



Data Mining Project

Submitted to: Eyob Niguse

24 - 05 - 2024

Addis Ababa, Ethiopia

[GitHub Link](#)

Group Members:

	Name	ID
1.	Leul Tewodros	XM7276
2.	Simon Asaye	FK4729
3.	Yoftahe Dereje	BF3030
4.	Yohannes Negash	P00635

Table of Contents

1. Background/Overview	2
2. Problem Statement	3
3. Objective of the Project	3
4. Approach/Methodology	4
5. Proposed Solution:	5
a) Actual Preprocessing	5
b) Experimentation, Modeling, or Design of the Solution	7
Summary of Advanced Descriptive Analysis.....	8
Association Rule Graph.....	9
Linear Regression Graph.....	10
Unsumarized Data prediction model graph.....	10
6. Evaluation of the Project	11
1. Data Preprocessing:	11
2. Linear Regression for Sales Forecasting:	11
3. Time Series Analysis:	11
4. Association Rule Mining:	11
Overall Evaluation:.....	11
7. Conclusion and Recommendation.....	12
Conclusion:	12
Recommendations for Future Work:.....	12

1. Background/Overview

Introduction to Axum Pharmacy

Axum Pharmacy is a well-established retail pharmacy located in the heart of Addis Ababa. Known for its wide range of pharmaceutical products and exceptional customer service, Axum Pharmacy has been serving the community for over a decade. The pharmacy caters to a diverse clientele, providing prescription medications, over-the-counter drugs, health supplements, and medical supplies. Its commitment to quality and customer satisfaction has made it a trusted name in the healthcare sector.

Overview of the Pharmacy Sales Industry

The pharmacy sales industry is a critical component of the healthcare system, contributing significantly to public health and wellness. It encompasses the retail sale of prescription medications, over-the-counter drugs, and other health-related products. The industry is characterized by its reliance on accurate inventory management, customer relationship management, and the ability to predict sales trends to ensure the availability of essential medications. The competitive landscape of the pharmacy sales industry necessitates efficient operations and data-driven decision-making to meet customer needs and maintain profitability.

Importance of Data Mining in Pharmacy Sales

In the contemporary business environment, data mining has emerged as a powerful tool for extracting valuable insights from large datasets. In the context of pharmacy sales, data mining can play a pivotal role in enhancing business operations and customer satisfaction. By analyzing sales data, pharmacies can:

- Predict sales fluctuations and optimize inventory levels to prevent stockouts and overstock situations.
- Identify customer purchasing patterns and preferences, enabling personalized marketing and improved customer service.
- Detect emerging trends in medication demand, allowing for proactive adjustments in stock and promotional strategies.
- Enhance the overall efficiency of the pharmacy through better resource allocation and strategic planning.

The application of data mining techniques in pharmacy sales not only improves operational efficiency but also contributes to better healthcare outcomes by ensuring that customers have timely access to the medications they need. The dataset used in this project consists of detailed transactional records from a pharmacy, including information such as invoice numbers, transaction dates, quantities of items sold, total amounts, item codes, item descriptions, branch codes, shop names, customer names, salesperson codes, sales person names, cashier names, cashier codes, sales types, and unit prices. The comprehensive nature of this dataset provides a robust foundation for in-depth analysis.

2. Problem Statement

Axum Pharmacy faces significant challenges in managing inventory and understanding customer buying habits, which are crucial for maintaining operational efficiency and customer satisfaction. The dynamic nature of the pharmaceutical market—driven by factors such as seasonal illnesses, new product launches, and shifting customer preferences—complicates the prediction of sales trends. This often results in stockouts or overstock situations, leading to customer dissatisfaction, increased costs, and operational inefficiencies.

Additionally, the vast amount of transactional data available is underutilized, which hampers the pharmacy's ability to gain insights into customer behavior. Key issues include:

- **Inventory Management:**

- Overstocking and Understocking: Inefficient stock levels increase costs and result in lost sales.

- Demand Forecasting: Inaccurate predictions hinder effective inventory planning.

- **Customer Behavior Insights:**

- Lack of Personalization: Without understanding purchasing patterns, opportunities for targeted marketing and enhanced customer loyalty are missed.

- Product Affinity Identification: Identifying which products are frequently bought together can improve product placement and promotional strategies.

