CENTRAL TOOL ROOM & TRAINING CENTER

B-36, CHANDAKA INDUSTRIAL AREA, NEAR INFOCITY, BBSR,(O.D) 751024

Department Of PGDTD & CC

CERTIFICATE

Certified that Major Project Work entitled "**DATA ANALYSIS OF AMAZON TOP SELLING BOOKS**" is a bonafide work carried out by MCCAI 1st Batch in particular fulfillment of the requirements for the award of MASTER CERTIFICATE COURSE IN ARTIFICIAL INTELIGENCE from CENTRAL TOOL ROOM & TRAINING CENTER, Bhubaneswar, during the year 2019-2020. It is certified that all the correction/suggestion indicated for internal assessment have been incorporated in the report. The major project report has been approved as is satisfied the academic requirements in the respect of project work prescribed for PGDTD & CC, CTTC, Bhubaneswar (O.D).

.....	.....	.....
Workshop In-Charge (Dept. Of MCCAI)	Course Coordinator (Dept. Of MCCAI)	Examiner

ACKNOWLEDGEMENT

It is our proud privilege and duty to acknowledge the kind of help and guidance received from several people in preparation for this report. It would not have been possible to prepare this report in this form without their valuable help, cooperation, and guidance.

First and foremost, we wish to record our sincere gratitude to Management of this Training Center and to our beloved Principal, Mr. L. Rajasekhar, and Sr. Manager, Mr. Rajan K.M, CTTC, Bhubaneswar. For his constant support and encouragement in preparation of this report and for making the available library and laboratory facilities available needed to prepare this report. Our sincere thanks to Sr. Engineer Mrs. Ritu Maity, CTTC, BBSR. For his valuable suggestions and guidance throughout the period of this report. We express our sincere gratitude to our supervisor Er. Md Shamaun Alam, Er. Ashutosh Pati & Er. Aldrin Kerketta, Department of MCCAI, CTTC, Bhubaneswar for guiding us in investigations for this major project and in carrying out experimental work. Our numerous discussions with him were extremely helpful. We hold him in esteem for guidance, encouragement, and inspiration received from his.

The project on " **DATA ANALYSIS OF AMAZON TOP SELLING BOOKS** " was very helpful for us in giving the necessary background information and inspiration in choosing this Project. Our sincere thanks to Er. Ashutosh Pati, for having supported the work related to this project. Their contributions and technical support in preparing this report are greatly acknowledged. Last but not the least, we wish to thank our parents for financing our studies in this college as well as for constantly encouraging us to learn to engineer. Their personal sacrifice in providing this opportunity to learn engineering is gratefully acknowledged.

Place: Bhubaneswar MCCAI –2ND

DECLARATION

We the batch of MCCAI 1ST studying in the final semester of of MASTER CERTIFICATE COURSE IN ARTIFICIAL INTELIGENCE from CENTRAL TOOL ROOM & TRAINING CENTER at CTTC (Central Tool Room & Training Center), Bhubaneswar, hereby declare that this major project work entitled "**DATA ANALYSIS OF AMAZON TOP SELLING BOOKS**" which is being submitted by us in the partial fulfillment for the award of the degree of MCCAI, from CTTC, Bhubaneswar is an authentic record carried out during the Training year 2019-2020, under the supervision of Er. Md Shamaun Alam, Er. Ashutosh Pati & Er. Aldrin Kerketta, Department of MCCAI, CTTC, Bhubaneswar.

MCCAI **2nd** BATCH

ABSTRACT

Book Sales tracking is vital to all organizations irrespective of size and industry. What was considered the most advanced system for registering the employees' book sales, has become not only redundant, but also lethal with any kind of human body touch based contact carries enough risk of people getting contracted with the deadly virus.

This project proposes a method for Book Sales system using facial recognition technique by using python programming and from OpenCv library Haar cascade method.Facial recognition is an easy and secure way of taking down book sales. It is a biometric identification method using a face-scanning mechanism. The device captures the facial impression of employees and processes the information into a secure database. The scanned images are stored and mapped into a face coordinate structure.

Key Words: Excel, Power BI, Tableau, Python, Database, Came, Data Processing, Python, Encoding, Standard Scaler.

