SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

BACHELOR IN COMPUTER SCIENCR

MAY 2024 – SEMESTER 5

INDIVIDUAL ASSIGNMENT TASK 1 (15%)

Module Name:	**Big Data Technologies**
Module Code:	**ITS 66904**
Due Date:	**11th July, 2024, 23:59 PM (NPT)**
Platform:	**myTIMes**

Data Mining Tool in Retail Customer Analytics Application

Student Declaration: We Declare That –

- ✓ *I confirm my awareness about the university's regulations, governing cheating in tests and assignments, and form the guidance issued by the school of computing and it concerning plagiarism and proper academic practices, and the assessed work now submitted is in accordance with this regulation and guidance.*
- ✓ *I understand that, unless already agreed with the school of computing and it, that the assessed work has not been previously submitted, either in whole or in part, in this or any other institution.*
- ✓ *I recognize that should evidence emerge that my work fails to comply with either of the above declarations, then i may be liable to proceeding under regulation.*

Lecture B | Laboratory 3 | Group 5

Student Full Name	University ID	Signatures	Date, Day	Scores
Sujal Ratna Tuladhar	**036 2483**		**9th July 2024, Tuesday**	

TABLE OF CONTENTS

TABLE OF FIGURES

1. Introduction

1.1. Selected Tool and Application

R programming language was chosen as the data mining tool to do this project for Retail Customer Analysis. It provides a rich environment of statistical computing and graphics. With abundance of libraries, providing extensive capabilities for data manipulation, visualization, statistical modelling, machine learning, tidying, transformation and communication. It is also ideal for analyzing large amount of data and datasets being generated from business environments.

1.2. Installation

To run R programming language, we can download the software from CRAN website, which is Comprehensive R Archive Network. It will be easy and productive for the users to download the IDE (Integrated Development Environment) called R-Studio, which allow multiple in-built tools and interactive features from posit site

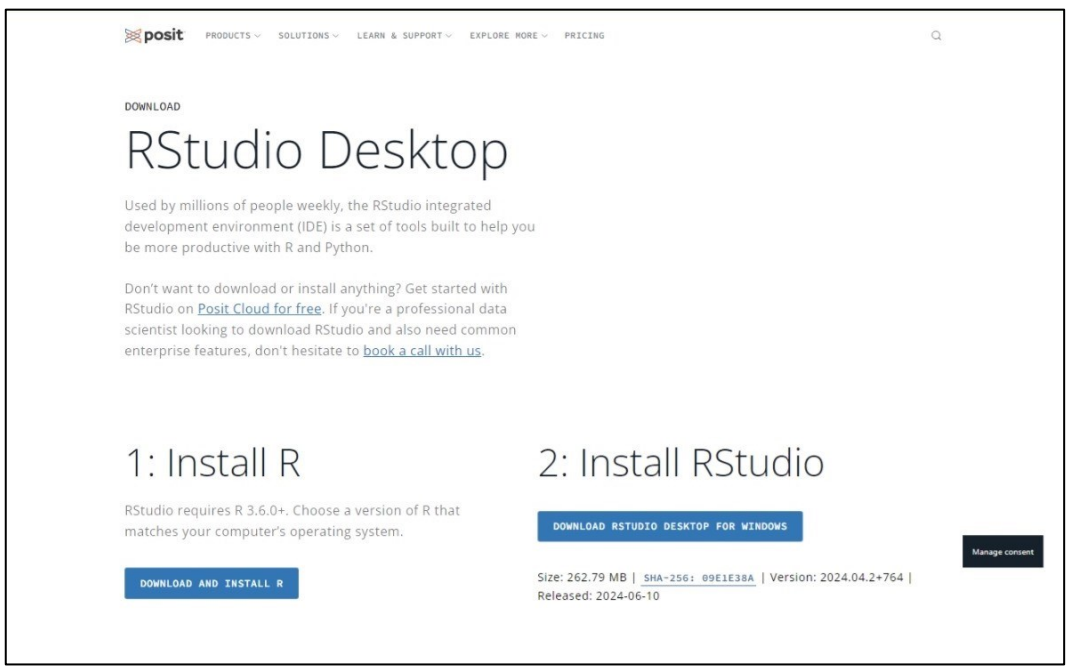


FIGURE 1. R AND R-STUDIO WEBSITE AND INSTALLATIONS

1.3. Configuration

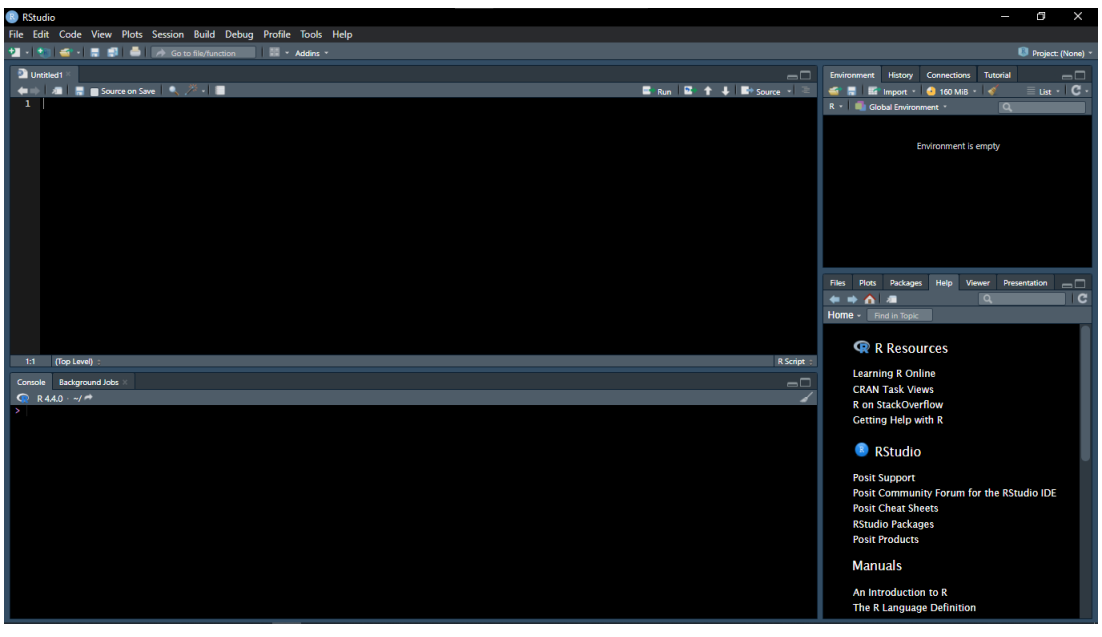


FIGURE 2. R STUDIO INTERFACE

It is not enough to just start coding. So, we need to install packages for function and to make a user-friendly graphical interface for the dashboard application. We download shiny which is an application builder pack and can be written with less lines of code containing an interface object, server function, and call function in single source code script.

1.4. Retail Customer Analysis

R language provide tools for advance data research with analytics to enhance statistical models. It is widely in use as it is open-source and licensed of general public (GNU) working across multiple operating systems. It is a valuable for finance and business application with process like Data Mining, Exploratory Data Analysis and Machine Learning.

1.4.1. Usage

R can be seen being used by Organizations like Google, Amazon, Meta. With roles like Scientists, Analysts, Architect, Statistician, who have advance skills to test hypothesis and proficient with their ability to understand behaviors, catch up to trends, make decisions based on forward insights, with algorithms like clustering, churn metrics and improve recommendations.

1.4.2. Addressed

Handling of big data in retail shops can be easily done with R as it can work with various and large volumes of data from multiple sources and formats (CSV, XLSX, JSON, APIs) be structured or unstructured. Facilitated by many essential techniques such as cleaning, transforming, aggregations, imputations and normalizations for efficiencies. It can also be used with support of penalizations and in distributed systems to have horizontal scaling.

