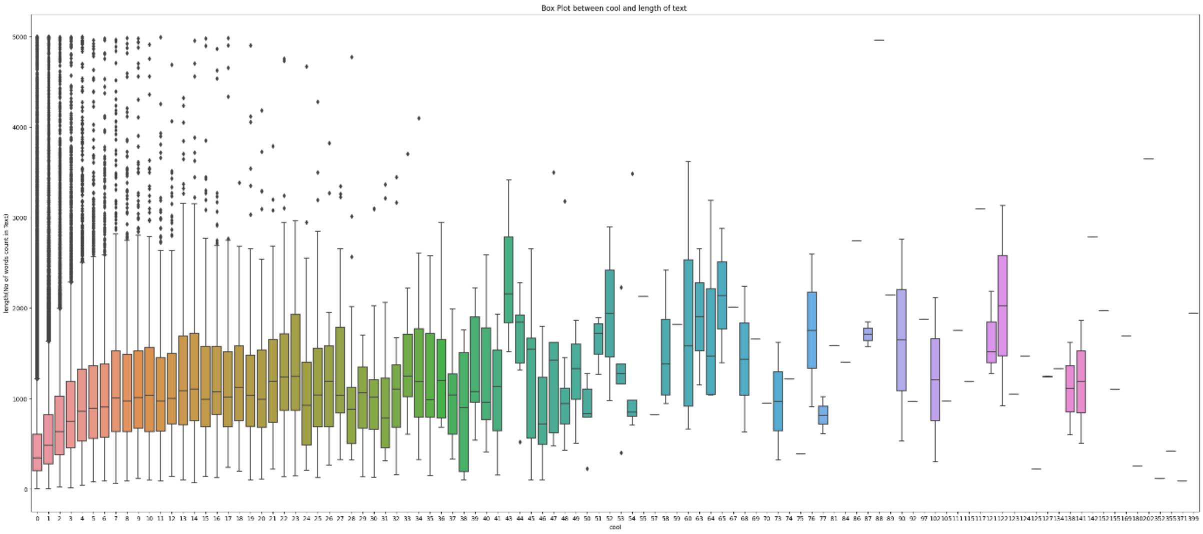


Figure 27: Box plot of cool data

Boxplot information of useful and length of text column data is shown below figure.

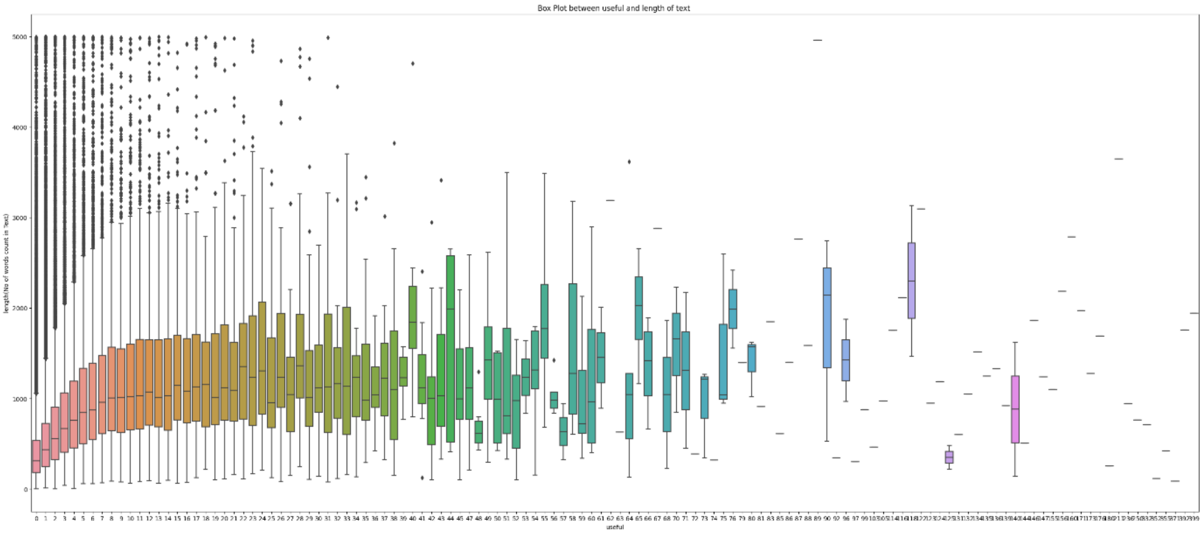
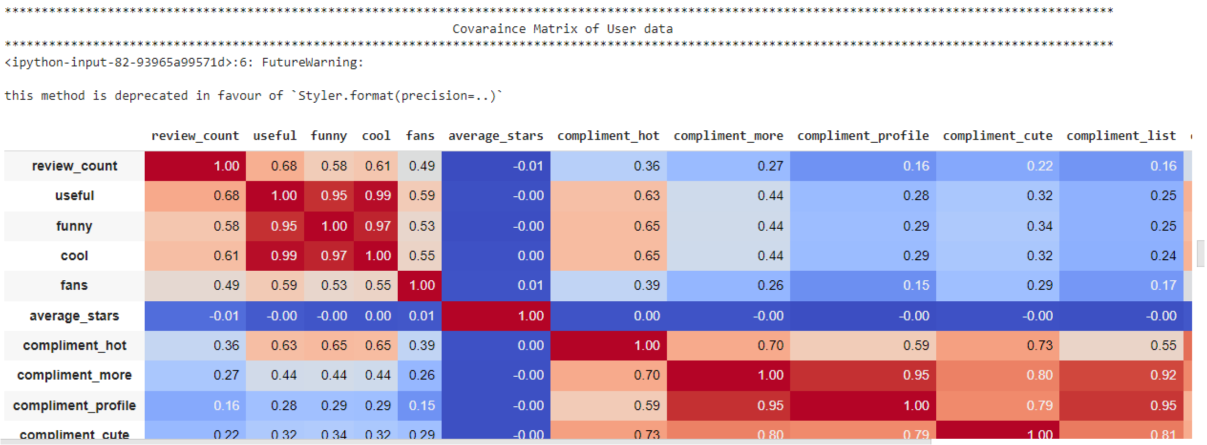


Figure 28: Box plot of useful data

I had plotted covariance distribution between user column data through a coolwarm map. The below figure shows the covariance information of user's data.



```
**********************************************************************************************************************
                                            Covaraince Matrix of User data
**********************************************************************************************************************
<ipython-input-82-93965a99571d>:6: FutureWarning:

this method is deprecated in favour of `Styler.format(precision=..)`
```

	review_count	useful	funny	cool	fans	average_stars	compliment_hot	compliment_more	compliment_profile	compliment_cute	compliment_list
review_count	1.00	0.68	0.58	0.61	0.49	-0.01	0.36	0.27	0.16	0.22	0.16
useful	0.68	1.00	0.95	0.99	0.59	-0.00	0.63	0.44	0.28	0.32	0.25
funny	0.58	0.95	1.00	0.97	0.53	-0.00	0.65	0.44	0.29	0.34	0.25
cool	0.61	0.99	0.97	1.00	0.55	0.00	0.65	0.44	0.29	0.32	0.24
fans	0.49	0.59	0.53	0.55	1.00	0.01	0.39	0.26	0.15	0.29	0.17
average_stars	-0.01	-0.00	-0.00	0.00	0.01	1.00	0.00	-0.00	-0.00	-0.00	-0.00
compliment_hot	0.36	0.63	0.65	0.65	0.39	0.00	1.00	0.70	0.59	0.73	0.55
compliment_more	0.27	0.44	0.44	0.44	0.26	-0.00	0.70	1.00	0.95	0.80	0.92
compliment_profile	0.16	0.28	0.29	0.29	0.15	-0.00	0.59	0.95	1.00	0.79	0.95
compliment_cute	0.22	0.32	0.34	0.32	0.29	-0.00	0.73	0.80	0.79	1.00	0.81

Figure 29: Covariance Matrix

Below figure shows the information about the data distribution of average review ratings of yelp business data.

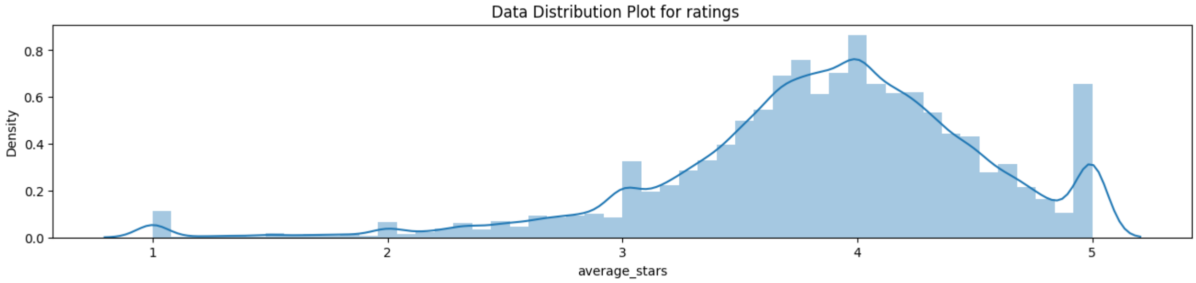


Figure 30: Data distribution of ratings

This figure shows the information about distribution of fans data and elite_count data. The data is represented in terms of density distribution curve.

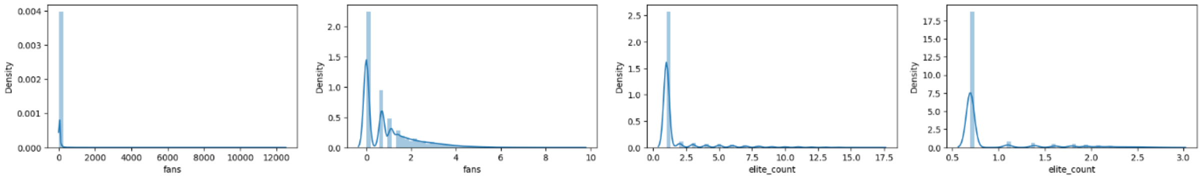


Figure 31: Data distribution of fans and elite_count

Density distribution of friend_count column data was shown in the below figure.

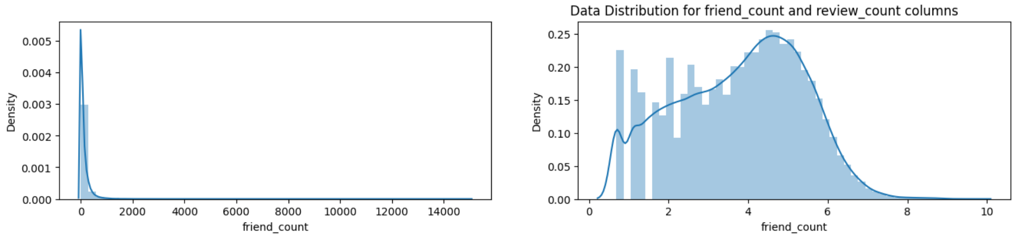


Figure 32: Data distribution of friend_count

Below figure furnishes the information about the review_count distribution data.

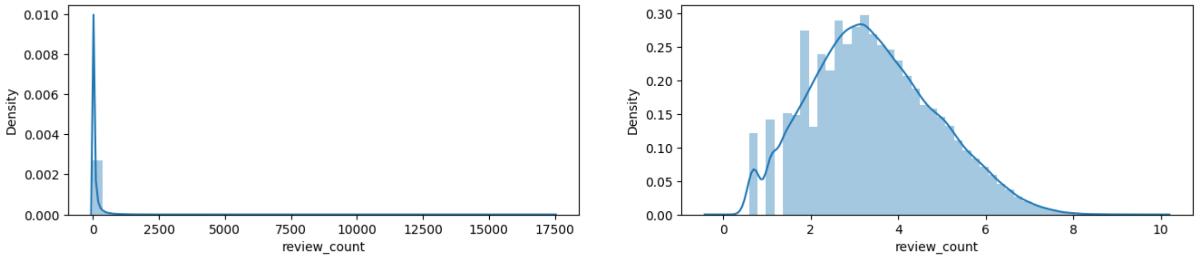
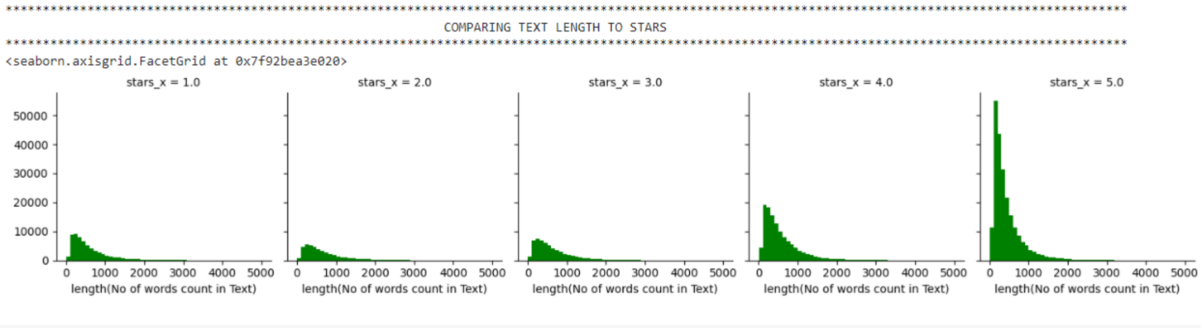


Figure 33: Data distribution of review_count

Distribution of review ratings with length of review text was shown in the below figure.



```
************************************************************************************************************************************
                                                   COMPARING TEXT LENGTH TO STARS
************************************************************************************************************************************
<seaborn.axisgrid.FacetGrid at 0x7f92bea3e020>
```

Figure 34: Distribution of review ratings

Distribution of useful column data is shown in the below figure.

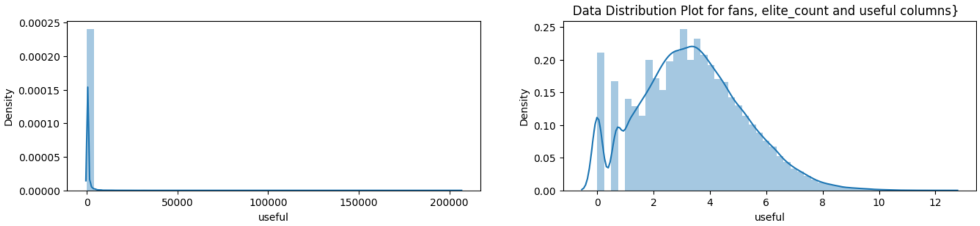


Figure 35: Data distribution of useful data

Box plot distribution of review ratings is shown in the below figure.

