CREATING A SUMMARY TABLE FOR PRODUCT SALES

A PROJECT REPORT

Submitted by

REXCIA A (2303811724322089)

in partial fulfillment of requirements for the award of the course

AGI1252 – FUNDAMENTALS OF DATA SCIENCE USING R

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112

JUNE- 2025

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY
(AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report on **" CREATING A SUMMARY TABLE FOR PRODUCT SALES"** is the bonafide work of **REXCIA A (2303811724322089)** who carried out the project work during the academic year 2024 - 2025 under my supervision.

SIGNATURE

Dr.T. AVUDAIAPPAN, M.E.,Ph.D.,

HEAD OF THE DEPARTMENT

PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.

SIGNATURE

Ms.S.Murugavalli., M.E.,(Ph.D).,

SUPERVISOR

ASSISTANT PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.

Submitted for the viva-voce examination held on **02.06.2025**

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION

I declare that the project report on "**CREATING A SUMMARY TABLE FOR PRODUCT SALES**" is the result of original work done by me and best of my knowledge, similar work has not been submitted to "**ANNA UNIVERSITY CHENNAI**" for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This project report is submitted on the partial fulfillment of the requirement oof the completion of the course **AGI1252 - FUNDAMENTALS OF DATA SCIENCE USING R**

Signature

REXCIA A

Place: Samayapuram

Date: 02/06/2025

ACKNOWLEDGEMENT

It is with great pride that I express our gratitude and indebtedness to our institution, **"K. Ramakrishnan College of Technology (Autonomous)",** for providing us with the opportunity to do this project.

I I glad to credit honourable chairman **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of our study in college.

I would like to express our sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

I would like to thank **Dr. N. VASUDEVAN, M.TECH., Ph.D.,** Principal, who gave the opportunity to frame the project to full satisfaction.

I whole heartily thanks to **Dr. T. AVUDAIAPPAN, M.E.,Ph.D.,** Head of the department, **ARTIFICIAL INTELLIGENCE** for providing his encourage pursuing this project.

I I express our deep expression and sincere gratitude to our project supervisor **Ms.S.Murugavalli., M.E.,(Ph.D).,** Department of **ARTIFICIAL INTELLIGENCE,** for her incalculable suggestions, creativity, assistance and patience which motivated us to carry out this project.

I render our sincere thanks to the Course Coordinator and other staff members for providing valuable information during the course.

I wish to express our special thanks to the officials and Lab Technicians of our departments who rendered their help during the periodthe work progress.

INSTITUTE

Vision:

- To serve the society by offering top-notch technical education on par with global standards.

Mission:

- Be a center of excellence for technical education in emerging technologies by exceeding the needs of industry and society.
- Be an institute with world class research facilities.
- Be an institute nurturing talent and enhancing competency of students to transform them as all – round personalities respecting moral and ethical values.

DEPARTMENT

Vision:

- To excel in education, innovation, and research in Artificial Intelligence and Data Science to fulfil industrial demands and societal expectations.

Mission

- To educate future engineers with solid fundamentals, continually improving teaching methods using modern tools.
- To collaborate with industry and offer top-notch facilities in a conducive learning environment.
- To foster skilled engineers and ethical innovation in AI and Data Science for global recognition and impactful research.
- To tackle the societal challenge of producing capable professionals by instilling employability skills and human values

PROGRAM EDUCATIONAL OBJECTIVES (PEO)

- **PEO1:** Compete on a global scale for a professional career in Artificial Intelligence and Data Science.
- **PEO2:** Provide industry-specific solutions for the society with effective communication and ethics.
- **PEO3** Enhance their professional skills through research and lifelong learning initiatives.

.

PROGRAM SPECIFIC OUTCOMES (PSOs)

- **PSO1:** Capable of finding the important factors in large datasets, simplify the data, and improve predictive model accuracy.
- **PSO2:** Capable of analyzing and providing a solution to a given real-world problem by designing an effective program.

