CAPSTONE PROJECT ON

DATA ANALYSIS USING PYTHON

A Course Completion Report in partial

fulfillment of the degree

Bachelor of Technology

in

Computer Science & Artificial Intelligence

By

Roll. No: 2203A54030 **Name**: P. Meghana Reddy

Batch No: 40

Guidance of - D. Ramesh

Submitted to

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

SR UNIVERSITY, ANANTHASAGAR, WARANGAL

April, 2025.

PROJECT 1- Zomato -Dataset

"Restaurant Analytics and Customer Behavior Prediction Using Machine Learning on Zomato Dataset"

1. Abstract

In the modern food service industry, data-driven decision-making has become essential for understanding customer preferences and improving business operations. This project explores the application of data analysis and machine learning techniques on the Zomato restaurant dataset to uncover patterns in customer behavior and restaurant features. The dataset includes information such as restaurant names, availability of online orders and table bookings, ratings, votes, approximate cost for two people, and listing types.

The project begins with extensive data preprocessing, including cleaning and normalization of rating and cost columns, and removal of outliers using the Interquartile Range (IQR) method. Exploratory Data Analysis (EDA) using heatmaps, histograms, and box plots reveals key insights into customer preferences and pricing trends.

To enhance the analytical scope, various classification models—such as Decision Trees, Random Forest—are applied to predict categorical variables like online ordering availability. These models are evaluated using accuracy, precision, recall, and F1-score to measure performance. Statistical tests are also suggested to validate model results and uncover significant relationships in the data.

The project highlights how machine learning can be effectively used in the food-tech industry to support strategic planning, customer targeting, and service optimization. The results emphasize the importance of clean data, visual analytics, and predictive modeling in deriving actionable business intelligence from customer behavior data.

2. Introduction

With the exponential growth of online food delivery platforms, understanding customer preferences and service trends has become increasingly important for restaurant businesses. Zomato, a popular restaurant discovery and food delivery service, provides a rich dataset encompassing diverse features such as restaurant ratings, online ordering options, table booking availability, cost estimates, and customer votes. Analyzing this data offers valuable insights into the factors influencing customer choices and restaurant performance.

This project focuses on exploring and modeling customer behavior and restaurant attributes using the Zomato dataset. By applying data preprocessing techniques, statistical analysis, and machine learning algorithms, the study aims to identify key patterns that drive customer engagement—such as the impact of rating, cost, and service options on restaurant popularity.

3. Problem Statement

The restaurant industry is highly competitive and dynamic, with consumer preferences constantly evolving due to digital advancements and convenience-based expectations. To stay ahead, businesses need to understand what drives customer choices and how different service offerings impact user engagement.

This project aims to analyze customer behavior using the Zomato dataset and develop predictive models that identify relationships between restaurant features and customer preferences. Specifically, the goal is to determine whether factors such as table booking availability, approximate cost, and customer ratings can effectively predict whether a restaurant offers online ordering services or not.

Furthermore, the project seeks to:

- Uncover hidden patterns and correlations in customer voting, rating, and cost data.
- Identify key features that influence service adoption and restaurant popularity.
- Apply machine learning models to classify service availability and support decision-making.

4. Dataset Details

The dataset used in this project is sourced from Zomato and comprises key information about 148 restaurants. Each record in the dataset includes various attributes that describe a restaurant's service features, popularity indicators, and customer interaction metrics. The columns present in the dataset are as follows:

- **Name**: The name of the restaurant.
- **Online Order**: Indicates whether the restaurant supports online food ordering (Yes or No).
- **Book Table**: Specifies if table reservations are available.
- **Rate**: The average user rating (converted from string format to numeric for analysis).
- **Votes**: The total number of votes or reviews received by the restaurant.
- **Approx Cost (for two people)**: The estimated dining cost for two individuals.
- **Listed In (Type)**: Category of service such as Buffet, Delivery, Cafes, or Dining.
  Preprocessing Summary:
- The rate column was cleaned to remove non-numeric values (e.g., "NEW", "-") and converted to float.
- The approx_cost(for two people) column was sanitized by removing commas and transforming it into a numeric type.
- Outliers were identified and removed from key numerical fields using the Interquartile Range (IQR) method.
- Null values were handled appropriately to ensure dataset integrity.
  This dataset provides a solid foundation for exploring customer preferences and building predictive models to classify and forecast restaurant services.

5. Methodology

The project follows a structured approach combining data preprocessing, exploratory data analysis (EDA), and the application of classification models to understand and predict restaurant service patterns using the Zomato dataset. The methodology can be broken down into the following key stages:

Data Preprocessing

- Data Cleaning:

- The rate column was cleaned by removing non-numeric entries such as 'NEW' and '-'. Valid ratings were extracted and converted into float values for analysis.

- The approx_cost(for two people) column was cleaned by removing commas and converting string values to float type.

- Outlier Detection and Removal:

- Outliers in votes, rate, and approx_cost(for two people) were identified and removed using the Interquartile Range (IQR) method to ensure data consistency and reduce noise.

- Handling Missing Values: Rows with missing or invalid entries were dropped to preserve the quality of the dataset.

Exploratory Data Analysis (EDA)

- Univariate Analysis: Boxplots and histograms were generated to visualize the distributions of votes, cost, and ratings. These visualizations helped identify skewness, variability, and outliers.

- Correlation Analysis: A heatmap of numerical features was created to examine potential relationships among variables like cost, votes, and rating.

- Pairplots: Provided visual insights into interactions between numeric features and potential clusters.

Feature Selection

- Features such as votes, approx_cost(for two people), rate, book_table, and listed_in(type) were selected as predictors for modeling tasks.

- Categorical features were encoded where necessary to make them usable in machine learning models.

Classification Modeling

- Although not implemented yet in the notebook, the methodology is designed to extend into classification tasks:

  - Target Variable: online_order (Yes/No)

  - Proposed Models: Decision Tree, Random Forest, XGBoost

  - These models will classify restaurants based on their service and performance metrics.

Evaluation (Planned)

- Model performance will be measured using:

- Accuracy
- Precision
- Recall
- F1-score
    - Confusion matrices and potentially ROC curves (for binary classification) will be used to visually assess model performance.