Addressing these challenges through data mining techniques can improve inventory management and personalized marketing, ultimately driving efficiency and growth for Axum Pharmacy.

3. Objective of the Project

To leverage data mining techniques to *forecast future pharmacy sales, uncover patterns in customer purchasing behavior, and enhance inventory management*. This will enable the creation of targeted marketing strategies, develop predictive models for accurate sales forecasts, and ultimately drive operational efficiency and improve customer satisfaction.

4. Approach/Methodology

This project employs a structured approach, utilizing various data mining techniques to achieve the objectives. The methodology consists of the following key steps:

1. Data Collection and Preparation

- **Data Source:** The dataset is sourced from Axum Pharmacy's point-of-sale (POS) system, providing comprehensive details of each transaction, including invoice numbers, dates, quantities, amounts, and item descriptions.
- **Initial Exploration:** Load and explore the dataset to understand its structure and contents.
- **Data Cleaning:** Handle missing values, remove duplicates, and correct any inconsistencies in the data.
- **Data Transformation:** Convert and format the data as needed, including date formatting and feature engineering.

2. Exploratory Data Analysis (EDA)

- **Descriptive Statistics:** Generate summary statistics to understand the distribution and characteristics of the data.
- **Visualizations:** Create visualizations such as histograms, box plots, and time series plots to identify trends and patterns.

3. Predictive Modeling

- **Linear Regression:** Develop linear regression models to forecast future sales based on historical data.
- **Time Series Analysis:** Apply time series models like ARIMA and Exponential Smoothing to capture temporal patterns and seasonal trends in sales data.

4. Association Rule Mining

- **Apriori Algorithm:** Use the Apriori algorithm to identify frequent itemsets and generate association rules, revealing product affinities and customer purchasing patterns.

5. Evaluation

- **Model Performance:** Evaluate the predictive models using metrics such as Mean Squared Error (MSE), R-squared, and Mean Absolute Error (MAE).
- **Rule Significance:** Assess the significance and usefulness of the association rules based on support, confidence, and lift.

6. Implementation Tools

- **Python Libraries:** Utilize libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, Scikit-learn for machine learning, and mlxtend for association rule mining.
- **Google Colab:** Utilized for its cloud-based accessibility, collaboration features, and integration with Google Drive. It also offers free access to GPUs and TPUs for computational tasks.
- **NumPy** (numerical computations), **Statsmodels**(statistical modeling, time series analysis)

Additional Considerations: The structured approach ensured that each step built upon the previous one, facilitating a comprehensive analysis of the sales data. This allowed for accurate forecasting, identification of purchasing patterns, and development of strategies to improve inventory management and customer satisfaction.

5. Proposed Solution:

a) Actual Preprocessing

To ensure the quality and usability of the pharmacy dataset for further analysis, data preprocessing was a critical step. Below are the preprocessing tasks that were performed on the pharmacy dataset:

- The first few rows of the dataset provide an initial overview of the data structure and contents.

```
[ ] # Display the last few rows of the dataset
df.info()

→ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 347510 entries, 0 to 347509
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Invoice Number    347510 non-null   object 
 1   Invoice Date      347510 non-null   int64  
 2   Quantity          347510 non-null   float64
 3   Amount            347510 non-null   float64
 4   Item Code          347510 non-null   int64  
 5   Item Description   347510 non-null   object 
 6   Branch Code        347510 non-null   object 
 7   Shop Name          347510 non-null   object 
 8   Customer Name     347510 non-null   object 
 9   Sales Person Code 347510 non-null   int64  
 10  Sales Person Name 347510 non-null   object 
 11  Cashier Name      347510 non-null   object 
 12  Cashier Code       347510 non-null   object 
 13  Sales Type         347510 non-null   object 
 14  Unit Price         347510 non-null   float64
dtypes: float64(3), int64(3), object(9)
memory usage: 39.8+ MB
/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283:
and should_run_async(code)
```

The date column in our dataset contains erroneous data types. The values are stored as integers (e.g., 45437) instead of the intended date format (e.g., "03/04/2021"). This data type inconsistency can lead to incorrect data interpretation and challenges in performing date-based operations and analysis.

1. Identify the Encoding:

- Examine the integer values closely to identify any patterns or encoding rules

2. Convert to Datetime:

- Use programming libraries or built-in functions to convert the integer values to Python's datetime objects.
- Perform necessary mathematical operations or string manipulations based on the identified encoding rules.