INDEX

INTRODUCTION

In the modern era of e-commerce, online marketplaces like Amazon generate massive volumes of data that can reveal customer preferences, product trends, and market behaviors. Among these, book sales data provides valuable insight into reading trends, genres in demand, and pricing strategies. However, due to the large and dynamic nature of such datasets, it becomes challenging to manually track patterns and extract actionable insights. To address this, several analytical tools and techniques have emerged to streamline data exploration and decision-making.

This project focuses on **Data Analysis of Amazon's Top Selling Books**, applying modern data analytics and machine learning techniques to uncover patterns and make predictions. The dataset includes attributes such as book title, author, genre, rating, reviews, and price. By leveraging tools like Excel, Power BI, Tableau, and Python, we transform raw data into meaningful insights.

In general, book sales analysis can be approached in different ways, including:

- **Manual Analysis in Spreadsheets** – Data is reviewed and processed manually using filters, pivot tables, and charts. This is time-consuming and prone to human error for large datasets.
- **Business Intelligence Dashboards** – Platforms like Power BI and Tableau create interactive dashboards that enable dynamic filtering and visualization of trends.
- **Statistical and Exploratory Data Analysis (EDA)** – Using Python libraries such as Pandas, Matplotlib, and Seaborn to identify correlations, distributions, and outliers.
- **Predictive Modeling** – Applying machine learning models such as Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors, and Support Vector Machines to predict categories like genre popularity or high/low sales.
- **Data Preprocessing Automation** – Automating encoding, scaling, and splitting datasets to improve modeling efficiency.

- **Recommendation Systems** – Advanced approaches like collaborative filtering and content-based recommendation for suggesting similar books to customers.

By combining visualization tools with predictive analytics, this project aims to provide a comprehensive understanding of Amazon's top-selling books, enabling better forecasting and strategic planning for publishers, sellers, and marketers.

POWER BI

The **Power BI dashboard** developed for this project serves as a comprehensive and interactive platform for exploring Amazon's **Top-Selling Books** dataset. It transforms complex, multi-dimensional data into clear and actionable insights through intuitive visualizations, making it accessible to both technical and non-technical users.

The dashboard is composed of multiple interconnected modules, each designed to address a specific analytical objective:

- **Sales Overview** – Displays total sales distribution segmented by genre, author, and price range, allowing quick identification of high-performing categories.
- **Ratings & Reviews Analysis** – Examines the average ratings alongside customer review counts to reveal how reader engagement correlates with book popularity.
- **Genre Popularity Trends** – Tracks changes in demand for various genres over time, helping identify emerging and declining categories.
- **Price vs. Sales Insights** – Analyzes the relationship between pricing strategies and sales volume, aiding in optimal pricing decisions.
- **Top Authors & Bestsellers** – Highlights the most successful authors and titles, ranked by sales and customer engagement metrics.

Using **Power BI's interactive slicers, filters, and drill-down capabilities**, users can customize their exploration by focusing on specific genres, time periods, or

price segments. This flexibility enables targeted analysis and supports better decision-making for publishers, marketers, and analysts.

Functioning as the **visual architecture** of the project, the dashboard seamlessly integrates results from data preprocessing, exploratory data analysis, and predictive modeling into a single platform. It not only simplifies complex interpretations but also serves as a strategic decision-support tool, maximizing the practical value of the analysis.

EXCEL

The **Excel dashboard** designed for this project provides a compact yet powerful platform for analyzing Amazon's **Top-Selling Books** dataset. Leveraging Excel's advanced features such as pivot tables, slicers, conditional formatting, and dynamic charts, the dashboard delivers a clear, user-friendly view of key performance indicators and trends.

The dashboard is structured into interconnected analytical components, each serving a distinct purpose:

- **Sales Overview** – Summarizes sales performance by genre, author, and time period, enabling quick comparisons across categories.

- **Ratings & Reviews Analysis** – Combines average ratings with the number of customer reviews to assess both popularity and reader satisfaction.
- **Genre Distribution** – Uses donut charts to visualize the proportion of Fiction and Non-Fiction sales, providing an immediate genre split.
- **Price vs. Sales Insights** – Presents the relationship between book prices and sales figures to highlight the impact of pricing strategies.
- **Top Authors & Bestsellers** – Displays the highest-performing authors and titles using bar charts and treemaps for intuitive ranking.