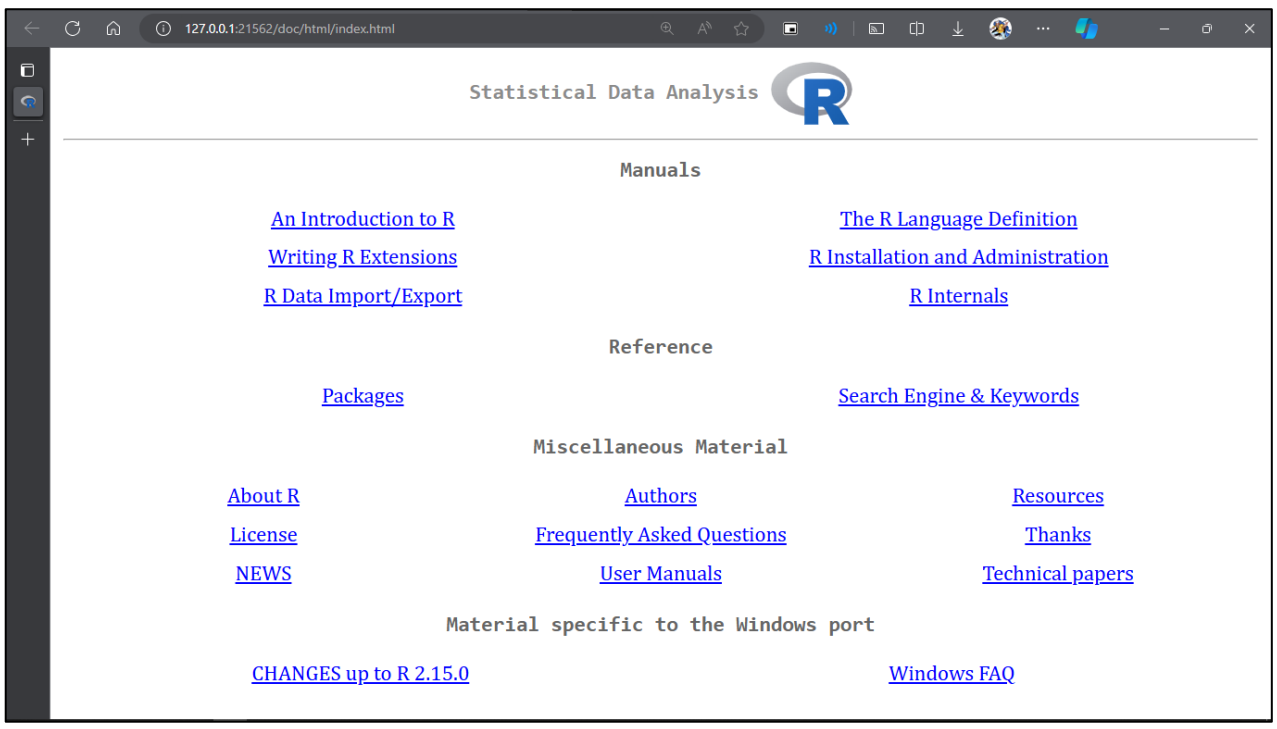


FIGURE 3. R DOCUMENTATION WITH HELP.START()

2. Objective

- Use R language to make a functional dashboard for a retail customer behavior analysis
- Extracting meaningful insights from the large dataset with multiple visual outputs
- Enhance on sound decision making and modification on strategies

3. Purpose of Application

- use multiple libraries to develop an application that calculate sales, profit, discounts, time series
- develop, extract, understand the behavior and preferences with performance data
- understand the advance technique used by businesses to optimize and improvise

4. Architecture

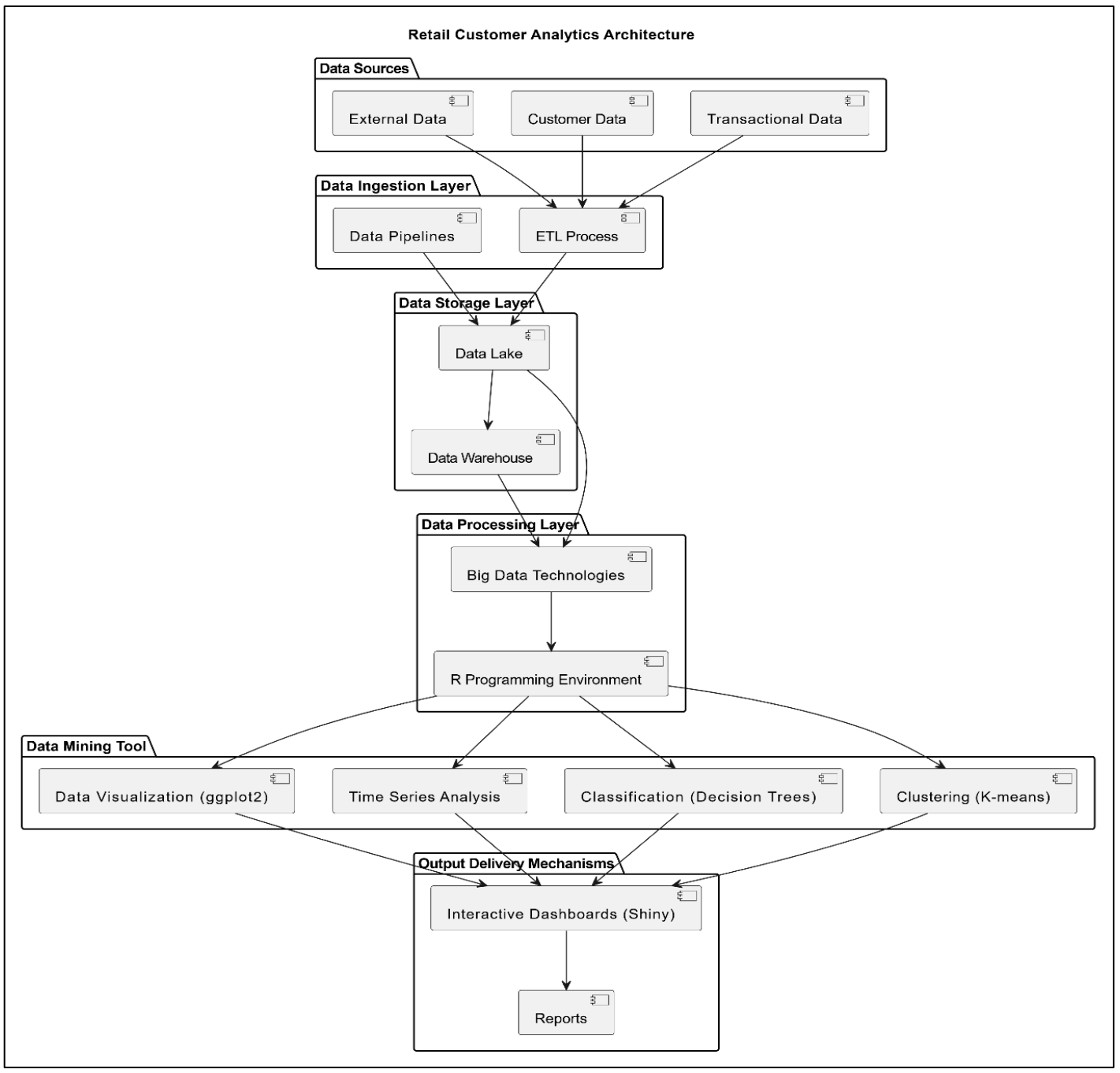


FIGURE 4. ARCHITECTURE

4.1. Big Data Techniques

4.1.1.Data Sources

- Sales and Transactional Record, Details of Order, with Required Customer Information's.
- SWOT Analysis of Competition, Social Media Metrics, Market Trends and Preferences
- while collecting we can automate script or tool to web scrap and extract from websites,
- from various sources like in secure cloud location, sensors, transactions, in frequent intervals or real-time IoT
- physical backup appliance like external hard drives in a Network Attached Storage (NAS)

4.1.2.Data Storage

- Lake like reservoir or repositories which contain vast amount of raw data in unstructured format
- on site Warehouse-like structures and labeled for efficient processing and reliable performance
- other optimized large scale non/relational databases(no/SQL) or (hadoop)distributed file systems (mongodb, cassandra)

4.2. Data Mining Tool

4.2.1. Data Ingestion

- Batch Processing, Periodic Extraction, Loading ETL, large volume in scheduled intervals (mapruduce and spark)
- Stream Processing, Real-Time Processing for Immediate Insights (kafka, flink)
- importing from the sources and storage systems, simultaneous to process parallel in multiprocessors