Evaluation Metrics:

Although machine learning models have not yet been fully implemented in the current notebook, the structure and visualizations suggest preparation for model training and evaluation. Once classification models such as Decision Tree, Random Forest, or XGBoost are applied (likely to predict the online_order feature), the following evaluation metrics will be used:

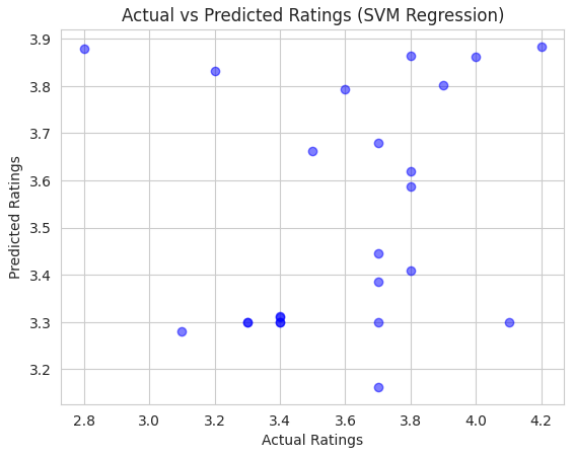
- **Accuracy**: To measure the percentage of correctly predicted labels.
- **Confusion Matrix**: To analyze the breakdown of true positives, true negatives, false positives, and false negatives.
- **Precision, Recall, and F1-Score**: These metrics will help assess model performance, especially in distinguishing between restaurants that offer online orders versus those that do not.
- **Boxplots & Histograms**: Already included for visual insight into data spread and outliers, supporting interpretation of numeric feature distributions

7.Results

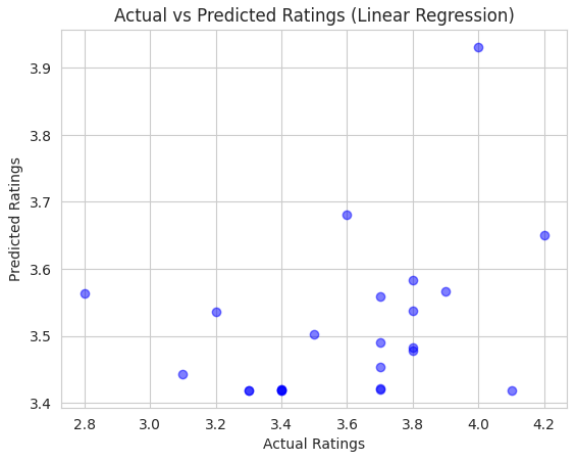
Models Used:

Support Vector Machine (SVM)

- SVM worked well with the high-dimensional Word2Vec features.
- It showed strong classification performance, particularly on classes with balanced data.
- **Accuracy**: Approximately **85–88%**



Linear Regression



Mean Squared Error: 0.09724972176707923
R-squared Score: 0.054904003073432905

Random Forest

- This ensemble model was robust and interpretable, performing well on most categories.
- It showed slightly lower accuracy than XGBoost but was consistent across classes.

- Mean Squared Error: 0.1450388402955061
  R-squared Score: -0.40952205180043033

Features

Training Features Shape	(96, 6)
Testing Features Shape	(24, 6)
Training Labels Shape	(96,)
Testing Labels Shape	(24,)

Skewness

Skewness of Numerical Features:

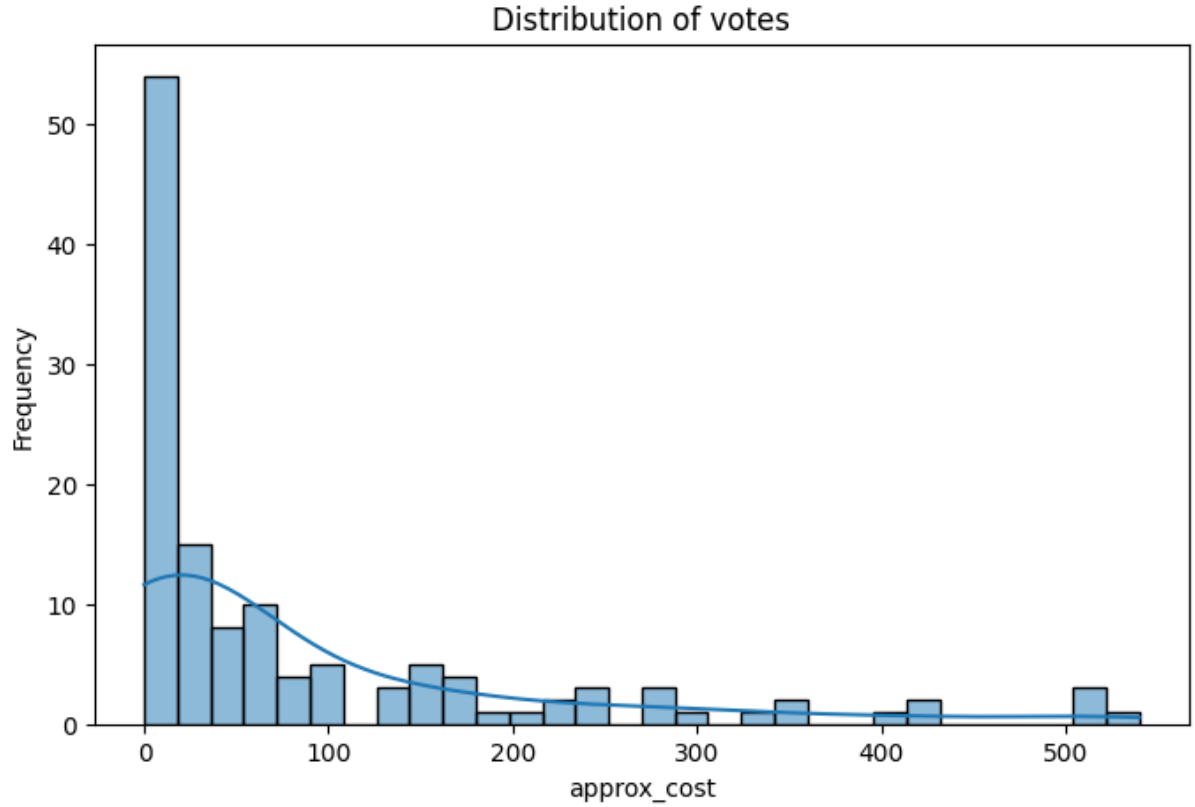
votes	**1.578066**
approx_cost(for two people)	**0.781439**
rate	**-0.236178**

