PUNE INSTITUTE OF COMPUTER TECHNOLOGY

Department of Computer Engineering
(2024-2025)

LP-VI

Batch:- R3

BI Mini Project

Performing Data Mining Tasks on Adventure Works 2019 Dataset

Submitted By -

41352 - Advait Naik

41353 - Kaustubh Netke

41363 - Rhea Shah

Guided By- Prof. Hema Kumbhar

TABLE OF CONTENTS

1. ABSTRACT

This mini project presents a comprehensive application of Business Intelligence (BI) and data mining principles using the Adventure Works 2019 Data Warehouse. The primary objective is to explore how various data mining tasks—such as summarization, frequent item set discovery, and data cleaning—can be effectively applied to extract meaningful insights from a large, multi-domain dataset. Adventure Works 2019 simulates data from a fictional manufacturing company and includes detailed information across multiple dimensions such as sales, products, customers, employees, and time.

By leveraging tools like SQL Server for data querying and Power BI for visualization, this project illustrates how structured and semi-structured data can be transformed into intuitive, actionable reports. Key insights derived from the dataset include trends in internet sales over time, high-performing products, salesperson-based performance comparisons, and budget evaluations. These insights not only help understand the functionality of BI tools but also simulate how real-world business decisions are supported by data.

The results of this mini project validate the effectiveness of BI in uncovering hidden patterns and trends that may not be apparent through traditional data inspection. This hands-on implementation bridges theoretical knowledge and practical application, offering a valuable learning experience in the realm of data analysis, visualization, and business intelligence.

2. INTRODUCTION

In today's data-driven economy, organizations are increasingly relying on Business Intelligence (BI) to gain competitive advantages by making informed decisions. BI encompasses a wide array of technologies, tools, and methodologies aimed at collecting, analyzing, and presenting business data in an actionable format. This mini project introduces the application of BI and data mining tasks on the Adventure Works 2019 Data Warehouse—a rich, multi-dimensional dataset that represents operations of a fictional bicycle manufacturing company called Adventure Works Cycles.

The dataset is widely recognized as an industry-standard sample used for learning and demonstration purposes. It consists of several fact and dimension tables capturing data across sales, production, customer demographics, marketing campaigns, and financial transactions. The project uses this dataset to identify insights that could be used by companies to improve their strategic planning and operational performance.

Using Microsoft SQL Server and Power BI, the project focuses on practical implementations such as summarizing sales data, identifying frequent itemsets, and cleaning inconsistent data entries. The project also emphasizes visual storytelling through dashboards and interactive charts, which aid in intuitive decision-making.

3. PROBLEM STATEMENT

Organizations today generate massive volumes of data from various sources such as sales transactions, customer interactions, marketing campaigns, and employee activities. However, the mere availability of this data does not ensure meaningful insights. The real challenge lies in converting this vast amount of raw, unstructured, and often noisy data into information that can guide strategic decisions. This is where Business Intelligence and data mining come into play.

The problem addressed in this mini project is to apply essential data mining tasks on a comprehensive dataset—the Adventure Works 2019 Data Warehouse—in order to extract actionable insights. Specifically, we aim to tackle the following challenges:

- Identify and clean inconsistent or missing data entries.

- Perform summarization tasks to provide an overview of sales, customers, and employees.

- Discover frequently occurring product combinations using frequent itemset mining.

- Generate interactive visualizations for better understanding of sales trends, performance metrics, and budget tracking.

4. MOTIVATION

The rapid advancement of digital technologies has led to a data explosion across industries. From customer preferences and purchasing patterns to supply chain metrics and financial data, businesses now collect more information than ever before. However, without effective tools to analyze and interpret this data, valuable insights remain hidden. This situation presents an opportunity and a necessity to explore Business Intelligence (BI) and data mining technologies, which transform raw data into meaningful narratives.

The motivation behind this project stems from the need to bridge the gap between data collection and actionable decision-making. The Adventure Works 2019 Data Warehouse was chosen because of its realistic structure and diversity of data, simulating a real enterprise environment. It allows learners and professionals to engage with complex relationships across sales, marketing, human resources, and production—all within a controlled setting.