Interactive elements such as **slicers** for year, genre, and author allow users to filter the dataset dynamically, ensuring tailored analysis without altering the core data.

The Excel dashboard acts as a **lightweight, offline-friendly analytics solution**, making it ideal for quick presentations, reporting, and situations where business intelligence tools like Power BI or Tableau are not available. It brings together data visualization and analysis in an accessible format, ensuring that insights can be extracted quickly and effectively.

TABLEU

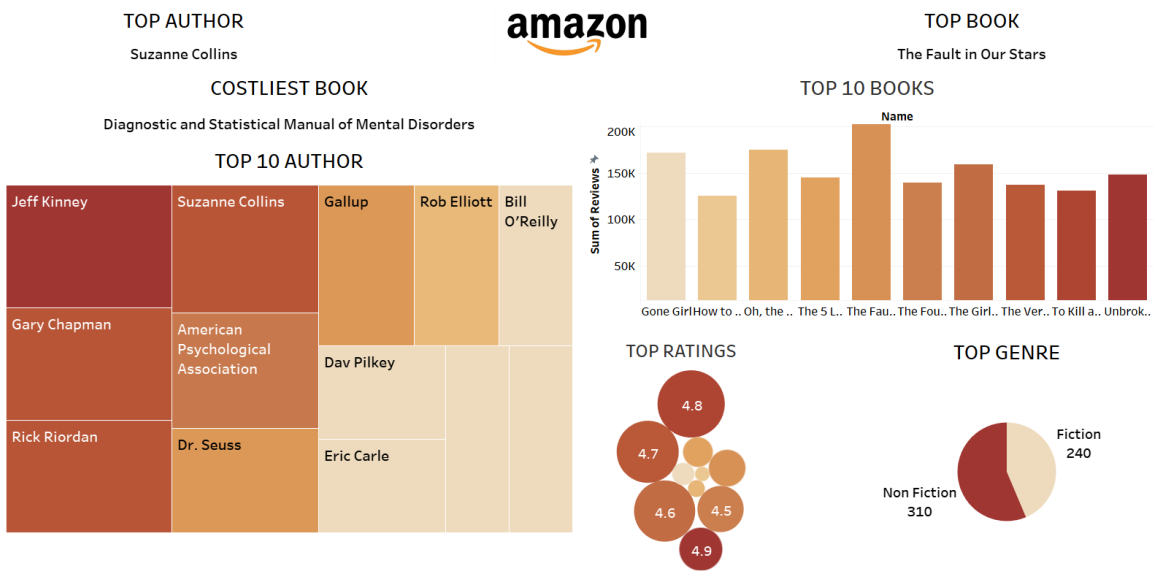
The **Tableau dashboard** developed for this project offers an aesthetically refined and highly interactive platform for analyzing Amazon's **Top-Selling Books** dataset. Tableau's drag-and-drop functionality and powerful visualization engine enable rich storytelling through data, making complex relationships easy to understand.

The dashboard is organized into distinct analytical components, each focused on a specific aspect of the dataset:

- **Sales Overview** – Highlights total sales and distribution patterns across genres, authors, and time periods.
- **Ratings & Reviews Analysis** – Integrates average ratings with customer review counts to reveal the relationship between reader feedback and sales performance.
- **Genre Popularity Insights** – Uses pie charts to visualize Fiction vs. Non-Fiction shares, providing an immediate understanding of market composition.
- **Top 10 Books & Authors** – Presented through bar charts and treemaps for clear ranking and proportional comparisons.
- **Price vs. Sales Visualization** – Offers insights into how price variations impact demand, aiding in strategic pricing decisions.
- **Ratings Distribution** – Displays rating patterns using bubble and scatter plots for quick interpretation of customer sentiment.

Interactive filters and parameters allow users to explore the dataset by author, genre, rating range, or price segment. This flexibility supports targeted exploration while maintaining the integrity of the underlying data.

The Tableau dashboard serves as a **visually rich and insight-driven analytics environment**, merging storytelling with statistical depth. Its design prioritizes user engagement and clarity, making it an effective tool for both quick decision-making and in-depth analysis.