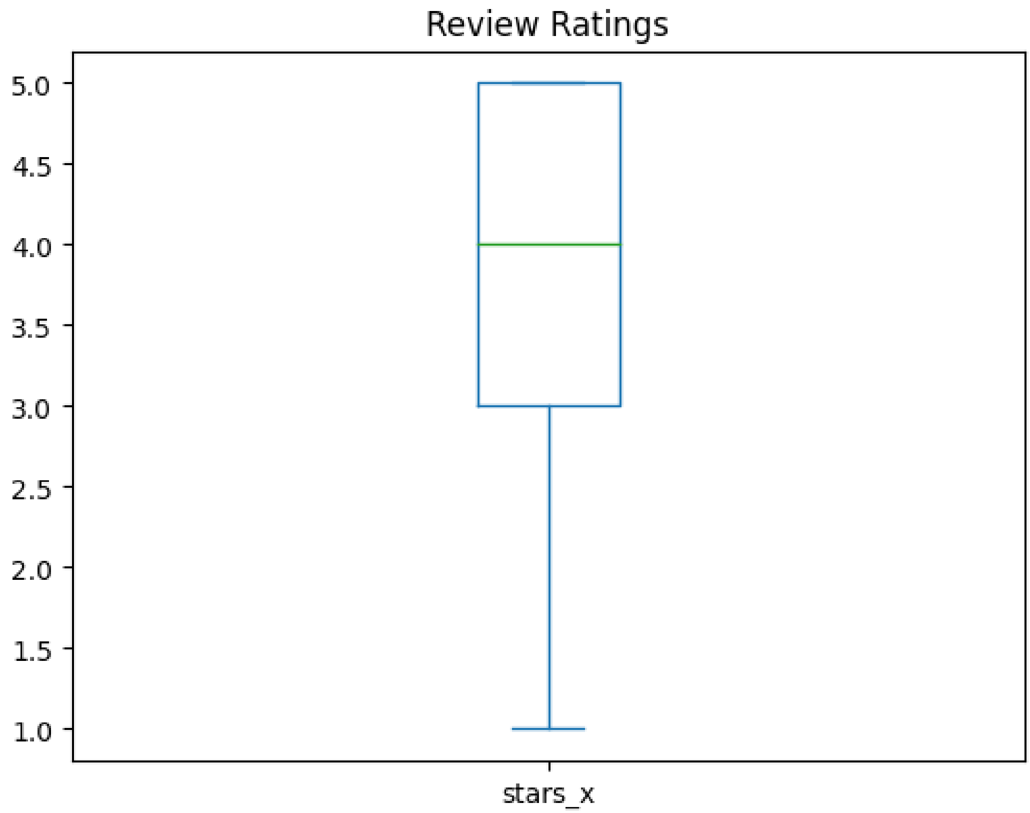
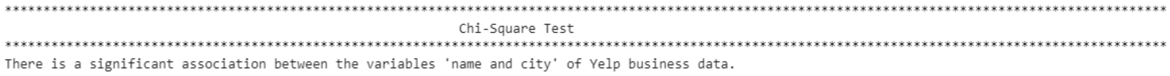


Figure 36: Box plot of ratings

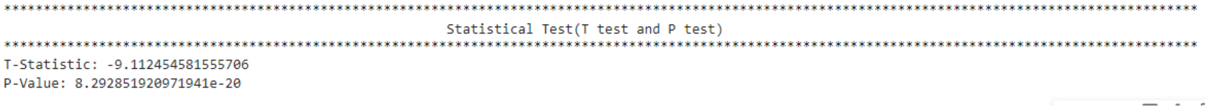
Below figure shows the information about Chi- square test results of 'name and city' columns data.



```
*****************************************************************************************************************************************
                                                    Chi-Square Test
*****************************************************************************************************************************************
There is a significant association between the variables 'name and city' of Yelp business data.
```

Figure 37: Chi-square test

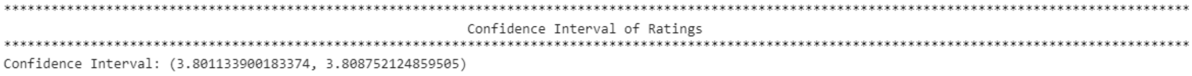
Below figure shows the information about P-Test and T-Test results.



```
*****************************************************************************************************************************************
                                                Statistical Test(T test and P test)
*****************************************************************************************************************************************
T-Statistic: -9.112454581555706
P-Value: 8.292851920971941e-20
```

Figure 38: P-Test & T-Test

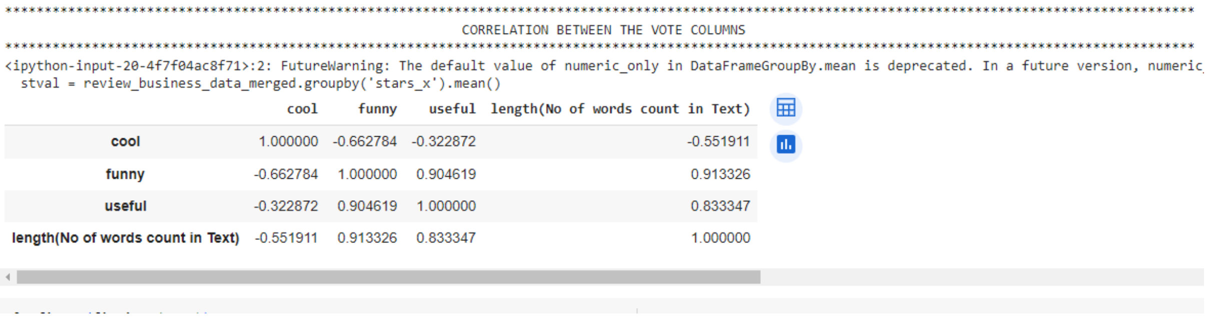
The below figure furnishes the information about confidence interval.



```
*****************************************************************************************************************************************
                                                    Confidence Interval of Ratings
*****************************************************************************************************************************************
Confidence Interval: (3.801133900183374, 3.808752124859505)
```

Figure 39: Confidence interval of ratings

Below figure shows the corelation details.



```
*****************************************************************************************************************************************
                                                  CORRELATION BETWEEN THE VOTE COLUMNS
*****************************************************************************************************************************************
<ipython-input-20-4f7f04ac8f71>:2: FutureWarning: The default value of numeric_only in DataFrameGroupBy.mean is deprecated. In a future version, numeric
  stval = review_business_data_merged.groupby('stars_x').mean()
```

	cool	funny	useful	length(No of words count in Text)
cool	1.000000	-0.662784	-0.322872	-0.551911
funny	-0.662784	1.000000	0.904619	0.913326
useful	-0.322872	0.904619	1.000000	0.833347
length(No of words count in Text)	-0.551911	0.913326	0.833347	1.000000

Figure 40: Correlation of Vote columns

Below picture displays the information about box plot of cool column.

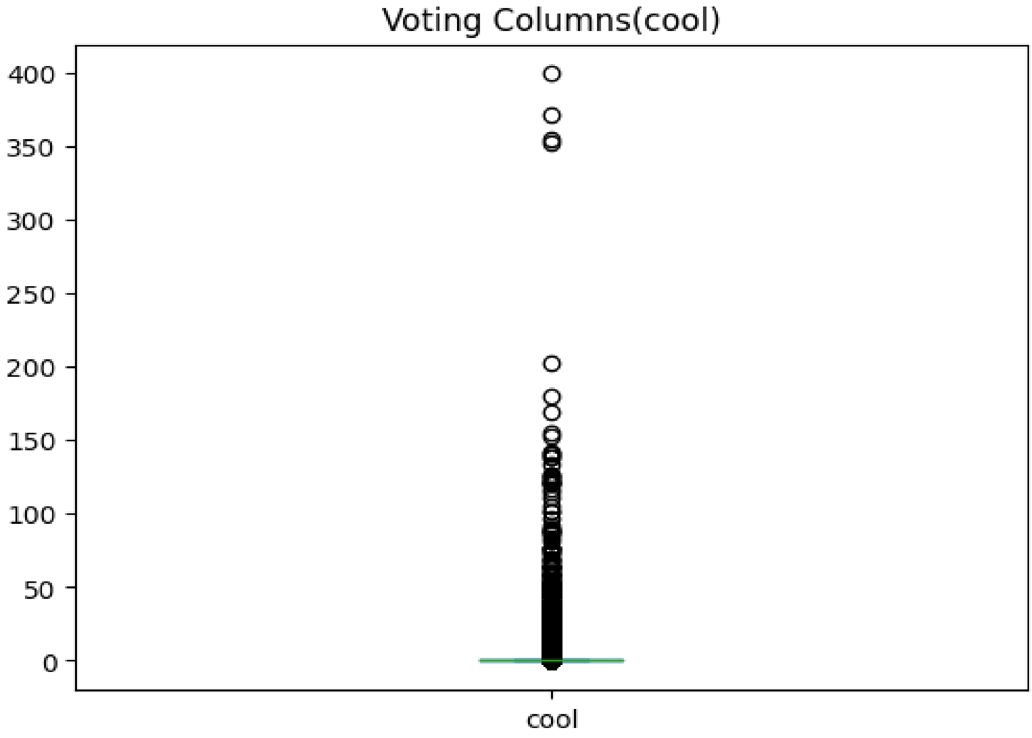


Figure 41: Box plot of cool column

Below picture displays the information about box plot of funny column.

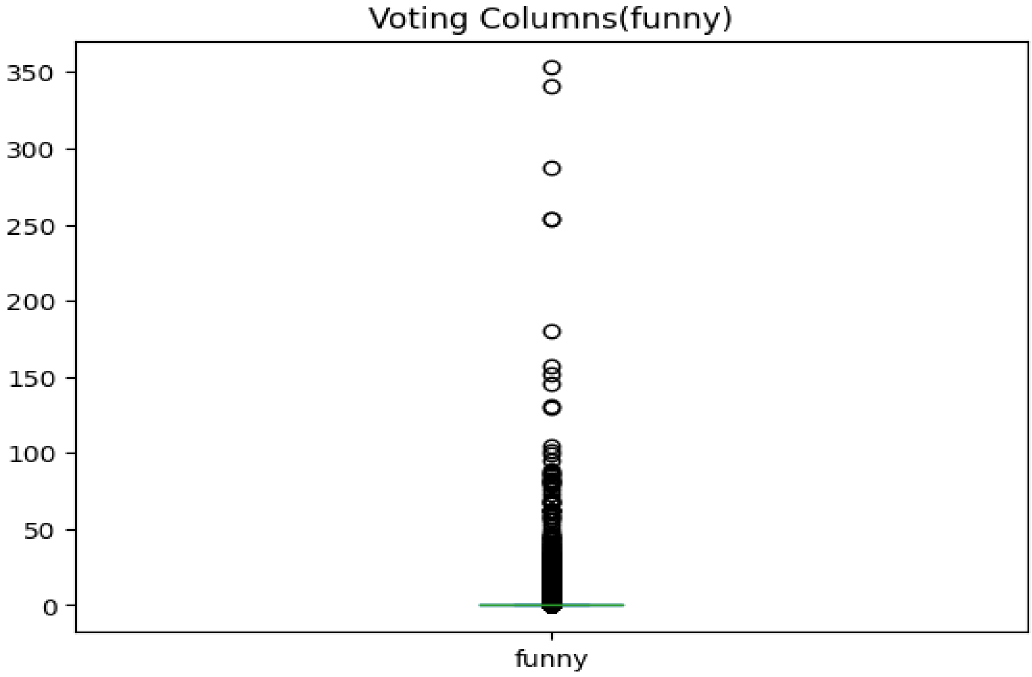


Figure 42: Box plot of funny column

Below picture displays the information about box plot of stars column.

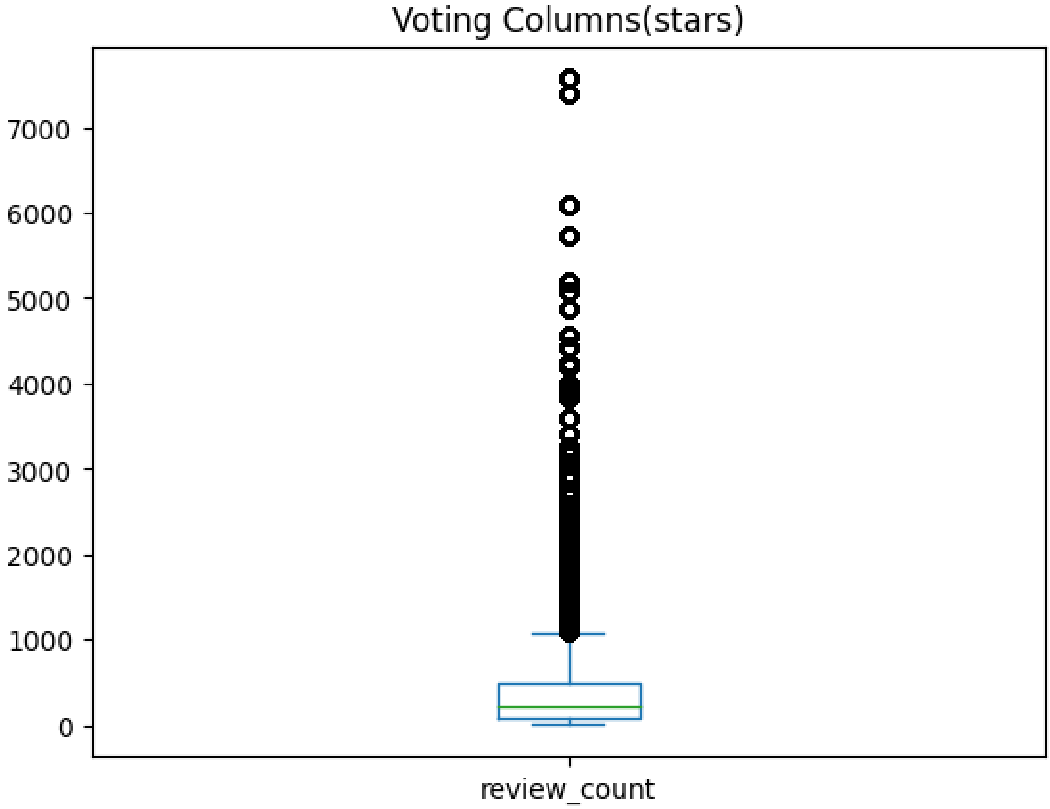


Figure 43: Box plot of review_count

Below picture displays the information about box plot of useful column.