4.2.2. Manipulation Layers

- Cleaning removal of inconsistencies, handling of missing values, and ensuring quality data
- correcting inaccurate, errors, inconsistencies,
- Integration and Transformation like Aggregations, Summarizations, Structuring, Tidying

4.3. Related Algorithms

4.3.1.Analytics Engine

- Uses of Packages and Libraries for various analysis
- Clustering, Classification, Time Series, Decision Tree,
- predictive: provide future outcomes with historical data
- prescriptive: having insights of behaviors and recommending path actions

4.3.2.Output Delivery

- Visualization, Interactive Dashboards with Plots with key metrics, and trends, (tableau, power bi)
- Reporting, Outlines, Guidelines

4.3.3.Security and Privacy

- protecting data by secure conversions and permissions only to authorized users
- implement policies and restriction based on roles, complicate process for unauthorized user

5. Critical Analysis

5.1. Advantages

- using r and shiny, we are able to dynamically see real-time updates in visualizations inside the dashboard, and have sound decision-making when the user input/update the data set
- it is able to handle large volumes of data generated by retail business field and its architecture handles batch and stream processing well of analytics
- anyone is able to integrate multiple techniques like clustering, classification, time series and can illustrate them with libraries like ggplot2, even with non-technical background
- Stakeholder and business managers can analyze and identify multiple areas of profitable, scalable, optimized strategies for sales, market, different types of customers segmented, and targeted with pricing, discounts.

5.2. Disadvantages

- analysts can only work with the data that has been given to them, so the accuracy and reliability depends on the input data collected, any inaccuracy can have a negative impact.
- Although it may work on extremely large dataset for data analysis, but may fall short for highly complex computations and may require big data technology like Hadoop or spark.
- advanced machine learning may be a problem as python is mainly used for such task, although there maybe libraries out there it may not be able to provide deep insights. apart from basic clustering, classification, prediction.

6. Functional Analysis

6.1. R Programming Language and its Packages

```
# install necessary packages ####
install.packages(c("shiny", "ggplot2", "dplyr", "readr", "lubridate", "tidyr", "shinythemes", "GGally", "cluster", "rpart"))
```

FIGURE 5. INSTALL NECESSARY PACKAGES

I choose R for retail customer analytics as it is widely popular and powerful statistical programming language which is open-source with communities developing many packages that can be installed and are useful for data mining, included below provide comprehensive functionalities to perform range of functionality. analogy for installing package and calling library would be like installing a light bulb once and turning the switch on/off when every required.

```
# import libraries ####
library(shiny)       # for creating interactive web applications
library(ggplot2)     # for creating graphics and visualizations
library(dplyr)       # for data manipulation
library(readr)       # for reading csv files
library(lubridate)   # for working with dates and times
library(tidyr)       # for tidying data
library(shinythemes) # for applying themes to shiny apps
library(GGally)      # for creating scatter plot matrices
library(cluster)     # for clustering analysis
library(rpart)       # for recursive partitioning and regression trees
```

FIGURE 6. IMPORT LIBRARIES

the codes that utilize the functions from different libraries would be: function call: function_name(argument_passing)

'readr' reading rectangular data like CSV files. '<-' is an assignment operator value of right to variable of left.

```
# load the data set ####
data <- read_csv("C:/Users/Tuladhar/Desktop/BD/individual/superstore_dataset.csv")
```

FIGURE 7. LOAD CSV DATA

'lubridate' hands functions related to wrangling, arithmetic date and time with series. access 'order_date', 'ship_date' from dataframe. 'format argument' specify character format to date object

```
data$order_date <- as.Date(data$order_date, format = "%d-%m-%y")
data$ship_date <- as.Date(data$ship_date, format = "%d-%m-%y")
```

FIGURE 8. CONVERT DATA FORMAT

‘dplyr’ for data manipulation functions like select, filter, mutate, summarize.

check for ‘profit_margin’ is not ‘!’ a column ‘colnames’. ‘%in%’ check membership, in ‘if’ statement when {true}

‘%>%’ is pipe operator of ‘dplyr’ to pass dataframe, ‘mutate’ will create new column named ‘profit_margin’ with value of ‘profit/sales’

```r
# calculate profit margin ####
if(!"profit_margin" %in% colnames(data)) {
  data <- data %>% mutate(profit_margin = profit / sales)
}
```

FIGURE 9. CALCULATE PROFIT MARGIN

the reactive expression to filter the user input

‘reactive’ is expression function in shiny for whenever ‘inputs&__’ changes it automatically re-evaluate, ‘filter’ subset rows, ‘input$__ == “___”’ check for match in the column. ‘>= | <=’ greater and less than equal to

```r
# reactive expression to filter data ####
filtered_data <- reactive({
  data %>%
    filter(
      (input$category == "All Categories" | category == input$category), # filter by category
      (input$region == "All Regions" | region == input$region),        # filter by region
      order_date >= input$dateRange[1],                                # filter by start date
      order_date <= input$dateRange[2]                                 # filter by end date
    )
})
```

FIGURE 10. INTERACTIVE FILTERS

‘ggplot2’ for creating advanced graphics, also known as “grammar of graphic”.

‘renderPlot({…})’ reactive and automatic plot output when updated ‘aes’ aesthetic. ‘stat = identity’ height of bar is value

```r
# render plot for total sales by region
output$salesByRegion <- renderPlot({
  filtered_data() %>%
    group_by(region) %>%
    summarise(sales = sum(sales, na.rm = TRUE)) %>%
    ggplot(aes(x = region, y = sales, fill = region)) +
    geom_bar(stat = "identity") +
    labs(title = "Total Sales by Region", x = "Region", y = "Sales")
})
```

FIGURE 11. TOTAL SALES BY REGION

‘GGally’ is an extension of ggplot2 with extended functionality and reduced complexity of plotting matrix

‘select’ specified columns, ‘ggpairs’ scatterplot & correlation coefficient & density plot matrix

```r
# render scatter plot matrix of sales, profit, discount, and quantity
output$scatterMatrix <- renderPlot({
  filtered_data() %>%
    select(sales, profit, discount, quantity) %>%
    ggpairs() +
    labs(title = "Scatterplot Matrix of Sales, Profit, Discount, and Quantity")
})
```

FIGURE 12. SCATTERPLOT MATRIX OF SALES, PROFIT, DISCOUNT, QUANTITY

‘cluster’ to find groups in the data. ‘kmeans’ cluster, ‘geon_point’ add points

```r
# sales profit ####
output$kmeansPlot <- renderPlot({
  cluster_data <- filtered_data() %>%
    select(sales, profit) %>%
    na.omit() # remove rows with missing values
  kmeans_result <- kmeans(cluster_data, centers = 3) # apply k-means clustering with 3 clusters
  cluster_data <- cluster_data %>%
    mutate(cluster = as.factor(kmeans_result$cluster)) # add cluster assignment to data
  ggplot(cluster_data, aes(x = sales, y = profit, color = cluster)) +
    geom_point() +
    labs(title = "K-means Clustering of Sales and Profit", x = "Sales", y = "Profit")
})
```

FIGURE 13. K-MEANS CLUSTERING OF SALES AND PROFIT

‘rpart’ is useful for classification, regression and survival tree. ‘tree_model’ result decision tree, ‘category ~ __+__’ is response and other is predictor

```r
# category classification ####
output$decisionTreePlot <- renderPlot({
  tree_data <- filtered_data() %>%
    select(sales, profit, category) %>%
    na.omit() # remove rows with missing values
  tree_model <- rpart(category ~ sales + profit, data = tree_data, method = "class") # build decision tree model
  rpart.plot::rpart.plot(tree_model, main = "Decision Tree for Category Classification")
})
```

FIGURE 14. DECISION TREE FOR CATEGORY CLASSIFICATION

tidyr’ for tidying data, in a table like format, of rows, columns, cells, pivot, hierarchy. “geom_line”c connect points

```r
# monthly sales over time ####
output$timeSeriesPlot <- renderPlot({
  filtered_data() %>%
    mutate(month = floor_date(order_date, "month")) %>%
    group_by(month) %>%
    summarise(sales = sum(sales, na.rm = TRUE)) %>%
    ggplot(aes(x = month, y = sales)) +
    geom_line() +
    labs(title = "Monthly Sales Over Time", x = "Date", y = "Sales")
})
```

FIGURE 15. MONTHLY SALES OVER TIME

‘shiny’ is useful to build an interactive web application, dashboards and markdown documents.