Furthermore, tools like Power BI and SQL Server are industry-standard platforms that offer robust capabilities for querying, transformation, and visualization. By working hands-on with these tools, students gain practical experience in managing large datasets, performing key data mining tasks, and creating intuitive dashboards.

5. LITERATURE SURVEY

1. Han, Kamber, and Pei (2011) – In their book *Data Mining: Concepts and Techniques*, the authors outline core data mining tasks such as data cleaning, summarization, and frequent pattern mining. These concepts were applied in our project to extract insights from the Adventure Works 2019 dataset.

2. W. H. Inmon (2005) – Known as the "father of data warehousing," Inmon's work *Building the Data Warehouse* discusses top-down warehouse design and schema modeling. His principles helped us understand how the Adventure Works dataset uses star and snowflake schemas.

3. Kimball and Ross (2013) – In *The Data Warehouse Toolkit*, the authors explain dimensional modeling and business-focused data marts. Their bottom-up approach complements Inmon's and provided clarity for understanding fact-dimension relationships within the dataset.

4. Microsoft Adventure Works Dataset – The Adventure Works 2019 dataset is a sample data warehouse used for demonstrations and learning. It offers realistic business data across departments like sales, production, and finance. It is widely used in academic and industry training.

5. Microsoft Documentation – Official documentation on SQL Server and Power BI was used to guide the technical implementation, including querying, data modeling, and dashboard creation.

6. Eckerson (2007) – In *Performance Dashboards*, Eckerson emphasizes the value of visual BI tools to track business metrics. This aligns with our use of Power BI to present sales trends, budget summaries, and product performance.

6. SUMMARY OF PROPOSED APPROACH

The proposed approach involves conducting a structured BI and data mining process using the Adventure Works 2019 dataset to extract valuable business insights. The approach is divided into well-defined stages starting from data acquisition and understanding, followed by cleaning and preprocessing, and finally data analysis and visualization.

The dataset used in this project is a comprehensive SQL Server-based data warehouse comprising fact and dimension tables that span multiple business functions such as sales, marketing, and production. The initial step involves restoring the .bak file into SQL Server Management Studio (SSMS), followed by exploratory data analysis to understand the structure, relationships, and anomalies in the data.

Once familiarized with the dataset, data mining tasks such as summarization, frequent itemset identification, and data cleaning are performed. Summarization is done using SQL aggregation functions and pivot tables in Power BI. Frequent itemsets are analyzed to uncover patterns in product purchases and customer preferences. Data cleaning is also critical, involving the identification and handling of null values, duplicates, and inconsistent data entries.

Power BI is then used to build interactive dashboards representing internet sales trends, customer distribution, sales personnel filtering, and budget comparisons over time. This visualization layer provides a high-level overview of the business data and allows for deeper insight into various KPIs.

This systematic and modular approach ensures the effective application of theoretical concepts in a real-world-like setting.

7. METHODOLOGY

The methodology for this mini project is designed to ensure a step-by-step implementation of data mining and business intelligence tasks using industry-standard tools. The process is broken down into several stages:

1. Data Source Setup:
   The first step involves downloading and restoring the AdventureWorksDW2019.bak file into SQL Server Management Studio. This sets up the data warehouse environment, enabling access to various fact and dimension tables.

2. Exploratory Data Analysis (EDA):
   EDA is performed using SQL queries to understand the structure of tables such as FactInternetSales, DimProduct, DimCustomer, and DimEmployee. Relationships among tables are reviewed to plan appropriate joins and data aggregations.

3. Data Cleaning:
   This includes identifying null values, removing duplicates, and correcting data inconsistencies. SQL functions such as ISNULL, GROUP BY, and HAVING are used for data validation.

4. Data Mining Tasks:

   ○ Summarization: Aggregating data to get total sales, revenue trends, and customer demographics.

   ○ Frequent Itemset: Identifying commonly purchased products or services using count-based filtering and frequency distribution.

   ○ Transformation: Creating calculated columns and measures to support analytical operations.

5. Data Visualization in Power BI:

Cleaned and transformed datasets are loaded into Power BI, where various charts (bar, pie, line), slicers, and filters are applied to create dynamic dashboards. These dashboards enable business users to interact with data in real time.