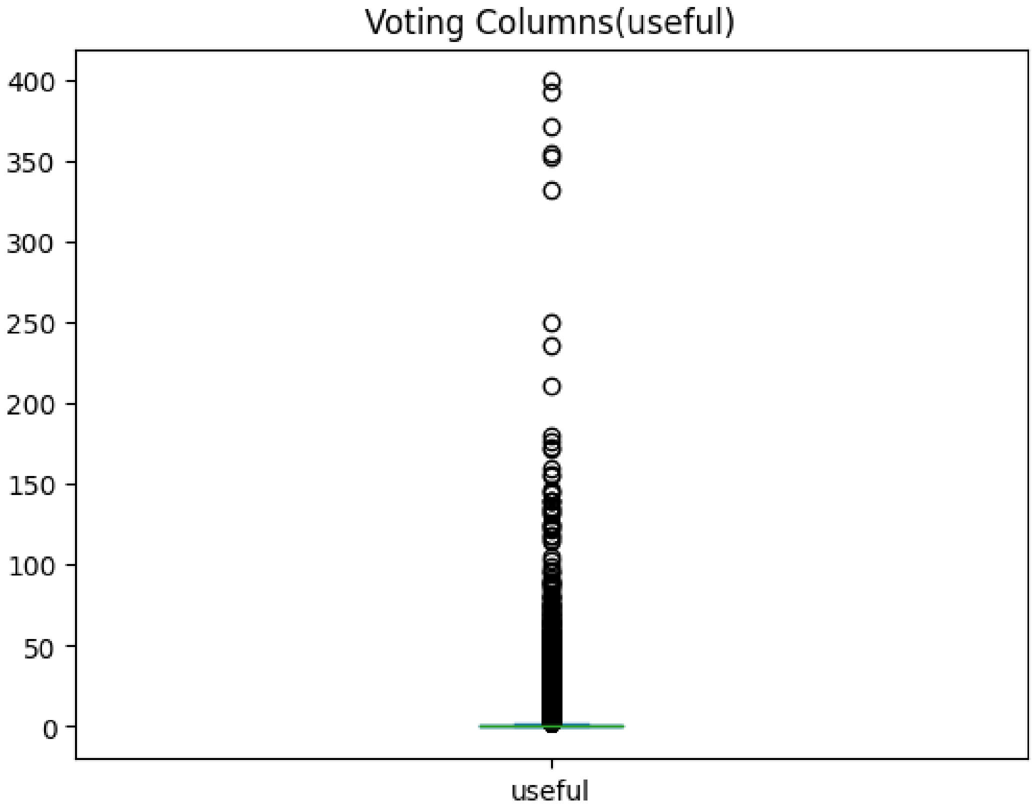


Figure 44: Box plot of useful column

The below picture depicts the information about rain plot of ratings.

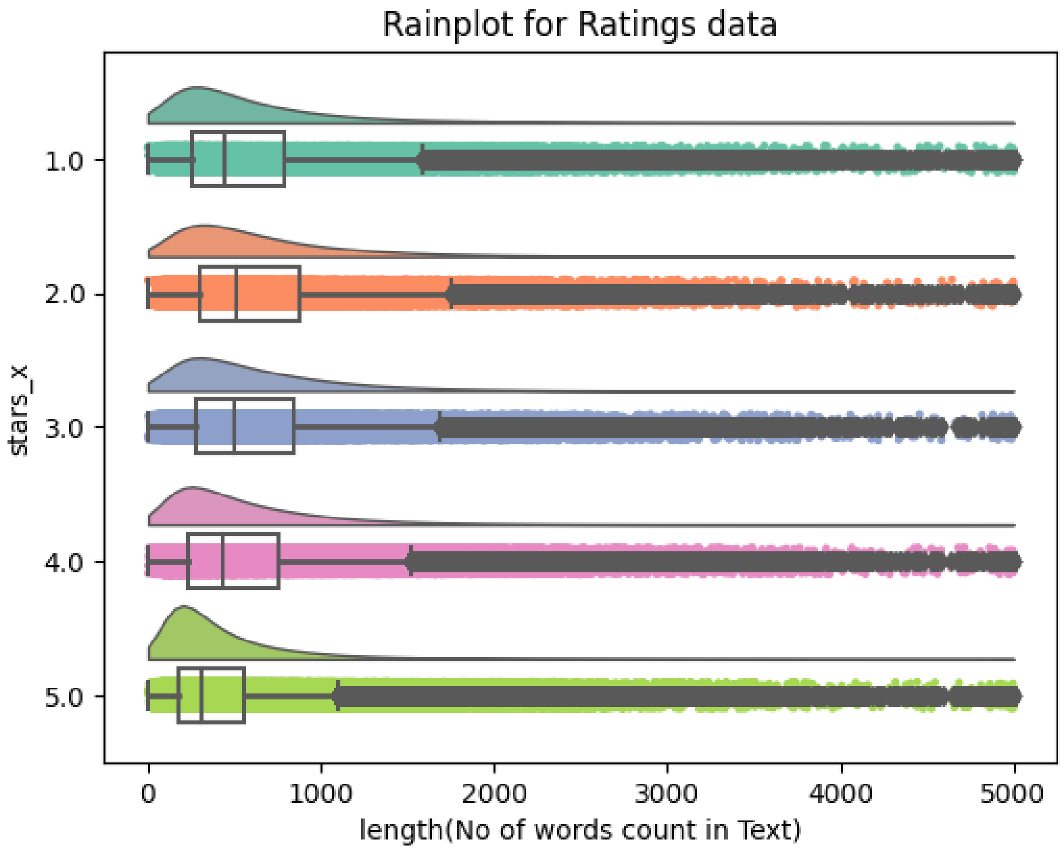
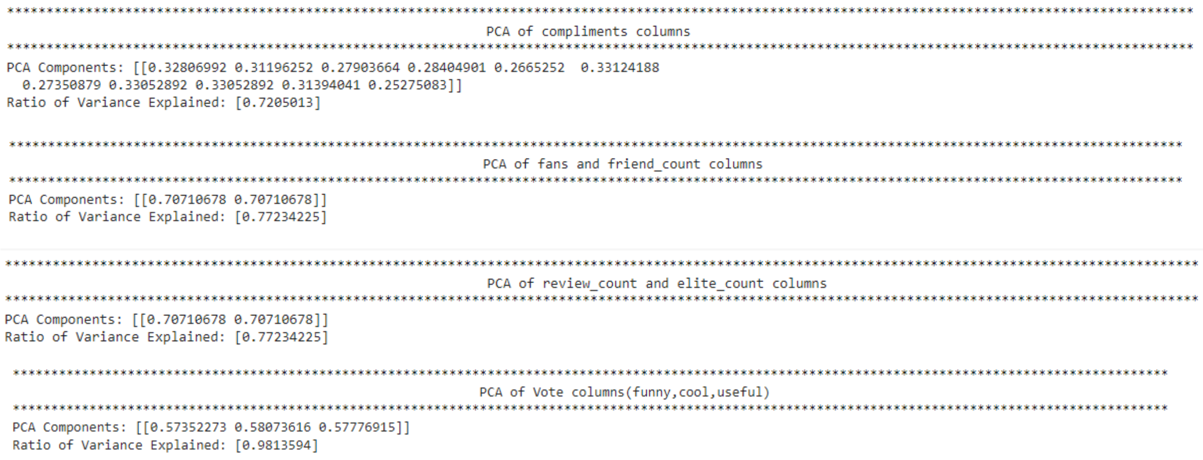


Figure 45: Rain Plot of Ratings data

Below pictures shows the information about Principal component analysis of various columns.



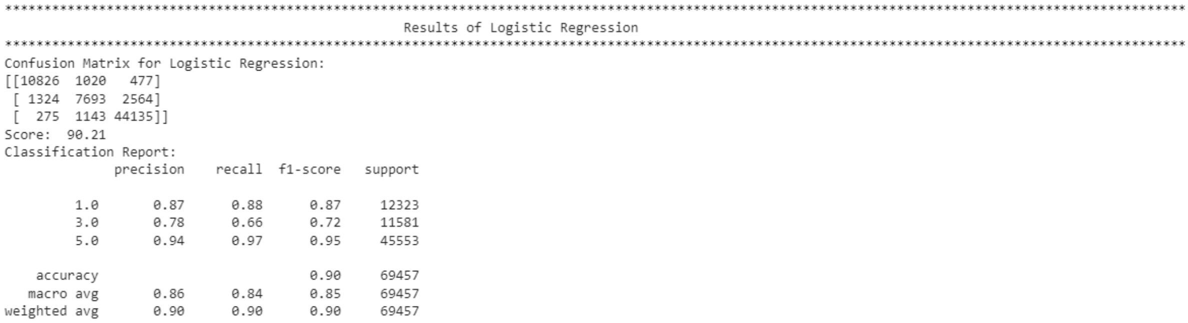
```
****************************************************************************************************
                                     PCA of compliments columns
****************************************************************************************************
PCA Components: [[0.32806992 0.31196252 0.27903664 0.28404901 0.2665252  0.33124188
  0.27350879 0.33052892 0.33052892 0.31394041 0.25275083]]
Ratio of Variance Explained: [0.7205013]

****************************************************************************************************
                               PCA of fans and friend_count columns
****************************************************************************************************
PCA Components: [[0.70710678 0.70710678]]
Ratio of Variance Explained: [0.77234225]

****************************************************************************************************
                            PCA of review_count and elite_count columns
****************************************************************************************************
PCA Components: [[0.70710678 0.70710678]]
Ratio of Variance Explained: [0.77234225]

****************************************************************************************************
                              PCA of Vote columns(funny,cool,useful)
****************************************************************************************************
PCA Components: [[0.57352273 0.58073616 0.57776915]]
Ratio of Variance Explained: [0.9813594]
```

Figure 46: Results of PCA analysis

Machine Learning Results:

Machine learning is a subfield of artificial intelligence (AI) that focuses on developing algorithms and models that allow computer systems to learn from and make predictions or decisions based on data, without being explicitly programmed. In essence, it enables machines to learn patterns and make informed decisions or predictions without being explicitly programmed for each task. In my research I had used six different machine learning algorithms in-order to achieve my project aim. Logistic Regression, Multinomial Navie Bayes, Gradient Boosting Classifier. The results of each machine learning technique were discussed in the below sections. Logistic Regression Algorithm performs better when compared with other two algorithms.

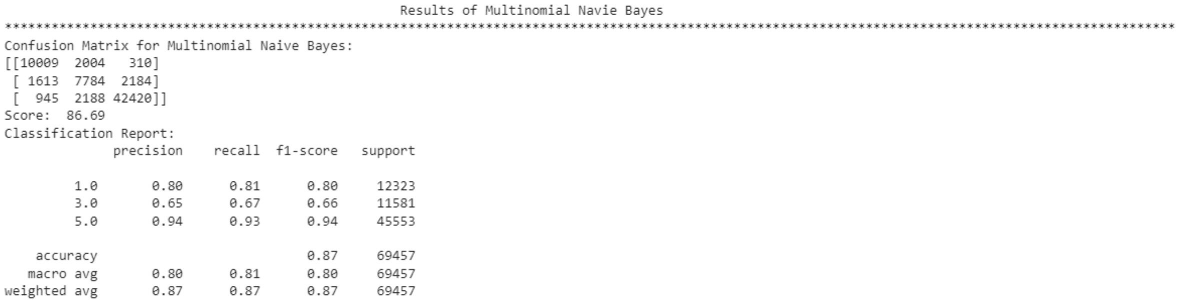
Logistic Regression: Logistic regression is a statistical method and a popular algorithm in machine learning used for binary classification tasks. It is used when the dependent variable is categorical, and it predicts the probability of an observation belonging to one of two classes. Unlike linear regression, which predicts a continuous outcome, logistic regression predicts the probability that an observation falls into one of two categories (e.g., yes/no, 1/0, true/false). In my research logistic regression achieved an accuracy of about 90%. The confusion matrix and classification report of logistic regression is shown in the below figure.



```
****************************************************************************************************************************************
                                                    Results of Logistic Regression
****************************************************************************************************************************************
Confusion Matrix for Logistic Regression:
[[10826  1020   477]
 [ 1324  7693  2564]
 [  275  1143 44135]]
Score:  90.21
Classification Report:
              precision    recall  f1-score   support

         1.0       0.87      0.88      0.87     12323
         3.0       0.78      0.66      0.72     11581
         5.0       0.94      0.97      0.95     45553

    accuracy                           0.90     69457
   macro avg       0.86      0.84      0.85     69457
weighted avg       0.90      0.90      0.90     69457
```

Figure 47: Logistic regression results

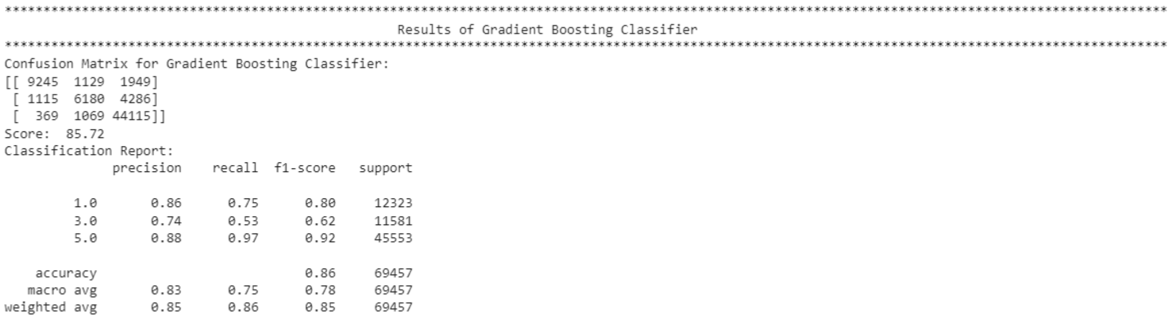
Multinomial Navie Bayes: Multinomial Naive Bayes is a probabilistic classification algorithm that is used for text classification and document categorization tasks, particularly when dealing with text data where the features (words or terms) are discrete and represent counts, frequencies, or occurrences. It is an extension of the Naive Bayes algorithm, which assumes that the features are conditionally independent given the class label. Multinomial Naive Bayes, in particular, is well-suited for problems involving text data, such as spam detection, sentiment analysis, and topic classification. Multinomial Navie Bayes achieved an accuracy of 86.69%. Below picture shows the results of multinomial navie bayes.



```
****************************************************************************************************************************************
                                                    Results of Multinomial Navie Bayes
****************************************************************************************************************************************
Confusion Matrix for Multinomial Naive Bayes:
[[10009  2004   310]
 [ 1613  7784  2184]
 [  945  2188 42420]]
Score:  86.69
Classification Report:
              precision    recall  f1-score   support

         1.0       0.80      0.81      0.80     12323
         3.0       0.65      0.67      0.66     11581
         5.0       0.94      0.93      0.94     45553

    accuracy                           0.87     69457
   macro avg       0.80      0.81      0.80     69457
weighted avg       0.87      0.87      0.87     69457
```

Figure 48: Multinomial Navie Bayes results

Gradient Boosting Classifier: A Gradient Boosting Classifier is a powerful ensemble machine learning algorithm used primarily for classification tasks. It is an ensemble of decision trees, and it builds predictive models by combining the predictions of multiple weak learners (usually decision trees) to create a strong predictive model. Gradient boosting is known for its high accuracy and robustness and is often used in competitions and real-world applications. It achieved an accuracy of 85.72%. Below figure depicts the information about the classification report of gradient boosting classifier.