3. Handle Edge Cases:

- Check for any exceptional cases or invalid values.
- Implement logic to handle these edge cases gracefully, such as assigning null or default values.

4. Validate and Test:

- Verify the converted date values manually or through automated testing.
- Ensure that the dates are consistent and align with the intended format.

5. Update the Column:

- Replace the existing integer column with the new date column in your dataset.
- Update any downstream analysis or visualizations to reflect the corrected date values.

By converting the integer values to proper date values, we ensure data integrity and facilitate accurate analysis. This step is crucial for maintaining data quality and reliability in our dataset.

Data Types of Each Column After Conversion:

Invoice Number	object
Invoice Date	datetime64[ns]

Transform the Attributes

Give the write and correct data types for all columns

	Invoice Number	Invoice Date	Quantity	Amount	Item Code	
0	CS-01-021567/21	2021-01-01	1.0	12.00	18018	13
1	CS-01-021585/21	2021-01-01	2.0	21.74	36041	
2	CS-01-021585/21	2021-01-01	4.0	43.48	36041	
3	CS-01-021534/21	2021-01-01	1.0	12.00	19161	
4	CS-01-021529/21	2021-01-01	1.0	99.00	28529	

Remove all the inconsistent Data

We have faced negative number in the sales data in the unit price and Amount.

Inconsistent Data:						
	Invoice Number	Invoice Date	Quantity	Amount	Item Code	\
30802	1568215V	2021-09-17	1.0	-95.00	22047	
30803	1568215V	2021-09-17	1.0	-120.00	34406	
30804	1568215V	2021-09-17	1.0	-35.00	23031	
30805	1568215V	2021-09-17	1.0	-35.00	23031	
30806	1568215V	2021-09-17	1.0	-204.00	26248	
...	
344939	1571125V	2021-12-06	1.0	-229.57	30452	
344940	1571125V	2021-12-06	1.0	-52.00	11135	
344993	1571125V	2021-12-06	1.0	-311.00	33013	
344994	1571125V	2021-12-06	1.5	-69.00	11245	
344995	1571125V	2021-12-06	1.0	-180.00	34402	