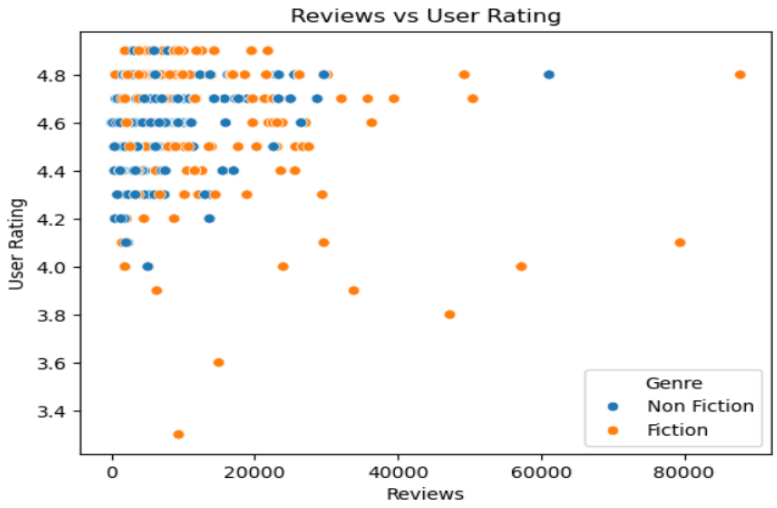


PYTHON DATA ANALYSIS

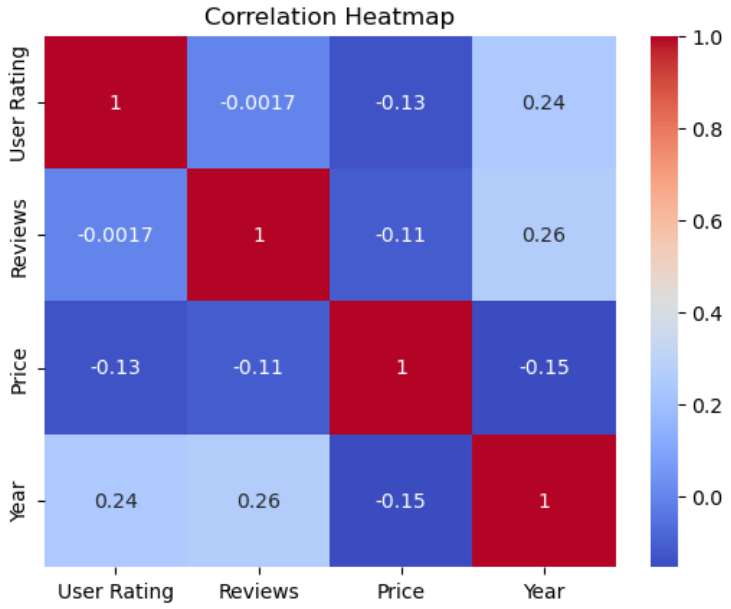
The **Python-based data analysis** component of this project focuses on extracting, visualizing, and interpreting meaningful patterns from the Amazon **Top-Selling Books** dataset. Python's flexibility and powerful libraries such as **Pandas**, **NumPy**, **Matplotlib**, and **Seaborn** provide the foundation for in-depth exploratory data analysis (EDA) and statistical visualization.

The analysis process is divided into key stages, each supported by visual outputs:

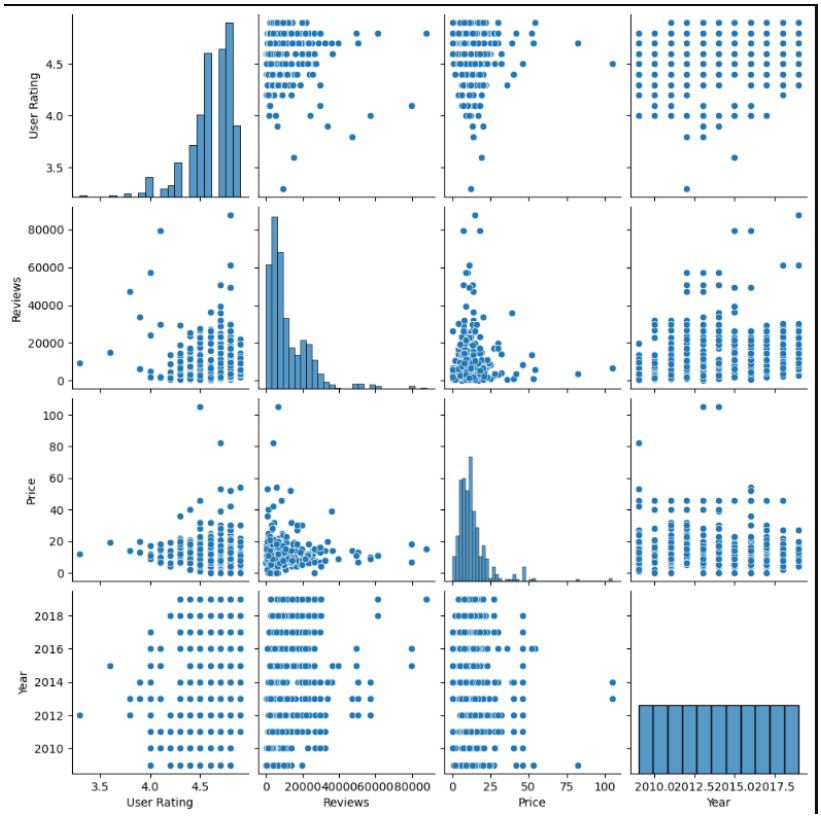
- **Scatterplot Analysis** – Visualizes the relationship between two numerical variables, such as *Price vs. Ratings* or *Number of Reviews vs. Ratings*, to detect potential correlations or trends. *(Insert Scatterplot Image Here)*



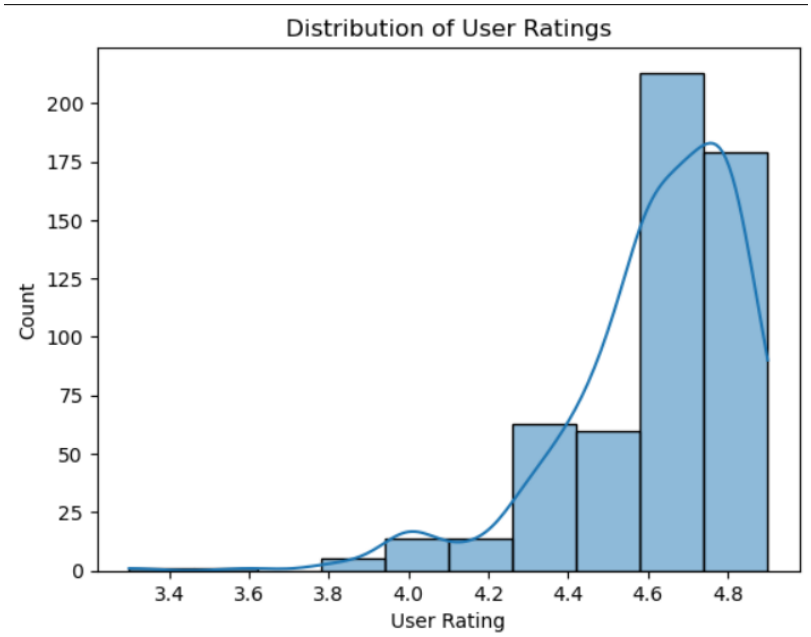
- **Heatmap** – Displays the correlation matrix of numerical features, highlighting the strength and direction of relationships between variables. This helps identify features that might be most influential in predicting book sales. *(Insert Heatmap Image Here)*



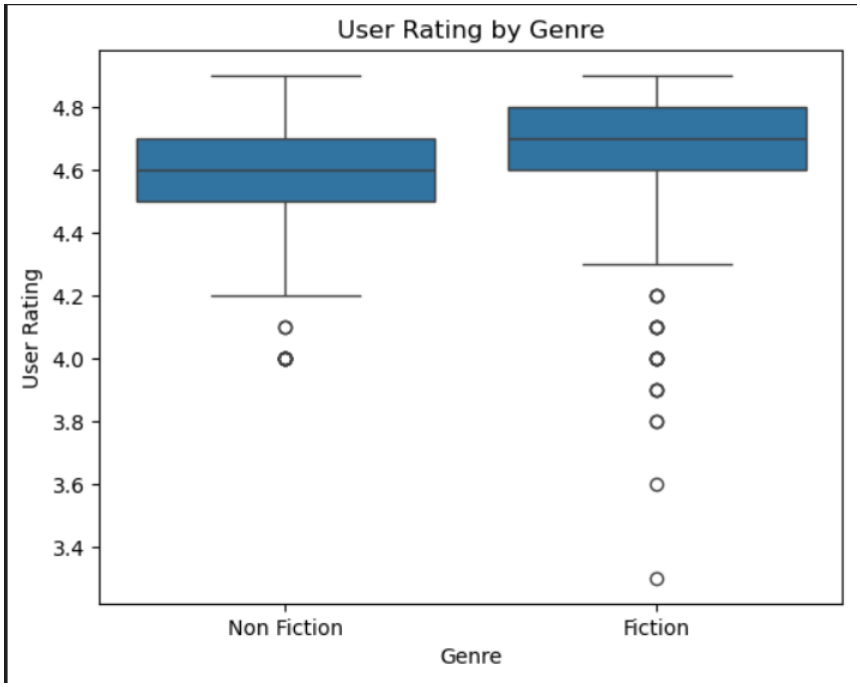
- **Pairplot** – Provides multi-variable visualization, allowing simultaneous examination of pairwise relationships between multiple numerical features while also showing their distributions. *(Insert Pairplot Image Here)*