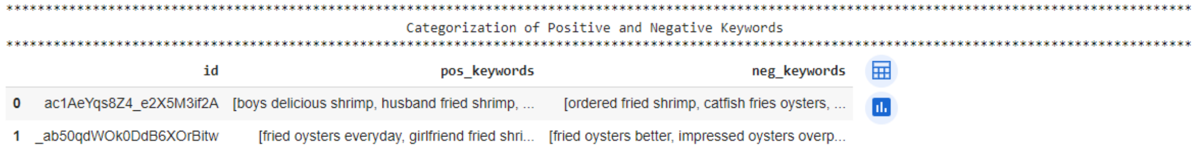


```
****************************************************************************************************************************
                                             Results of Gradient Boosting Classifier
****************************************************************************************************************************
Confusion Matrix for Gradient Boosting Classifier:
[[ 9245  1129  1949]
 [ 1115  6180  4286]
 [  369  1069 44115]]
Score:  85.72
Classification Report:
              precision    recall  f1-score   support

         1.0       0.86      0.75      0.80     12323
         3.0       0.74      0.53      0.62     11581
         5.0       0.88      0.97      0.92     45553

    accuracy                           0.86     69457
   macro avg       0.83      0.75      0.78     69457
weighted avg       0.85      0.86      0.85     69457
```

Figure 49: Gradient Boosting results

Sentimental Analysis Results: Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine and classify the sentiment or emotional tone expressed in text data. The goal of sentiment analysis is to understand the opinions, attitudes, or emotions conveyed by individuals or groups of people in written or spoken language. Sentiment analysis can be applied to various forms of text data, including social media posts, product reviews, news articles, customer feedback, and more.[6][7] In my research I had used sentimental analysis to plot the distribution based on the sentiments, categorized the positive and negative review comments, distribution of word clouds etc. The results of the sentimental analysis are discussed below.

Below figure depicts the information about the categorization of positive and negative words of a business id.

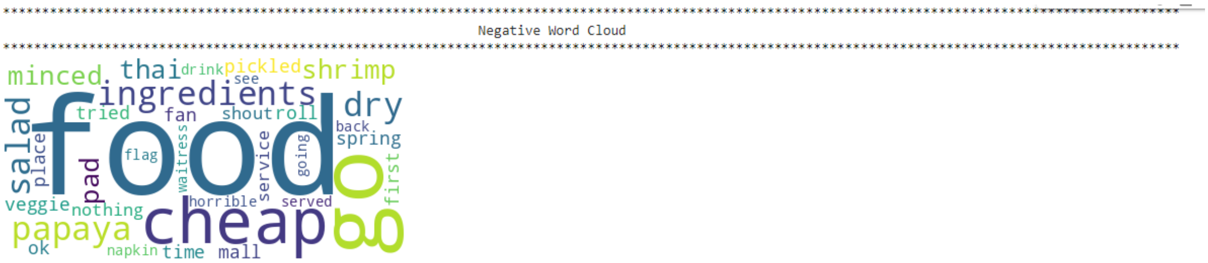


```
****************************************************************************************************************************
                                         Categorization of Positive and Negative Keywords
****************************************************************************************************************************
```

	id	pos_keywords	neg_keywords
0	ac1AeYqs8Z4_e2X5M3if2A	[boys delicious shrimp, husband fried shrimp, ...	[ordered fried shrimp, catfish fries oysters, ...
1	_ab50qdWOk0DdB6XOrBitw	[fried oysters everyday, girlfriend fried shri...	[fried oysters better, impressed oysters overp...

Figure 50: Categorization of positive and negative keywords

I had categorized positive word clouds and negative clouds of a business. The below figure shows the information about the negative word clouds of a particular business.



```
****************************************************************************************************************************
                                                    Negative Word Cloud
****************************************************************************************************************************
```

Figure 51: Negative word cloud

The below figure depicts the information about the distribution of sentiments of a particular business_id.

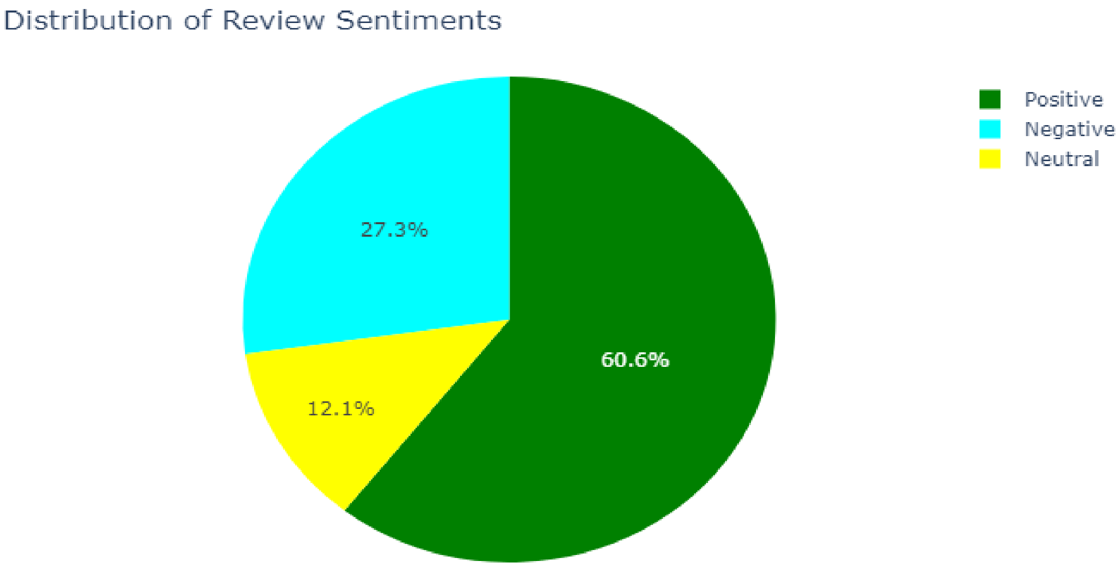
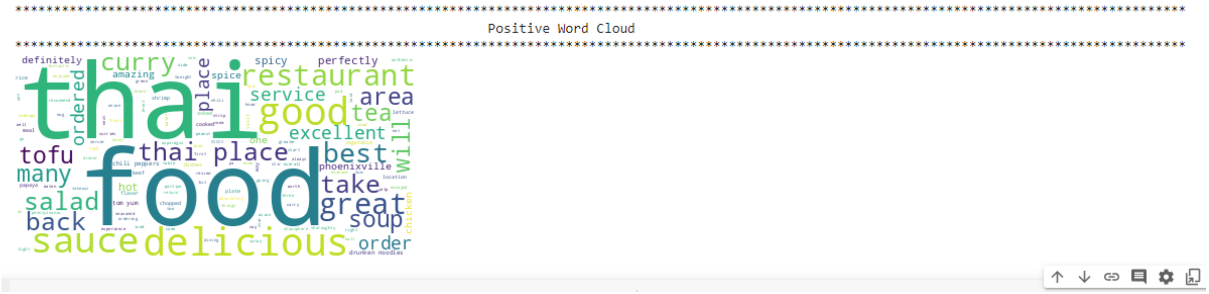


Figure 52: Distribution of ratings

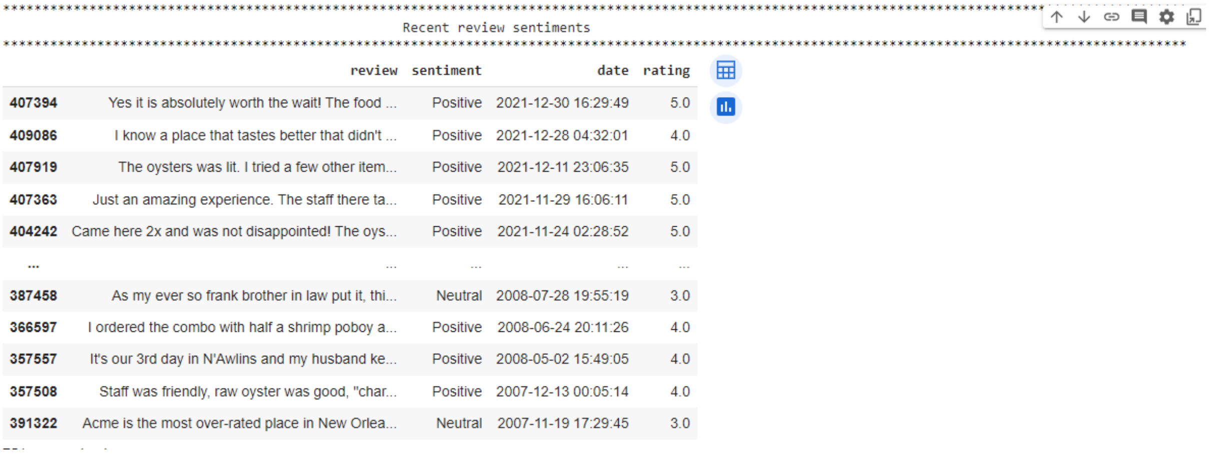
The below figure shows the information about the positive word clouds of a particular business.



```
****************************************************************************************************
                                        Positive Word Cloud
****************************************************************************************************
```

Figure 53: Positive word cloud

Below picture furnishes recent review ratings.



```
*****************************************************************************************************
                               Recent review sentiments
*****************************************************************************************************
```

	review	sentiment	date	rating
407394	Yes it is absolutely worth the wait! The food ...	Positive	2021-12-30 16:29:49	5.0
409086	I know a place that tastes better that didn't ...	Positive	2021-12-28 04:32:01	4.0
407919	The oysters was lit. I tried a few other item...	Positive	2021-12-11 23:06:35	5.0
407363	Just an amazing experience. The staff there ta...	Positive	2021-11-29 16:06:11	5.0
404242	Came here 2x and was not disappointed! The oys...	Positive	2021-11-24 02:28:52	5.0
...	...	...	...	...
387458	As my ever so frank brother in law put it, thi...	Neutral	2008-07-28 19:55:19	3.0
366597	I ordered the combo with half a shrimp poboy a...	Positive	2008-06-24 20:11:26	4.0
357557	It's our 3rd day in N'Awlins and my husband ke...	Positive	2008-05-02 15:49:05	4.0
357508	Staff was friendly, raw oyster was good, "char...	Positive	2007-12-13 00:05:14	4.0
391322	Acme is the most over-rated place in New Orlea...	Neutral	2007-11-19 17:29:45	3.0

Figure 54: Recent review sentiments

Below picture shows the word cloud of starbucks company

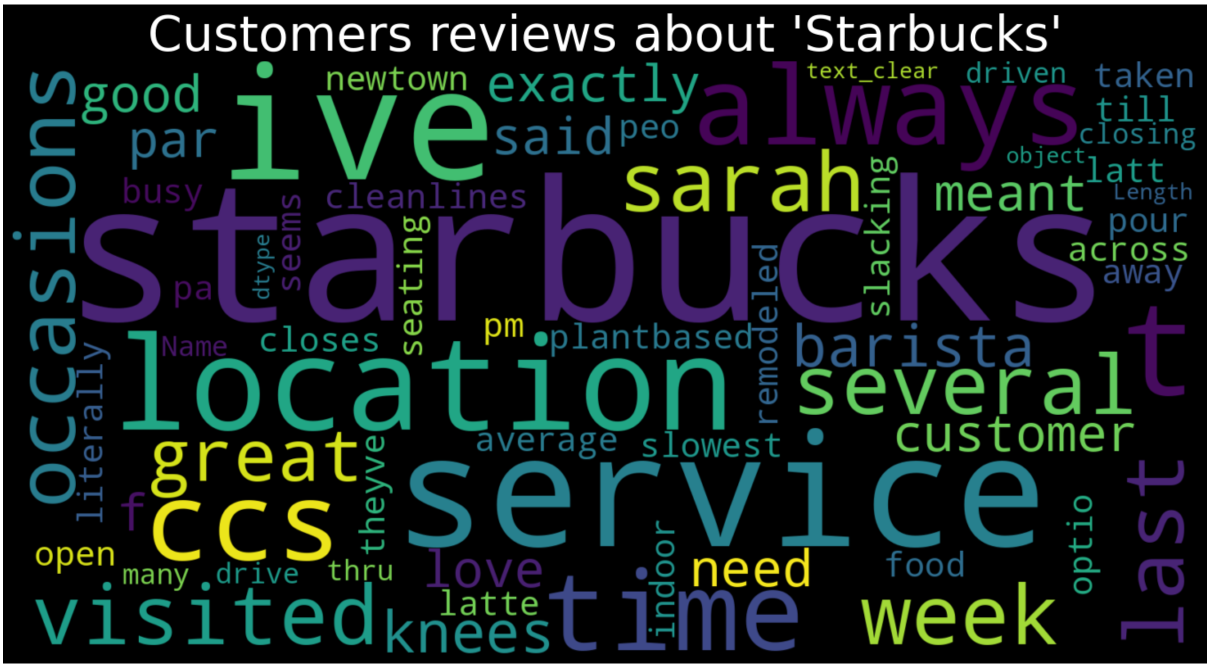


Figure 55: Word cloud of Starbucks

Below picture shows the word cloud of Taco Bell Business.

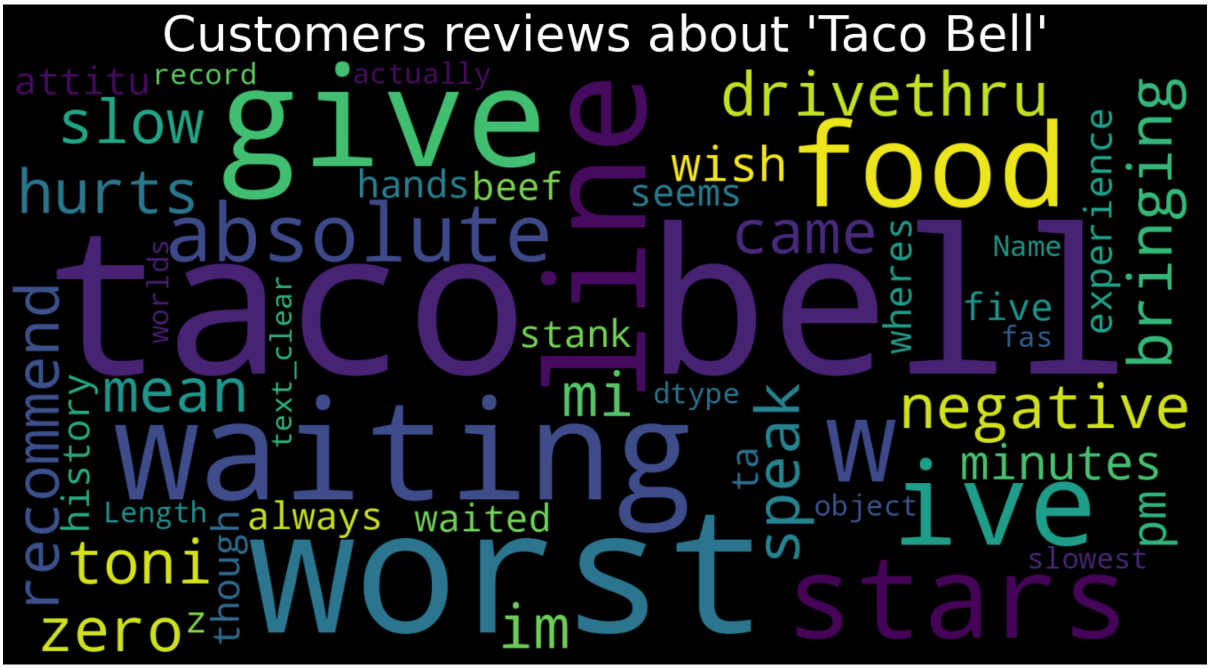


Figure 56: Word cloud of Taco Bell

Exploratory Data Analysis (EDA): Exploratory Data Analysis (EDA) is a critical initial step in the data analysis process. It involves examining and visualizing data to gain a better understanding of its characteristics, uncover patterns, identify outliers, and generate hypotheses. EDA helps data analysts and scientists make informed decisions about data pre-processing, feature selection, and modeling. In my research I had analysed a lot of exploratory data analysis to get useful insights for the business. The results of Exploratory Data Analysis are discussed in the sections.