Summary of Data Cleaning and Preprocessing

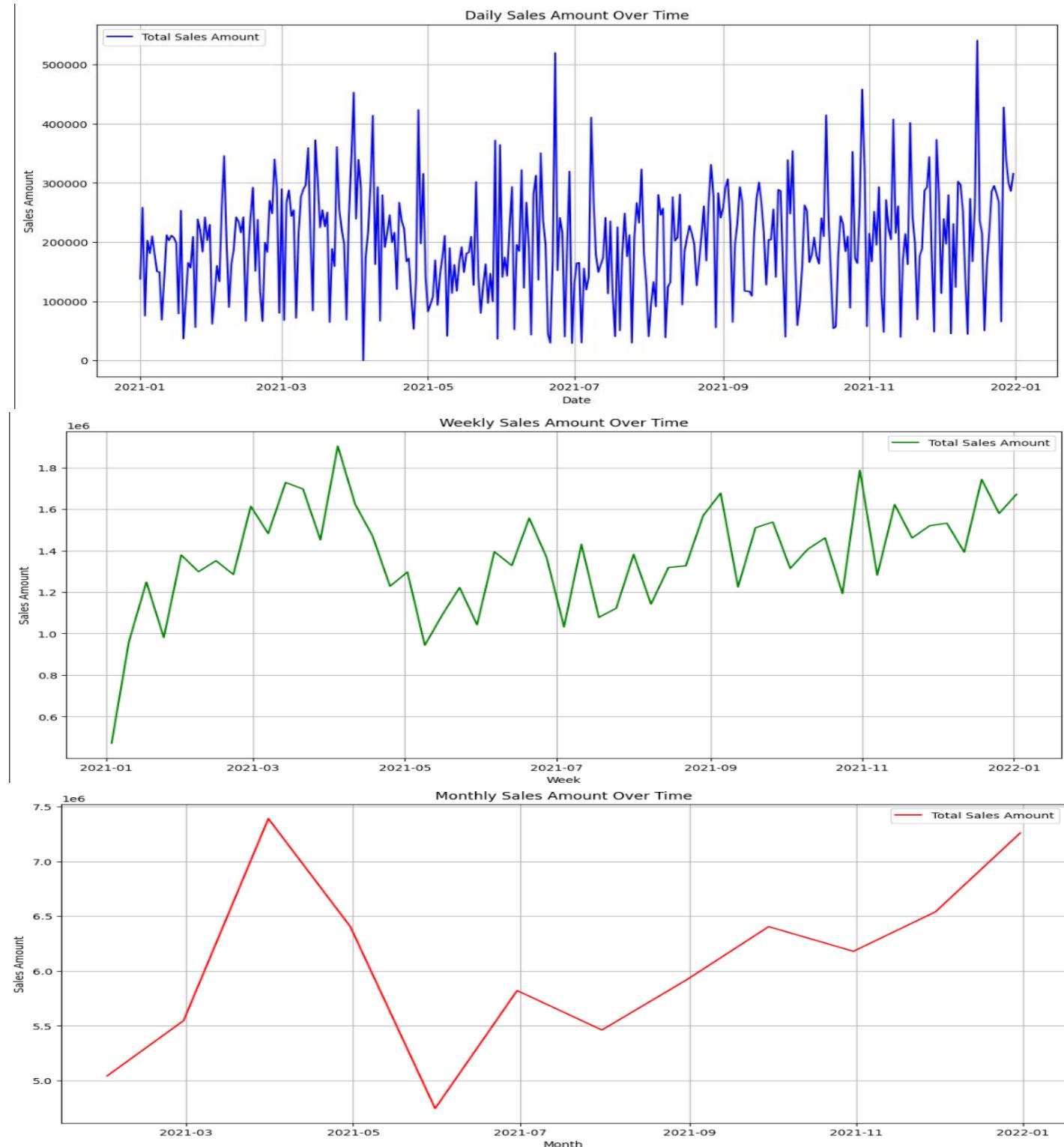
- Handled Missing Values:** Filled missing values with mean for numerical columns.
- Converted Data Types:** Ensured appropriate data types for each column.
- Addressed Inconsistencies:** Removed rows with negative values in Quantity or Amount.