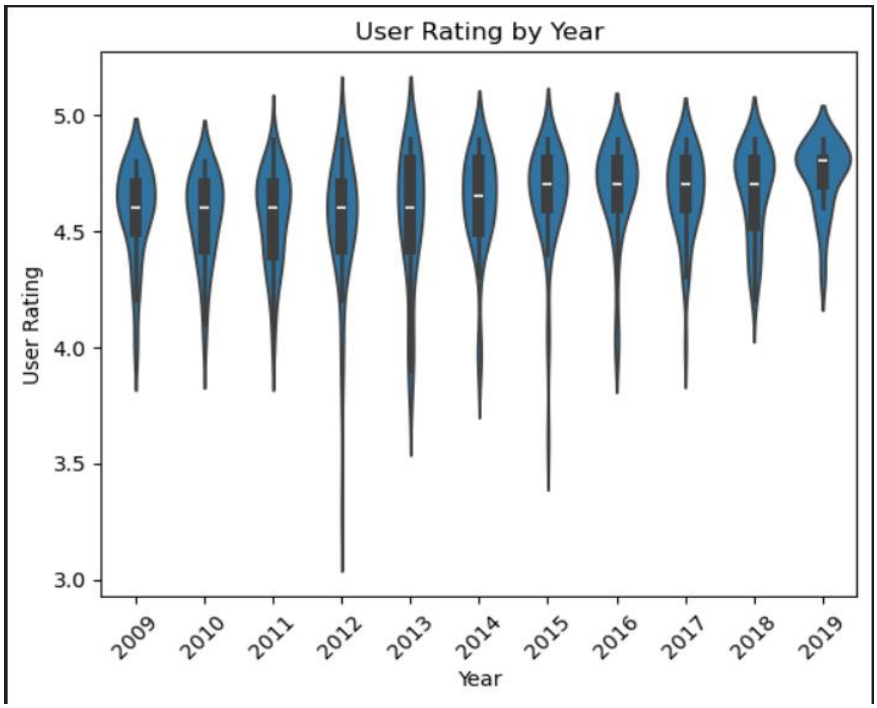
- **Histplot (Histogram)** – Shows the frequency distribution of key numerical features, such as *Book Prices* or *Ratings*, providing insights into data skewness and spread. *(Insert Histplot Image Here)*



