MC DONALD’S CUSTOMER REVIEW ANALYSIS

# 1. ABSTRACT:

This study uses machine learning techniques to analyze McDonald's customer happiness. With sentiment analysis and predictive modeling, we look into what affects satisfaction based on online reviews. Two examples of preliminary data preprocessing are feature extraction and text normalization. Predictive modeling identifies important data such as store details and time-related components, while sentiment analysis offers insights into customer attitudes. Accuracy, F1 score, and MSE are some of the performance evaluation measures used to validate the effectiveness of the model. More machine learning techniques and sensitivity analysis are considered during subsequent research to enhance forecasting abilities.   
  
I hope that this study project will provide insightful information to McDonald's stakeholders, such as management, marketing departments, and customer service representatives. I aim to offer practical suggestions for improving customer happiness and streamlining operational methods by utilizing machine learning approaches. In the end, this research adds to the larger knowledge of customer sentiment analysis in the fast-food sector and emphasizes the significance of data-driven decision-making for enhancing the general customer experience.

## Table Of Contents:

[1. ABSTRACT 1](#_Toc166269026)

[2. INTRODUCTION WITH RESEARCH QUESTION: 3](#_Toc166269028)

[2.1 Initial Research Question: 3](#_Toc166269029)

[2.2 Dataset Attributes 3](#_Toc166269030)

[2.3 Final Research Question 4](#_Toc166269031)

[2.4 Overview of Analysis and Objectives 4](#_Toc166269032)

[3. PLAN FOR ANALYZING DATA AND EVALUATION 5](#_Toc166269033)

[4. EXPLORATORY DATA ANALYSIS (EDA) 6](#_Toc166269034)

[4.1 Opening and Reading Data 7](#_Toc166269035)

[4.2 Data Cleaning and Preprocessing 7](#_Toc166269036)

[4.3 Feature Engineering: 8](#_Toc166269037)

[4.4 Summary Statistics: 11](#_Toc166269038)

[4.5 Visualizations Of EDA: 12](#_Toc166269039)

[4.6 Sentiment Analysis: 15](#_Toc166269040)

[4.7 Statistical Analysis: 20](#_Toc166269041)

[4.8 Performance Metrics: 20](#_Toc166269042)

[5. FINAL ANALYSIS AND CONCLUSION 22](#_Toc166269043)

[6. REFERENCES 24](#_Toc166269044)

# 2. INTRODUCTION WITH RESEARCH QUESTION:

In a highly competitive fast-food business, keeping customers happy is the most important thing that can be done. As a worldwide leader in fast food, **McDonald's** knows how important it is to understand and improve customer happiness. To do this, this analysis study will use a large set of data from online reviews to investigate the factors that affect how happy McDonald's customers are. The goal is to find ideas that can be turned into strategies that can be used to improve operational processes, make the customer experience better, and make **McDonald's** more competitive.

## **2.1 Initial Research Question:**

**Research Question: "**What factors influence customer satisfaction in McDonald's based on online reviews, and how can these insights be leveraged to enhance the overall dining experience?**"**

The study question seeks to apply the potential of data science and machine learning to address important business difficulties for McDonald's. A thorough dataset is built by examining the elements that influence customer satisfaction through online evaluations, which includes both quantitative ratings and qualitative insights. Data science approaches like statistical analysis and predictive modeling can be used to find trends, correlations, and important drivers of happiness. Machine learning algorithms can predict client preferences and behaviors, allowing for tailored recommendations and focused improvement strategies. The dataset's rich textual content enables sentiment analysis, providing a deep knowledge of client sentiments. Finally, the study question serves as a platform for adopting data-driven solutions that optimize operational processes, improve customer experiences, and boost McDonald's competitiveness in the fast-food business.

## **2.2 Dataset Attributes:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Id** | **Column Name** | **Data Type** | **Description** |
| 1 | reviewer\_id | int | Number for each reviewer |
| 2 | store\_name | String | Name of the McDonald's store |
| 3 | category | String | Category or type of the store |
| 4 | store\_address | String | Address of the store |
| 5 | latitude | Float | Latitude coordinate of the store's location |
| 6 | longitude | Float | Longitude coordinate of the store's location |
| 7 | rating\_count | Integer | Number of ratings/reviews for the store |
| 8 | review\_time | Timestamp | Timestamp of the review |
| 9 | review | Text | Textual content of the review |
| 10 | rating | Float | Rating provided by the reviewer |

## **2.3 Final Research Question:**

**Final Research Question:** "Through the application of machine learning techniques, specifically utilizing predictive modeling and sentiment analysis, what specific attributes—including store details (store\_name, category, store\_address, latitude, longitude), temporal aspects (review\_time), and textual content (review)—significantly influence customer satisfaction, as measured by the rating in online reviews, for McDonald's stores?"

