SAVITRIBAI PHULE PUNE UNIVERSITY

A MINI PROJECT REPORT ON

“Data Mining on Amazon Prime Movies and TV Shows Dataset using BI Tools”

Submitted by

Name: Tejas Hirurkar	**Roll no: A-59**
Name: Shreyash Jadhav	**Roll no: A-62**
Name: Utkarsh Khalkar	**Roll no: A-77**
Name: Pratik Patil	**Roll no: B-36**

CLASS: BE **DIV: A**

Under the Guidance of

Prof. Vanita Babanne

Sinhgad Institutes

DEPARTMENT OF COMPUTER ENGINEERING

RMD SINHGAD SCHOOL OF ENGINEERING

WARJE, PUNE 411058

2023 - 2024

DEPARTMENT OF COMPUTER ENGINEERING

RMD SINHGAD SCHOOL OF ENGINEERING

WARJE, PUNE 411058

CERTIFICATE

This is to certify that the project report entitles

“Data Mining on Amazon Prime Movies and TV Shows Dataset using BI Tools”

Submitted by

Name: Shreyash Jadhav PRN No: 72218241E

is a bonafide work carried out by them under the supervision of Prof. Vanita Babanne And it is submitted towards the partial fulfillment of the requirement of University of Pune for Fourth Year.

(Prof. Vanita Babanne)
Guide
Department of Computer Engineering

(Dr. Vina M. Lomte)
Head,
Department of Computer Engineering

(Dr. V. V. Dixit)
Principal,
RMD Sinhgad School of Engineering Pune – 58

Certificate by Guide

This is to certify that Mr. Shreyash Jadhav has completed the MINI Project work under my guidance and supervision and that, I have verified the work for its originality in documentation, problem statement, implementation and results presented in the Project. Any reproduction of other necessary work is with theprior permission and has given due ownership and included in the references.

Signature of Guide

Prof. Vanita Babanne

ACKNOWLEDGEMENT

It is our pleasure to acknowledge sense of gratitude to all those who helped us in making this project.

We thank our Mini Project Guide **Prof. Vanita Babanne** for helping us and providing all necessary information regarding our project.

We are also thankful to **Dr. Vina M. Lomte (Head - Department of Computer Engineering)** for providing us the required facilities and helping us while carrying out this project work.

Finally, we wish to thank all our teachers and friends for their constructive comments, suggestionsand criticism and all those directly or indirectly helped us in completing this project.

Tejas Hirurkar
Shreyash Jadhav
Utkarsh Khalkar
Pratik Patil

CONTENTS

ABSTRACT

Data mining on the Amazon Prime Movies and TV Shows dataset using Business Intelligence (BI) tools involves extracting valuable insights from a vast collection of media content available on the Amazon Prime platform. This process integrates data analysis techniques with BI tools to uncover patterns, trends, and relationships within the dataset. Key steps include data preprocessing, exploratory data analysis, feature engineering, and applying machine learning algorithms to derive meaningful information. The goal is to enhance decision-making processes, improve content recommendations, and optimize user experience on the Amazon Prime platform. Additionally, the analysis may focus on viewer preferences, content popularity, genre trends, and content performance metrics to drive strategic content creation and platform growth strategies. Moreover, by leveraging BI tools such as data visualization dashboards, interactive reports, and predictive analytics, stakeholders can gain actionable insights to refine marketing strategies, personalize content offerings, and enhance customer satisfaction, ultimately leading to increased user engagement and business success for Amazon Prime.

1. INTRODUCTION:

This report explores data mining and BI tools on Amazon Prime's media dataset. The goal is to extract insights for decision-making and enhance user experience. As streaming platforms grow, they gather extensive data on viewer behavior. This report discusses how data mining and BI tools uncover trends for growth and satisfaction.

We'll cover data preprocessing, analysis, and predictive modeling techniques on Amazon Prime's dataset. Also, we'll highlight BI tools like dashboards and machine learning. The report aims to provide recommendations for content optimization and user engagement.