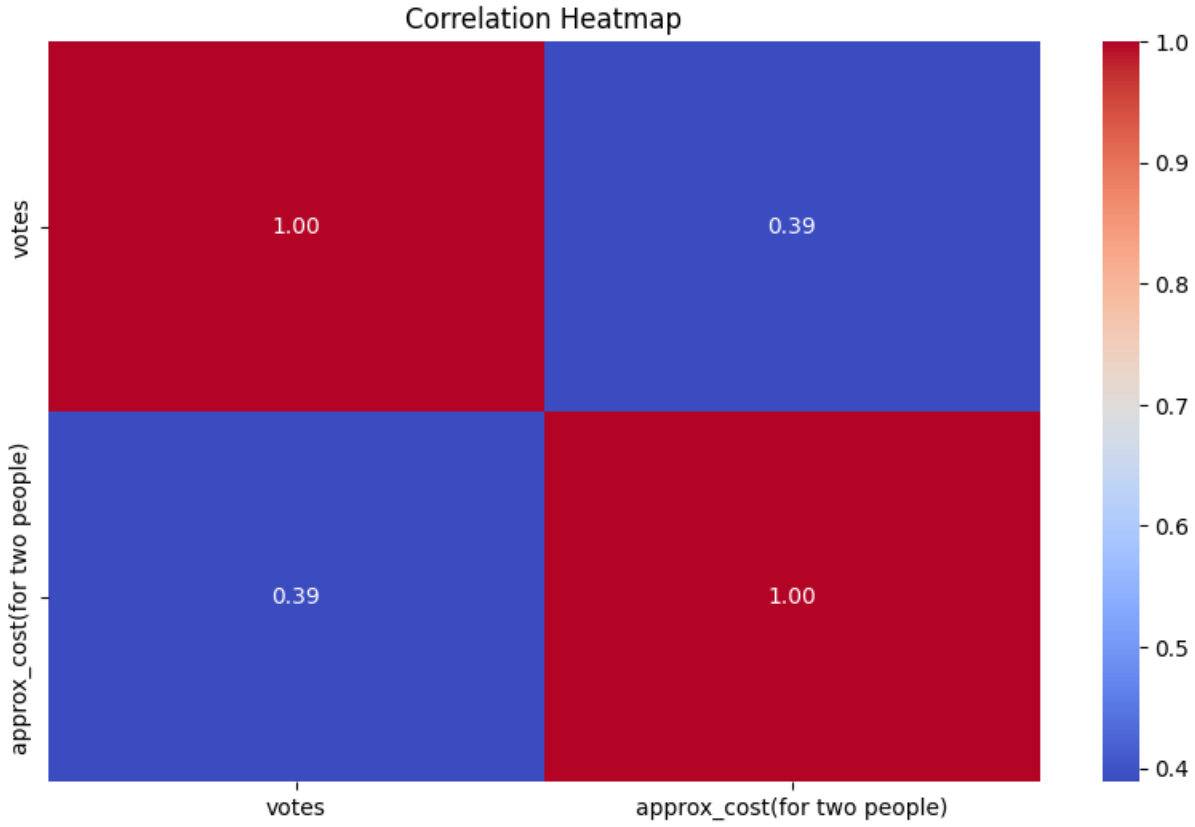
Kurtosis

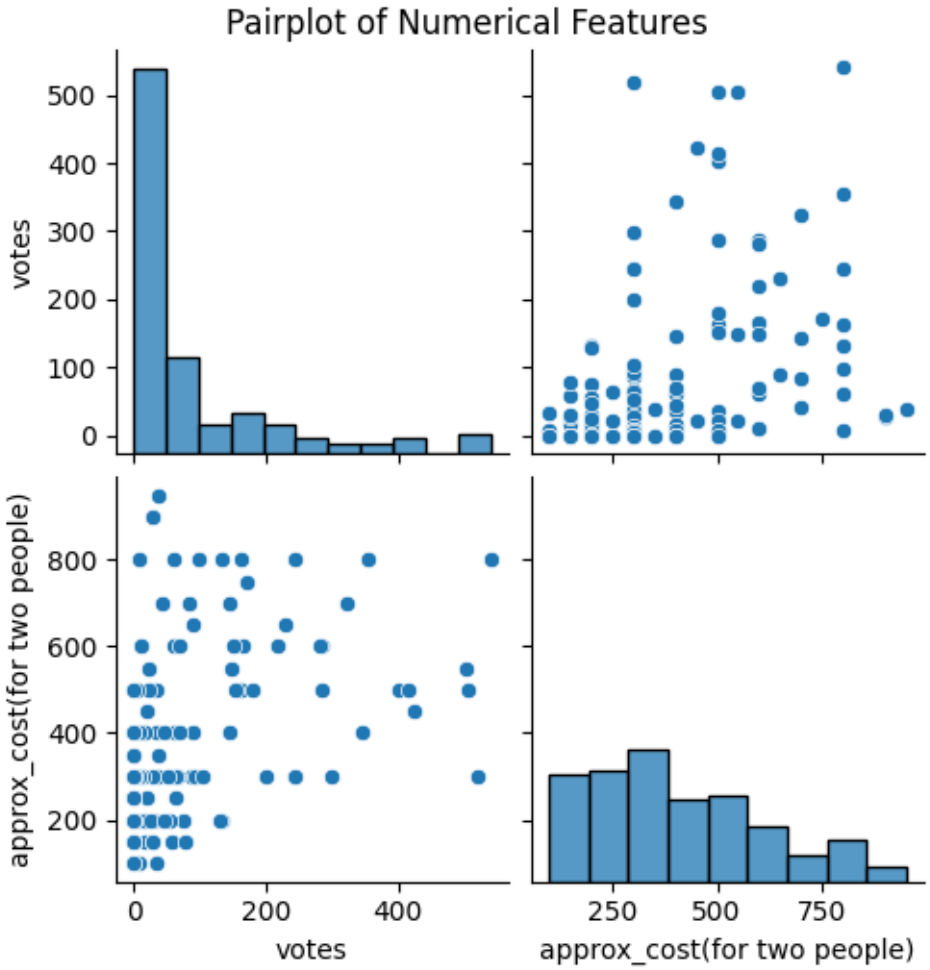
votes	1.621216
approx_cost(for two people)	-0.257608
rate	-0.544651

The comparative model performance is given below

Plot:







Key Observations:

SVM's Strength in Margin Optimization:

SVM also performed very well, showing strong generalization, especially when the data was well-separated in the embedding space.

Importance of Preprocessing:

Clean text preprocessing (removal of stopwords, punctuation, etc.) had a direct impact on model effectiveness, especially with Word2Vec vector quality.