PROGRAM OUTCOMES (POs)

Engineering students will be able to:

1. **Engineering knowledge:** Apply knowledge of mathematics, natural science, computing, engineering fundamentals, and an engineering specialization to develop solutions to complex engineering problems.
2. **Problem analysis:** Identify, formulate, review research literature and analyze complex engineering problems reaching substantiated conclusions with consideration for sustainable development.
3. **Design/development of solutions:** Design creative solutions for complex engineering problems and design/develop systems/components/processes to meet identified needs with consideration for the public health and safety, whole-life cost, net zero carbon, culture, society and environment as required.
4. **Conduct investigations of complex problems:** Conduct investigations of complex engineering problems using research-based knowledge including design of experiments, modelling, analysis & interpretation of data to provide valid conclusions.
5. **Engineering Tool Usage:** Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems.
6. **The Engineer and The World:** Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment.
7. **Ethics:** Apply ethical principles and commit to professional ethics, human

8. Values, diversity and inclusion; adhere to national & international laws.
9. **Individual and Collaborative Team work:** Function effectively as an individual, and as a member or leader in diverse/multi-disciplinary teams.
10. **Communication:** Communicate effectively and inclusively within the engineering community and society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations considering cultural, language, and learning differences.
11. **Project management and finance:** Apply knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, and to manage projects and in multidisciplinary environments.
12. **Life-long learning:** Recognize the need for, and have the preparation and ability for i) independent and life-long learning ii) adaptability to new and emerging technologies and iii) critical thinking in the broadest context of technological change.

ABSTRACT

Creating a summary table for product sales addresses key challenges faced by businesses in analyzing and utilizing sales data effectively. Many organizations struggle with inconsistent reporting, time-consuming manual processes, and a lack of predictive insight, which hampers strategic planning and operational efficiency.This approach leverages R programming to automate the generation of comprehensive sales summary tables. It includes essential metrics such as total sales, average sales, growth rates, and performance by product category. Additionally, time series forecasting is integrated to provide forward-looking insights into future sales trends. By combining historical analysis with predictive modeling, the summary table becomes a powerful decision-support tool that enhances data accuracy, reduces reporting effort, and supports more informed business planning.

ABSTRACT WITH POs AND PSOs MAPPING

CO 5 : BUILD DATA SCIENCE USING R PROGRAMMING FOR SOLVING REAL-TIME PROBLEMS.

`

ABSTRACT	POs MAPPED	PSOs MAPPED
Creating a summary table for product sales addresses key challenges faced by businesses in analyzing and utilizing sales data effectively. Many organizations struggle with inconsistent reporting, time-consuming manual processes, and a lack of predictive insight, which hampers strategic planning and operational efficiency.This approach leverages R programming to automate the generation of comprehensive sales summary tables. It includes essential metrics such as total sales, average sales, growth rates, and performance by product category.	**PO1 -3** **PO2 -3** **PO3 -3** **PO4 -3** **PO5 -3** **PO6 -3** **PO7 -3** **PO8 -3** **PO9 -3** **PO10 -3** **PO11-3**	**PSO1 -3** **PSO2 -3**

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 OBJECTIVE

The primary objective of this work is to design and implement an automated solution for summarizing product sales data using R programming. Businesses often face challenges in maintaining consistent and timely sales reports due to manual methods and fragmented data sources. This system aims to overcome those limitations by providing a structured and efficient approach to sales data analysis.

- Calculating key metrics such as total sales, average sales, and growth rates.
- Categorizing sales performance by product.
- Integrating time series forecasting to predict future sales.
- Simplifying reporting to support faster, data-driven business decisions

1.2 OVERVIEW

Efficient analysis of product sales is essential for informed business decisions. However, manual reporting processes often lead to inconsistent data and time-consuming efforts. Traditional sales summaries focus only on historical data, offering limited value for future planning.This work uses R programming to automate the creation of summary tables that include key metrics such as total sales, average sales, and growth rates. It also

integrates forecasting to predict future sales, enabling better planning and streamlined decision-making.To address these issues, the development of an automated system using R programming offers a practical and scalable solution. By generating summary tables that not only consolidate historical sales data but also include key metrics such as average sales, total revenue, growth rates, and category-wise performance, businesses gain a clearer understanding of their operations.

1.3 DATA SCIENCE RELATEDCONCEPTS

The proposed system incorporates several core data science concepts to automate sales analysis and forecasting effectively. These concepts form the foundation of the analytical pipeline and contribute to generating actionable insights. The key data science concepts involved are:

1. Data Collection and Preprocessing

Data Cleaning: Removing missing, duplicate, or inconsistent records to ensure quality data.