**Evolution of Research Question:** The final research question evolved from the initial research question by incorporating specific methodologies and variables based on the analysis conducted. While the initial research question was broader and focused on identifying factors influencing customer satisfaction, the final question became more precise, outlining the utilization of machine learning techniques and specifying the attributes of interest, such as store details, temporal aspects, and textual content. This evolution reflects a refinement and narrowing down of the research scope to facilitate a more focused and actionable analysis.

## **2.4 Overview of Analysis and Objectives:**

**1. The purpose:** The purpose of the research question is to find out what makes McDonald's customers happy by reading reviews online and using that information to make the whole eating experience better. The goal of the analysis report is to give McDonald's suggestions for what they can do after a full analysis of the dataset that includes both quantitative and qualitative factors.

**2. Type of Report on Analysis:** Full Report on Customer Satisfaction Analysis: The report's goal is to look at the dataset in depth by mixing statistical, machine learning, and sentiment analysis methods to learn more about what makes customers happy.  
  
**3. Target Audiences:** McDonald's managers, including the marketing, operations, and customer experience teams. Data analysts, industry researchers, and other people with an interest in the fast-food business and customer satisfaction trends are the secondary audience.

**4. Uses of the Analysis Report:** Strategic Decision-Making: Help with the processes of making choices about how to improve operations, create marketing plans, and make the customer experience better. Operational Optimization: Figure out what can be done to make certain things better, like the quality of the food, the speed of service, the cleanliness, and the value for the money.   
Competitive Edge: Make McDonald's more competitive in the fast-food business by fixing problems with customer happiness.

# 3. Plan for analyzing data and evaluation:

**1. Opening and Reading Data:** In this stage, the dataset is loaded into memory, and its contents are read. For our research, the McDonald's review dataset must be read from a CSV file using pandas.  
  
**2. Data Cleaning and Preprocessing:** In data cleaning, duplicates are eliminated, missing values are located and handled, and dataset discrepancies are addressed. In order to get the data ready for analysis, our project involves steps like eliminating stopwords, lowercasing text, and deleting whitespace.  
  
**3. Feature Engineering:** Feature engineering is the process of adding new features or altering current ones in order to increase the model's capacity for prediction. Features like review time, store location information, and numerical ratings are among the features that will be extracted for our project.  
  
**4. Summary Statistics:** Measures of central tendency and dispersion, among other numerical aspects of the dataset, are summarized in summary statistics. This part of our study entails computing summary statistics for numerical variables such as ratings.  
  
**5. Visualizations:** Understanding the distribution of data, trends, and correlations between variables is aided by visualizations. In our project, we examine the dataset and find insights using visualizations like word clouds, box plots, and histograms.  
  
**6. Sentiment Analysis:** Sentiment analysis is the process of identifying the sentiment—whether positive, negative, or neutral—expressed in textual data. Our project involves sentiment analysis of McDonald's reviews utilizing natural language processing techniques.  
  
**7. Statistical Analysis:** Using statistical techniques, statistical analysis investigates links and patterns within the data. Our project uses methods to comprehend the link between variables, such as correlation analysis.  
  
**8.Evaluation of Performance Metrics:** This process evaluates the effectiveness of predictive models by calculating metrics such as F1 score, accuracy, precision, recall, and Mean Squared Error (MSE). As part of our project, we'll be measuring the efficacy and correctness of sentiment analysis models through performance evaluation.

# 4. Exploratory Data Analysis (EDA)

**Dataset Purpose:** This information is meant to be carefully analyzed so that we can get a full picture of the things that affect McDonald's customers' happiness. The dataset allows for a thorough study of the dining experience because it includes both numerical ratings and qualitative comments.

The dataset has **33396** records. I have performed the following steps to my dataset as part of Exploratory Data Analysis (EDA)

1. Opening and Reading Data
2. Data Cleaning and Preprocessing
3. Feature Engineering
4. Summary Statistics
5. Visualizations
6. Sentiment Analysis
7. Statistical Analysis
8. Performance Metrics Evaluation

## **Opening and Reading Data:**

Using the Pandas library, the McDonald's Reviews dataset is loaded into a DataFrame named df. The dataset is read from a CSV file named "McDonald\_s\_Reviews.csv", with the encoding specified as 'latin1' to handle any special characters present in the data.