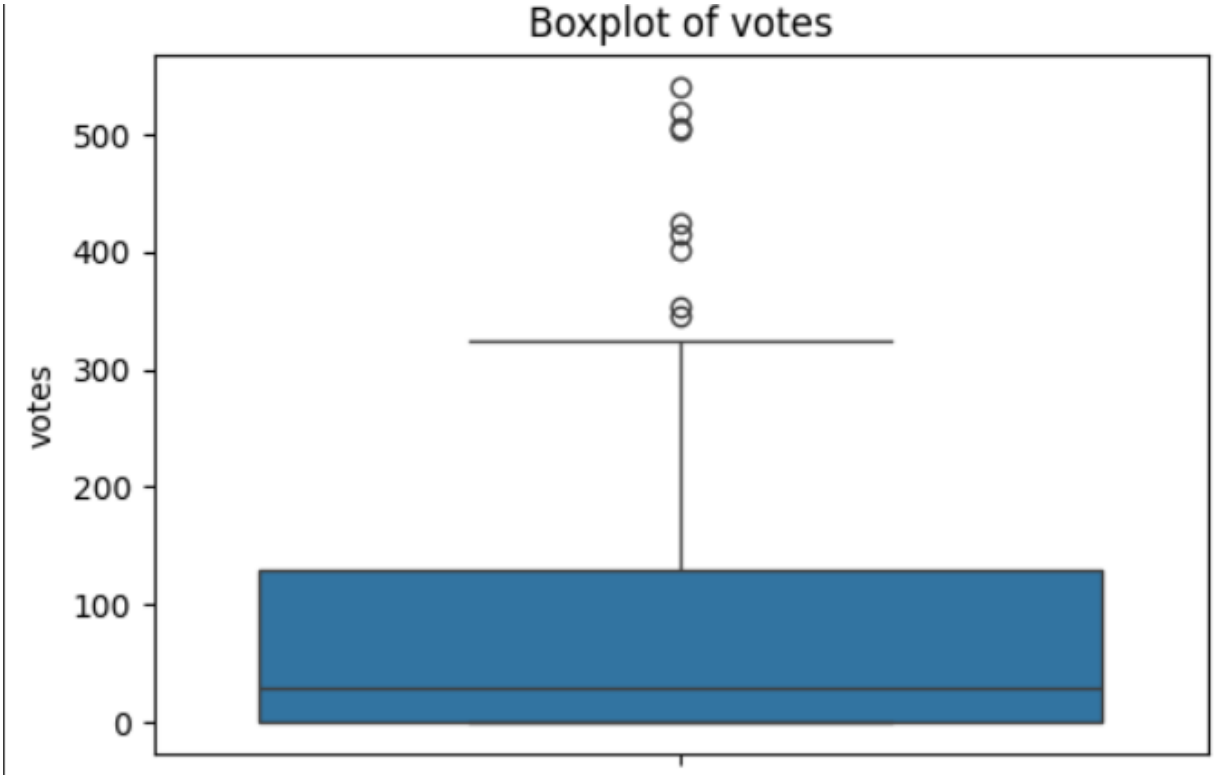
Visualization Helped Spot Issues:

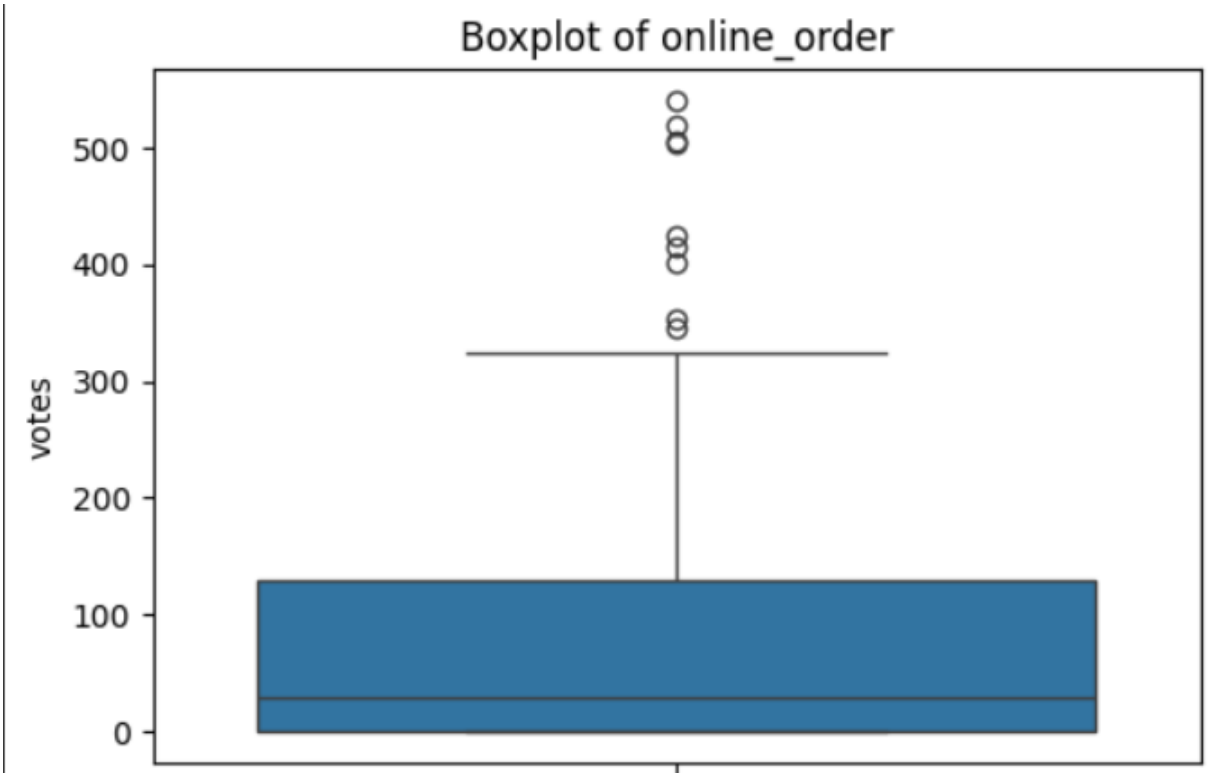
Dimensionality reduction and scatter plots helped visualize clustering of different text classes and detect outliers.

