MILESTONE 2: A CASE STUDY ON MCDONALD'S CUSTOMER REVIEWS

Name: ManikantaReddy Annem
Student Id: @01450370

INTRODUCTION AND BACKGROUND

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.

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.

Plan for analyzing data and evaluation:

1. Data Cleaning and Preprocessing: To ensure a clean dataset for analysis, eliminate duplicates, manage absent values, standardize data, and convert timestamps.

2. Exploratory Data Analysis (EDA): Derive meaningful insights by visualizing rating and review distributions, investigating variable correlations, analyzing sentiment trends, and exploring patterns in rating_count.

3. Sentiment Analysis: Employ NLP to extract sentiment, visualize the distribution of sentiment, and identify significant positive and negative phrases present in customer reviews.

4. Statistical Analysis: Perform correlation and significance tests, investigate potential variations according to store categories or locations, and evaluate the influence of rating count on the overall satisfaction of the customers.

5. Predictive Modeling: Divide data into training and testing sets, choose machine learning algorithms to assess prediction accuracy, train models, assess performance, and identify influential features to ensure accurate predictions.

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

RESEARCH QUESTION:

Fine-Tuned Research Question: “To what extent do specific attributes such as " store details (store_name, category, store_address, latitude, longitude), temporal aspects (review_time), and textual content (review) influence customer satisfaction, as measured by the " rating " in online reviews, for McDonald's stores?”

Key Components:
Independent Variables (Predictors):

- store_name
- category
- store_address
- latitude
- longitude
- rating_count
- review_time
- review

Dependent Variable (Response Variable):

- rating

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.

Data preparation for analysis:

- **Handling Missing Values:** Find and rate the McDonald's dataset's missing values in each property.
  Choose the best way to deal with missing values by thinking about how they will affect the study.

  For example, if "Location" has values that are missing, you could choose to fill them in with the mode (most common value) or get rid of the rows that have missing "Location" values.

- **Data Types and Conversion:** Look at the data types of each property in the McDonald's dataset.
  For temporal analysis, change "Date" to "datetime" format so that it is shown in a normal way that computers can read.
  To make sure the data is shown correctly, check to see if any other traits need to be converted, such as numerical ratings.

- **Encoding categorical variables:** Look through the McDonald's dataset for categorical factors like "Location." Pick an encoding method, like label encoding or one-hot encoding, to change categorical variables into a style that machine learning algorithms can understand.
  As a result of this step, categorical data is properly included in the study.

- **Feature Engineering:**
  Based on the specific goals of the analysis, look into feature engineering possibilities.
  By adding new features, you can get more information. For instance, getting the day of the week from the "Date" attribute can add a temporal layer to the study.
  The goal of feature engineering is to add useful information to a dataset so that it can be used in a more complete study.

EXPLORATORY DATA ANALYSIS (EDA):

This EDA section we can find in the following link.
https://github.com/manikantareddy12/McDonalds-Review-System contains two files:
1. Mc Donalds Review EDA.ipynb (notebook)
2. McDonald_s_Reviews.csv (dataset)

CORRELATIONS

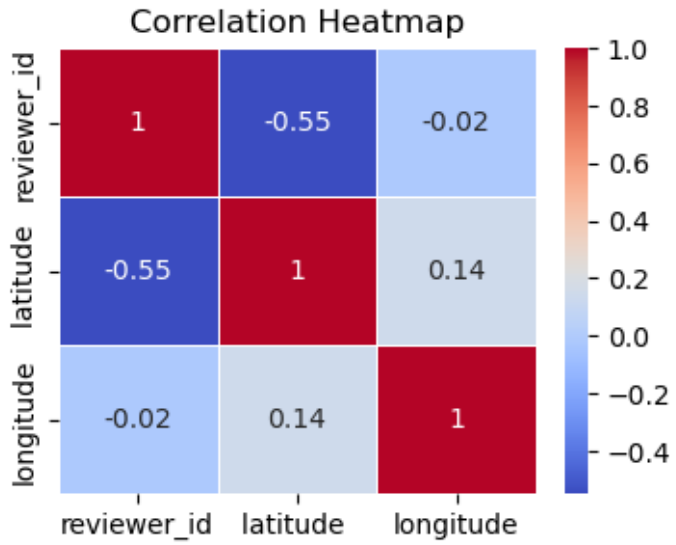
Correlation Heatmap for the dataset:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Assuming 'df' is your DataFrame
# Drop non-numeric columns for correlation analysis
numeric_df = df.select_dtypes(include=['float64', 'int64'])

# Calculate the correlation matrix
correlation_matrix = numeric_df.corr()

# Create a heatmap using seaborn
plt.figure(figsize=(4, 3))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```



DATA UNDERSTANDING:

A synopsis of the insights gained from the exploratory data analysis (EDA):

1. Data Quality:
When values were absent, rows were discarded.
Standardizing the text data in the'review' column involved converting it to lowercase, eliminating stopwords, and removing numerals.