Below figures shows the information about the top performing categories of the customer feedback.

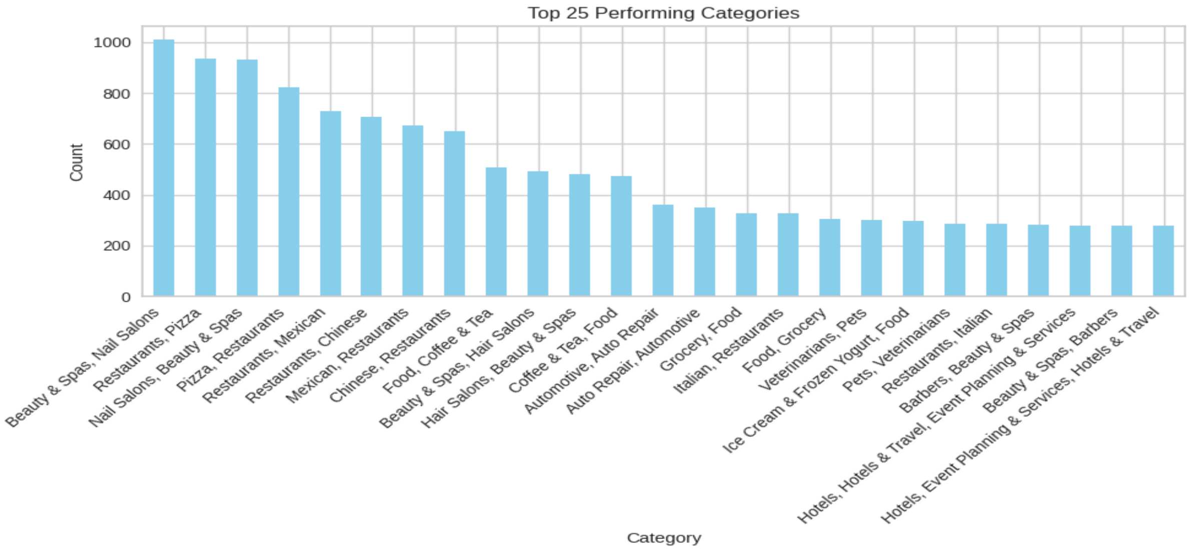


Figure 57: Top performing business

The below bar chart shows the information about the distribution of reviews that are performing best.

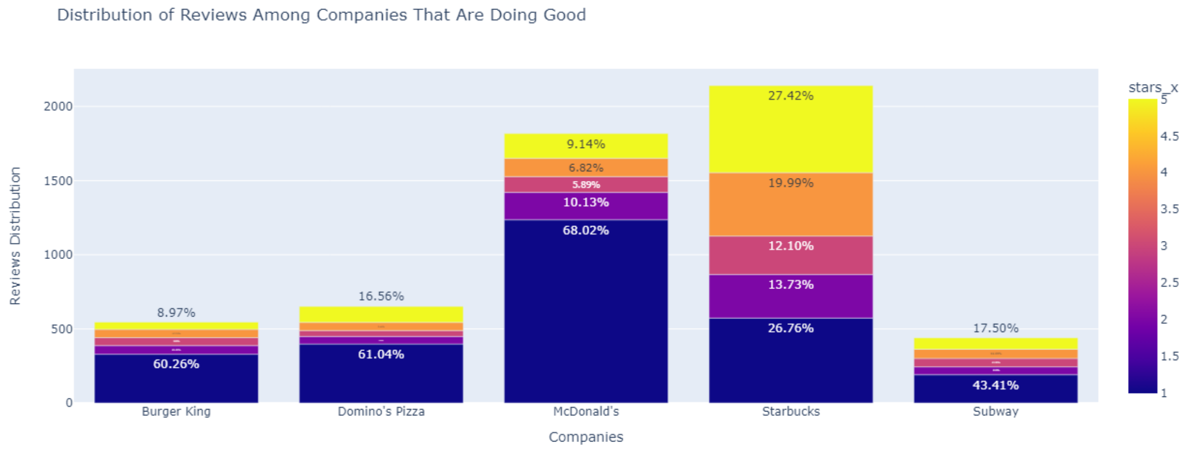


Figure 58: Ratings distribution of top companies

Below bar chart depicts the information about the top performing cities in terms of business.

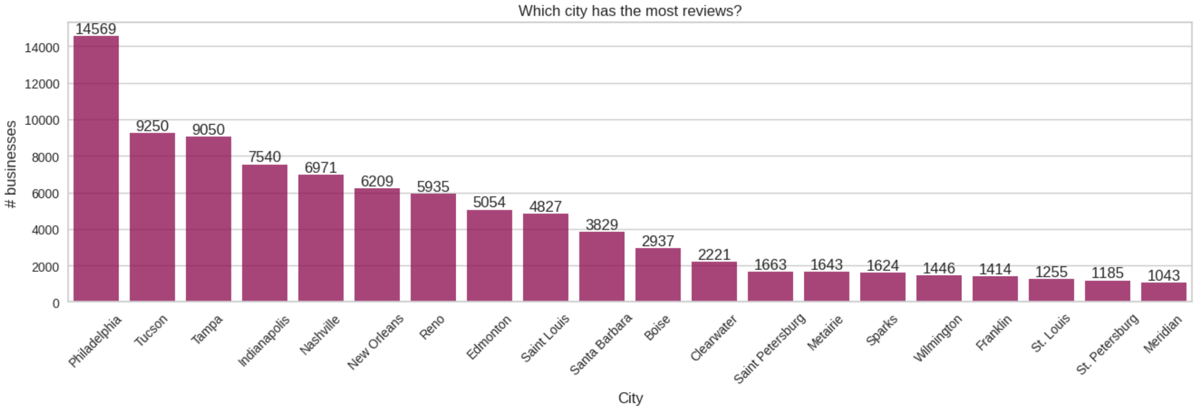


Figure 59: Top performing business cities

Below figure shows the information about the KMeans clustering to find out best number of clusters.

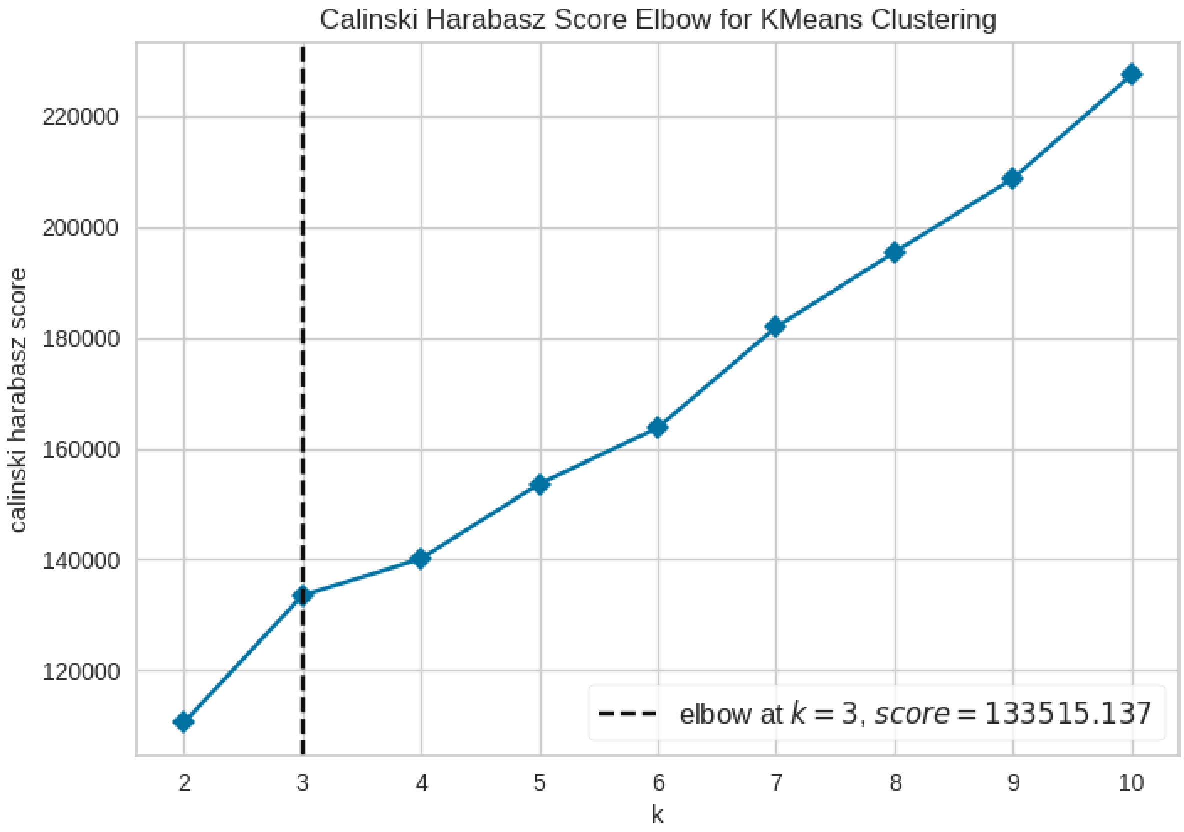


Figure 60: KMeans clustering

The below figure shows the information about the active feedback and average stars based on the number of clusters.

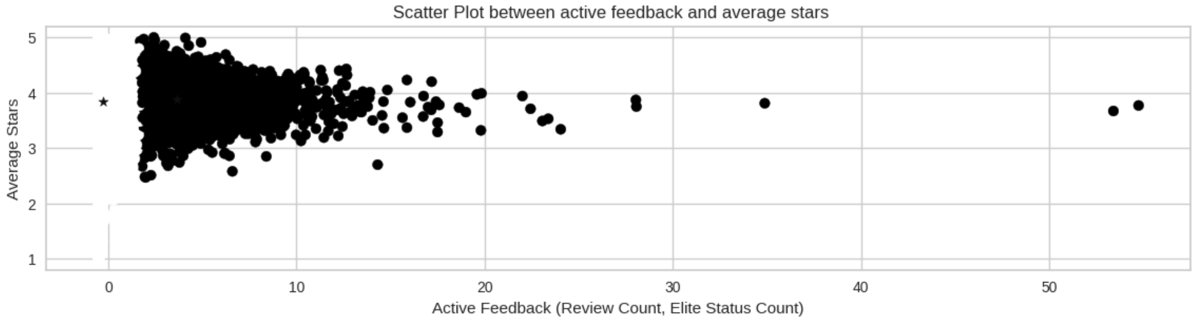


Figure 61: Scatter plot between active feedback and ratings

Below figure generates the information about compliments feedback and average stars based on the number of clusters.

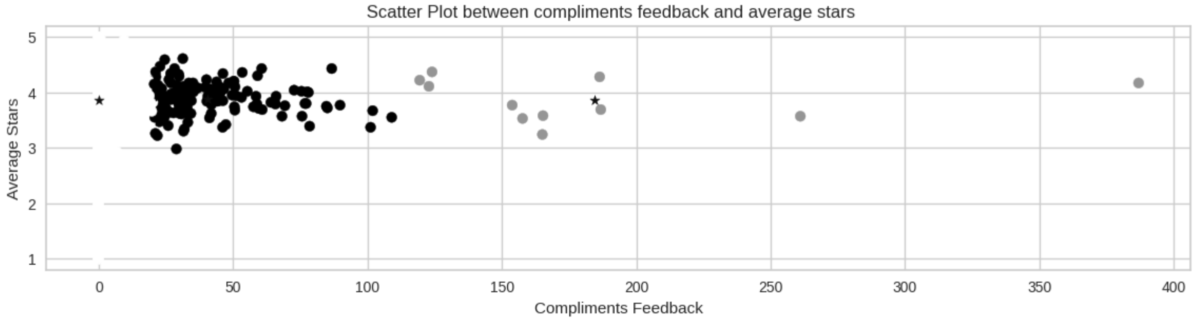


Figure 62: Scatter plot between compliment feedback and ratings

The below figure shows the information about the popularity feedback and average stars based on the number of clusters.

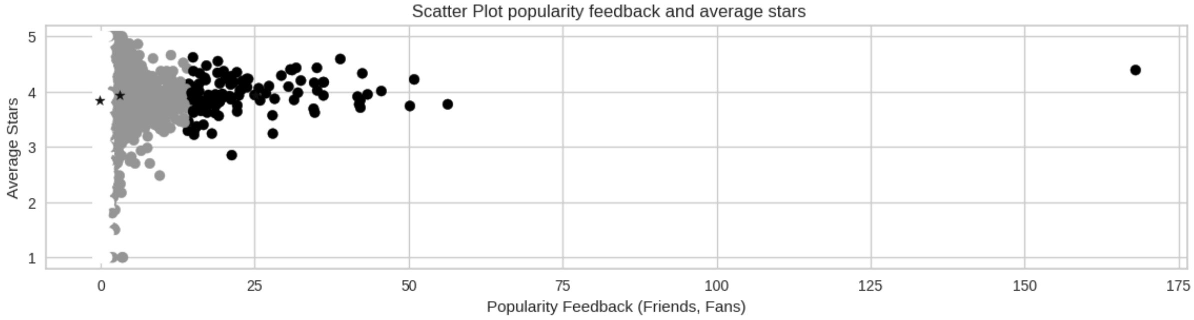


Figure 63: Scatter plot between popularity feedback and ratings

Below figure depicts the information about the Review feedback and Active feedback of stars based on the number of clusters.