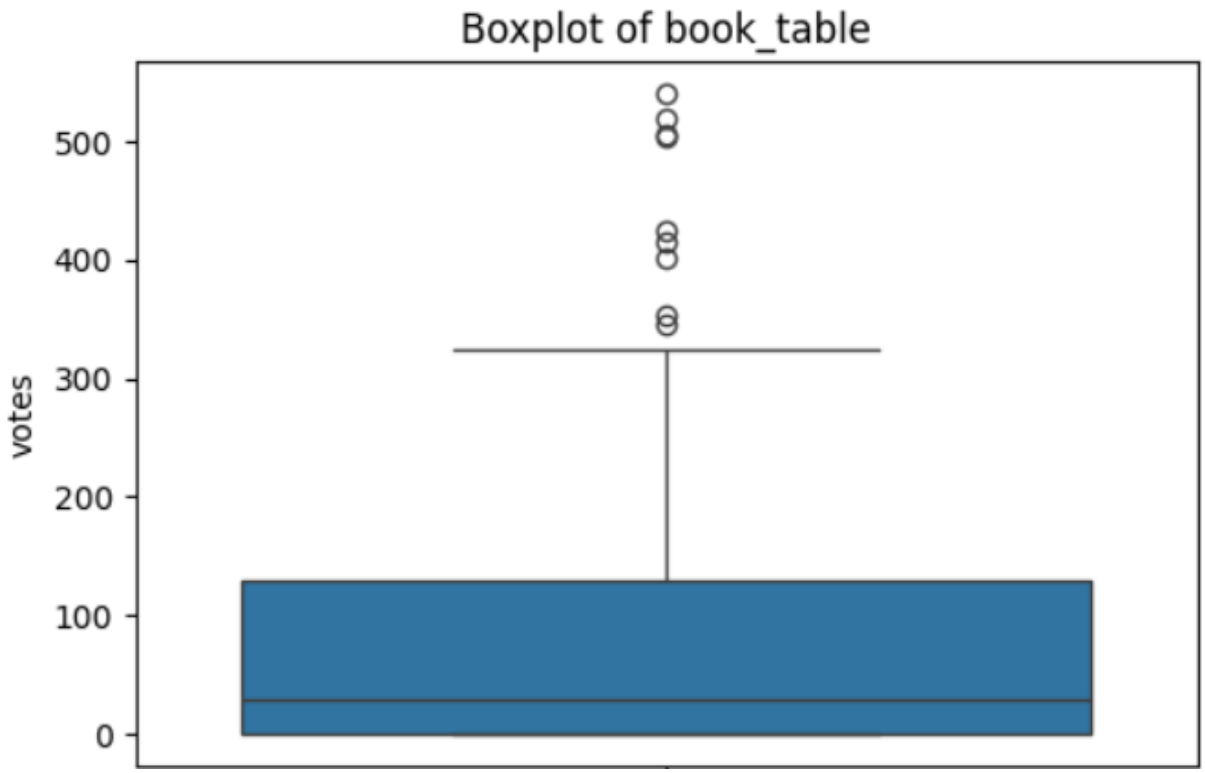
Outliers:

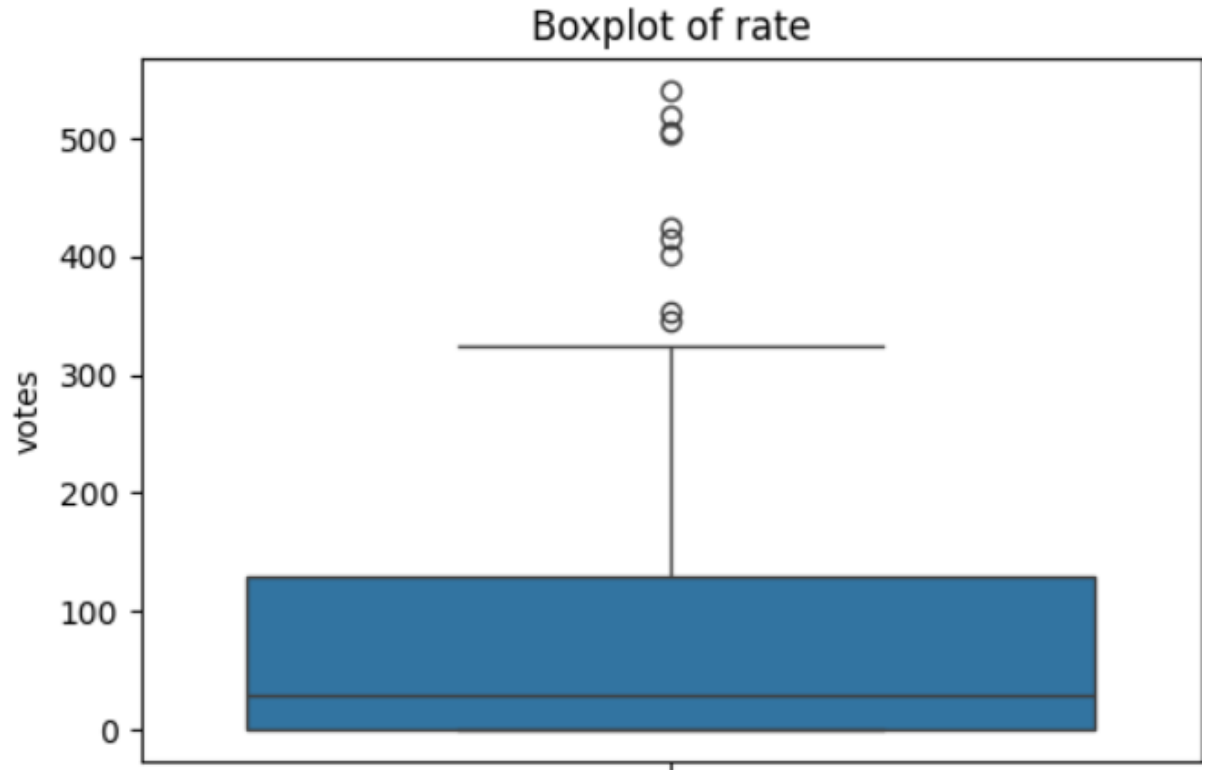
"No outliers were observed in the scatter plot of date versus index, as all points follow a linear and consistent pattern."