‘shiny themes’ to apply themes to Shiny apps. I have gone with a darker tone using “superhero”.

```r
# define ui for the shiny app ####
ui <- fluidPage(
  theme = shinytheme("superhero"), # apply superhero theme
  titlePanel("Superstore Sales and Profits Analysis"), # title panel for the app
```

FIGURE 16. USER INTERFACE, THEME, TITLE INITIALIZATION

this sets the title

this defines the sidebar layout for input and output is on the main panel, also consist of input controls

```r
sidebarLayout(
  sidebarPanel(
    selectInput("category", "Select Category:",
                choices = c("All Categories", unique(data$category))), # drop-down for category selection
    selectInput("region", "Select Region:",
                choices = c("All Regions", unique(data$region))),     # drop-down for region selection
    dateRangeInput("dateRange", "Select Date Range:",
                   start = min(data$order_date), end = max(data$order_date)), # date range input for selecting date range
```

FIGURE 17. SIDEBAR, DROPDOWNS, CALENDAR

```r
    # additional text below ####
    tags$hr(),  # horizontal line for separation
    tags$p("Module Code: ITS66904", style = "color: #ffffff; text-align: left;"), # module code
    tags$p("Module Name: Big Data Technologies", style = "color: #ffffff; text-align: left;"), # module name
    tags$p("Task 1: Individual Assignment", style = "color: #ffffff; text-align: left;"), # task information
    tags$hr(),  # horizontal line for separation
    tags$p("Full Name: Sujal Ratna Tuladhar", style = "color: #ffffff; text-align: left;"), # full name
    tags$p("Section, Group: B, 5", style = "color: #ffffff; text-align: left;"), # section and group
    tags$p("Student ID: 22013 230", style = "color: #ffffff; text-align: left;"), # student id
    tags$p("University ID: 036 2483", style = "color: #ffffff; text-align: left;") # university id
  ),
```

FIGURE 18. ADDITIONAL TEXT TO SHOW INFORMATION

```r
mainPanel(
      plotOutput("salesByRegion"),     # plot output for total sales by region
      plotOutput("profitByCategory"),  # plot output for profit margin by category
      plotOutput("discountDistribution"), # plot output for discount distribution
      plotOutput("profitVsSales"),     # plot output for profit vs sales
      plotOutput("quantityByState"),   # plot output for quantity sold by state
      plotOutput("profitMarginPie"),   # plot output for profit margin pie chart
      plotOutput("timeSeriesPlot"),    # plot output for time series plot
      plotOutput("scatterMatrix"),     # plot output for scatter plot matrix
      plotOutput("profitPlot")         # plot output for total profit by region
    )
  )
)
```

FIGURE 19. MAIN PANEL, POLT OUTPUTS

a server functions with reactive expression and rendering functions

```r
# define server logic for the shiny app
server <- function(input, output) {
```

FIGURE 20. SERVER FUNCTION

runs shiny app using the above defined ui and server logics

```r
# run the shiny app
shinyApp(ui = ui, server = server)
```

FIGURE 21. SHINY APP RUN CODE

7. Tool Evaluation Overview

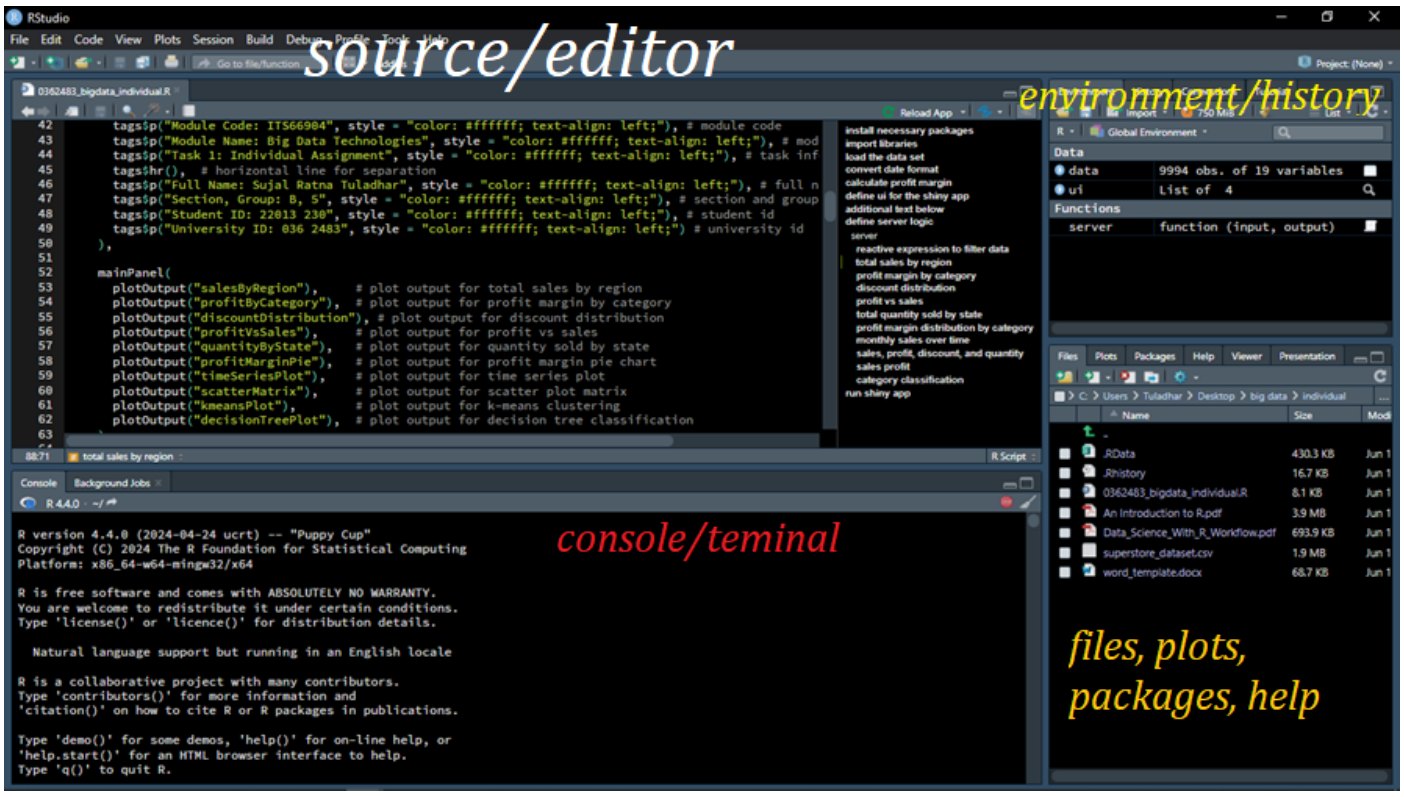


FIGURE 222. R STUDIO INTERFACE

7.1. Techniques like clustering, classification, or time series analysis

- K-means Clustering: we can segment sales and profit data then distinctly identify customers based on their purchasing behavior and tailor market strategies. to high-value and priority customers, discount-seekers, and infrequent buyers' engagement and loyalty.
- Decision Trees Classification: for inventory management, based on their purchase pattern, prediction can be made on features and categories they are most likely to buy. working on their past purchases, it will personalize the recommendation.
- Time Series Analysis: we can identify monthly sales trends, seasonal patterns, then future performance as well as forecasting. these demand pattern can help retailers to avoid overstocking or understocking or being sold out, such techniques will improve inventory planning for supply chains.

7.2. How the dashboard application leverage to process

- This shiny app helps to create an interactive dashboard, and allow users to filter data based on date range, categories of items and regions where the stores branches are located. the users can see detailed insight for any subset of data.
- manipulations and transformations are really effective in r application due to packages that takes care of cleaning, formatting, restructuring, summarizing.
- each reload or refresh the shiny application ensure that the visual feedbacks shown are up to date with the user input, this real time updates and filtration are crucial.
- performance of sales as per categories and analysis of behavior classification, will make advances in making quick and sound data-driven decisions.
- r's ability to have support and adapt to/from both batch and real-time processing, makes it fairly dependent on scalable ever-growing volume of data generated in retail stores.

