**Final Assessment – Question Paper**

| **Semester** | **202310** | | **Division** | **Business** |
| --- | --- | --- | --- | --- |
| **Assessment title in Syllabus** | **Final Assessment** | | **Program** | **Business Analytics** |
| **Version** | **1** | | **Number of pages** | **5** |
|  |  | |  |  |
| **Course Code** | **BNA 3003** | | | |
| **Course Title** | **Business Analytics Application** | | | |
| **Assessment Weight** | **30%** | | **Date** |  |
|  | | | | |
| **Student Declaration**: **Academic Integrity Statement**  In accordance with the HCT Academic Integrity Policy  • Students are required to refrain from all forms of academic integrity breaches as defined and explained by HCT.  • A student found guilty of having committed acts of academic integrity breach(es) will be subject to the relevant sanctions as outlined by HCT.  إفادة النزاهة الأكاديمية  **وفقًا لسياسة كليات التقنية العليا للنزاهة الأكاديمية**  **• على الطلبة الإلتزام بلوائح وقواعد النزاهة الأكاديمية، كما هو مبيّن وموضح في السياسات والإجراءات الخاصة بكليات التقنية العليا.**  **• في حالة ارتكاب الطالب أي شكل من أشكال الإخلال بالنزاهة الأكاديمية، سيتعرض الى العقوبات الموضحة في السياسات ذات الصلة.**  This assignment is entirely my own work except where I have duly acknowledged other sources in the text and listed those sources at the end of the assignment. I have not previously submitted this work to the HCT, or any other entity. I understand that I may be orally examined on my submission.  **Student Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | | | | |
|  | | | | |
| **INSTRUCTIONS:**   * DO NOT open this question paper until you are instructed to do so. * This is a closed-book exam. * All questions must be answered. | | | | |
| * **MATERIALS PERMITTED:** * Pen, pencil, eraser, ruler, color pencils | | * **MATERIALS PROVIDED:** | | |
| **Student Name :** |  | | | |
| **Student HCT ID :** |  | | | |

**For Examiner’s Use Only**

| **Section No.** |  |  |  | **Total** | **%** |
| --- | --- | --- | --- | --- | --- |
| **Marks Allocated** |  |  |  |  |  |
| **Marks Obtained** |  |  |  |  |  |

**Chatbot ( e.g. ChatGPT ) statement**

**This submission does not allow you to use any AI tool to complete. Any breach of this requirement will be treated as a plagiarism attempt and shall be subject to the HCT Academic Integrity Policy.**

INSTRUCTION: Use the R-Studio the following tasks:

• Answer all questions

• Paste R script within the answer and all results should be shown.

• Submit your answer as a pdf file in the link available in blackboard ASD-non-FWA-Project.

**Strategic Analysis for SNS Hypermarket's Market Expansion**

Introduction: SNS Hypermarket, a prominent player in the retail industry, is facing challenges posed by an expanding market and intensified competition. To make informed decisions in this dynamic environment, they have enlisted our expertise to analyze their business data and identify the best strategies to pursue. This case study delves into the analysis of the provided dataset, 'superstoreXX.csv,' containing 200 records of sales data. Our objective is to determine which products, geographical regions, product categories, and consumer segments SNS Hypermarket should target, and conversely, which they should avoid.

Dataset Description: The 'superstoreXX.csv' dataset comprises sales records with various attributes that we will analyze to make data-driven recommendations. Key attributes include:

1. Order ID: A unique identifier for each sales transaction.
2. Product ID: A unique identifier for each product.
3. Product Name: The name of the product sold.
4. Category: The product category (e.g., electronics, furniture, office supplies).
5. Sub-Category: A subcategory that further classifies the product.
6. Sales: The total sales revenue generated by the transaction.
7. Quantity: The quantity of the product sold.
8. Discount: The discount applied to the transaction.
9. Profit: The profit generated from the transaction.
10. Customer Name: The name of the customer.
11. Region: The geographical region where the sale occurred.
12. Segment: The customer segment (e.g., consumer, corporate, home office).

Analysis and Recommendations: To assist SNS Hypermarket in making strategic decisions, you will conduct a comprehensive analysis of the dataset.