A screenshot of a computer

Description automatically generated

## **Data Cleaning and Preprocessing:**

* **Handling Missing Values:** The code first checks for missing values in the DataFrame ‘df’ using the ‘isnull()’ method, which returns a Series with the count of missing values for each column. Then, the ‘dropna()’ method is applied to remove rows with missing values, effectively handling them by eliminating any observations with NaN values. After this operation, the DataFrame ‘df’ is updated to exclude rows containing missing values, ensuring that the dataset is free from incomplete entries. This approach helps maintain data integrity and avoids potential issues during analysis caused by missing or incomplete data. Finally, the variable ‘missing\_values’ stores the count of missing values for each column, providing insight into the extent of missing data in the original dataset.

A screenshot of a computer

Description automatically generated

* **Standardizing Text Data:** To improve analysis accuracy and guarantee uniformity, the text data is standardized. To standardize the format of the text throughout the dataset, it is first transformed to lowercase. After that, any extraneous characters are eliminated from the text by removing the leading and trailing whitespaces. In order to concentrate on the textual information that is pertinent to the study, numbers are also eliminated from the text. In order to extract relevant information for sentiment analysis and other text-based activities, stopwords—common words that might not add value to the analysis—are finally eliminated from the text. The text data is summarized and made ready for additional analysis and modeling with the aid of these preparation procedures.

**A screenshot of a computer program

Description automatically generated**

## **Feature Engineering:**

* During the feature engineering process, we initiated by duplicating the original dataset to carry out modifications without modifying the original data. Consequently, we eliminated extraneous columns such as 'store\_name', 'category', 'latitude', 'longitude', and 'rating\_count' in order to concentrate on crucial attributes for our research. This phase assured that our dataset exclusively consisted of pertinent information directly relevant to our research subject.   
    
  Subsequently, we created supplementary attributes from the available data, specifically the 'City' and 'State' fields obtained from the 'store\_address' attribute. This enabled us to investigate geographical patterns in customer satisfaction.

A screenshot of a computer

Description automatically generated

* In addition, we conducted an analysis of the 'review\_time' column to examine temporal patterns in customer feedback. This allowed us to gain insights into any recurring patterns or trends that may impact customer satisfaction. Additionally, we converted the 'rating' column into a more comprehensible format by extracting the star ratings as a distinct attribute called 'Star', which simplifies the analysis and understanding of customer ratings. By performing feature engineering, we were able to enhance our dataset by incorporating significant information, making it ready for further modeling and analysis.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* In feature engineering, we implemented a function to clean text data by converting it to lowercase, removing non-alphabetic characters, and stripping extra whitespace. We then applied this function to the 'review' column to generate 'clean\_reviews', eliminating stopwords and ensuring consistency for sentiment analysis.

A screenshot of a computer program

Description automatically generated

* We performed a thorough study to evaluate the distinct values in specific columns and gain insights into the variety and spread of data across important characteristics. We calculated the total number of distinct values for each column, including 'City', 'State', 'review\_time', and 'Star'. This analysis gives us insights into the level of variability in the dataset. Furthermore, through the analysis of the highest values and their corresponding frequencies, we obtained more detailed information on the significance of particular categories or timestamps. This allowed us to develop a more sophisticated comprehension of the dataset's structure and spread.

A screenshot of a computer code

Description automatically generated

## **4.4 Summary Statistics:**

Summary statistics offer a concise overview of important attributes within the dataset, facilitating comprehension of its distribution and variability. Descriptive statistics, such as the mean, standard deviation, and quartile ranges, provide information on the average value, variability, and spread of numerical variables. Categorical variables are summarized by calculating frequency counts, which provide information on the prevalence of each category and their distribution in relation to each other. Temporal variables are displayed as a range, highlighting the earliest and latest timestamps to encompass the temporal extent of the dataset.

A screenshot of a computer

Description automatically generated

## **Visualizations Of EDA:**

* Visualizations are excellent resources for identifying and deciphering patterns in data. Graphical representations facilitate the easy interpretation and communication of intricate linkages and trends. Diverse kinds of visualizations have been used in the offered code snippets to examine unique features of the dataset. Pie charts and bar plots, which are examples of **univariate analysis**, show the distribution and frequency of categorical data like star ratings, states, and cities. These graphics provide information about the general distribution and frequency of categories in the dataset.