2. Feature Engineering:
Additional columns labeled 'City', 'State', and 'Star' were generated in accordance with the'store_address' and 'rating' parameters.
Additional filtration was performed on the textual evaluations, and sentiment scores were generated utilizing the SentimentIntensityAnalyzer.

3. Summary Statistics:
The description of numerical variables such as 'Star' offered valuable insights into their distribution. 'Category' and similar categorical variables were evaluated for their value counts.

4. Univariate Analysis:
The pie chart illustrated the spatial arrangement of 'Star' ratings.

5.Sentiment Analysis:
The sentiment labels ('positive', 'negative', 'neutral') were designated in accordance with the sentiment scores that were derived from the compound scores.

6. Data Information:
The dataset, after preprocessing, included columns like 'Star', 'clean_reviews', 'sentiment_score', and 'sentiment'.
Information on data types and non-null counts was provided.

7.Geographical Particulars:
The 'City' and 'State' columns displayed distinct values along with their corresponding counts.

8. Rating Distribution:
The pie chart illustrated the dispersion of 'Star' ratings, providing valuable insights into the patterns of consumer sentiment.

MEASURABLE METRICS AND ML ALGORITHMS/UTILITIES

1. Measurable Metrics for Evaluation:

The subsequent Measurable Metrics for Evaluation will be implemented on the dataset.

a. Accuracy: Determines the proportion of sentiments that were accurately classified.

b. Precision, Recall, and F1 Score:
Precision: The ratio of accurately predicted positive sentiments to the total number of positive sentiments predicted.
Recall, or sensitivity, is the percentage of true positive sentiments that were accurately predicted.
Equalizing false positives and false negatives, the F1 Score is the harmonic mean of recall and precision.

c. Cross-validation: Cross-validates the model's generalizability by dividing the dataset into numerous subsets for training and testing.

d. Sentiment Score: This parameter specifies the sentiment score that is compounded using the SentimentIntensityAnalyzer.

The performance of the sentiment analysis model is comprehensively assessed through the utilization of various metrics. These metrics include accuracy, precision, recall, F1 score, cross-validation for generalization, and sentiment score, which assesses the model's correspondence with sentiment patterns identified in exploratory data analysis (EDA).

2. ML Algorithms/Utilities:

Support Vector Machine (SVM): The EDA classifies sentiments utilizing the Support Vector Machine (SVM) algorithm.
Supervised learning algorithms, such as SVM, are employed for classification tasks. It exhibits strong performance with linear and non-linear data and is widely employed in the classification of texts, including sentiment analysis.

Software Implementations:
1. Python and Pandas: The code is executed in the Python programming language, employing libraries like Pandas to facilitate the manipulation and analysis of data.
Python is widely utilized for machine learning and data analysis duties due to its extensive ecosystem of libraries catering to these domains.

2. Scikit-Learn: also known as Sklearn, is a Python library designed for machine learning. It offers instruments for constructing and assessing machine learning models.
The code incorporates various machine learning-related functionalities, including cross-validation and the SVM classifier, by utilizing Scikit-Learn.

3. Text Processing and the Natural Language Toolkit (NLTK): The NLTK is employed to execute text processing operations, including stopword removal and tokenization.
The NLTK library is a robust natural language processing tool implemented in Python.

4.Matplotlib and Seaborn: Matplotlib and Seaborn are used for data visualization. Matplotlib is a popular plotting library, while Seaborn provides a high-level interface for statistical graphics.
The code uses these libraries to create various plots for exploratory data analysis (EDA).

5.Plotly: Plotly facilitates the development of dynamic and interactive visualizations.

The code creates a bar chart for frequent words in text data using Plotly Express.

6. Colorama: The Colorama library is designed to augment terminal text with color. It is utilized to format console output text.
Its utilization within the code is to augment the console output by including colored text.

7. Sentiment Intensity Analyzer and TextBlob: For sentiment analysis, TextBlob and SentimentIntensityAnalyzer are implemented.
TextBlob offers a straightforward API for routine natural language processing tasks, whereas NLTK's SentimentIntensityAnalyzer computes sentiment scores.

8. Fake Data Generation: The Faker library is employed to produce fabricated data. Although not explicitly associated with machine learning, it may prove beneficial in the context of testing and development.

ML METHODS:

The following machine learning techniques can be used to look at the collection of McDonald's customer reviews:

Support Vector Machines (SVM): (used in EDA)
SVM is a strong supervised learning method that is used to sort things into groups. Its goal is to find the hyperplane that best divides the different classes. This makes it useful for sentiment analysis, which uses features taken from textual reviews to figure out whether a mood is positive, negative, or neutral.

And we can also use,

Random Forest:
Random Forest is a type of ensemble learning that makes many decision trees and then combines the estimates they make. It can be used for both classification and regression tasks, which means it can be used to guess how satisfied customers will be based on different factors. Random Forest is famous for being strong and able to deal with complicated data connections.

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://github.com/manikantareddy12/McDonalds-Review-System