**Part 1**

**20 marks**

Before working with the dataset, make sure to load the necessary libraries for data manipulation. Install these libraries if you haven't already using install.packages(). Explore the SNS dataset and its attributes attached by answering the followings:

1. Read the attached dataset csv file into R. Ensure that you specify the correct file path to access the data. To understand the data, perform the following tasks: (10 marks)

a. Provides a quick **overview of the data records**.(2 marks)

> # load libraries

> library(tidyverse)

>

> #read the data into r

> #read the data from the current working directory

> Superstore<- read\_csv("./Superstore.Latifa.csv",

+ col\_types = cols(Sales= col\_double(),

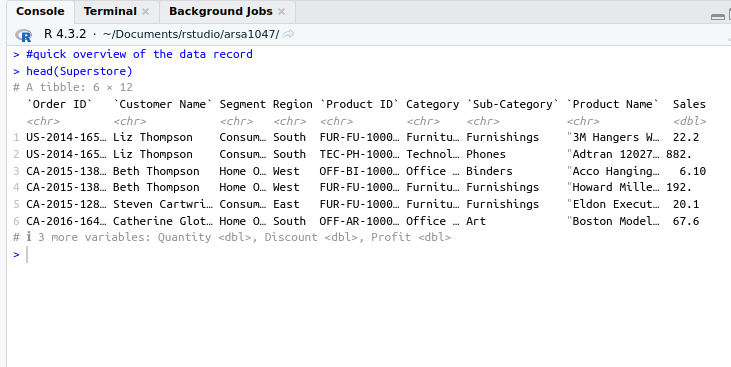
+ Quantity= col\_double(),

+ Discount=col\_double(),

+ Profit= col\_double()))

> #quick overview of the data record

> head(Superstore)

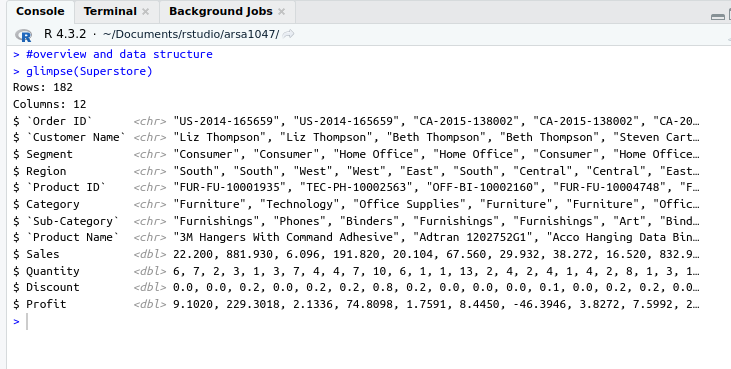


The screenshot shows the first 6 rows of the data

b. Examine **data structure**.(2 marks)

> #overview and data structure

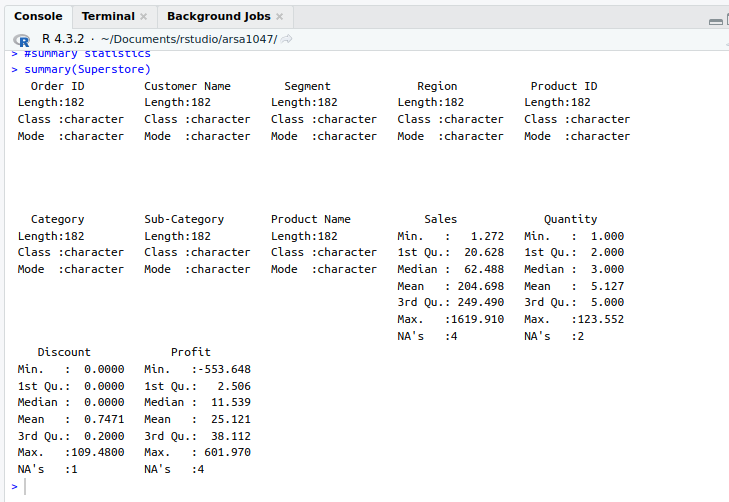
> glimpse(Superstore)



c. **Summary Statistics.(2 marks)**

**> #summary statistics**

**> summary(Superstore)**

****

d. **Describe and evaluate your findings**.(4 marks)