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

A graph of different colored bars

Description automatically generated with medium confidence

* In contrast, **bivariate analysis** examines the connections between two variables, like star ratings and geographic locations (cities, states). Trends and differences in consumer satisfaction can be found by displaying the distribution of star ratings among various states and cities. The visualizations are more readable and clearer when color schemes and ordering strategies are used, which makes it easier for viewers to see patterns. Given the circumstances, these graphics offer a thorough summary of the major variables impacting customer happiness and point out areas that require more research and study.

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

## **Sentiment Analysis:**

Sentiment analysis uses natural language processing (NLP) techniques to interpret textual reviews, which is a critical component in understanding customer feedback. Evaluation of each review's sentiment through the use of NLP makes it possible to derive sentiment scores or categories that offer information about the reviews' overall sentiment polarity (positive, negative, or neutral).   
  
A stakeholder's ability to spot trends and patterns in customer sentiment can be utilized to target interventions and make improvements in areas where customer satisfaction may be low by visualizing the sentiment distribution across stores and categories. Furthermore, by making it easier to identify words or phrases linked to both positive and negative attitudes, sentiment analysis offers useful information about the elements influencing consumer satisfaction and discontent. Businesses can customize their approaches to answer customer issues and improve the customer experience by identifying particular linguistic trends.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**4.6.1 Visualizing Sentiment Distribution:**

The distribution of sentiments gleaned from customer evaluations is visualized as part of the sentiment analysis process. Sentiments like positive, negative, and neutral are represented using a pie chart, with each color designated for clarity. A thorough grasp of the sentiment distribution is made possible by the size of each segment, which reflects the percentage of reviews that exhibit that particular sentiment.  
  
Sentiment percentages are shown in each segment of this graphic, which adds more context regarding the frequency of each sentiment category. The ability to highlight particular sentiments is made possible by the inclusion of explode parameters, which improves the visualization's ability to identify sentiment trends. In general, this procedure helps decision-making processes targeted at enhancing customer happiness and experience by providing stakeholders with insights into the overall sentiment distribution.

A screenshot of a computer

Description automatically generated

**4.6.2 Common Words in Text Data:** Using text data taken from customer evaluations, the snippet creates a word cloud visualization that shows frequently occurring words. Larger words indicate higher occurrence, and word sizes are proportionate to word frequencies. With the help of this graphic breakdown of the most often used terms in the reviews, users can gain insight into the themes or subjects that consumers find most interesting. In order to improve customer happiness, WordCloud aids in identifying important facets of consumer attitude and experiences, directing subsequent analysis and decision-making procedures.

**A screenshot of a computer

Description automatically generated**

**4.6.3 Words in Positive Sentiment:** With word splitting and counting, the system finds the most frequent terms in reviews that are linked to good emotion. The process generates a DataFrame with a green color gradient for emphasis, showing the frequencies of the top ten commonly used positive sentiment terms.

**A screenshot of a computer program

Description automatically generated**

**4.6.4 Words in Neutral Sentiment:** The following snippet extracts the top 10 most common words from reviews categorized as neutral sentiment and presents them along with their frequencies in a DataFrame.

A computer screen shot of a computer screen

Description automatically generated

**4.6.5 Words in Negative Sentiment:** The sample finds the 10 most often occurring terms in reviews that are labeled as having a negative sentiment and shows them in a DataFrame with their frequencies applied.

A white rectangular object with text

Description automatically generated with medium confidence

## **4.7 Statistical Analysis:**

Correlation analysis is used in this process to find correlations between the dataset's numerical variables. To compute the correlation matrix, it first chooses just the Data Frame’s numerical columns. The direction and intensity of correlations between pairs of variables are quantified by the correlation matrix. Ultimately, seaborn is used to create a heatmap visualization that visually represents the correlations, making it simpler to understand and recognize important linkages.

A screenshot of a computer program

Description automatically generated

## **Performance Metrics:**

To assess the success of the machine learning model in forecasting customer happiness for McDonald's locations, lets summarize the results of performance measures and metrics:

**Support Vector Machines (SVM)** is a strong supervised learning technique that is widely used for classification applications, including sentiment analysis. Here is analysis of SVM in the context of sentiment analysis:

Let's interpret each evaluation metric:  
  
**4.8.1. Mean Squared Error (MSE):** A value of 10.48 represents the average squared difference between real sentiment labels and anticipated sentiment scores. A lower MSE suggests greater model performance, however the absolute number must be interpreted in relation to the sentiment score scale.