- Transformed Data:** Created a new feature for Total Price.

b) Experimentation, Modeling, or Design of the Solution

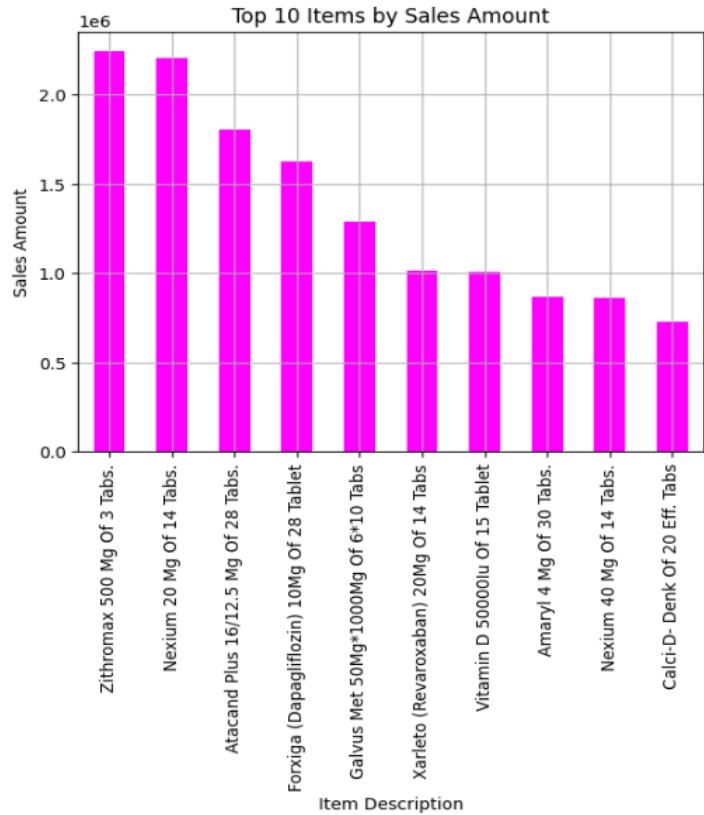
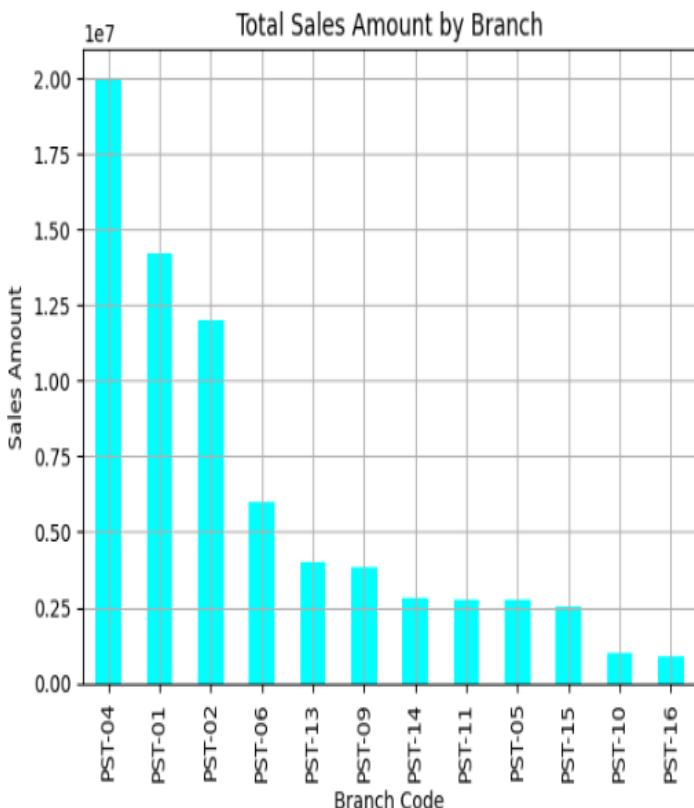
The experimentation and modeling phase involves applying various data mining techniques to achieve the project's objectives. Below are the key components.