The data was loaded using the read\_csv function. The head function shows the first few rows of the data and helps show an overview. The glimpse function from the library tidyverse shows how the data is structured. From the results, we see that we have data with 182 observations and 12 features. Most the columns are of character data type aprtfrom sales, Discount,Quantity and profit which are double or numeric as shown from the results. Finally, the summary function shows the summary statistics for the data. In our case there are a few missing values under columns, Sales, quantity, discoun and profit. All variables, except for ‘Sales’, ‘Quantity’, ‘Discount’, and ‘Profit’, are of character type. ‘Sales’ has a mean of 204.698 with a range from 1.272 to 1619.910 and 4 missing values. ‘Quantity’ averages at 5.127, ranging from 1 to 123.552, with 2 missing values. ‘Discount’ has a mean of 0.7471, a range from 0 to 109.48, and 1 missing value. Lastly, ‘Profit’ averages at 25.121, with a range from -553.648 to 601.97 and 4 missing values.

1. **Divide** the data into sub-data frames based on the geographical characteristics by breaking down the dataset into smaller, more focused subsets based on geographical attributes. Then, **analyze and describe** the data characteristics, allowing for more targeted analysis or exploration of specific regions or areas within the data. **The result should be evaluated and analyzed within the context of the business domain** (10 marks, 4 marks for code, 6 marks for results evaluation)

> #divide the data into regions

> #check the unique regions in the data

> unique(Superstore$Region)

[1] "South" "West" "East" "Central"

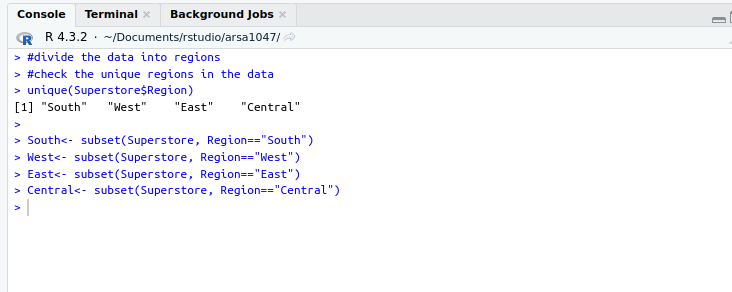
>

> South<- subset(Superstore, Region=="South")

> West<- subset(Superstore, Region=="West")

> East<- subset(Superstore, Region=="East")

> Central<- subset(Superstore, Region=="Central")



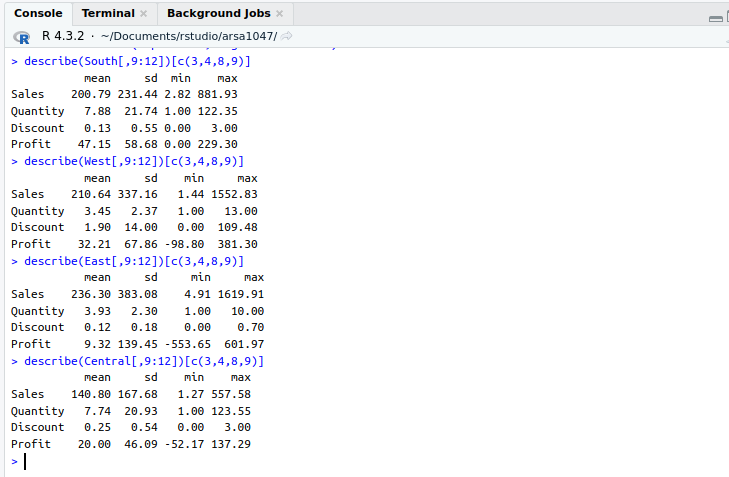
#Decriptive

describe(South[,9:12])[c(3,4,8,9)]

describe(West[,9:12])[c(3,4,8,9)]

describe(East[,9:12])[c(3,4,8,9)]

describe(Central[,9:12])[c(3,4,8,9)]



Notably, the South region stands out with the highest mean sales of $200.79, indicating robust revenue generation. It is closely followed by the East region, which records a mean sales figure of $236.30. In contrast, the West and Central regions exhibit mean sales values of $210.64 and $140.80, respectively, highlighting variation in sales performance among the regions. Furthermore, when it comes to the mean quantity of items sold, the South region leads with a mean of 7.88, closely followed by Central with 7.74.

On the subject of discounts, the West region offers the highest mean discount at 1.90, while Central provides a substantially lower mean discount of 0.25. Notably, Central has the highest maximum discount at 109.48, indicating that some products receive significant reductions. In terms of profitability, the South region once again excels with a mean profit of $47.15, followed by the West region with $32.21. However, the East region exhibits substantial variability in profit, with a broad range from a minimum of -$553.65 to a maximum of $601.97.

