

**SENG8080-23F-Sec2: Case Studies Big Data**

Analyzing Walmart Sales

Group 5

Abhishek Srivastava (8899988)

Karmvir Singh (8899422)

Amandeep Kaur (8900569)

December 9th, 2023

# Abstract:

This report focuses on the analysis and forecasting of sales data for Walmart, one of the largest retail chains in the world. The study aims to understand sales trends and make accurate forecasts to help optimize inventory and sales strategies. We utilize data from Walmart's historical sales records, encompassing various stores, departments, and other relevant attributes. Our analysis incorporates data preprocessing, exploratory data analysis, regression models, and time series forecasting to provide valuable insights into Walmart's sales patterns.

Table of Contents

[Abstract: 2](#_Toc150030563)

[Introduction: 4](#_Toc150030564)

[Data Research and Integration 4](#_Toc150030565)

[Data Collection 4](#_Toc150030566)

[Data Storage and Maintenance 4](#_Toc150030567)

[Data Quality 4](#_Toc150030568)

[Data Analysis and Visualization 5](#_Toc150030569)

[Sales Forecasting 5](#_Toc150030570)

[Tools and Libraries 5](#_Toc150030571)

[Proposed Allocation Project Team Roles 6](#_Toc150030572)

[Project Timeline 6](#_Toc150030573)

[References: 7](#_Toc150030574)

# Introduction:

The retail industry is highly competitive, and understanding sales patterns is crucial for making informed decisions. In this case study, we focus on analyzing the sales data of Walmart. Walmart operates numerous stores, and its sales data is rich with information that can help optimize inventory management, marketing strategies, and more. Our objective is to perform a comprehensive analysis of this data and create accurate sales forecasts.

# Data Research and Integration

For this assignment, we will use historical sales data provided by Walmart. The data will be collected from Walmart's database or data sources provided for analysis. Additionally, external data sources like economic indicators, holidays, and weather data may be integrated to enhance the accuracy of the forecasts.

# Data Collection

We will collect sales data from Walmart's database using SQL queries or other appropriate methods. The data will be cleaned and preprocessed to handle missing values and outliers.

# Data Storage and Maintenance

The data will be stored in a suitable format, such as a relational database or a data warehouse, to ensure efficient data retrieval and analysis. Regular maintenance of the data storage system will be performed to keep the data up-to-date.

# Data Quality

To maintain data quality, we will use Python libraries such as Pandas for data cleaning, transformation, and merging. This ensures that the data used for analysis is of high quality, consistent, and suitable for in-depth analysis and visualization.

# Data Analysis and Visualization

We will use Python libraries such as Pandas for data manipulation, Matplotlib and Seaborn for data visualization, and NumPy for numerical operations. These tools will help us gain insights into sales data trends and patterns.

# Sales Forecasting

Sales forecasting is a crucial aspect of this assignment. We will use tools and libraries like Scikit-Learn for regression models, Stats models for time series analysis, and Facebook Prophet for time-dependent pattern forecasting. These tools will enable us to make accurate sales forecasts.

# Tools and Libraries

The following tools and libraries will be used for this assignment:

1. Python: The primary programming language for data analysis and modeling.

2. Pandas: Used for data manipulation and analysis.

3. NumPy: For numerical operations and mathematical functions.

4. Matplotlib and Seaborn: Helpful for data visualization.

5. Scikit-Learn: Provides tools for building regression models and time series forecasting models.

6. Statsmodels: Useful for statistical modeling and hypothesis testing.

7. Prophet: A forecasting tool designed for time series data.

8. Jupyter Notebook: An interactive environment for data analysis and code execution.

9. SQL (Structured Query Language): If needed for working with databases.

# Proposed Allocation Project Team Roles

1. Data Analysis and Visualization:

2. Data Storage and Maintenance:

3. Data Quality:

4. Data Research and Integration:

# Project Timeline

|  |  |  |
| --- | --- | --- |
| **Date** | **Deliverable** | **Responsible** |
| Oct 20 | Data Collected and planned | All members |
| Nov 4 | 1st Draft Circulated to Team | All members |
| Nov 11 | 1st Draft of Presentation Circulated | All members |
| Nov 11 | User testing by the team and errors/refinements identified. | All members |
| Nov 17 | Final Adjustments made and checked | All members |
| Nov 18 | Process and Report Due at 10pm | All members |

GitHub Repository Link: <https://github.com/Amandeep0569/SENG8080-23F-Sec2-Final-Project>

# References:

1. Pandas Documentation: [*https://pandas.pydata.org/docs/*](https://pandas.pydata.org/docs/)
   * The official documentation for the Pandas library, which is a valuable resource for learning how to work with data in Python.
2. Matplotlib Documentation: [*https://matplotlib.org/stable/contents.html*](https://matplotlib.org/stable/contents.html)
   * The official documentation for the Matplotlib library, which is essential for creating data visualizations.
3. Prophet Documentation:[*https://facebook.github.io/prophet/docs/*](https://facebook.github.io/prophet/docs/)
   * The official documentation for Prophet, a time series forecasting library developed by Facebook.
4. Scikit-Learn Documentation:[*https://scikit-learn.org/stable/documentation.html*](https://scikit-learn.org/stable/documentation.html)
   * The official documentation for Scikit-Learn, which is used for machine learning and evaluating forecasting models.
5. Jupyter Notebook Documentation: [*https://jupyter-notebook.readthedocs.io/en/stable/*](https://jupyter-notebook.readthedocs.io/en/stable/)
   * Documentation for Jupyter Notebook, a popular tool for interactive data analysis and code presentation.
6. Walmart Dataset: [*https://www.kaggle.com/datasets/yasserh/walmart-dataset?rvi=1*](https://www.kaggle.com/datasets/yasserh/walmart-dataset?rvi=1)

* This Kaggle competition provides a real-world dataset and challenges related to Walmart sales forecasting.