Transformation: Converting raw sales data into a structured format suitable for analysis (e.g., date formatting, data normalization).

Feature Engineering: Creating new variables such as monthly sales totals, growth rates, or category-wise aggregates to enrich the dataset.

2. Descriptive Analytics

Summarizing historical data using statistical metrics like **mean**, **sum**, **min/max**, and **standard deviation**.

Creating **summary tables** that provide insight into sales trends across different products and time periods.

3. Data Visualization

Using graphical tools such as **bar charts**, **line graphs**, and **time series plots** to illustrate trends, seasonality, and performance variations.

Helps stakeholders easily interpret complex datasets.

4. Time Series Forecasting

Applying **ARIMA** (AutoRegressive Integrated Moving Average) and **Prophet** models to predict future sales based on past patterns.

Key concepts include **trend**, **seasonality**, and **residual analysis**.

Forecasting helps in proactive inventory planning and strategy development..

5. Automation and Reporting

Automating the entire process—from raw data input to final output—using scripting in R.

Final output includes **interactive dashboards or reports** that combine historical summaries with future forecasts.

CHAPTER 2

PROJECT METHODOLOGY

2.1 PROPOSED WORK

The proposed Work in this context refers to the systematic plan to collect, process, and analyze raw sales data to generate meaningful summary reports. It involves designing methods to extract relevant sales information, organizing the data efficiently, and presenting it in a concise format that highlights key metrics such as total sales, trends, and performance indicators. The proposed work ensures accurate, timely, and automated generation of sales summaries to support business decision-making.

1. Collecting Raw Sales Data

- **Identify All Data Sources:**

  Determine all possible data sources including point-of-sale (POS) systems, online sales platforms, ERP systems, and manual logs to ensure no sales data is missed during collection.

- **Automate Data Extraction:**

  Use tools like SQL scripts, APIs, or data connectors to automatically extract data at regular intervals, improving consistency and minimizing manual effort.

2. Data Preprocessing and Cleaning

- **Handle Missing and Null Values:**

Identify incomplete records and fill missing values using techniques like imputation, default substitution, or row removal, depending on data importance.

- **Remove Duplicate Records:**

Detect and eliminate duplicate entries that could skew summary results, ensuring each transaction is counted only once.

- **Standardize Data Formats:**

Unify inconsistent formats for dates, currency, and product codes to maintain data integrity and compatibility across systems.

3. Creating Sales Summary Tables

- **Aggregate Sales Metrics:**

Compute key indicators such as total sales value, number of transactions, and average transaction size across daily, weekly, or monthly periods.

- **Group by Product and Region:**

Organize summary tables by product category, individual SKUs, and sales region to support targeted performance evaluation.

- **Track Time-Based Trends:**

Summarize data over time to capture seasonal variations, growth rates, or declines, which are essential for strategic planning.

4. Applying Forecasting Models

- **Train Models on Historical Data:**

Feed cleaned, chronological sales data into the models and optimize parameters to improve the accuracy of forecasts.

- **Generate Future Sales Predictions:**

Use the trained models to forecast future sales, identifying expected volumes per product, region, or time period.

5. Integrating Historical and Forecasted Data

- **Merge Forecasted and Actual Data:**

Combine both datasets into one structured format to allow for side-by-side comparisons between actual and expected sales figures.

- **Enable Filtering and Drill-Down:**

Implement capabilities to filter data by product, time, or geography to allow detailed exploration and customized analysis.

6. Visualizing Sales Trends and Forecasts

- **Create Interactive Dashboards:**

Use tools like Power BI, Tableau, or Python's Plotly to build dashboards that visually represent key sales insights and predictions.

- **Highlight Key KPIs and Anomalies:**

Display indicators such as top-selling products, sales dips, or forecast deviations to quickly draw attention to important insights.

2.2 BLOCK DIAGRAM

CHAPTER 3
R PROGRAMMING CONCEPTS

3.1 Data Import and Export

- Use read.csv() to load sales data from CSV files, and read_excel() (from the readxl package) to read Excel-based sales reports.
- Save processed or summarized sales data using write.csv() or write.xlsx() from the openxlsx package for further reporting or dashboard tools.