South region is the best-performing region in terms of sales.

**Part 2**

**20 marks**

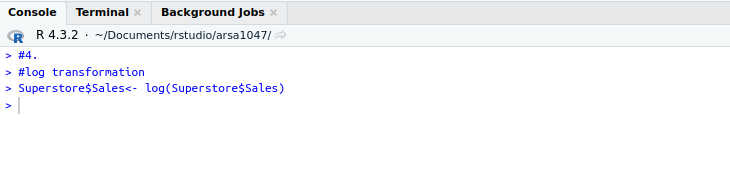
1. Utilizing the same dataset, recommend and apply two pre-processing methods. These methods can involve transforming data conversion from one data type to another. Suggest **TWO (2) attributes** that can be transformed for a better data analytics process. Provide a rationale for each proposed and executed technique **with results evaluation**. (10 marks: 5 marks each)

In our case, the sale might contain outliers as indicated by the large difference between the mean (204.698) and the median (62.488). We can apply a log transformation to reduce the impact of extreme values as shown below.

> #4.

> #log transformation

> Superstore$Sales<- log(Superstore$Sales)

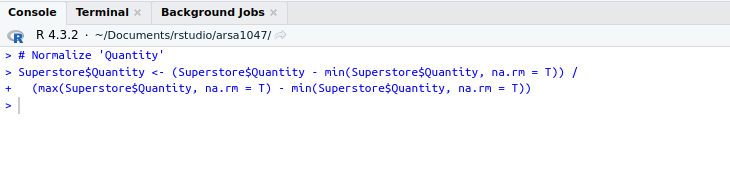


The Quatntity variables seems to have a wide range (from 1 to 123.552). We can normalize this attribute to a range between 0 and 1, which can be useful for certain types of analyses (like clustering or classification) where it’s important that all variables are on the same scale.

> # Normalize 'Quantity'

> Superstore$Quantity <- (Superstore$Quantity - min(Superstore$Quantity, na.rm = T)) /

+ (max(Superstore$Quantity, na.rm = T) - min(Superstore$Quantity, na.rm = T))



1. Within the context of the same dataset, recommend and apply **TWO (2) data cleansing strategies**, such as validating values and handling missing data. Please provide a detailed rationale for each suggested and specify the use of R for the implementation, making it clear which programming language to utilize, **with results evaluation** (10 marks: 5 marks each)

#impute missing values

If the data has missing values they can lead to incorrect analysis or errors. One of the most used strategies is to fill the missing values with the mean (for numerical data) or mode (for categorical data) of the respective column. We only have missing in the numeric variables i.e sales therefore, we can replace the missing values with mean.

> #4.

> #impute missing values

> Superstore$Sales<- impute(Superstore$Sales, mean)

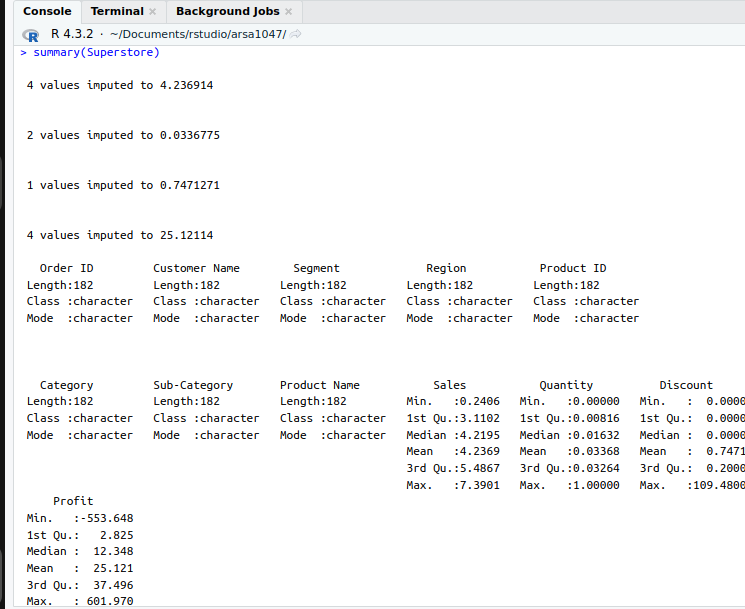
> Superstore$Quantity<- impute(Superstore$Quantity, mean)