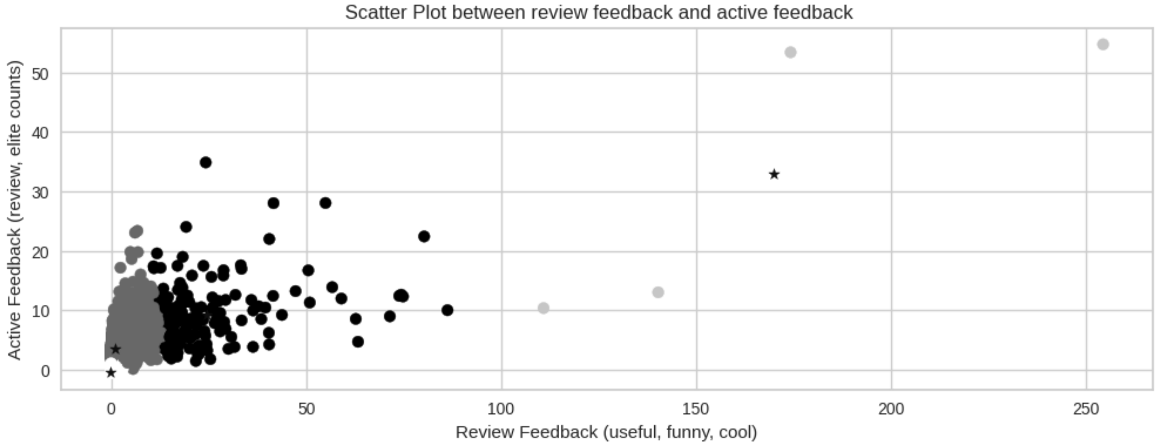


Figure 64: Scatter plot between review feedback and ratings

Below figure depicts the information about the Review feedback and Popularity feedback of stars based on the number of clusters.

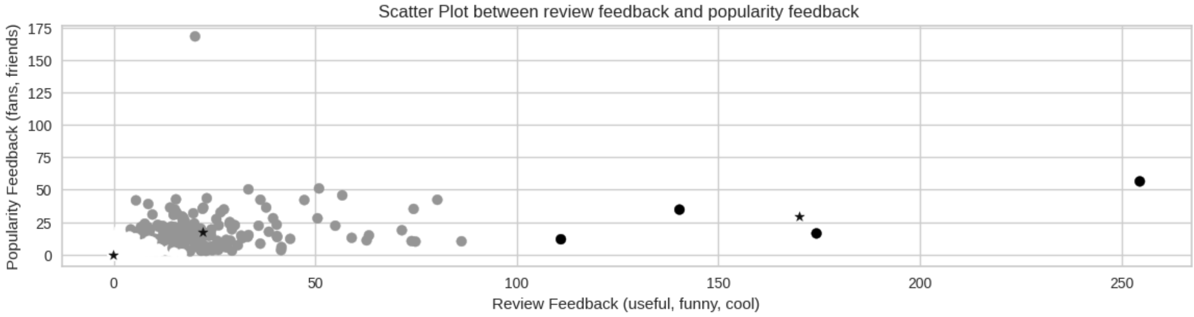


Figure 65: Scatter plot between review feedback and popularity feedback

Below figure generates the information about the distribution of review ratings.

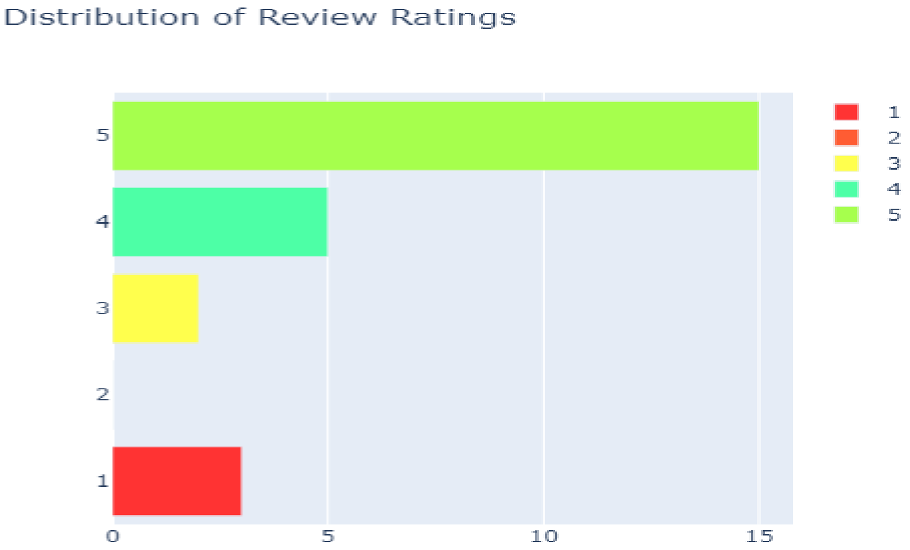
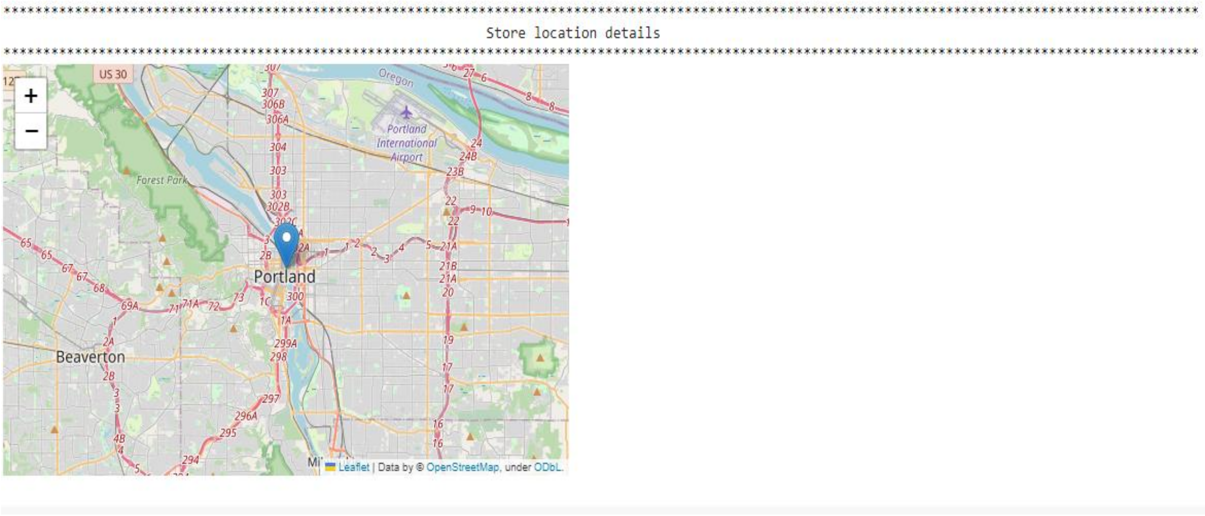


Figure 66: Interactive ratings distribution

The below figure shows the location of a particular business store.

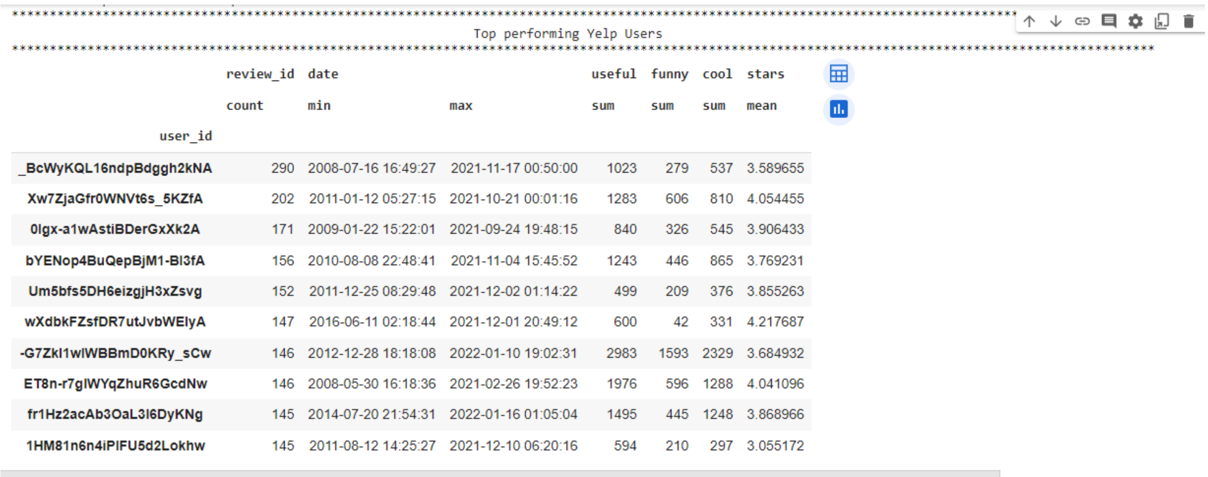


```
************************************************************************************************************************************
                                                   Store location details
************************************************************************************************************************************
```

US 30
Oregon
Portland International Airport
Forest Park
Portland
Beaverton
+
−
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 67: Store location details

The below figure generates the information of top performing users.



```
************************************************************************************************************************************
                                                   Top performing Yelp Users
************************************************************************************************************************************
```

	review_id	date		useful	funny	cool	stars
	count	min	max	sum	sum	sum	mean
user_id							
_BcWyKQL16ndpBdggh2kNA	290	2008-07-16 16:49:27	2021-11-17 00:50:00	1023	279	537	3.589655
Xw7ZjaGfr0WNVt6s_5KZfA	202	2011-01-12 05:27:15	2021-10-21 00:01:16	1283	606	810	4.054455
0lgx-a1wAstiBDerGxXk2A	171	2009-01-22 15:22:01	2021-09-24 19:48:15	840	326	545	3.906433
bYENop4BuQepBjM1-Bl3fA	156	2010-08-08 22:48:41	2021-11-04 15:45:52	1243	446	865	3.769231
Um5bfs5DH6eizgjH3xZsvg	152	2011-12-25 08:29:48	2021-12-02 01:14:22	499	209	376	3.855263
wXdbkFZsfDR7utJvbWElyA	147	2016-06-11 02:18:44	2021-12-01 20:49:12	600	42	331	4.217687
-G7Zkl1wlWBBmD0KRy_sCw	146	2012-12-28 18:18:08	2022-01-10 19:02:31	2983	1593	2329	3.684932
ET8n-r7glWYqZhuR6GcdNw	146	2008-05-30 16:18:36	2021-02-26 19:52:23	1976	596	1288	4.041096
fr1Hz2acAb3OaL3l6DyKNg	145	2014-07-20 21:54:31	2022-01-16 01:05:04	1495	445	1248	3.868966
1HM81n6n4iPIFU5d2Lokhw	145	2011-08-12 14:25:27	2021-12-10 06:20:16	594	210	297	3.055172

Figure 68: Top performing yelp users

Below figure depicts the information about the distribution of sentimental trends.

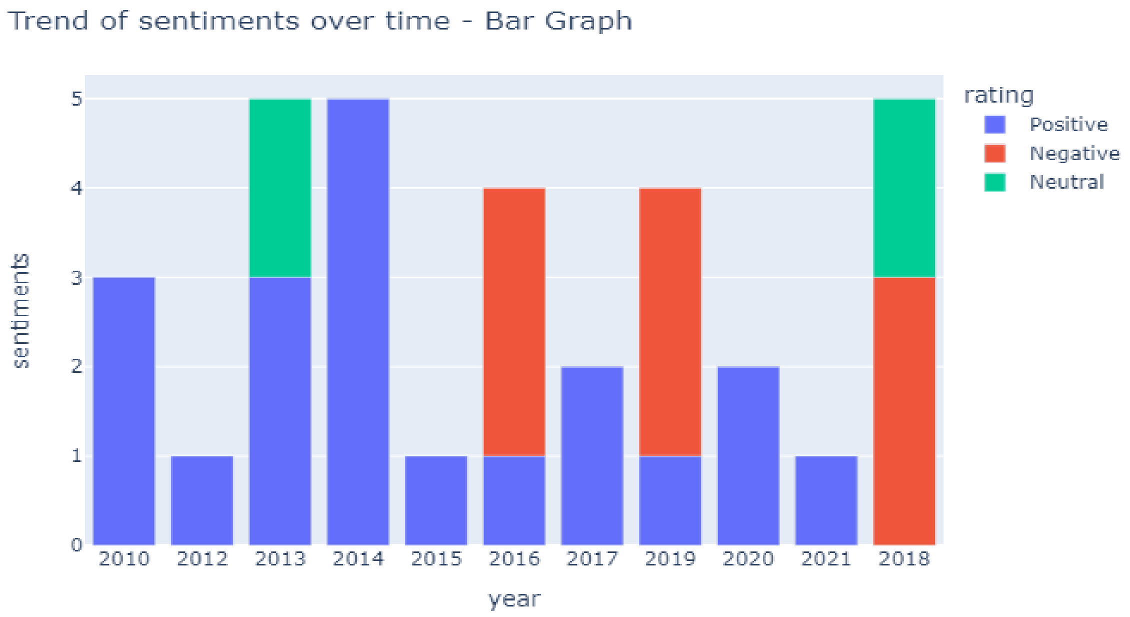


Figure 69: Sentiments distribution through bar graph

The below figure shows the information about the distribution of sentiments over years.

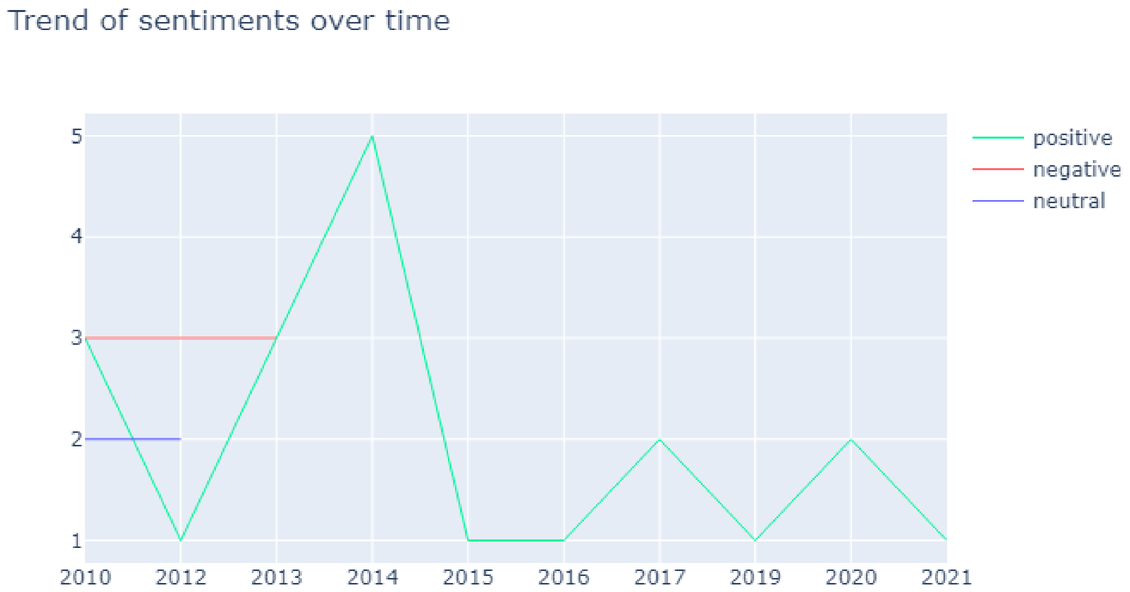


Figure 70: Trend of Sentiments over time

Below figure shows the information about the distribution of usage of yelp based on year.