7.3. Evaluation of Analysis

bar graph for “total sales by 4 regions of America” and “profit margin of categories”

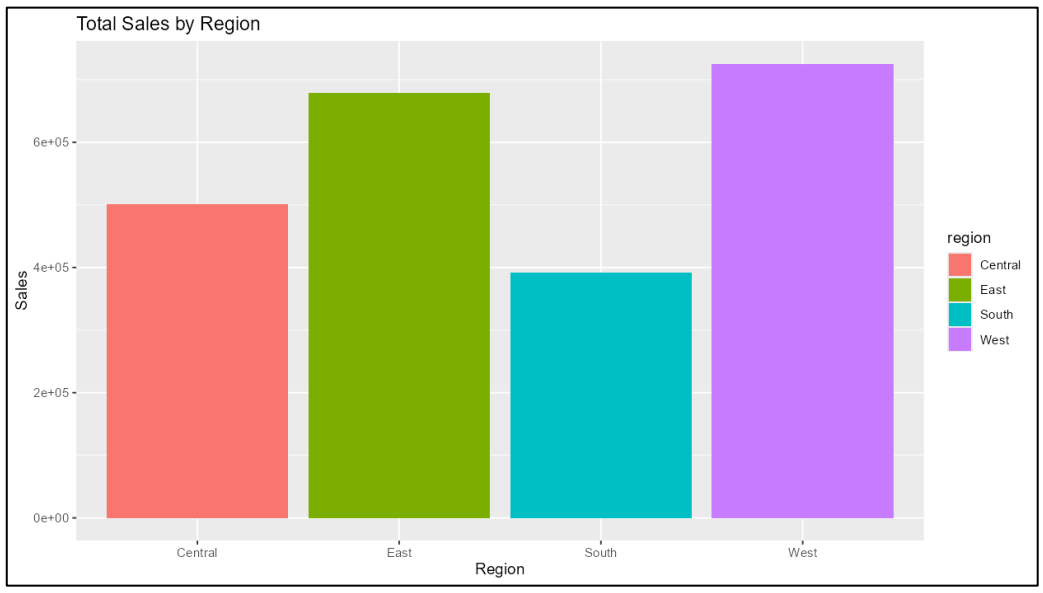


FIGURE 23. TOTAL SALES BY REGIONS (ALL)

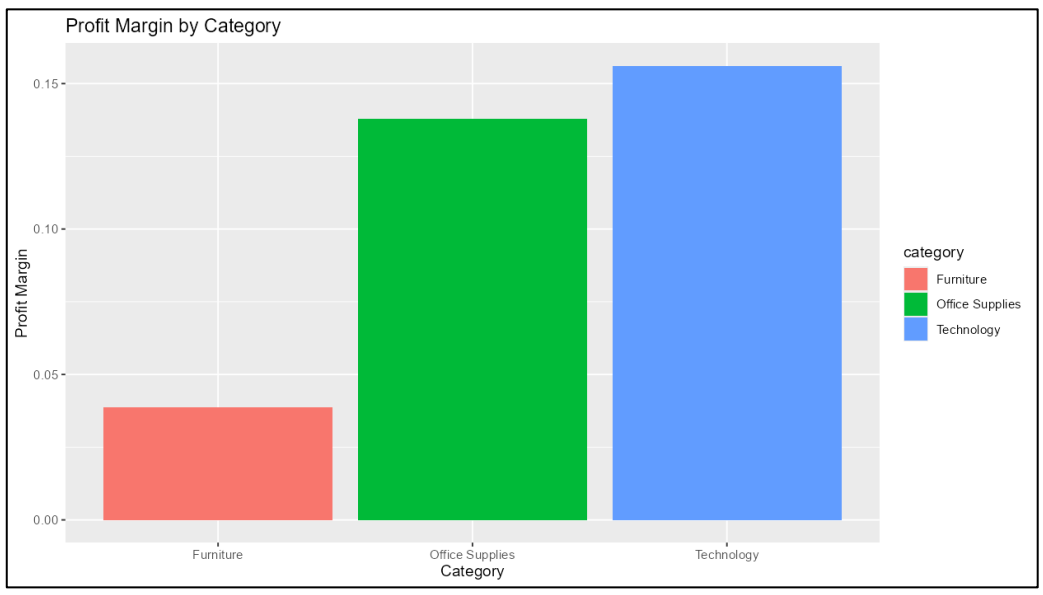


FIGURE 24. PROFIT MARGIN BY CATEGORY

“total quantity sold by each states”

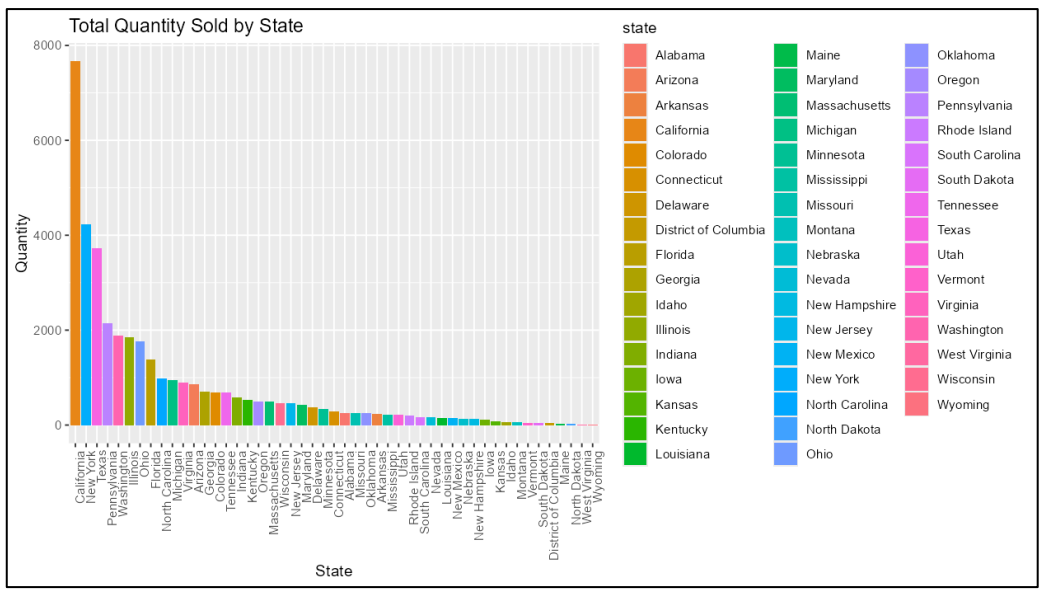


FIGURE 25. TOTAL QUANTITY SOLD BY STATE (ALL)