3.1.1 Data Cleaning and Preprocessing

- Detect and manage missing sales entries using is.na(), na.omit() for removal, or replace() to impute with mean or zero..
- Convert columns like Date to as.Date() and Region to as.factor() for proper analysis.
- Use scale() to standardize sales figures for clustering or modeling; normalization helps compare regions with different sales ranges.

3.2 Descriptive Statistics

- Summary functions: summary(), mean(), median(), sd(), var(), quantile().
- Data exploration: table(), cor(), aggregate().

3.3 Data Visualization

- Base R plotting: plot(), hist(), boxplot(), barplot().
- ggplot2 package: ggplot(), geom_point(), geom_bar(), geom_boxplot().

3.3.1 K-Means Clustering

- Function: kmeans() for clustering analysis.

- Finding optimal clusters using the Elbow Method or Silhouette Score.
- Visualization with fviz_cluster() from the factoextra package.

3.3.2 Regression Modeling

- Linear regression using lm() function.
- Model evaluation: summary(), residual plots, R^2, and RMSE.
- Diagnostic checks for multicollinearity using vif() from car package.

3.3.3 Classification Models

- Decision Tree using rpart() and rpart.plot().
- Logistic Regression using glm(family = "binomial").
- Splitting data: sample(), createDataPartition() from caret.
- Model evaluation: confusion matrix with table() or confusionMatrix().

3.3.4 Data Manipulation

- Using dplyr functions: filter(), select(), mutate(), group_by(), summarise().
- Using tidyr for reshaping data: pivot_longer(), pivot_wider().

3.4 Model Evaluation Techniques

- Performance metrics: Accuracy, Precision, Recall, F1-score.
- Cross-validation using train() from the caret package.

CHAPTER 4

MODULE DESCRIPTION

4.1 Data Collection & Preprocessing Module

This foundational module is responsible for acquiring raw sales data from various sources, including spreadsheets (e.g., Excel), relational databases (e.g., MySQL), or APIs (e.g., e-commerce platforms or ERP systems). Once collected, the data undergoes a series of preprocessing steps to ensure it is clean and analysis-ready. These steps include:

- **Handling missing values** using imputation or removal strategies.
- **Correcting data types** to ensure numerical and date fields are properly formatted.
- **Removing duplicates** to avoid data redundancy and bias.
- **Standardizing formats** such as product IDs, date formats, and currency fields.

  This module ensures data consistency and quality, which is critical for reliable downstream analytics and forecasting.

4.2 Sales Summary Table Module

This module transforms the cleaned dataset into meaningful and structured summary tables. It computes and displays key sales performance metrics such as:

- **Total revenue**, **units sold**, **average unit price**, and **return rates**.

Categorization by multiple dimensions such as **product**, **time period** (daily, monthly, yearly), and **region**.

- **Growth rate calculations** over time for trend analysis.

  The summary tables enable stakeholders to quickly identify high-performing products, seasonal trends, and underperforming areas. Advanced techniques like pivoting and grouping are used to generate these summaries efficiently.

4.3 Forecasting & Reporting Module

This module applies advanced statistical and machine learning models to identify trends and predict future sales. The forecasting process includes:

- **Model selection**: Using models like ARIMA, Holt-Winters, and Prophet based on data characteristics (trend, seasonality).
- **Trend and seasonality detection**: Identifying cyclic behaviors, peak periods, and sales drops.
- **Forecast generation**: Producing future sales estimates for each product or category.
- **Error metrics**: Evaluating forecast accuracy using MAE, RMSE, and MAPE.

  These predictions guide inventory control, budget planning, and marketing efforts, helping organizations make proactive decisions.

4.4 Automated Reporting Module

To reduce manual effort and ensure timely access to insights, this module automates the generation of comprehensive sales reports. Key features include:

- Use of **R Markdown** for generating PDF/HTML reports and **Shiny dashboards** for interactive visualizations.
- **Scheduled or event-triggered report generation**, ensuring up-to-date insights.
- **Graphical outputs** including line charts for trends, bar graphs for product comparisons, and heatmaps for regional sales.
- **Customizable templates** for different stakeholders (e.g., sales, finance, marketing).