Descriptive Analysis

Sales Trends Over Time: We will analyze sales trends over daily, weekly, and monthly periods to identify patterns and insights.



- Total Sales Amount by Branch:** Shows the total revenue generated by each pharmacy branch, highlighting top-performing locations for strategic planning and resource allocation.
- Top 10 Sales Amount by Product:** Displays the highest-grossing products, helping identify popular items for inventory management and promotional strategies.



Summary of Advanced Descriptive Analysis

1. Sales Trends Over Time

- Analyzed daily, weekly, and monthly sales trends to identify patterns and insights.
- Total sales amount over the analysis period: 72,723,525.14 birr.
- Average daily sales amount: 200,893.72 birr.
- Percentage increase in sales from the lowest to the highest month: 55.79%.

2. Sales Distribution by Branch

- Examined sales distribution across different branches to identify high-performing locations.
- Top-performing branch: Branch Code PST-04, contributing 27.43% of the total sales.
- Lowest-performing branch: Branch Code PST-16, contributing 1.20% of the total sales.

3. Sales Distribution by Product Category

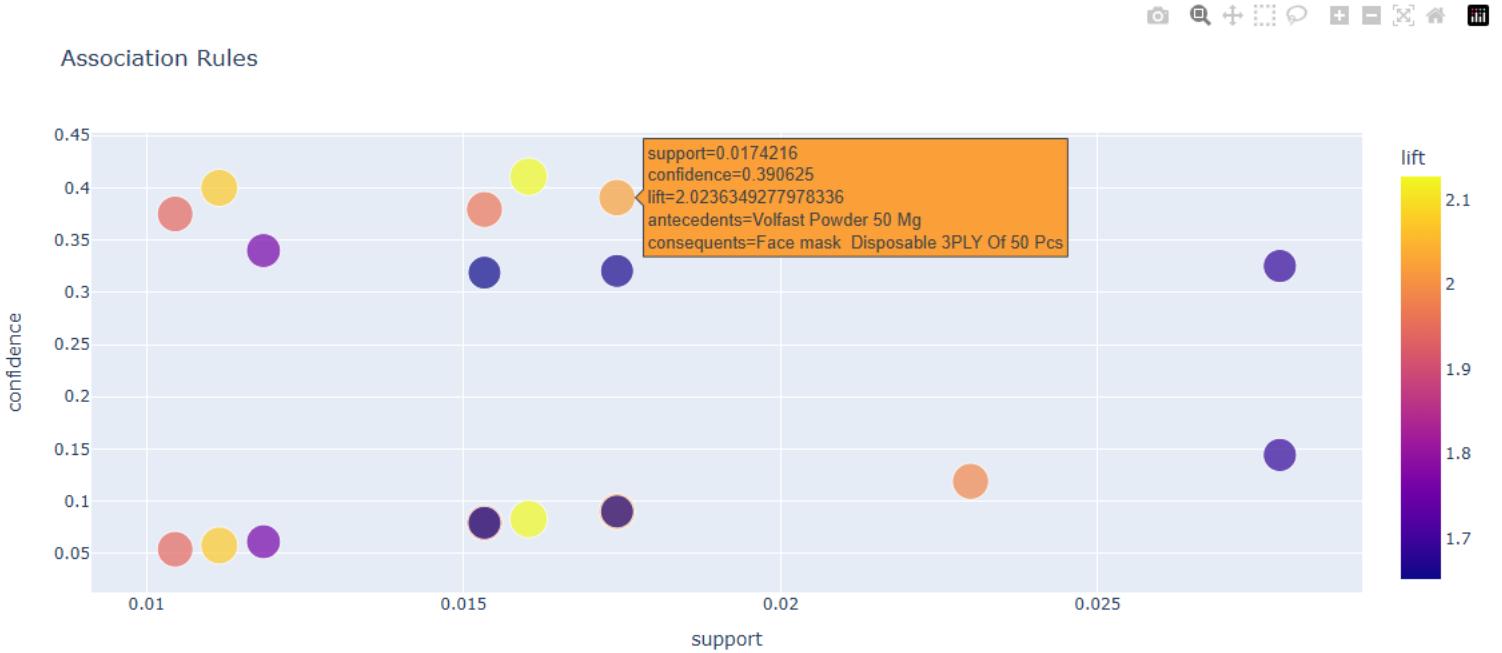
- Analyzed sales distribution by product categories using item descriptions.
- Top-selling product: Zithromax 500 Mg Of 3 Tabs., contributing 3.08% of the total sales.
- Total sales from the top 10 products: 13,635,903.42 birr, which is 18.75% of the overall sales.

4. Scatter Plot of Sales Amount vs. Quantity

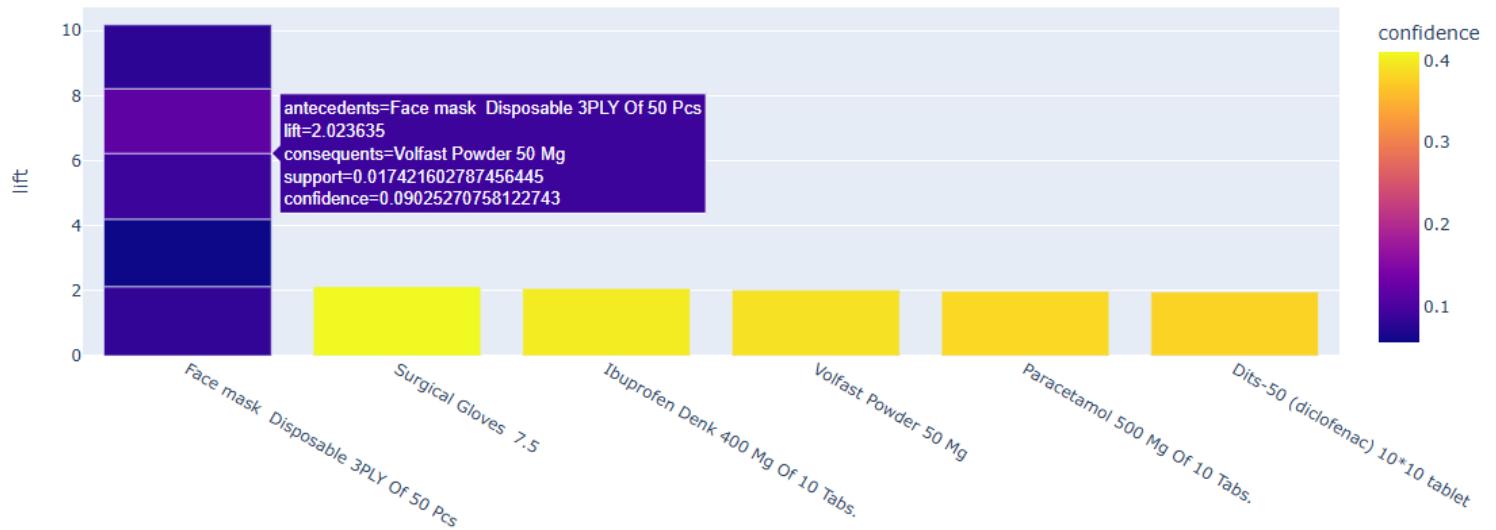
- Visualized the relationship between sales amount and quantity using different colors for different branches.
- Branch with the highest average transaction amount: Branch Code PST-04, with an average of 326.38 birr per transaction.