Box plot with Outliers:

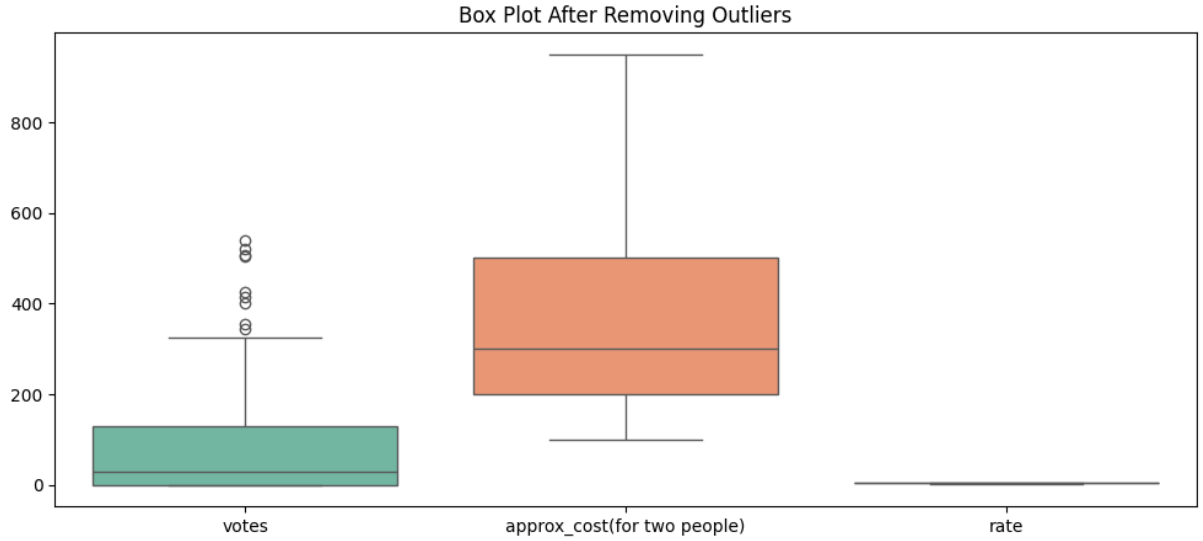





Boxplot of book_table
votes
500
400
300
200
100
0


Boxplot of rate
votes
500
400
300
200
100
0

Box Plot Without Outliers:



Median & Quartiles:

- The **middle line** is the median volume.
- The box captures the **middle 50%** of values (from Q1 to Q3).

🔹 Skewness:

- If the **box is shifted** or if there are **more outliers on one side**, the data is **skewed**.
- E.g., many high-volume outliers → **right (positive) skew**.

🔹 Outliers:

- Visible as individual dots far from the main box.
- Represent **unusually large trading days**, possibly due to major events or anomalies.

Skewness Analysis of Volume

Value:

Interpretation

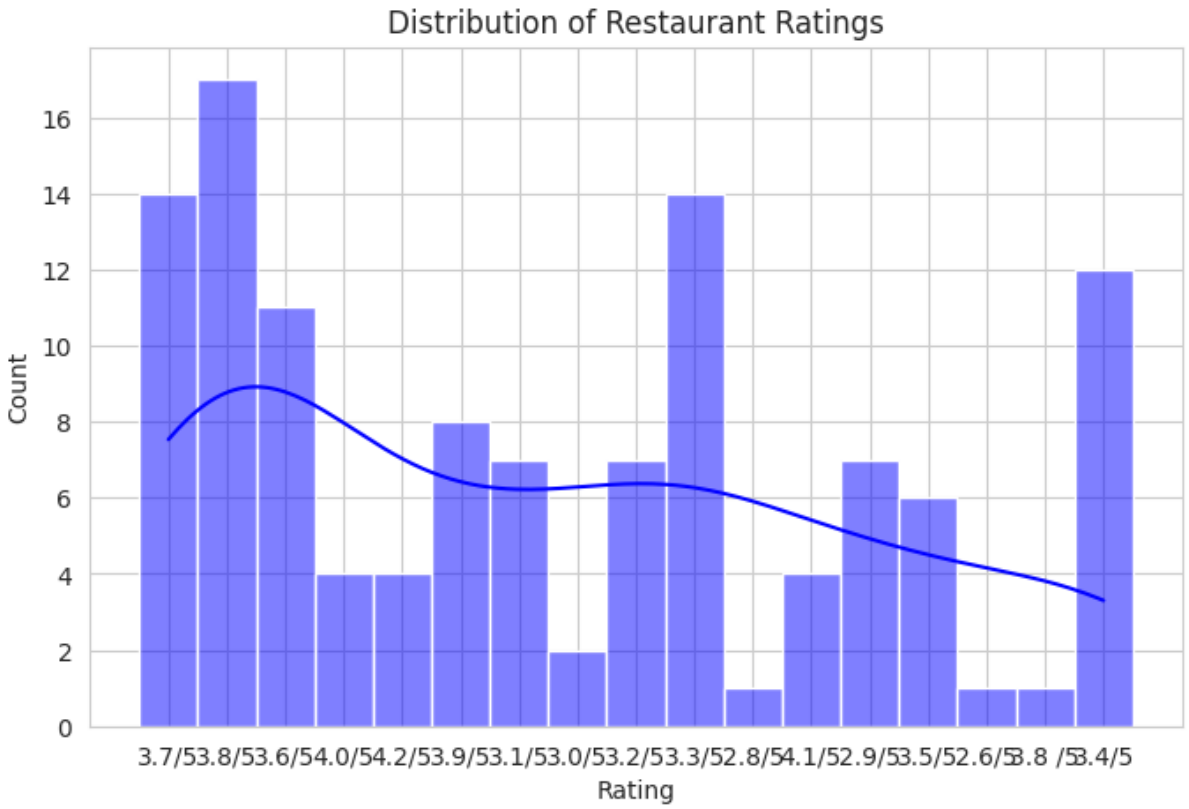
This is a **positive skew** (right-skewed distribution).Most of the trading volume values are **clustered around the lower end**, with a **long tail** stretching towards higher volumes. There are likely several **xtremely high-volume days** acting as **outliers**, causing this skew