> Superstore$Discount<- impute(Superstore$Discount,mean)

> Superstore$Profit<- impute(Superstore$Profit,mean)

> #Data summary

> summary(Superstore)



#Valiadate the data

We can ensure that the values in the dataset are valid and make sense. For example, if you have a ‘Discount’ column, the values should be between 0 and 1

> #validate the data

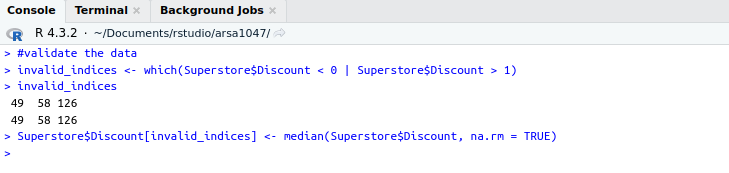
> invalid\_indices <- which(Superstore$Discount < 0 | Superstore$Discount > 1)

> invalid\_indices

49 58 126

49 58 126

> Superstore$Discount[invalid\_indices] <- median(Superstore$Discount, na.rm = TRUE)



**Part 3**

**60 marks**

1. The senior management is keen on exploring sales patterns based on customer demographics. Please suggest and execute **TWO (2) descriptive analytics** and insights that include appropriate data distribution analysis and ggplot visualizations. The results should be shown, explained and evaluated within the business context. (20 marks: 10 marks each)

Sales Distribution by Segment: The first analysis analysis will provide insights into which customer segment contributes the most to the sales.

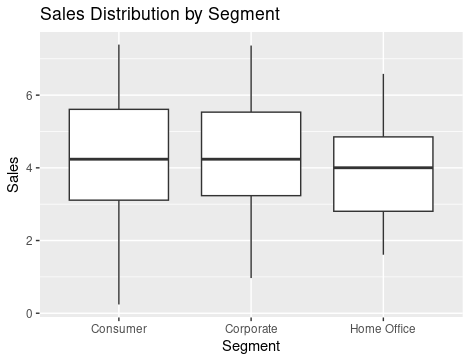
> # Sales Distribution by Segment

> ggplot(Superstore, aes(x=Segment, y=Sales)) +

+ geom\_boxplot() +

+ labs(title="Sales Distribution by Segment", x="Segment", y="Sales")

Don't know how to automatically pick scale for object of type <impute>. Defaulting to continuous.



The boxplot shows no major difference interms of sales between consumers and corporate. The sales for office segments are slightly low than those of consumer and corporate. There are no outliers in the data.

The second comparison is Sales vs Profit by Region: The analysis shows the relationship between sales and profit for each region.

> #Sales vs Profit by Region

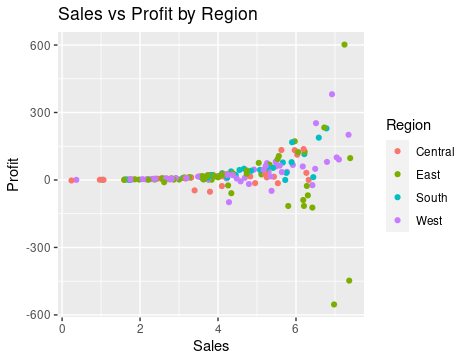
> ggplot(Superstore, aes(x=Sales, y=Profit, color=Region)) +

+ geom\_point() +

+ labs(title="Sales vs Profit by Region", x="Sales", y="Profit")

Don't know how to automatically pick scale for object of type <impute>. Defaulting to continuous.

Don't know how to automatically pick scale for object of type <impute>. Defaulting to continuous.



Most of the sales in the southern region are profitables. East regions seems to have a good number of sales that made loses.

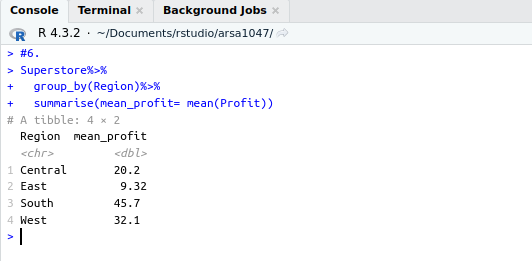
1. Propose and execute **TWO (2) descriptive analytics** and analytical insights providing a relationship between data attributes with class levels. Provide a comprehensive description of the performed analysis.The results should be shown, explained and evaluated within the business context. (20 marks: 10 marks each)

#Average profit by region  
> #6.