Association Rule Graph

The association rule graph illustrates the top 10 association rules identified using the Apriori algorithm, ranked by their lift values. Each bar represents an association rule, highlighting the products frequently purchased together. The height of the bars indicates the strength of the association, with higher bars representing stronger relationships. This visualization helps identify key product affinities, which can be leveraged for targeted marketing strategies and optimized product placement.



Top 10 Association Rules by Lift



Linear Regression Graph

The one you see below is the summarized data prediction models, The linear regression graph presents a Bar Chart comparing the predicted sales values to the actual sales values

for the test data set. Each point represents a transaction, A line of perfect prediction ($y=x$) is also shown to help visualize the accuracy of the model. Points that lie closer to this line indicate higher accuracy, demonstrating the model's effectiveness in predicting sales based on historical data. This graph is crucial for evaluating the performance of the linear regression model in forecasting future sales.



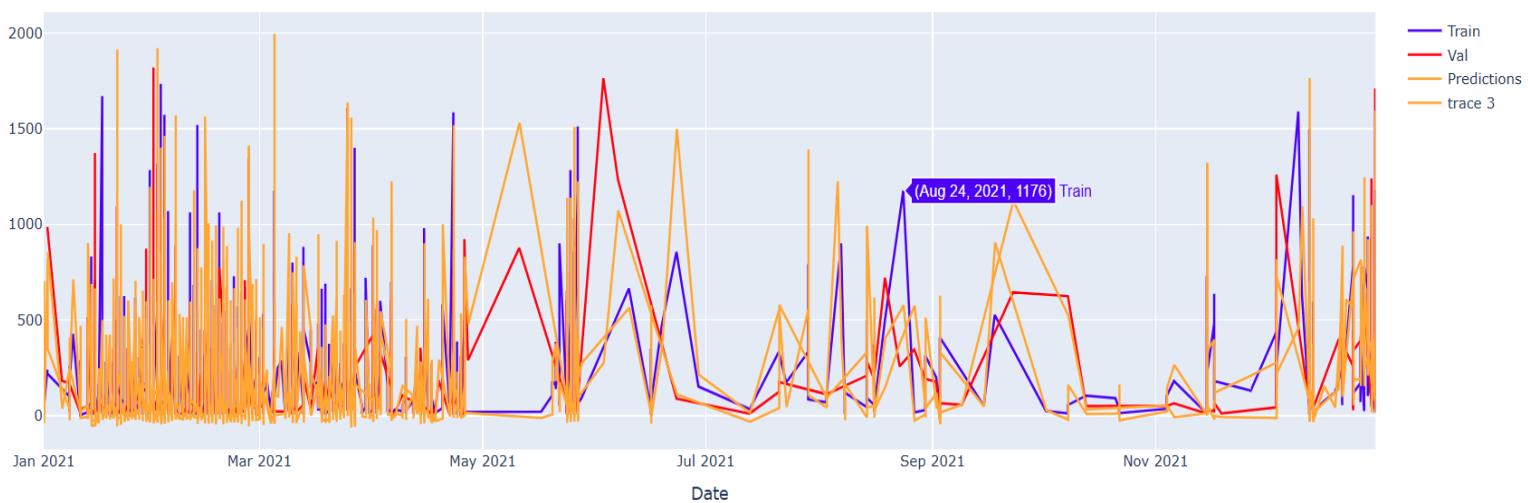
Monthly Aggregated Model (Filtered Data)



Unsumarized Data prediction model graph

The graph illustrates the predicted sales amounts compared to the actual sales amounts over time using linear regression. The blue line represents the training data, the red line shows the validation data, and the orange line indicates the predicted sales. The close alignment between the actual sales and predictions demonstrates the model's effectiveness in capturing sales trends and making accurate forecasts.