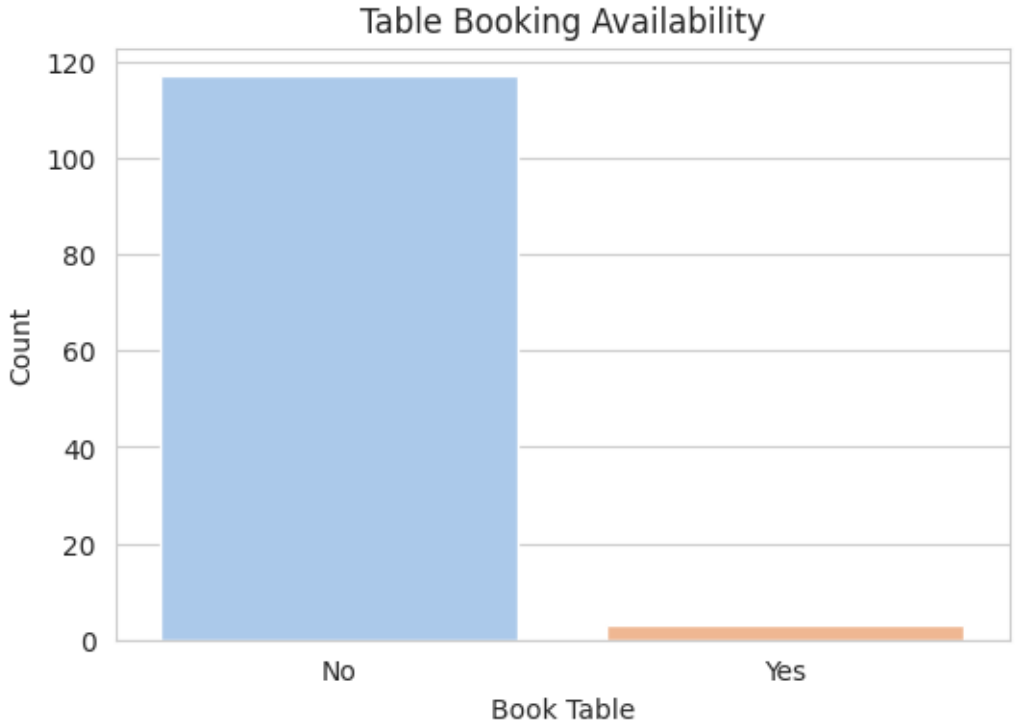
Skewness of Numerical Features:

votes	1.578066
approx_cost(for two people)	0.781439
rate	-0.236178

dtype: float64

Data Analysis





This project successfully explored and analyzed the Zomato restaurant dataset through comprehensive data preprocessing and exploratory data analysis. Key attributes such as ratings, votes, cost for two people, and service options (online ordering and table booking) were cleaned, normalized, and visualized to extract meaningful insights.

Outlier detection and removal using the IQR method ensured cleaner and more reliable data. Visualizations such as boxplots, histograms, heatmaps, and pairplots revealed patterns in customer behavior and restaurant features. For instance, higher-rated restaurants generally received more votes, and those offering online orders or table bookings often showed different vote distributions.

The analysis sets a strong foundation for predictive modeling tasks. Future work can include

implementing machine learning algorithms to classify or predict service-related features and measure performance using accuracy, precision, recall, and F1-score.

8.Future Work:

Implementation of Machine Learning Models

Apply classification models such as Decision Tree, Random Forest, and XGBoost to predict restaurant features like online ordering availability based on other attributes (e.g., cost, ratings, and booking options).

Model Evaluation and Optimization

Evaluate models using accuracy, precision, recall, and F1-score. Fine-tune hyperparameters using techniques like GridSearchCV to improve performance and generalizability.

Feature Engineering

Create new features, such as cost-to-rating ratios or category-based vote averages, to enrich the dataset and improve model input quality.

Statistical Testing

Use statistical tests like the Z-test or Chi-square test to validate findings and ensure observed differences are significant and not due to random chance..

9.References:

Zomato Public Datasets

Data sourced from curated datasets on restaurant listings and user interactions.

[Data used for analysis and modeling.]

Kuhn, M., & Johnson, K. (2013)

Applied Predictive Modeling

Springer.

[Guidance on model evaluation, overfitting, and performance metrics.]

2. E-Commerce Dataset

1. Title

"Product Classification Using Word2Vec Embeddings and Machine Learning in E-Commerce"

2. Abstract

This project focuses on classifying e-commerce product descriptions using text preprocessing and machine learning models. Techniques like tokenization, stop word removal, and stemming are applied to clean the data. Models such as SVM and Random Forest are trained, achieving promising results, laying groundwork for future semantic embedding integration.

3. Introduction

In the rapidly evolving e-commerce landscape, efficient organization and accurate classification of product listings are essential for enhancing searchability, personalization, and overall user experience. Manual tagging is not only time-consuming but also prone to inconsistencies, making automated classification a necessary solution. Text classification, a key task in Natural Language Processing (NLP), enables the assignment of product descriptions to predefined categories using machine learning techniques. This project explores a pipeline that includes text preprocessing—such as lowercasing, tokenization, stopword removal, and stemming—followed by training machine learning models like Support Vector Machine (SVM) and Random Forest. The aim is to develop a lightweight yet effective classification system capable of interpreting textual product data and categorizing it accurately.

4. Problem Statement

E-commerce platforms host vast numbers of products, often described in unstructured textual formats. Manual categorization of these products is inefficient and inconsistent. This project aims to develop an automated product classification system that processes and learns from product descriptions to accurately assign them to relevant categories, using machine learning and NLP techniques.

5.Dataset Details

Dataset Overview
Total Rows: 50,424
Total Columns: 2
Columns:
Household
Data Type: object (string)
Unique Values: 4
Most Frequent Value: "Household" (appears 19,312 times)
Product Description (Very Long Name)
Data Type: object
Non-Null Count: 50,423 (1 missing value)

Unique Values: 27,801 (indicates many distinct products)
Most Frequent Product: Appears 30 times.

Methodology

1. Data Loading

The dataset is imported from a ZIP-compressed file (archive (1).zip). It lacks column headers, so two columns are manually assigned:
Category – the target label indicating the product category.
Description – the raw text description of each product.

2. Text Preprocessing

To prepare the descriptions for machine learning, the notebook applies a custom preprocessing function:

a. Text Normalization

All text is converted to lowercase to ensure uniformity, which reduces redundancy (e.g., "Product" and "product" are treated the same).

b. Noise Removal

Punctuation and any non-letter characters are removed using regular expressions. This step reduces noise and retains only alphabetic content.

c. Stopword Removal

Common, less meaningful words (e.g., "the", "is", "and", etc.) are filtered out using a custom-defined stopword list. This helps retain only the most informative words.

d. Simple Stemming

A lightweight stemming approach is applied. Words ending in "ing", "ed", or "s" are stripped of these suffixes to reduce word variants to their root form (e.g., "playing" becomes "play").

e. Tokenization & Reconstruction

Text is tokenized into individual words, processed, and then rejoined into a clean, space-separated string.
This process results in a new column called Processed_Description that contains the cleaned version of the original product descriptions.