> Superstore%>%

+ group\_by(Region)%>%

+ summarise(mean\_profit= mean(Profit))



The mean profit varies by region. The Central region has a mean profit of 20.2, which is the second lowest among the regions. The East region has the lowest mean profit of 9.32. On the other hand, the South region has the highest mean profit of 45.7, followed by the West region with a mean profit of 32.1. This suggests that the South and West regions are more profitable on average compared to the Central and East regions.

# Correlation matrix

> #correlation martrix

> cormat<-cor(Superstore[,9:12])

> corPlot(cormat, cex = 1)



The correlation matrix above shows the correlation between the numeric variables i.e sales, quantity, discount, and profit. The values range from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation. Sales and Quantity have a correlation of approximately 0.051, indicating a very weak positive relationship. This suggests that an increase in sales is slightly associated with an increase in the quantity sold.

The correlation between Sales and Discount is approximately 0.00166, which is almost zero, implying that there is virtually no linear relationship between these two variables. Sales and Profit have a correlation of about 0.242, suggesting a weak positive relationship. This means that an increase in sales is somewhat associated with an increase in profit, but the relationship is not strong. Quantity and Discount have a correlation of about -0.0726, indicating a very weak negative relationship. This suggests that an increase in the quantity sold is slightly associated with a decrease in the discount. The correlation between Quantity and Profit is approximately 0.0241, indicating a very weak positive relationship. This suggests that an increase in the quantity sold is slightly associated with an increase in profit. Finally, Discount and Profit have a correlation of about -0.348, indicating a moderate negative relationship. This means that an increase in the discount is associated with a decrease in profit.

1. Propose and implement a **predictive** analytics insight using **TWO (2)**  algorithms that can be derived from the dataset, evaluate and explain them from a business perspective, emphasizing their practical significance. The results should be shown, explained and evaluated within the business context. (20 marks: 10 marks each)

# Linea regression, model.

We will build a linear regression with the following hypothesis.

H0 (Sales): There is no linear relationship between Sales and Profit. In other words, Sales is not a significant predictor of Profit.

H0 (Quantity): There is no linear relationship between Quantity and Profit. In other words, Quantity is not a significant predictor of Profit.

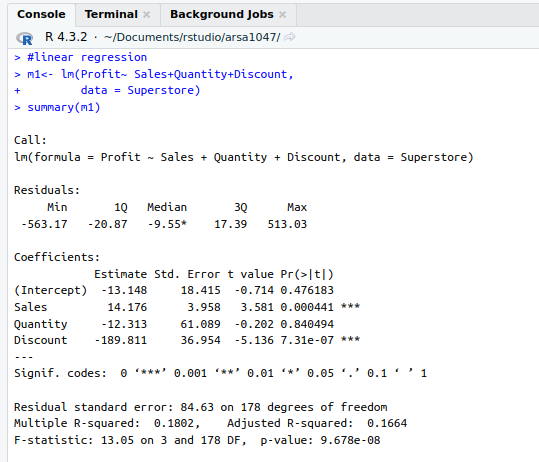
H0 (Discount): There is no linear relationship between Discount and Profit. In other words, Discount is not a significant predictor of Profit.

> #linear regression

> m1<- lm(Profit~ Sales+Quantity+Discount,

+ data = Superstore)

> summary(m1)



The coefficient for Sales is 14.176 and is statistically significant (p-value < 0.001). This suggests that for each unit increase in Sales, the Profit increases by approximately 14.176 units, holding all other variables constant. This is a valuable insight for the business as it emphasizes the importance of increasing sales to boost profits. The coefficient for Quantity is -12.313 but is not statistically significant (p-value = 0.840494). This implies that the Quantity sold does not have a significant impact on Profit. From a business perspective, this could mean that selling more units of a product does not necessarily lead to an increase in profit. The business might need to focus on selling higher-margin products rather than increasing the quantity of lower-margin products sold. The coefficient for Discount is -189.811 and is statistically significant (p-value < 0.001). This suggests that for each unit increase in Discount, the Profit decreases by approximately 189.811 units, holding all other variables constant. This is a critical insight for the business as it highlights the negative impact of discounts on profit. The business might need to reconsider its discount strategy to maximize profits. The model’s R-squared value is 0.1802, indicating that approximately 18.02% of the variation in Profit can be explained by Sales, Quantity, and Discount. This suggests that there are other factors not included in the model that could be influencing Profit.