  This module ensures that decision-makers receive professional, data-rich reports with minimal manual input.

4.5 Full System Integration Module

This module serves as the backbone that links all other modules into a cohesive, end-to-end system. It ensures seamless data flow and process automation, including

- **Pipeline orchestration** to automatically run tasks in sequence—from data collection to final reporting.
- **Error handling and logging mechanisms** for monitoring system health and troubleshooting.
- **Scalability** for handling increasing data volumes and multiple product categories.

- **Extensibility** allowing the addition of new forecasting models, data sources, or reporting templates with minimal code changes.

By integrating all modules, this system offers a unified solution that minimizes manual intervention, reduces operational costs, and enhances data reliability and decision-making efficiency.

CHAPTER 5
CONCLUSION

The development and implementation of the comprehensive sales summary system marks a significant advancement in how sales data is managed, analyzed, and utilized within an organization. This system efficiently handles the entire workflow starting from raw data collection and rigorous preprocessing to ensure accuracy and consistency, moving through the generation of detailed summary tables that highlight crucial sales metrics across different dimensions such as product categories, regions, and time periods. By incorporating sophisticated forecasting techniques, the system is capable of uncovering hidden patterns and trends within historical sales data, enabling the prediction of future sales with a high degree of reliability. These forecasts serve as invaluable tools for strategic decision-making, aiding inventory management, budgeting, and marketing efforts.y cleaning and transforming raw sales data into structured summaries and predictive insights, the system empowers businesses to make informed, data-driven decisions. Forecasting models help anticipate future trends, while automated reporting ensures timely access to key metrics and performance indicators. Overall, it enhances operational efficiency, improves strategic planning, and provides a scalable solution for continuous sales monitoring and analysis.

FUTURE SCOPE

The sales summary system presents a solid foundation for data-driven decision-making; however, there are numerous opportunities to enhance its capabilities further. In the future, integration with real-time data streams can be developed, enabling instantaneous sales tracking and more responsive forecasting that adapts dynamically to sudden market changes or promotional campaigns. Advanced machine learning models, such as deep learning or ensemble methods, could be incorporated to improve forecast accuracy and capture complex, nonlinear patterns in sales behavior.The automated reporting module could be enhanced by integrating with business intelligence platforms or mobile apps to provide on-the-go access to sales performance metrics. Finally, implementing robust data security and compliance measures will be essential as the system scales and handles increasingly sensitive business data. Overall, these advancements will transform the system into a highly intelligent, adaptive, and strategic tool, further driving business growth and agility.