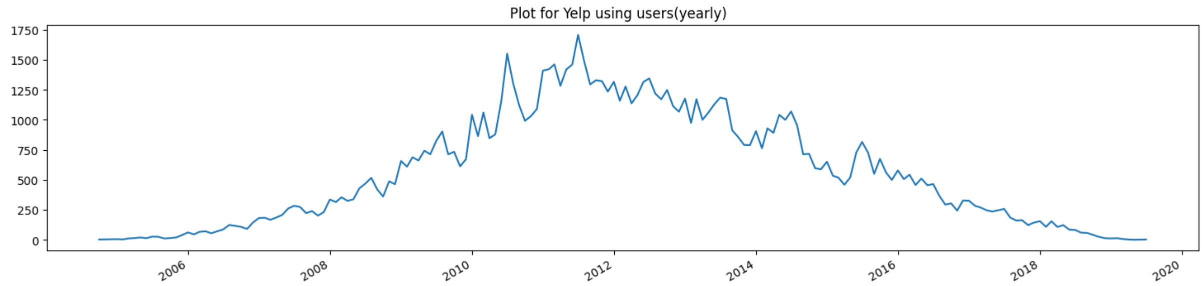


Figure 71: Yelp users yearly

The below line plot describes the usage of yelp based on the year.

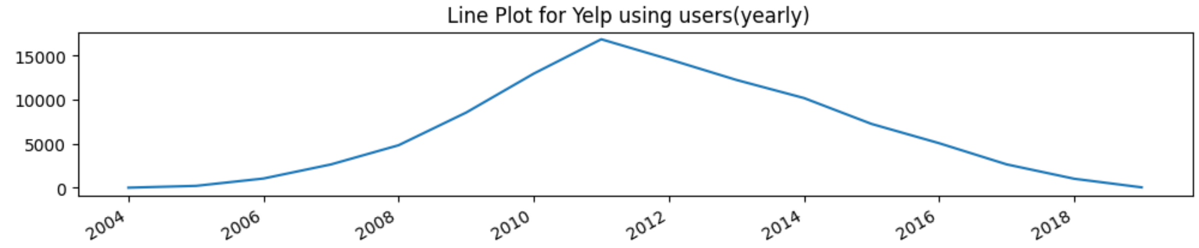


Figure 72: Line plot for yelp users yearly

Below figure shows the information about usage of yelp based on particular time period.

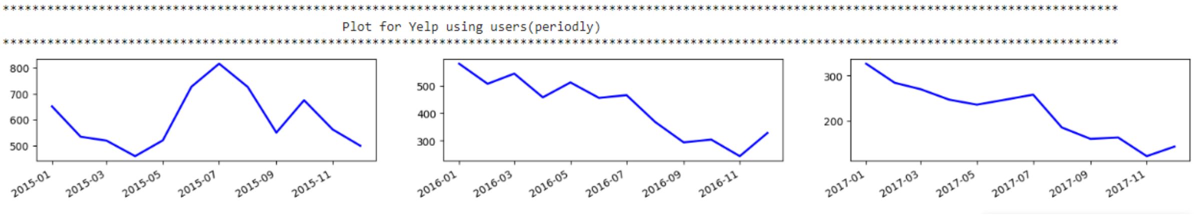
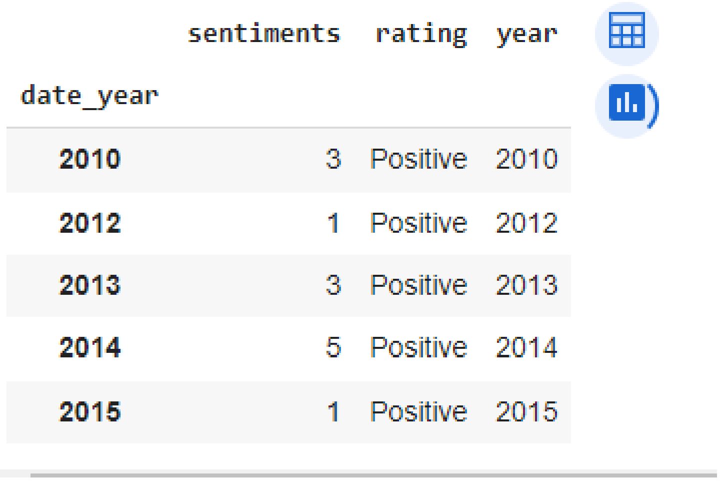


Figure 73: Yelp users on time period

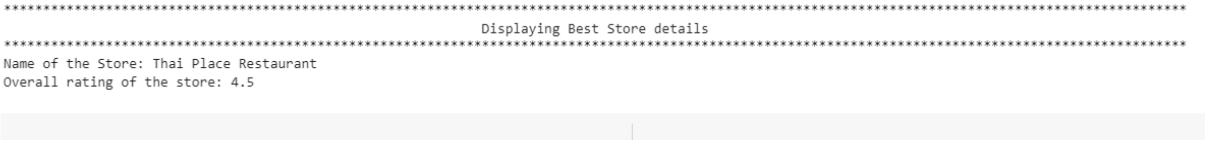
The below figure shows the rating sentiments on yearly basis.



	sentiments	rating	year
date_year			
2010	3	Positive	2010
2012	1	Positive	2012
2013	3	Positive	2013
2014	5	Positive	2014
2015	1	Positive	2015

Figure 74: Review Sentiments

The below figure shows the information about store details.



```
******************************************************************************************************************************************************
                                                                  Displaying Best Store details
******************************************************************************************************************************************************
Name of the Store: Thai Place Restaurant
Overall rating of the store: 4.5
```

Figure 75: Best store details

Below figure shows the information about distribution of ratings of each different category business.

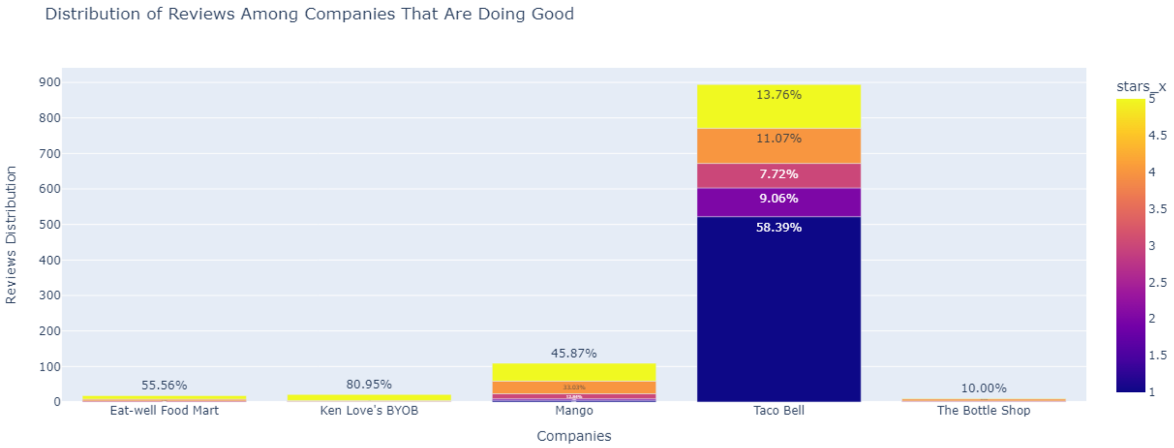
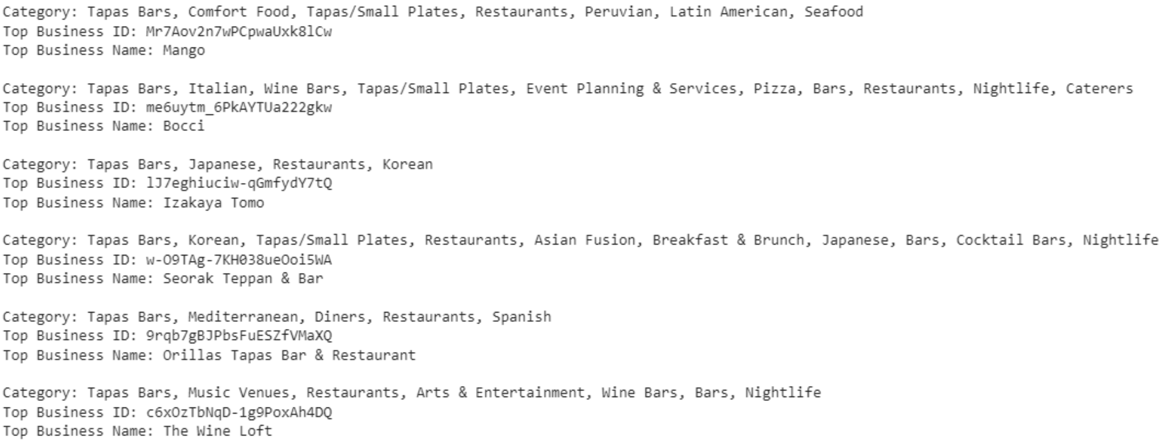


Figure 76: Top performing business of each category

The below picture depicts the information about top performing business of each category.



```
Category: Tapas Bars, Comfort Food, Tapas/Small Plates, Restaurants, Peruvian, Latin American, Seafood
Top Business ID: Mr7Aov2n7wPCpwaUxk8lCw
Top Business Name: Mango

Category: Tapas Bars, Italian, Wine Bars, Tapas/Small Plates, Event Planning & Services, Pizza, Bars, Restaurants, Nightlife, Caterers
Top Business ID: me6uytm_6PkAYTUa222gkw
Top Business Name: Bocci

Category: Tapas Bars, Japanese, Restaurants, Korean
Top Business ID: lJ7eghiuciw-qGmfydY7tQ
Top Business Name: Izakaya Tomo

Category: Tapas Bars, Korean, Tapas/Small Plates, Restaurants, Asian Fusion, Breakfast & Brunch, Japanese, Bars, Cocktail Bars, Nightlife
Top Business ID: w-O9TAg-7KH038ueOoi5WA
Top Business Name: Seorak Teppan & Bar

Category: Tapas Bars, Mediterranean, Diners, Restaurants, Spanish
Top Business ID: 9rqb7gBJPbsFuESZfVMaXQ
Top Business Name: Orillas Tapas Bar & Restaurant

Category: Tapas Bars, Music Venues, Restaurants, Arts & Entertainment, Wine Bars, Bars, Nightlife
Top Business ID: c6xOzTbNqD-1g9PoxAh4DQ
Top Business Name: The Wine Loft
```

Figure 77: Top performing business of each category

Evaluation

Evaluation of a model is necessary because it determines its performance and its effectiveness to work on new and unseen data. To test and evaluate my customer feedback analysis project, I will use different evaluation metrics to evaluate my overall aims of my project. In order to achieve Understand Customer Sentiment and Monitor Brand Reputation aim we will display it through the visualization graph. I will analyse the customer feedback of a particular company and display a proper visualization plot to display the customers sentiment towards that company. Coming to Application of different Machine Learning models to Customer Feedback Analysis I used different evaluation metrics to measure the performance of the model by calculating accuracy, precision, recall, F1 score, sentiment analysis. Regarding the next aim Improve Customer Experience and Enhance Decision Making through Visualization, is going to evaluate it by visualizing a proper visualization plot to understand about the customer feedback in terms of the customer feedback. We are going to visualize different visualization plots to evaluate about the complete analysis of the yelp business dataset to showcase various case studies of different business.

Regarding the first aim, understanding customer sentiment and monitoring brand reputation is evaluated by the customers feedback data. I had analysed the customers sentiments through sentimental analysis and categorized the review sentiments of a particular business. The reviews from the customer feedback can be used for improving the brand reputation of a company. The results are populated in the above sections.

Coming to the next aim, employing different machine learning algorithms on the same data and finding the best accurate model is evaluated by the application of six different machine learning algorithms. All the algorithms are evaluated by its accuracy and classification report. All the results are of different learning algorithms are displayed in the above sections.

Coming to the last aim, Improve Customer Experience and Enhance Decision Making through Visualization is evaluated by the visualization of different plots from the customer data. Different type s of visual representations is represented which can be easily understood by a normal person. The results of visualizations are populated in the above sections.

Each and every project aim was successfully evaluated and able to achieve all my project goals and objectives.

Related Work

Different kinds of customer feedback analysis are made on Yelp Business datasets but only a few things can be compared with other results. Some of the sentimental analysis and visualizations of data can be compared with other works. Word clouds of different users of a particular company will be almost same. Some kinds of data visualizations are almost similar because of the same data. Moreover, I tried to differentiate the results by using different plots other than the existing plots. All the related work from others which helped me to throws the light on visualization in an easy and effective manner to plot various graphs and to identify different sentiment approaches.

6. Conclusion

In conclusion, the overall outcome of the projects has been met and all the aims and objectives of the project has been evaluated with proper outcomes. Customer feedback analysis is analysed on Yelp business dataset. Yelp business dataset is best for analysing the customer reviews for a business. I The key outcomes achieved by this project are, done an extensive analysis on the dataset, applied different Statistical techniques and produced best outcomes, one of the other major outcomes of the project was employing different machine learning algorithms on the same data and finding the best machine learning algorithm. Finding the major sentiments of the customer feedback is other getting most recent reviews is other major achievement. An extensive Exploratory Data Analysis was made on the customer feedback and plotted different visualization outcomes after analysing the customer feedback.

Reflection

Customer feedback analysis is the process of collecting, examining, and deriving insights from feedback provided by customers regarding a product, service, or experience. It plays a crucial role in understanding customer opinions, improving products or services, and enhancing overall customer satisfaction. Customer feedback analysis is essential for businesses seeking to enhance customer satisfaction, identify opportunities for innovation, and address areas of improvement. By systematically analysing and acting upon customer feedback, organizations can build stronger customer relationships and drive business growth. Coming to my personal experience analysing customer feedback helps to enhance customer satisfaction, get new ideas for product development, enhance brand growth and reputation, helps in churn reduction etc. This research helps me to learn various new concepts, practical exposure of learned concepts, deeper understanding of some known concepts. It is clear from the analysis that my project has achieved its aims and objectives successfully without any deviation. Each and every aim has been successfully demonstrated and evaluated with valid outcomes. My entire project experience is discussed in the below section.