pie chart of profit margin distribution by category

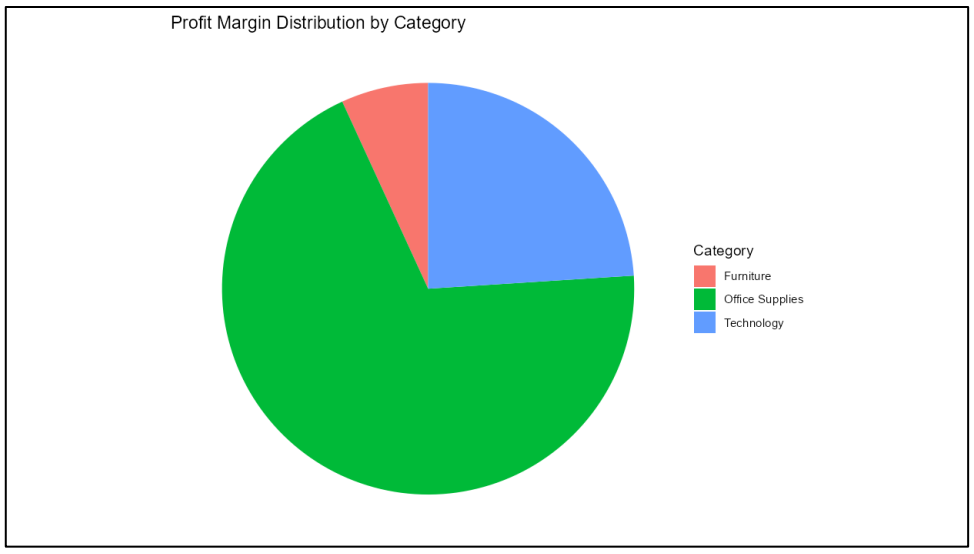


FIGURE 26. PROFIT MARGIN DISTRIBUTION BY CATEGORY

total quantity sold by states

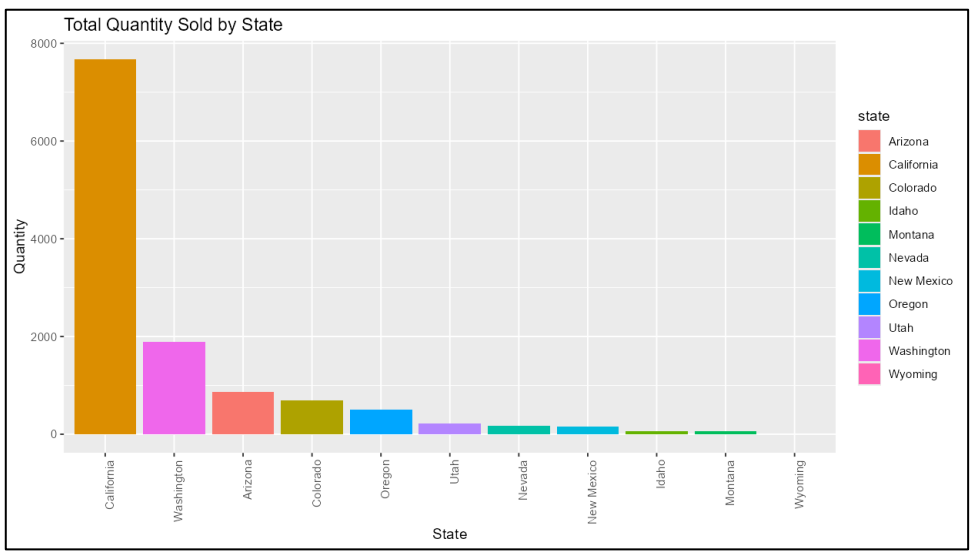


FIGURE 27. TOTAL QUANTITY SOLD BY STATE (WEST)

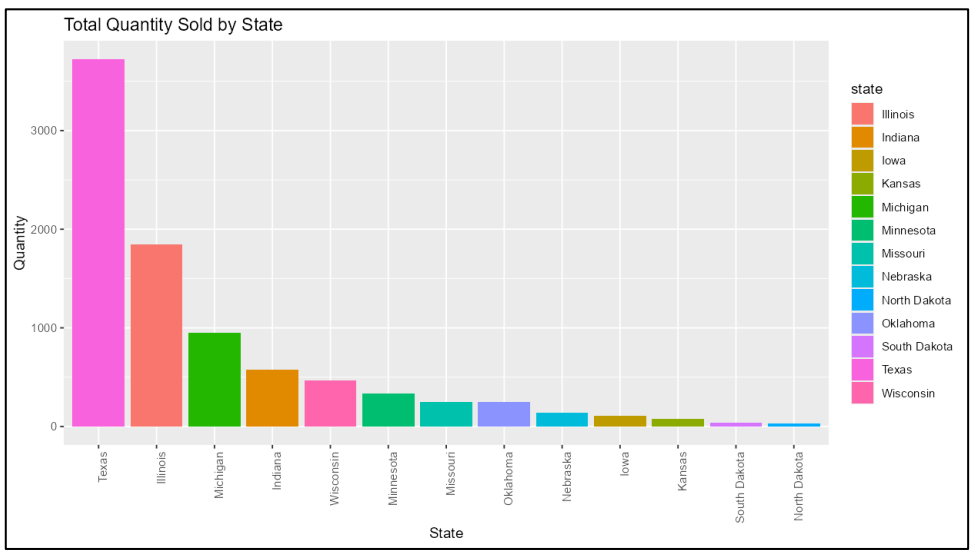


FIGURE 28. TOTAL QANTITY SOLE BY STATE (CENTRAL)

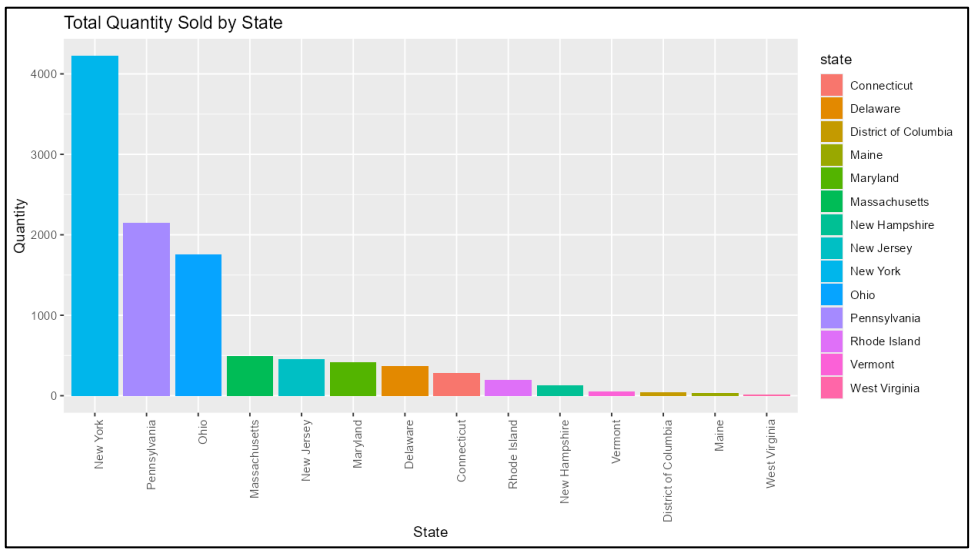


FIGURE 29. TOTAL QUANTITY SOLD BY STATE (EAST)

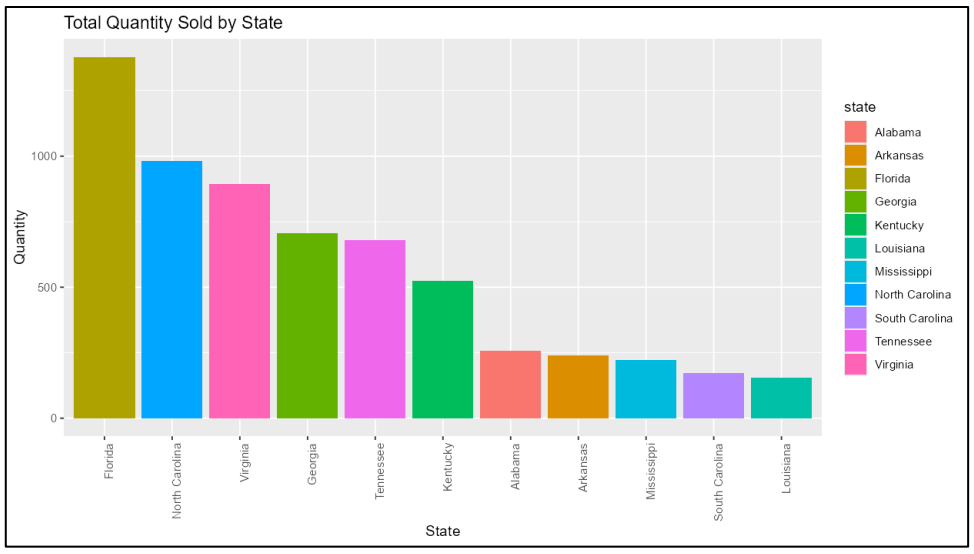


FIGURE 30. TOTAL QUANTITY SOLD BY STATE (SOUTH)

scatter plot and k-means of profit and sales identify positive, negative or correlation. and grouping data to cluster.