APPENDIX-A

SOURCE CODE

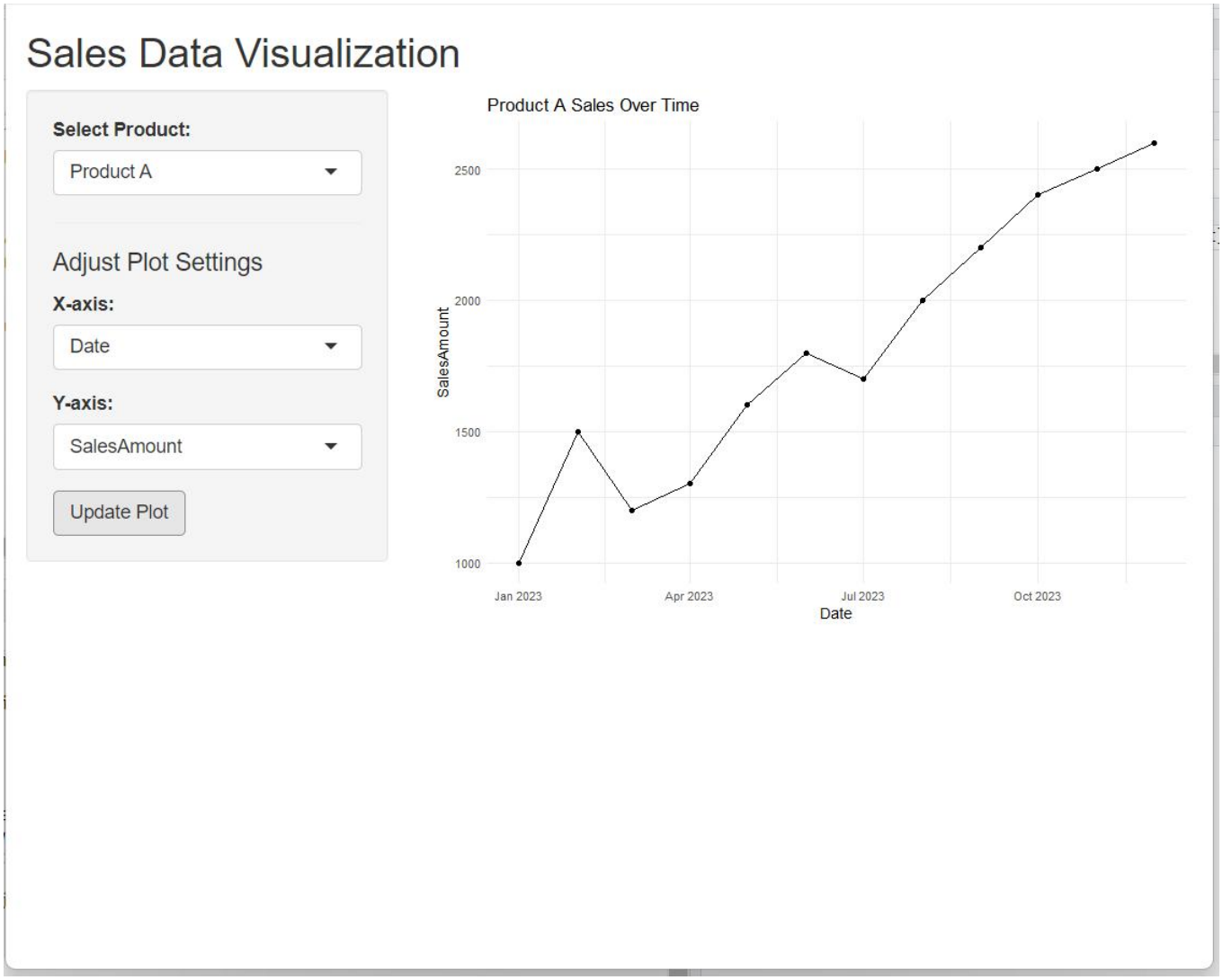
```
library(shiny)
library(ggplot2)
library(dplyr)
# Sample sales data with a "Product”column
sample_data <- data.frame(
  Date = rep(seq.Date(from =
as.Date("2023-01-01"), by =
"month", length.out = 12), 2),
  SalesAmount = c(1000, 1500,
1200, 1300, 1600, 1800, 1700,
2000, 2200, 2400, 2500, 2600,
1100, 1400, 1250, 1350, 1550,
1700, 1650, 1900, 2050, 2250,
2450, 2550),
  Product = rep(c("Product A",
"Product B"), each = 12)
)
# Define UI for the app
ui <- fluidPage(
  titlePanel("Sales Data
Visualization"),
```

```
  sidebarLayout(
    sidebarPanel(
      # Dropdown for selecting the
product
      selectInput("product", "Select
Product:", choices =
unique(sample_data$Product)),
      tags$hr(),
      h4("Adjust Plot Settings"),
      selectInput("x_var", "X-axis:",
choices = c("Date",
"SalesAmount")),
      selectInput("y_var", "Y-axis:",
choices = c("SalesAmount",
"Date")),
      actionButton("update",
"Update Plot")
    ),
    mainPanel(
      plotOutput("salesPlot")
    )
  )
)
# Define server logic
```