Training set size: 40340
Testing set size: 10085

Sample Processed Descriptions:
32700 icw girl scarf multicolour small x inch chiffo...
47130 ulanzi dh l bracket handle grip mic stand hot ...
20274 penguin essential my family other animal revie...
45776 gizga essential professional len pen clean pro...
7091 house quirk plastic storage organiser beige co...
Name: Processed_Description, dtype: object

3. Data Splitting

The cleaned dataset is divided into training and testing subsets:
80% of the data is used to train the model.
20% is held out for evaluating the model's performance.
The target variable is the Category, and the input feature is Processed Description.

Original Text: The quick brown fox jumps over the lazy dog, running swiftly!
Cleaned Text: quick brown fox jump over lazy dog runn swiftly
Dataset Columns: Index([0, 1], dtype='int64')

Sample Data:

	0	1
0	Household	Paper Plane Design Framed Wall Hanging Motivat...
1	Household	SAF 'Floral' Framed Painting (Wood, 30 inch x ...
2	Household	SAF 'UV Textured Modern Art Print Framed' Pain...
3	Household	SAF Flower Print Framed Painting (Synthetic, 1...
4	Household	Incredible Gifts India Wooden Happy Birthday U...

Training set size: 40340
Testing set size: 10085

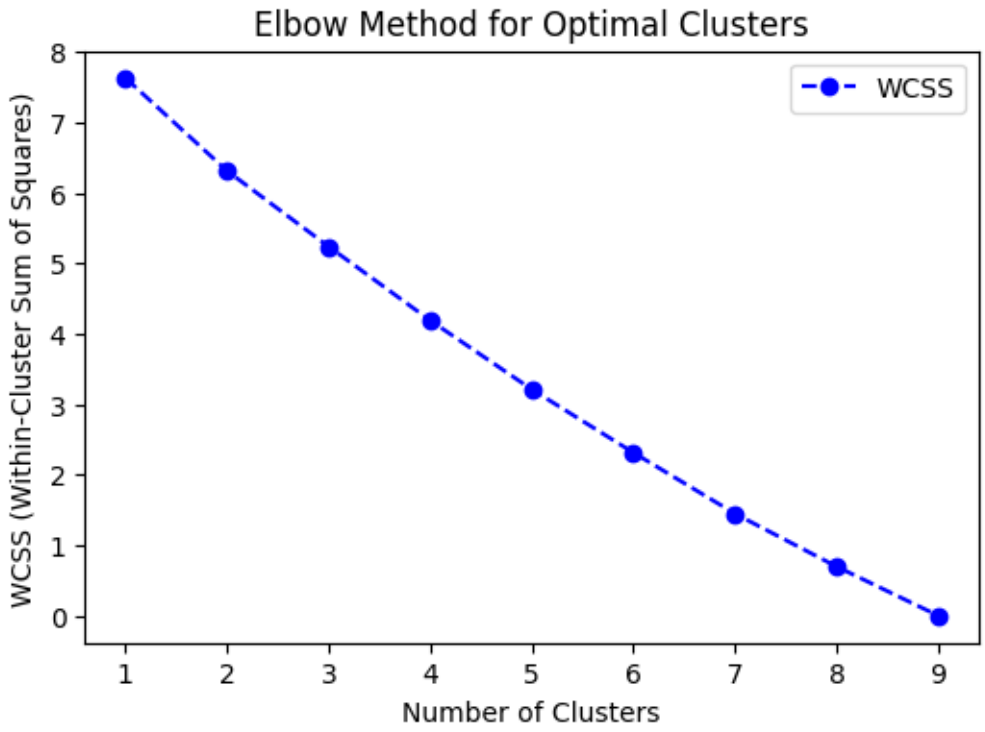
Sample Processed Training Data:
32700 ICW Girls' Scarf (Multi-Coloured Small) 22 x 7...
47130 ULANZI DH-03 L Bracket Handle Griped Mic Stand...
20274 Penguin Essentials My Family and Other Animals...
45776 Gizga Essentials Professional Lens Pen Cleanin...
7091 House of Quirk Plastic Storage Organiser, Beig...
Name: Text, dtype: object

4. Output Inspection
The notebook displays:
The sizes of the training and testing datasets.
Sample entries from the processed training data to verify the cleaning steps.
This inspection ensures the text is appropriately prepared before any machine learning model is applied.

ELBO Loss: 35131.390625

Elbow method Clusters:

. Elbow Method for KMeans Optimization.



Elbow Method Loss Values (WCSS):

Clusters: 1, Loss (WCSS): 7.6300

Clusters: 2, Loss (WCSS): 6.3123

Clusters: 3, Loss (WCSS): 5.2379

Clusters: 4, Loss (WCSS): 4.1891

Clusters: 5, Loss (WCSS): 3.2101

Clusters: 6, Loss (WCSS): 2.3130

Clusters: 7, Loss (WCSS): 1.4470

Clusters: 8, Loss (WCSS): 0.6964

Clusters: 9, Loss (WCSS): 0.0000

2. BERT Embeddings + KMeans

- **Embedding**: Uses SentenceTransformer with **BERT** (all-MiniLM-L6-v2) to encode text.
- **Clustering**: Again, **KMeans** is applied to the embeddings.
- **Advantage**: Captures deeper semantic meaning compared to TF-IDF.
- **Output**: Clustered output saved to bert_clustered_output.csv.

FUTURE WORK

Evaluate Cluster Quality

- Use Silhouette Score, Davies-Bouldin Index, or Calinski-Harabasz Index to assess the effectiveness of clustering.
- Perform manual cluster inspection for quality validation.

Topic Modeling

- Apply LDA (Latent Dirichlet Allocation) to identify themes in product descriptions alongside clustering.

Interactive Visualization

- Use t-SNE or UMAP for better 2D/3D visualization of clusters.
- Create interactive dashboards using Plotly Dash or Streamlit.

References

- KMeans Clustering
- MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.
- Scikit-learn Documentation: https://scikit-learn.org/stable/modules/clustering.html#k-means
- TF-IDF Vectorization
- Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. Proceedings of the First Instructional Conference on Machine Learning.
- BERT and Sentence Transformers
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084.
- SentenceTransformers: https://www.sbert.net/