A screenshot of a computer code

Description automatically generated

**4.8.2. Accuracy:** With a score of 99.17%, the model accurately predicts sentiment for nearly all reviews in the dataset. It measures the overall accuracy of the model's predictions.  
A screenshot of a computer code

Description automatically generated  
**4.8.3. Precision:** Precision is the percentage of accurately detected positive sentiment predictions among all positive forecasts. A Precision of 99.18% shows that the model is 99.18% correct when predicting a good review.  
  
**4.8.4. Recall:** Recall represents the fraction of correctly detected positive sentiment predictions out of all actual positive sentiments in the dataset. A recall of 99.17% indicates that the model accurately detects 99.17% of the positive feelings in the dataset.  
  
**4.8.5. F1 Score:** The F1 score balances precision and recall by taking the harmonic mean of both. A high F1 score (99.16%) suggests that the model performs well in both precision and recall.

**4.8.6. Cross-validation:**

A screenshot of a computer code

Description automatically generated  
Cross-validation scores show the model's performance across subsets of data. The average cross-validation score of 99.99% indicates that the model generalizes effectively to new data, since it performs consistently across different folds.

# 5. Final Analysis AND conclusion

Throughout this investigation, several significant revelations regarding the dataset and its consequences were made:  
  
**5.1. Understanding Data:**

* The dataset offers a thorough understanding of consumer attitude toward McDonald's restaurants, including store specifics, time-related features, and text from online reviews.
* We found patterns, distributions, and correlations in the data using exploratory data analysis (EDA), which laid the groundwork for further study.

**5.2. Addressing Research Questions:**

* We used machine learning techniques, including sentiment analysis and predictive modeling, to address the research question on factors influencing consumer happiness at McDonald's.
* We were able to deliver meaningful insights into customer sentiment by training models on features that were retrieved from the dataset and assessing their performance using metrics like accuracy, precision, recall, and mean squared error (MSE).

**5.3. Justification of Answers:**

* The sentiment analysis model's strong recall, accuracy, and precision scores support its efficacy in predicting customer satisfaction levels derived from online reviews.
* The model's performance was further validated by the mean squared error (MSE), which offered a quantifiable measure of prediction accuracy.
* Cross-validation scores confirmed the model's dependability across various data subsets by demonstrating its resilience and generalizability.

**5.4. Validation of Metrics:**

* A baseline evaluation of the model's performance was given by the first validation of metrics, which included accuracy, precision, recall, and MSE.
* These measures functioned as standards for assessing the sentiment analysis model's efficacy and pinpointing areas in need of development.

**5.5. Additional Machine Learning Algorithms:**

* Although the present study utilized Support Vector Machines (SVM) for predictive modeling, investigating alternative algorithms such as neural networks, decision trees, or ensemble techniques may provide different perspectives.
* Through evaluating various algorithms' performances and determining which ones work best with the dataset, we can improve our strategy and even find previously unnoticed patterns or relationships.

**In conclusion,** this investigation sheds light on the factors that influence McDonald's customer happiness while also demonstrating how well machine learning techniques work to extract useful information from internet reviews. Through the application of sentiment analysis and predictive modeling, we have determined the salient variables that impact customer sentiment, such as temporal aspects and store-specific characteristics. The effectiveness of these methods highlights how data-driven strategies may support strategic decision-making and enhance the fast-food industry's customer experience.   
  
Going forward, additional investigation into different machine learning techniques and performance metrics validation will help to improve the analysis's robustness and applicability. By continuously improving their predictive models and doing sensitivity analyses, McDonald's stakeholders will be able to successfully adjust to changing consumer preferences and market circumstances. In the end, this research emphasizes how critical it is to use sophisticated analytics to generate business results and how valuable data-driven insights are for enhancing operational strategies and raising customer happiness.   
  
McDonald's may obtain deeper insights into consumer preferences and behaviors by using machine learning approaches into sentiment analysis. This will enable the company to develop more targeted and efficacious strategies for augmenting client pleasure and loyalty. Adopting data-driven strategies will be essential for remaining competitive and satisfying the market's ever-evolving expectations as customer tastes continue to change.

# 6. References

<https://scikit-learn.org/stable/>

<https://www.nltk.org/>

<https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e>

<https://plotly.com/python/>

<https://github.com/MohamedhanySakr/DataScience-CheatSheets>

<https://faker.readthedocs.io/en/master/>

<https://scikit-learn.org/stable/modules/svm.html>

<https://scikit-learn.org/stable/modules/model_evaluation.html>

<https://scikit-learn.org/stable/modules/cross_validation.html>

<https://machinelearningmastery.com/k-fold-cross-validation/>