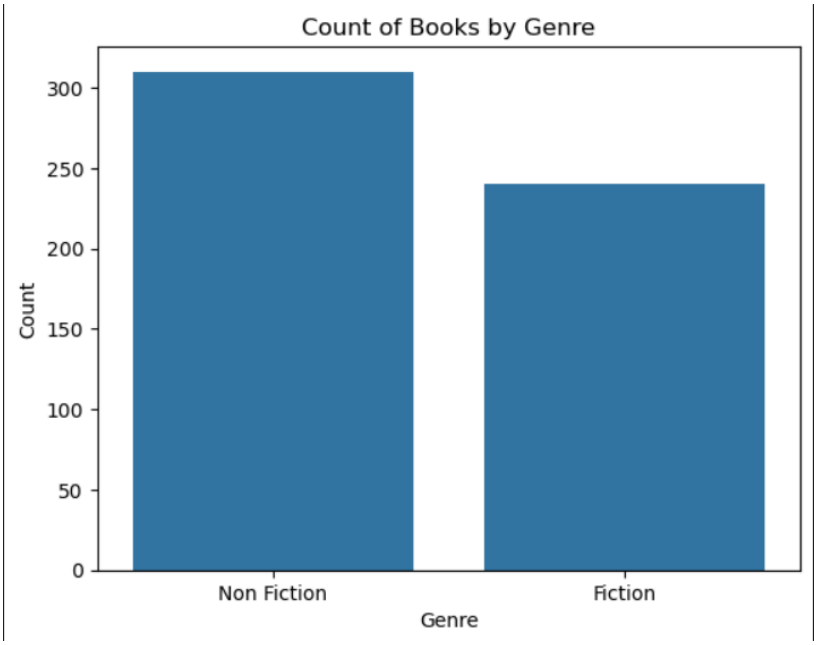
- **Boxplot** – Summarizes the distribution of numerical variables and identifies outliers. Useful for comparing price distributions across genres or rating categories. *(Insert Boxplot Image Here)*



- **Violinplot** – Combines boxplot and density estimation to visualize the full distribution of data, giving deeper insight into variation within each category. *(Insert Violinplot Image Here)*



- **Countplot** – Displays the count of categorical values, such as *Number of Books per Genre* or *Books per Author*, enabling quick comparisons of category sizes. *(Insert Countplot Image Here)*



By integrating these visualizations, the Python analysis not only reveals patterns and anomalies but also supports hypothesis formulation for predictive modeling. The visual outputs serve as an essential bridge between raw numerical data and actionable insights, guiding further stages such as feature selection and model training.

MODEL TRAINING AND CLASSIFICATION

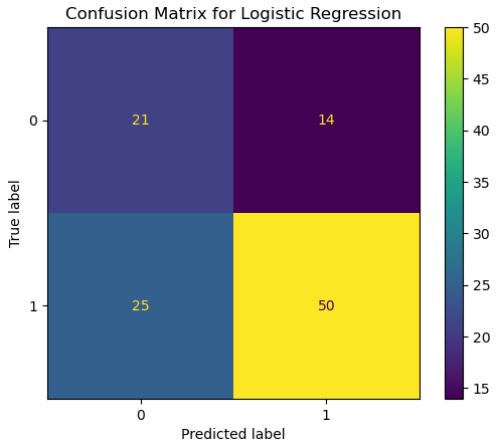
The **model training phase** builds on the foundation established during data preprocessing and exploratory analysis. The objective is to develop and evaluate multiple classification models to identify the most effective approach for predicting target outcomes in the Amazon **Top-Selling Books** dataset.

The process follows a systematic workflow:

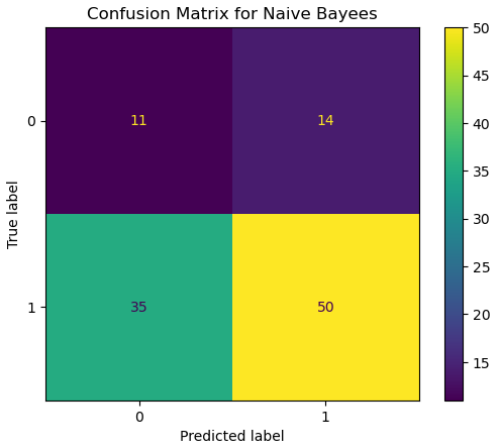
1. **Data Cleaning** –
    - Removed duplicate entries and irrelevant columns.
    - Handled missing values by either imputing with statistical measures (mean/median/mode) or removing incomplete records where appropriate.
2. **Data Preprocessing** –

- Ensured uniform formatting of categorical and numerical values.
- Removed inconsistencies such as extra spaces or capitalization errors in text features.