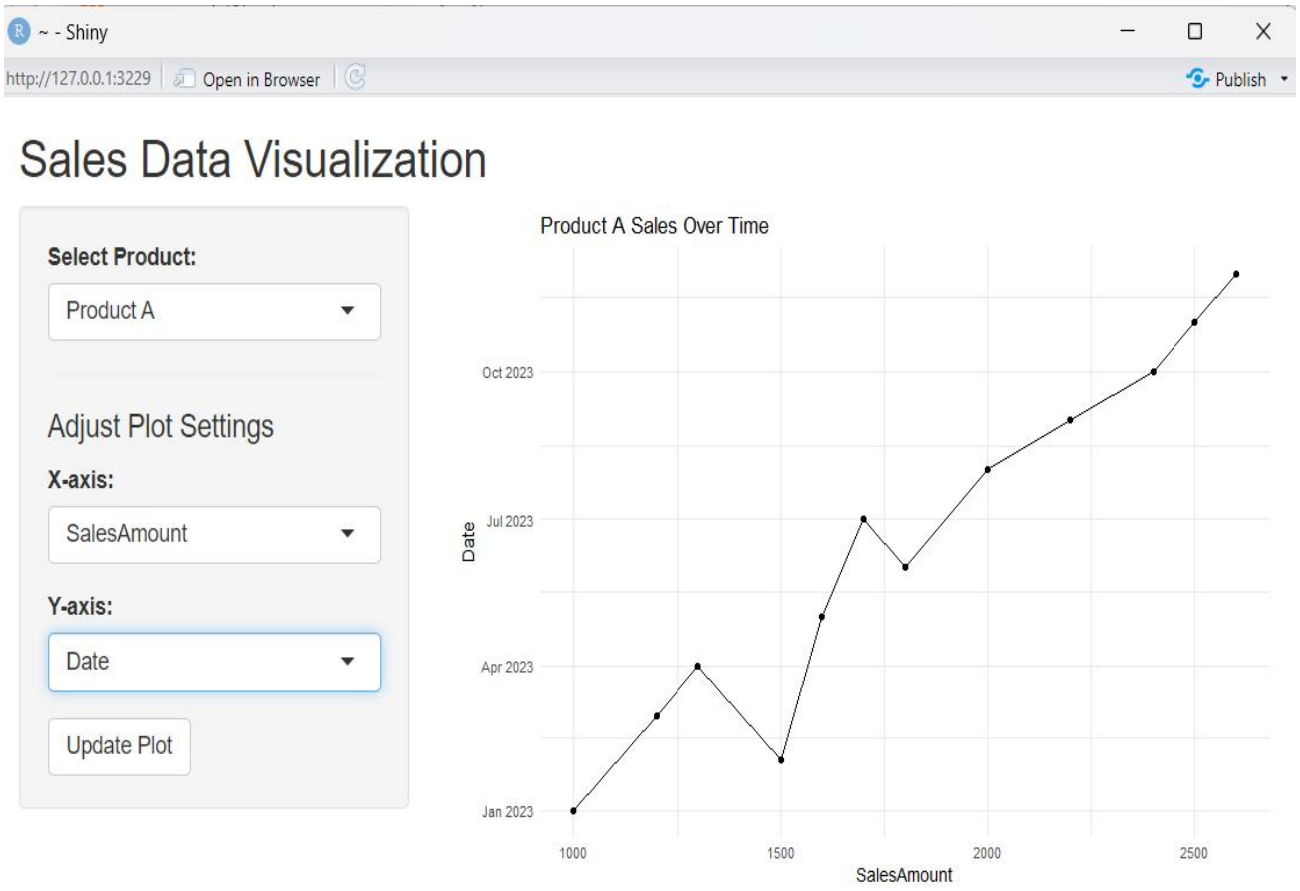
```
server <- function(input, output) {
  # Reactive value to store the
filtered data
  sales_data <- reactive({
    # Filter the data based on the
selected product
    sample_data %>%
filter(Product == input$product)
  })
  observeEvent(input$update, {
    req(sales_data())
    plot_data <- sales_data()
    # Convert 'Date' column to
Date type if necessary
    if ("Date" %in%
colnames(plot_data)) {
      plot_data$Date <-
as.Date(plot_data$Date)
    }
    # Create the plot
    output$salesPlot <-
renderPlot({
      ggplot(plot_data, aes_string(x
= input$x_var, y = input$y_var))
```

```
+
        geom_line() +
        geom_point() +
        ggtitle(paste(input$product,
"Sales Over Time")) +
        xlab(input$x_var) +
        ylab(input$y_var) +
        theme_minimal()
    })
  })
}
# Run the application
shinyApp(ui = ui, server = server)
```