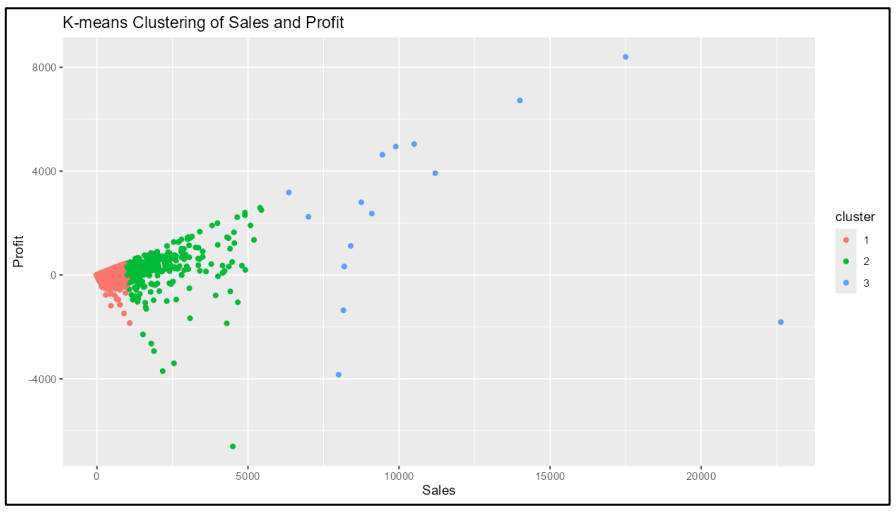


FIGURE 31. K-MEANS CLUSTERING OF PROFIT AND SALES CATEGORY

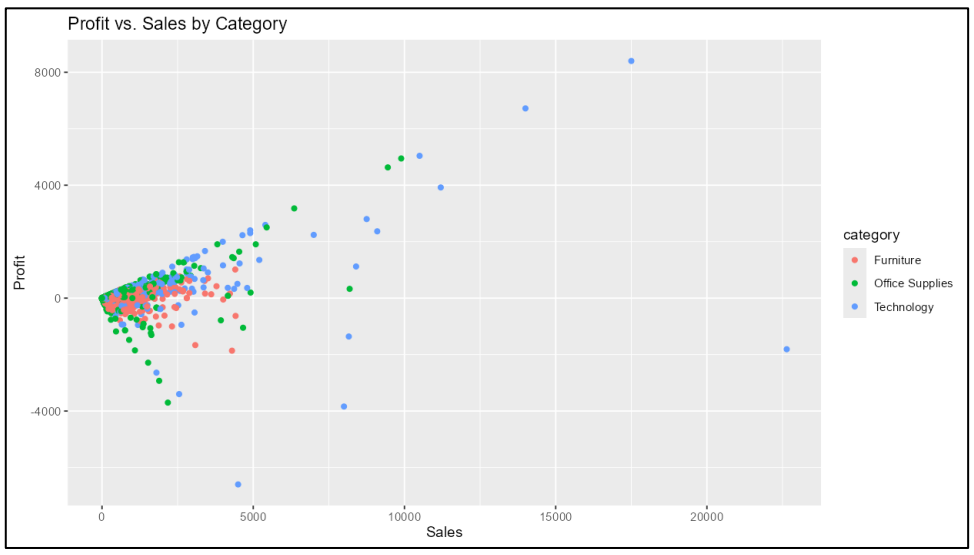
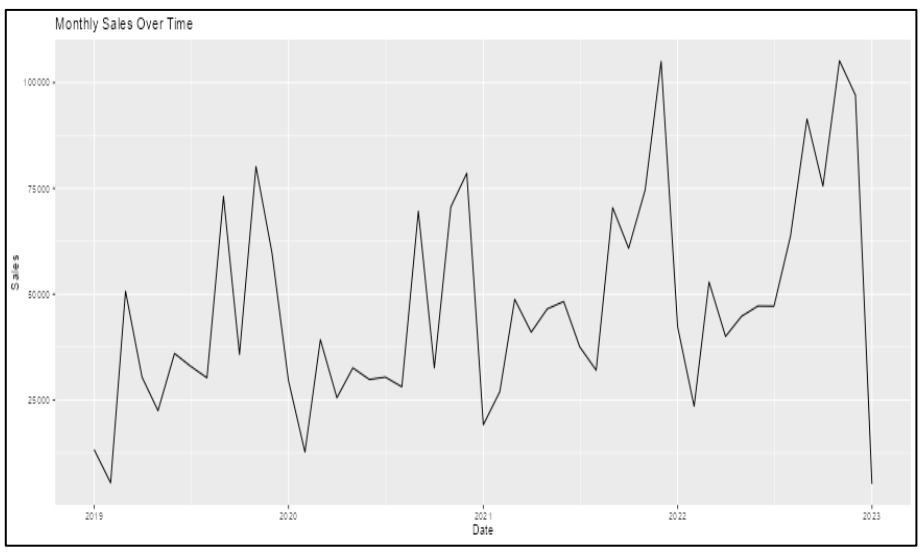


FIGURE 32. PROFIT VS SALES BY CATEGORY AND K-MEANS CLUSTER

time series of sales over dataset date range plotted overtime to see trends and seasonality



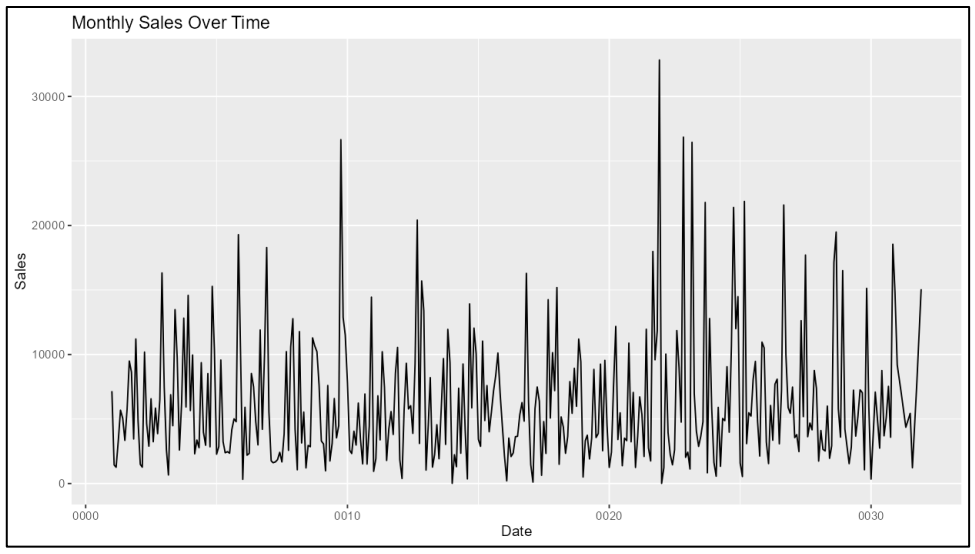


FIGURE33. MONTHLY SALES OVER TIME

scatterplot, correlation coefficient, density plot matrix of sales, profit, discount, quality. relation of pair, statical measurement, indicates strength and direction of linear realtion. off diagonal and pairwise relation for comprehensive review

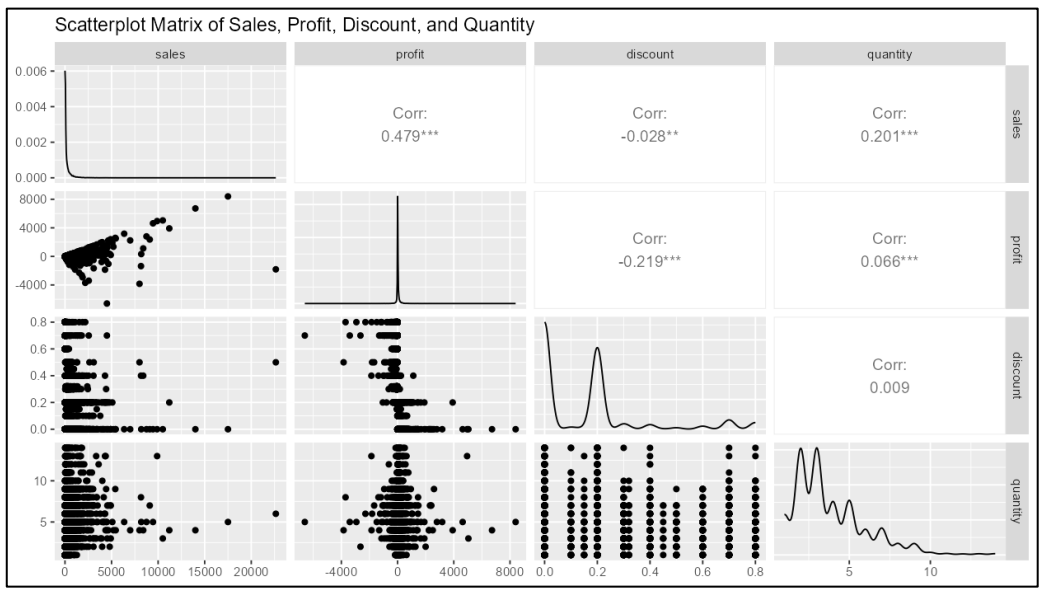


FIGURE 34. SCATTER-PLOT MATRIX OF SALES, PROFIT, DISCOUNT, QUANTITY

decision tree classification for categories, each split is a rule based variable, classify into high, medium , low on features