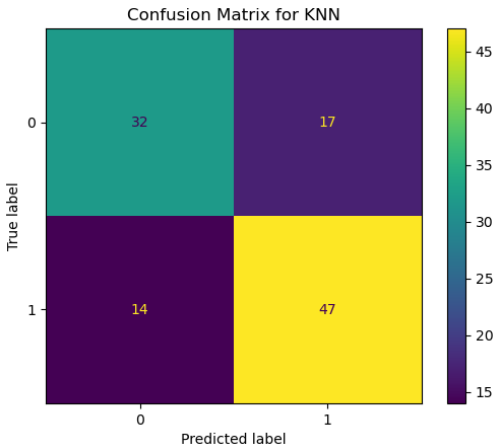
3. **Encoding** –
   - Applied **Label Encoding** or **One-Hot Encoding** to convert categorical variables (e.g., genre, author) into machine-readable numeric form.
4. **Feature Scaling** –
   - Used **StandardScaler** to normalize numerical features, ensuring equal weightage and improving algorithm performance.
5. **Input/Output Separation** –
   - **Input Variables (X)**: Selected relevant features such as price, ratings, reviews count, and genre indicators.
   - **Output Variable (y)**: Target classification label (e.g., category of sales performance or genre prediction).
6. **Train-Test Split** –
   - Divided the dataset into training and testing sets, typically in an **80:20** ratio, to evaluate model performance on unseen data.
7. **Model Training (Classification)** –
   Four supervised machine learning models were implemented using **scikit-learn**:
   - **Logistic Regression** – Applied as a baseline classification model due to its simplicity and interpretability.



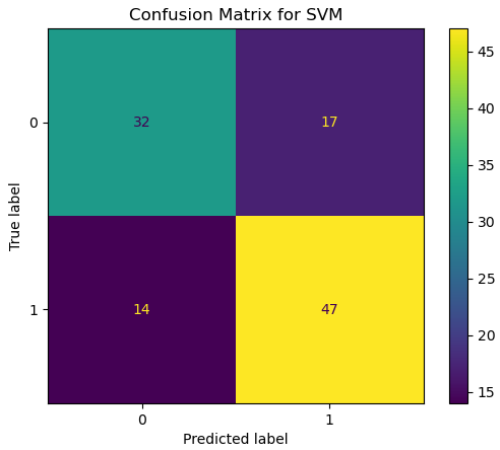
- **Gaussian Naive Bayes (GNB)** – Selected for its speed and effectiveness on smaller datasets with independent features.



- **K-Nearest Neighbors (KNN)** – Used for non-parametric classification based on feature similarity.



- **Support Vector Machine (SVM)** – Implemented with appropriate kernels to handle both linear and non-linear decision boundaries.



8. **Model Evaluation** –
    - Metrics used: **Confusion Matrix**, **Accuracy**, **Mean Squared Error (MSE)**, and **R^2 Score** (where applicable).
    - Visual evaluation: Confusion matrices for each model to assess classification accuracy and error distribution. *(Insert Confusion Matrix Images Here for Logistic Regression, GNB, KNN, and SVM)*

This structured approach ensures consistency, reproducibility, and clarity in comparing model performances. By evaluating multiple algorithms, the project identifies the model best suited for this dataset, balancing accuracy, computational efficiency, and generalization capability.

Model Performance Comparison

Model	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F1 Score (%)***
Logistic Regression	71.42	68.98	78.12	72.98
Gaussian Naive Bayes	65.45	66.45	78.12	67.11
K-Nearest Neighbors	71.81	77.09	73.43	75.20
Support Vector Machine	74.87	79.45	74.43	76.38

Conclusion

This project successfully analyzed the **Amazon Top-Selling Books** dataset using multiple tools and techniques, including **Excel**, **Power BI**, **Tableau**, and **Python-based exploratory data analysis**. The interactive dashboards in Power BI, Tableau, and Excel provided valuable visual insights into sales trends, genre distribution, top authors, and pricing patterns.

Python's data visualization capabilities, through scatterplots, heatmaps, pairplots, histograms, boxplots, violin plots, and countplots, enabled deeper exploration of correlations and distribution patterns. The preprocessing pipeline ensured data quality through cleaning, encoding, scaling, and splitting into training and testing sets.

Four machine learning models — **Logistic Regression**, **Gaussian Naive Bayes**, **K-Nearest Neighbors**, and **Support Vector Machine** — were trained and evaluated. Performance metrics such as **accuracy, precision, recall, and F1 score** allowed for an objective comparison of their effectiveness.

The results highlight that machine learning can be effectively applied to book sales data for classification tasks, enabling better prediction of trends and informed business decisions. The combination of interactive dashboards and predictive analytics demonstrates a complete, end-to-end approach to data analysis, offering a framework that can be adapted to similar e-commerce datasets in the future.