1.1 Objectives

- To analyze Amazon Prime Movies and TV shows.
- To perform various mining tasks on the data.
- To find insights regarding the data.

1.2 Problem Statement

To create a BI report outlining the following steps:

- Problem definition, identifying which data mining task is needed.
- Identify and use a standard data mining dataset available for the problem.

2. REQUIREMENTS:

- Data: Amazon Prime Titles
- Tools: Power BI, Excel.
- Data Mining Tasks: Summarization, Frequent Item Set, Data Cleaning.

3. THEORY

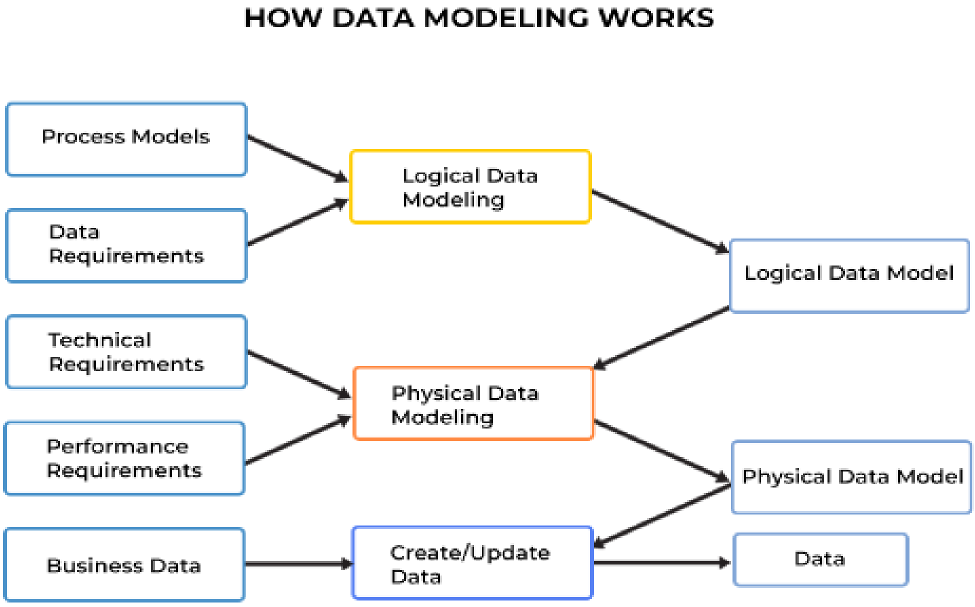
DATA MINING TASK:

To address the problem definition outlined earlier, several data mining tasks will be undertaken using the available sales dataset from the e-commerce platform. These tasks will involve extracting actionable insights and patterns from the data to inform decision-making and drive business growth. The specific data mining tasks include:

1. Descriptive Analysis: Conducting descriptive analysis to understand the basic characteristics of the dataset, such as summary statistics, data distribution, and missing values. This will provide an initial overview of the data and help identify any data quality issues that need to be addressed.
2. Segmentation Analysis: Performing segmentation analysis to group customers based on their purchasing behaviour, such as frequency of purchases, average order value, or preferred product categories. This will enable the identification of distinct customer segments for targeted marketing and personalized recommendations.
3. Association Rule Mining: Applying association rule mining techniques to uncover relationships between different products frequently purchased together. This will help identify cross-selling opportunities and inform product bundling strategies to maximize sales revenue.
4. Time Series Analysis: Conducting time series analysis to examine sales trends over time, including seasonality, trends, and cyclical patterns. This analysis will provide insights into the temporal dynamics of sales and help forecast future sales volumes to optimize inventory management and resource allocation.
5. Geospatial Analysis: Utilizing geospatial analysis techniques to visualize sales data on a map and identify regional variations in sales performance. This analysis will help pinpoint areas of high sales activity and potential expansion opportunities, as well as areas that may require targeted marketing efforts to boost sales.
6. Profitability Analysis: Analysing profitability metrics, such as profit margins and return on investment, to identify the most profitable product categories, customer segments, and marketing channels. This analysis will inform strategic decisions on resource allocation and pricing strategies to maximize profitability.