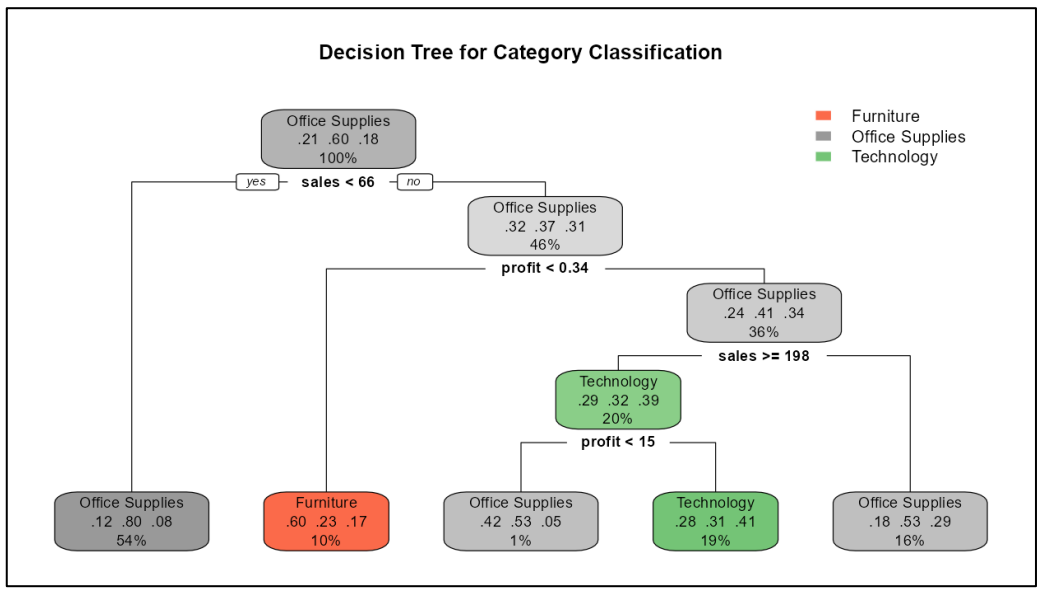


FIGURE 35. DECISION TREE CLASSIFICATION

the interface with interactive dropdown and date ramge

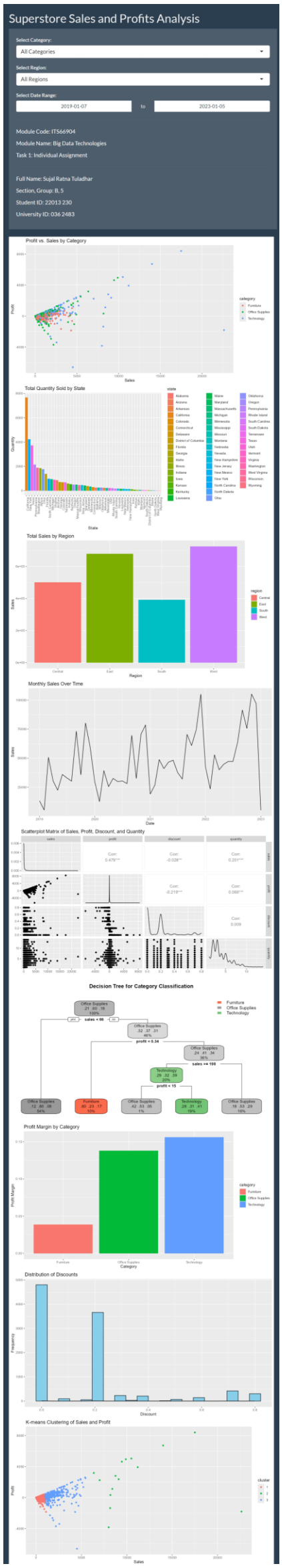


FIGURE 36. R SHINY DASHBOARD

8. Academic Papers

8.1. Article 1

Title: Development of an Application Software for Sales Prediction using Machine Learning Algorithms

Author: Ferjana Ahmed

Case Study: the goal was to evaluate the data mining, machine learning methods, for a prototype software in predicting sales. with the trend of current industries who are very keen on dealing with large volume of data as resource. author believe the processing power and time constraint were the limitations.

Methodologies: the sources she collected data is form open-source data repository like Kaggle it includes different item sales from supermarket and has displayed it in table separating training and testing datasets. she has used multiple analysis and processing methods like univariate, bivariate. cleaning her data, engineering her features, using predictive model with "five-fold cross validation"

Research Findings: the article emphasizes on concept of "feature importance" in predictive model and it is at use at all times. the scores reveal the crucial abilities and those with lower rating can be discarded. it is said that the application is not useful if it is not deployed and utilized. so, the author has used R Studio and its packages.

8.2. Article 2

Title: Does Information Provision Lead to a Better Performance of E-Marketplace Sellers? - An Empirical Analysis -

Author: Statistics Korea

Case Study: due to covid-19 e-marketplace has seen significant rise in popularity. and are very competitive and fierce. so, the support system can enhance performance and strategize. the user emphasize that dashboard and review system have increased sale effectiveness. they focus on online market of South Korea.

Methodologies: to seek the answer, the author had analyzed with empirical and statistical matching testing. to prevent bias, they try to match the covariate, the standard mean difference is balanced (closer to zero). a linear regression analysis showed difference in in percentage, and also substituted a logarithm value

Research Findings: the hypothesis was true and it demonstrates thar the sellers who utilizes the system can earn higher sales and profit from a dashboard. and a revenue also has effect due to reviews. and statistical analysis

8.3. Article 3

Title: Geospatial Operational Insights for ABC Retail: Applying the MCI Model in Shiny R

Author: Gauri Bhatnagar, Thavanesan S/O Sivananthan, Wang Tiantong (Singapore Management University)

Case Study: evaluation of outlet siting location, by analyzing through "prediction of catchment".

Methodologies: in pursuit of long term and infrequent moving decisions, and finding rhe distance from customers zone. the attractiveness is checked

Research Findings: although the paper is related to MCI such will help in improving customer relationship and churn analytics. giving providing proper insights

9. Reference

Ahmed, F. (2023, March 1). *DEVELOPMENT OF AN APPLICATION SOFTWARE FOR SALES PREDICTION USING MACHINE LEARNING ALGORITHMS*. Dspace.mist.ac.bd. https://dspace.mist.ac.bd/xmlui/handle/123456789/807

Bajaj, P., Ray, R., Shedge, S., Vidhate, S., & Shardoor, P. D. N. (2020). *SALES PREDICTION USING MACHINE LEARNING ALGORITHMS*. Semantic Scholar. https://www.semanticscholar.org/paper/SALES-PREDICTION-USING-MACHINE-LEARNING-ALGORITHMS-Bajaj-Ray/54d7b22ce8c266f67a1776e6231e8db4d2c3921b

CRAN packages by name. (2017). R-Project.org. https://cran.r-project.org/web/packages/available_packages_by_name.html

P. Shinde, P., S. Oza, K., & Kamat, R. K. (2017, October 5). *Big data predictive analysis: Using R analytical tool*. IEEE Xplore; 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). https://ieeexplore.ieee.org/document/8058297

Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: Methods, applications, and research opportunities. *Journal of Big Data*, *7*(1), 1–22. Springeropen. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00329-2