Before starting the project on Customer Feedback Analysis, I need to frame up the aims and objectives of my project. I had various things for my project aims and objectives and this create chaos what goals would be better for my project. By reading some research papers, blogs and internet articles my confusions are cleared and able to project my aims and objectives successfully. The next step was to research more customer feedback topics to know what researches are done till now. I had researched different published papers, journals, blogs etc. Finding the dataset is a major concern in this project, as effective dataset gives best results. I had found various customer feedback datasets and I found Yelp business data was best among all the other datasets. As yelp business data is readily available, it contains rich and diverse data, the data is real world data, it offers API's to access the data. All these considerations make me to choose Yelp business data for analysing the customer feedback. As the datasets are available in json format I need to convert into dataframe format for analysis. I used pyspark to covert the data into dataframe format. I don't have any knowledge about pyspark so tried to learn and used practically in my project work. The next step is to analyse the data for this I had chosen dataset analysis and statistical analysis. Analysing the dataset was bit easy as we had some direct functions to know about the dataset. Coming to statistical analysis, it took long time as I had made many numbers of statistical observations through various statistical approaches. I used box plot to find the outliers in the dataset, raincloud plot for extensive analysis, T-test, P-value, Covariance, Correlation, Principal Component Analysis between various columns of the dataset, applying normal distribution on the dataset, probability distribution of the ratings data. Some of the topics I need to put more light on those as I don't have any idea about those topics. Mathematics for Data Science and

Applications of Data Science courses in my curriculum helped me a lot to know about those topics. My job gets easier through these course knowledges. Coming to the next step, Pre-processing the data takes a lot of time for understanding the data what things need to be removed and added etc. Some of the important techniques used are removing punctuation, removing stop words, getting trending keywords and categorizing keywords etc. Some techniques I don't have any knowledge at all. So, it took time to learn these topics. The next step is to employ machine algorithms to this data as it contains a greater number of records it takes lots of time for training the records. So, I used parallel processing or batch processing techniques to complete the training in short period of time. Moreover, my project goal was to employ different machine learning algorithms on the same data and find out the efficient one. I had researched many algorithms which are efficient for analysing the customer feedback and I found six different machine learning algorithms that are good for customer feedback analysis. Logistic Regression, Multinomial Navie Bayes, Decision Tree Classifier, Random Forest Classifier, Support Vector Machines, Gradient Boosting Classifier are the six different algorithms employed on customer feedback data and I found that Support Vector Machines algorithm is performing better when compared with other algorithms. Most of the researches are analysing the customer feedback data and tries to visualize the data based on their findings. Sentimental analysis is also common in these customer feedback data but application of machine learning on these data are very rare and we less researches. My project will work a reference for other researches while employing machine learning for customer feedback data. I don't have knowledge about the Gradient Boosting algorithm so I learned in this research and applied practically. Regarding the next part, sentimental analysis on customer feedback data has progressed some good results while doing major part in sentimental analysis it takes me time to work it perfectly. The major things are getting trending keywords from customer feedback and getting most recent reviews from the customer data. All these things are researched, learned from other sources and executed perfectly. The Exploratory data analysis part is executed perfectly and processed better results when compared with other research results. The results of other researchers have been taken into reference and I had applied some new techniques to process the results. All the visualizations are in simply and easily understandable format even for a common man. So, this is my complete overview about the Customer Feedback Analysis Project.

Future Work

Future work refers to potential research of a project that can be undertaken in the future based on current findings and limitations. My research on Customer Feedback Analysis has been finished based on my project aims and objectives. Actually, analysing customer feedback on Yelp business dataset can be extended further by analysing more research on this. Some of the things that I want to implement on analysing the customer feedback are

1. Build a Tableau dashboard by finding the major observations.

2. Deploy the model as website for analysing the sentiment based on the review text.

The above-mentioned things are the major research work that I want to work in the future. Actually, I thought to implement these features in our project but as my MSc project is a time constraint, so its not possible to implement those features within short span of time. But in future I will extend this project as my personal project to achieve the above-mentioned objectives. The first objective is to build a tableau dashboard to showcase the major observations of Yelp business dataset. Through this dashboard anyone can find the important results from these observations and these findings can provide a complete picture about the Yelp business dataset. The other objective is to deploy the model as a website for the business to understand the sentiments of users based on their review text. Here I will deploy the model in a cloud environment there by accessing the model from cloud environment I will publish a website for the business to understand the sentiments of the user. In future we can extend more things to this website by getting automated results from customers feedback data. So, these are my future related work that I am going to implement on this yelp business dataset.

7. References

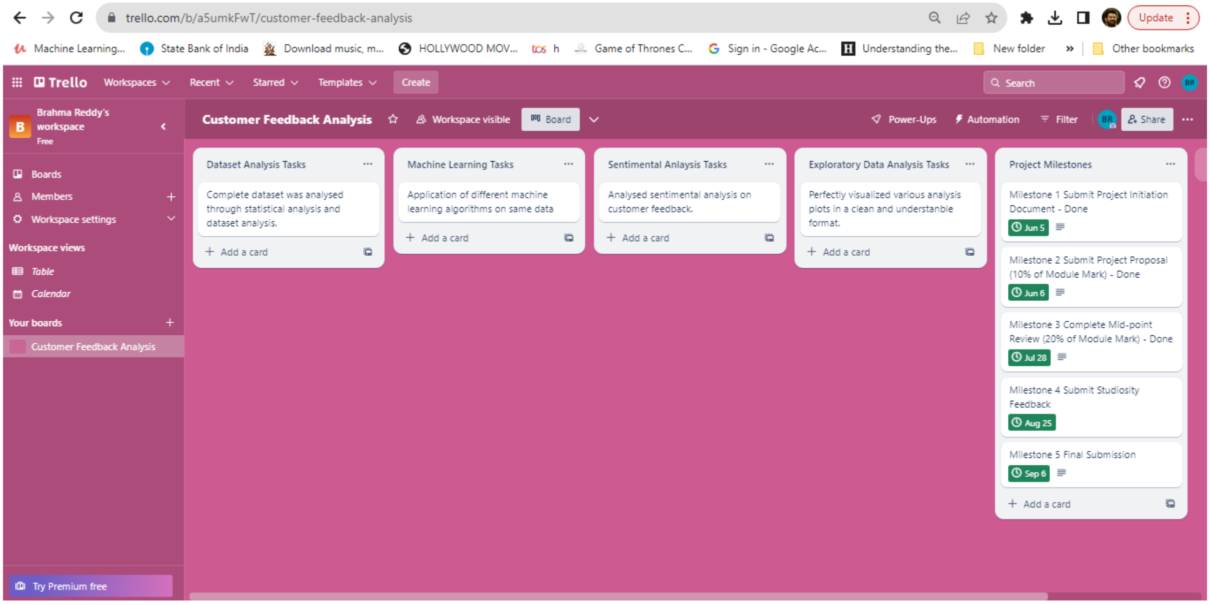
1. Akanksha Halde, Aditi Uttekar, Amit Vishwakarma 2022. SENTIMENT ANALYSIS ON AMAZON PRODUCT REVIEWS
2. Singh, U., Saraswat, A., Azad, H.K. *et al.* Towards improving e-commerce customer review analysis for sentiment detection.
3. Prof.Sonali J. Mane, Abhishek Parve, Bhavesh Patil, Akash Kamble, 2019. Customer Feedback Analysis Using Machine Learning.
4. Najma Sultana & Pintu Kumar & Monika Rani Patra & Sourabh Chandra and S.K. Safikul Alam (2019): Sentimental Analysis of product reviews.
5. Nikhita Mangaonkar and Sudarshan Sirsat, 2017, proposed a Neuro Lingusitic Programming Approach.
6. Yi, S., Liu, X. Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review.
7. Aamir Rashid and Ching-yu Huang 2021. Sentiment Analysis on Consumer Reviews of Amazon Products.
8. Kinga Edwards, "Customer feedback analysis," 22 August 2022. [Online]. Available: Customer Feedback Analysis: Steps and Use Cases (survicate.com) [Accessed 1 August 2023].
9. Federico Pascual, "Analysing customer feedback", 03 June 2019. [Online]. Available: Feedback Analysis: Know Your Customers (monkeylearn.com) [Accessed 5 August 2023].
10. Ben Goodey, "Analyse customer feedback manually". [Online]. Customer Feedback Analysis: Step-by-step + Template (sentisum.com) [Accessed 6 August 2023].
11. Miroslav Damyanov, "Analysis of customer feedback", 16 April 2023. [Online]. Available: How to Analyze Customer Feedback: Guide, Examples + Template (dovetail.com) [Accessed 4 August 2023].
12. Graham Maonaigh, "Customer Feedback strategy". [Online]. Available: 6 Keys to Getting High-Quality & Useful Customer Feedback (intercom.com) [Accessed 15 August 2023].
13. Alyona Medelyan, "Analysis of customer feedback". [Online]. Customer Feedback Analysis: How To Analyze Feedback (getthematic.com) [Accessed 10 August 2023].
14. Userpilot, "Feedback Analysis", 02 March 2023. [Online]. Available: Feedback Analysis: How To Analyze Customer Feedback? (userpilot.com) [Accessed 19 August 2023].
15. Dublin, Griffith & Joseph, Roshan. (2020). Amazon Reviews Sentiment Analysis: A Reinforcement Learning Approach. 10.13140/RG.2.2.31842.35523.
16. Dey, Sanjay and Wasif, Sarhan and Tonmoy, Dhiman and Sultana, Subrina and Sarkar, Jayjeet and Dey, Monisha(February 2020) A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews.
17. Haque, Tanjim & Saber, Nudrat & Shah, Faisal. (2018). Sentiment analysis on large scale Amazon product reviews.10.1109/ICIRD.2018.8376299.
18. Dave, K., Lawrence, S., and Pennock, D., 2017. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. WWW'03.
19. Bhumika, Prof Sukhjit Singh Sehra and Prof Anand Nayyar. A Review Paper On Algorithms Used For Text Classification- (Ijaiem)-2017.
20. Kowsari, Kamran, et al. Hdltex: Hierarchical deep learning for text classification. 2017 16th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2017.
21. Gaye, B., Zhang, D. & Wulamu, A. Sentiment classification for employees reviews using regression vector-stochastic gradient descent classifier (RV-SGDC). *PeerJ Computer Science* **7**, e712 (2021).

22. Bouazizi M, Ohtsuki T (2017) A pattern-based approach for multi-class sentiment analysis in twitter. IEEE Access 5:20617–20639.

8. Appendices

Appendix A: Customer Feedback Analysis project proposal form - https://drive.google.com/file/d/1zKTdBLJBNP8wb2QuWugDvjrtksNcuiiO/view?usp=drive_link

Appendix B: Trello Project management tool link - https://trello.com/b/a5umkFwT/customer-feedback-analysis



Customer Feedback Analysis

Dataset Analysis Tasks
- Complete dataset was analysed through statistical analysis and dataset analysis.
- + Add a card

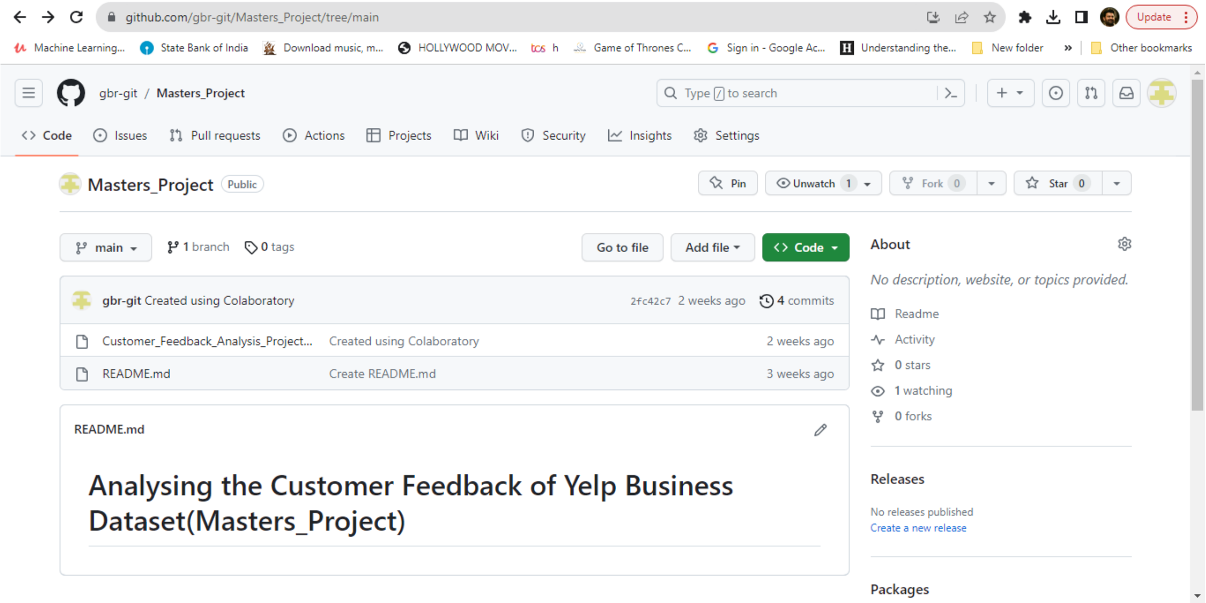
Machine Learning Tasks
- Application of different machine learning algorithms on same data
- + Add a card

Sentimental Anlaysis Tasks
- Analysed sentimental analysis on customer feedback.
- + Add a card

Exploratory Data Analysis Tasks
- Perfectly visualized various analysis plots in a clean and understanble format.
- + Add a card

Project Milestones
- Milestone 1 Submit Project Initiation Document - Done (Jun 5)
- Milestone 2 Submit Project Proposal (10% of Module Mark) - Done (Jun 6)
- Milestone 3 Complete Mid-point Review (20% of Module Mark) - Done (Jul 28)
- Milestone 4 Submit Studiosity Feedback (Aug 25)
- Milestone 5 Final Submission (Sep 6)
- + Add a card

Appendix C: GitHub repository link for accessing the technical code details – https://github.com/gbr-git/Masters_Project/blob/main/Customer_Feedback_Analysis_Project.ipynb



gbr-git Created using Colaboratory	2fc42c7 2 weeks ago	4 commits
Customer_Feedback_Analysis_Project...	Created using Colaboratory	2 weeks ago
README.md	Create README.md	3 weeks ago

README.md

Analysing the Customer Feedback of Yelp Business Dataset(Masters_Project)

About

No description, website, or topics provided.

- Readme
- Activity
- 0 stars
- 1 watching
- 0 forks

Releases

No releases published
Create a new release

Packages