By undertaking these data mining tasks, we aim to extract actionable insights from the e-commerce sales dataset that will drive informed decision-making and enable the e-commerce platform to optimize sales performance, enhance customer satisfaction, and achieve sustainable growth in a competitive market landscape.

MODEL BUILDING:



Before diving into the data mining tasks outlined in the previous section, it's essential to conduct thorough data exploration and preprocessing to ensure the quality and usability of the dataset. This involves several key steps:

1. Data Cleaning: Identify and handle missing values, outliers, and inconsistencies in the dataset. This may involve imputing missing values, removing outliers, and standardizing or normalizing data where necessary to ensure consistency and accuracy in the analysis.
2. Feature Selection: Evaluate the relevance and importance of each feature (e.g., sales amount, quantity, customer name, payment mode, etc.) in relation to the problem definition. Select the most relevant features for analysis while discarding irrelevant or redundant ones to simplify the dataset and improve model performance.

3. Data Transformation: Transform categorical variables into numerical representations using techniques such as one-hot encoding or label encoding. This allows categorical variables to be used in machine learning algorithms effectively.
4. Feature Engineering: Create new features or derive additional insights from existing features to enhance the predictive power of the dataset. This may involve aggregating or combining features, creating interaction terms, or extracting meaningful information from datetime variables (e.g., month, day of week).
5. Data Visualization: Visualize key aspects of the dataset using various graphical techniques such as histograms, box plots, scatter plots, and correlation matrices. This helps identify patterns, trends, and relationships within the data, guiding further analysis and model development.
6. Data Splitting: Split the dataset into training, validation, and testing sets to evaluate the performance of machine learning models accurately. This ensures that the model's performance is assessed on unseen data, reducing the risk of overfitting.
7. Handling Imbalanced Data (if applicable): Address any imbalance in the distribution of target classes by employing techniques such as oversampling, under sampling, or using algorithms specifically designed to handle imbalanced data.

By conducting thorough data exploration and preprocessing, we aim to prepare the dataset for analysis, ensuring that it is clean, well-structured, and ready for use in the data mining tasks outlined in the project. This process lays the foundation for accurate and reliable insights to be extracted from the e-commerce sales data.

POWER BI DASHBOARD DESIGN:

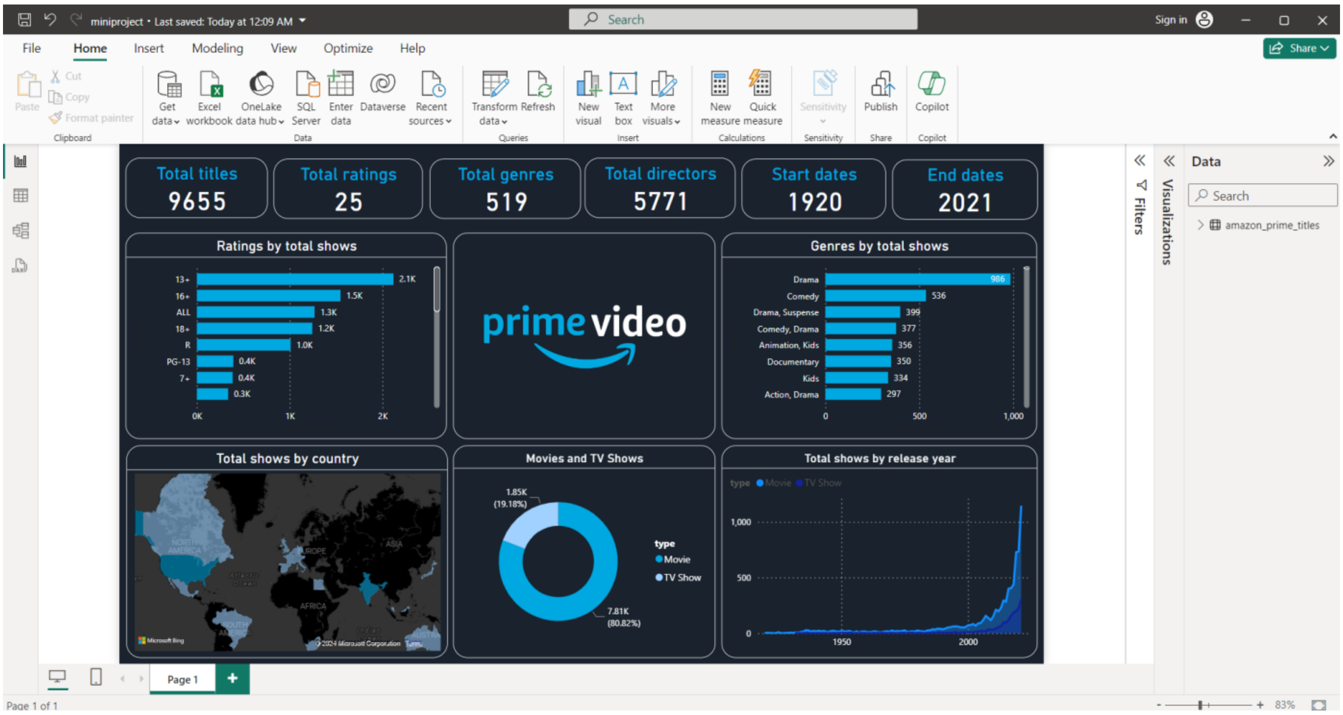
Creating a Power BI dashboard involves designing an interactive and visually appealing interface to present key insights and metrics derived from the e-commerce sales dataset. Here's an outline of the dashboard design process:

1. Data Source Connection: Connect Power BI to the e-commerce sales dataset to import the relevant data for analysis. This may involve connecting to a database, Excel file, or other data sources.
2. Data Preparation: Clean and preprocess the data within Power BI, including handling missing values, transforming data types, and creating calculated columns or measures as needed for analysis.
3. Dashboard Layout: Design the layout of the dashboard to include multiple visualizations that provide a comprehensive overview of sales performance. Arrange visualizations logically to guide the user's attention and facilitate intuitive navigation.

4. Key Performance Indicators (KPIs): Include KPI visualizations to highlight important metrics such as total sales revenue, average order value, conversion rate, and profitability. These KPIs serve as high-level indicators of the e-commerce platform's performance.
5. Charts and Graphs: Incorporate various types of charts and graphs to visualize different aspects of sales data, such as bar charts for sales by product category, line charts for sales trends over time, and pie charts for distribution of sales by region. Choose visually appealing and informative visualizations that effectively communicate insights.
6. Filters and Slicers: Implement filters and slicers to enable users to interactively explore the data and drill down into specific segments or time periods of interest. This allows users to customize their analysis and gain deeper insights based on their preferences.
7. Map Visualizations: Utilize map visualizations to display sales data geographically and identify regional sales patterns or hotspots. This helps identify opportunities for expansion or targeted marketing efforts in specific regions.
8. Dashboard Interactivity: Enable interactivity within the dashboard, such as cross-filtering and highlighting, to facilitate dynamic exploration of the data. Users should be able to interact with the visualizations seamlessly to gain insights and answer questions on-the-fly.
9. Customization and Branding: Customize the appearance of the dashboard to align with the e-commerce platform's branding guidelines. This may include using specific color schemes, logos, and fonts to maintain consistency with the platform's visual identity.
10. Testing and Optimization: Test the dashboard thoroughly to ensure functionality across different devices and screen sizes. Optimize performance by minimizing load times and maximizing responsiveness, especially for large datasets or complex visualizations.

By following these steps, you can design a compelling Power BI dashboard that effectively communicates insights from the e-commerce sales dataset and empowers users to make data-driven decisions to improve sales performance and drive business growth.

4. OUTPUT (Screenshot of Implementation)



5. CONCLUSION:

In conclusion, data mining and BI tools have uncovered valuable insights for decision-making on Amazon Prime. Leveraging these tools has led to recommendations for content optimization and user engagement. Moving forward, ongoing analysis will be crucial for staying competitive in the digital streaming market and enhancing customer satisfaction on the platform.