This methodology ensures reproducibility, clarity, and effectiveness in extracting meaningful patterns from large datasets.
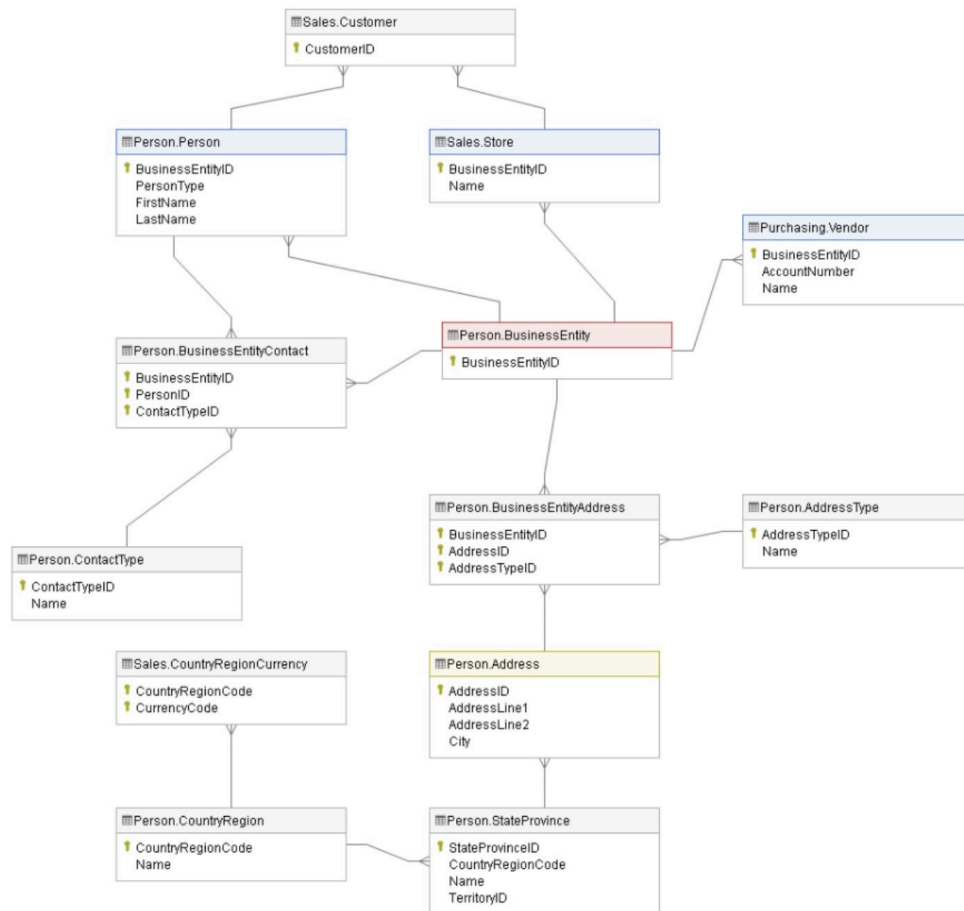
8. CONCLUSION

This mini project demonstrates the practical application of Business Intelligence and data mining concepts using a rich and structured dataset—Adventure Works 2019. Throughout the project, we have taken a stepwise approach starting from setting up the data warehouse, cleaning and exploring the data, to applying mining techniques and visualizing insights using Power BI.

The project achieved its objective of applying key data mining tasks like summarization, frequent itemset analysis, and data cleaning to uncover trends in internet sales, product preferences, and employee contributions. These findings were then visualized through interactive dashboards that simulate real-world business reporting environments.

9. REFERENCES

[1] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.

[2] Inmon, W. H. (2005). *Building the Data Warehouse*. Wiley.

[3] Microsoft Docs. (2024). SQL Server Documentation

[4] Microsoft Docs. (2024). Power BI Documentation

[5] GitHub – Microsoft. (n.d.). AdventureWorks Sample Databases

[6] Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley.

[7] Eckerson, W. (2007). *Performance Dashboards: Measuring, Monitoring, and Managing Your Business*. Wiley.