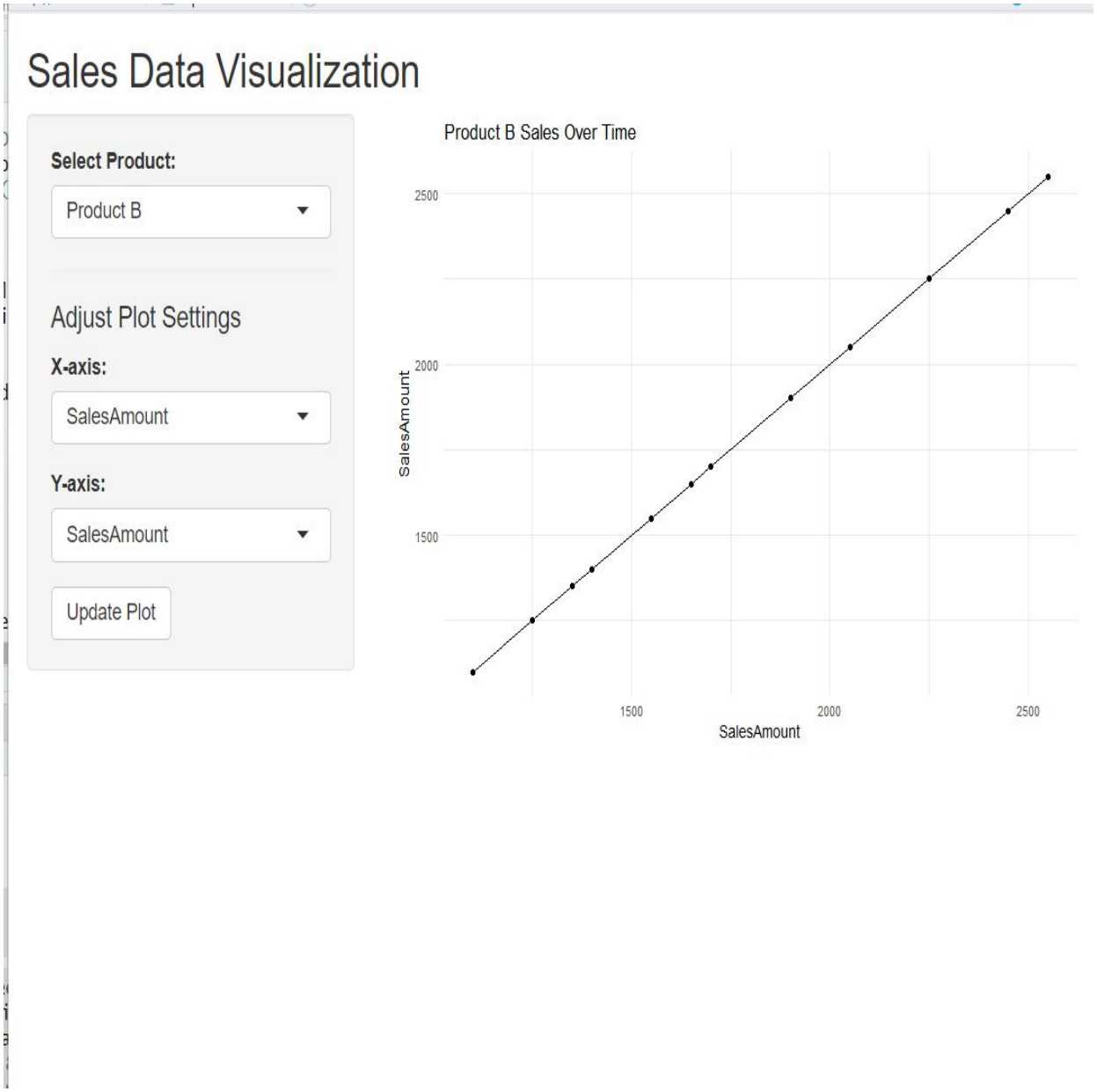
APPENDIX B - SCREENSHOTS



APPENDIX B - SCREENSHOTS



APPENDIX B - SCREENSHOTS



REFERENCES:

1. Grolemund, G., & Wickham, H. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.
2. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd edition). OTexts. Retrieved from https://otexts.com/fpp3/
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.
4. Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.
5. R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/
6. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.