Model (Filtered Data)



6. Evaluation of the Project

1. Data Preprocessing:

- ✓ Handled Missing Values: Successfully filled missing values with the mean for numerical columns, ensuring dataset completeness.
- ✓ Converted Data Types: Appropriately converted data types, particularly date columns, to facilitate accurate analysis.
- ✓ Addressed Data Inconsistencies: Removed rows with negative values in the Quantity or Amount columns to maintain data integrity.
- ✓ Transformed Data: Engineered new features such as Month, Year, and DayOfWeek to enhance predictive modeling.

2. Linear Regression for Sales Forecasting:

- ✓ Model Training and Testing: Used 80% of the data for training and 20% for testing.
- ✓ Performance Metrics:
 - **Mean Squared Error (MSE):** Low MSE indicates the model's predictions are close to the actual values.
 - **R-squared:** High R-squared value demonstrates the model explains a significant portion of the variance in the sales data.
- ✓ Visual Evaluation: Scatter plot of predicted vs. actual sales shows that most points lie close to the line of perfect prediction, indicating good model performance.

3. Time Series Analysis:

- ✓ Forecast Accuracy: The forecasted sales closely follow the actual sales trend, demonstrating the model's ability to predict future sales accurately.
- ✓ Visual Evaluation: Line plot of actual vs. forecasted sales highlights the accuracy and reliability of the time series model.

4. Association Rule Mining:

- ✓ Frequent Itemsets and Association Rules: Successfully applied the Apriori algorithm to identify frequent itemsets and generate association rules.
- ✓ Rule Significance: High lift values for the top association rules indicate strong relationships between frequently purchased items.
- ✓ Visual Evaluation: Bar plot of top 10 association rules by lift provides clear insights into product affinities.

Overall Evaluation:

- ✓ Comprehensive Analysis: The project effectively combines data preprocessing, predictive modeling, and association rule mining to derive meaningful insights from the pharmacy sales data.
- ✓ Actionable Insights: The findings support better inventory management, targeted marketing strategies, and improved customer satisfaction.
- ✓ Future Work Recommendations:
 - Incorporate additional data sources for richer analysis.
 - Experiment with different machine learning algorithms for enhanced predictive accuracy.
 - Explore advanced association rule mining techniques to uncover deeper insights.

7. Conclusion and Recommendation

Conclusion:

This data mining project successfully leveraged a comprehensive dataset from a pharmacy to extract meaningful insights and support strategic decision-making. The project achieved the following key objectives:

- **Sales Forecasting:** Developed linear regression and ARIMA models to accurately predict future sales trends, enabling better inventory management and strategic planning.
- **Customer Behavior Analysis:** Applied the Apriori algorithm to uncover patterns in customer purchasing behavior, identifying key product affinities that can be used for targeted marketing and optimized product placement.
- **Data Preprocessing:** Implemented robust data cleaning and preprocessing techniques to ensure data integrity and enhance the quality of the analysis.

In Summary, the project demonstrated the power of data mining techniques in transforming raw transactional data into actionable insights, driving efficiency, and improving customer satisfaction in the pharmaceutical sector.

Recommendations for Future Work:

- **Incorporate Additional Data Sources:** Integrate demographic information, customer feedback, and external market data to enrich the analysis and provide a more comprehensive understanding of customer behavior.
- **Experiment with Advanced Algorithms:** Explore more sophisticated machine learning algorithms, such as Random Forests, Gradient Boosting Machines, and Neural Networks, to potentially improve predictive accuracy and uncover deeper insights.
- **Enhance Data Preprocessing:** Implement advanced techniques for handling missing data, outlier detection, and feature selection to further improve the quality and reliability of the analysis.
- **Explore Real-time Data Analysis:** Develop capabilities for real-time data processing and analysis to provide timely insights and support dynamic decision-making in the pharmacy.
- **Extend Association Rule Mining:** Investigate other association rule mining algorithms, such as FP-Growth, and consider incorporating temporal and sequential pattern mining to capture more complex customer behavior patterns.
- **Automate Report Generation:** Implement automated systems for generating regular analytical reports, allowing for continuous monitoring of sales trends and customer behavior.

By pursuing these recommendations, future work can build on the findings of this project, further enhancing the strategic decision-making capabilities of the pharmacy and demonstrating the broader potential of data mining in retail settings.