#Logistic Regression

Next we build a logistic regression model to predict profitable order and non profitable order. The first step is to convert variable profit to a binary variable with profitable order being those above 0 and nonprofitable order being those below or equal to zero. The data will be partitioned into two i.e training and testing at a ration 80% to 20% for train and testing respectively. Finally we wil run a confusion matrix to determine the accuracy of the model.

> # Logistic regression

> # Load necessary libraries

> library(caret)

>

> Superstore<- Superstore[,9:12]

> # Convert Profit to a binary variable

> Superstore$Profit <- ifelse(Superstore$Profit > 0, 1, 0)

>

> # Partition the data into training and testing sets

> set.seed(123) # for reproducibility

> trainIndex <- createDataPartition(Superstore$Profit, p = 0.8, list = FALSE)

> trainSet <- Superstore[trainIndex, ]

> testSet <- Superstore[-trainIndex, ]

>

> # Build the model on the training set

> model <- glm(Profit ~ Sales + Quantity + Discount, data = trainSet, family = binomial)

>

> # Make predictions on the testing set

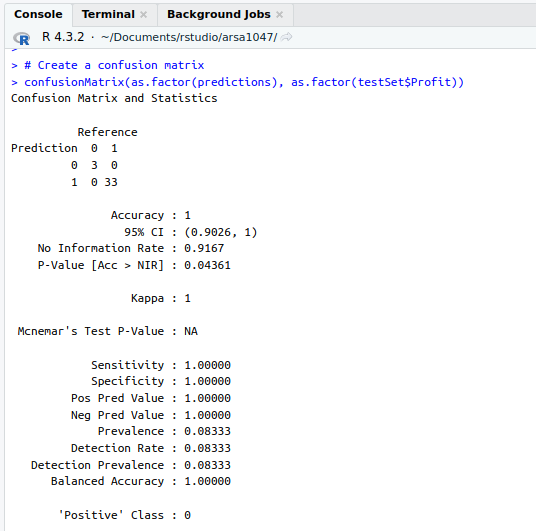
> predictions <- predict(model, newdata = testSet, type = "response")

> predictions <- ifelse(predictions > 0.5, 1, 0)

>

> # Create a confusion matrix

> confusionMatrix(as.factor(predictions), as.factor(testSet$Profit))



The model has an accuracy model has an accuracy of 100%. This means that the model performs extremely well in predicting profitable orders. In a business use case the model can be used to predict non-profitable orders and help make business decisions.

**END OF ASSESSMENT**

The marks/grades are based on the provided instrument, using the following descriptors:

| **Grade: 0-6** |
| --- |
| **Descriptor**: Achievement that does not meet requirements. |
| **Requirements**: |
| ·        Lacks clarity in explanation. |
| ·        Demonstrates little understanding of the report/project content. |
| ·        Cannot answer fundamental questions related to the report/project. |
| ·        Answer lacks structure and coherence. |
| ·        Provides no evidence or weak evidence to support statements. |
| **Grade: 6-7** |
| **Descriptor**: Achievement that minimally meets the course requirements but may not meet the GPA requirements. |
| **Requirements**: |
| ·        Provides basic explanations but with some inconsistencies or errors. |
| ·        Demonstrates an understanding of the main ideas but struggles with details or nuances. |
| ·        Can answer fundamental questions but struggles with more complex inquiries. |
| ·        Provides some evidence to support statements but may not always be relevant or strong. |
| **Grade: 7-8.5** |
| **Descriptor**: Achievement that satisfactorily meets the course and GPA requirements. |
| **Requirements**: |
| ·        Provides clear explanations of the content. |
| ·        Demonstrates a good understanding of both main ideas and supporting details. |
| ·        Can answer both fundamental and some complex questions related to the report/project. |
| ·        Answer provided is structured and coherent |
| ·        Uses relevant evidence consistently to support statements. |
| **Grade: 8.5-10** |
| **Descriptor**: Achievement that is outstanding relative to the course and GPA requirements. |
| **Requirements**: |
| ·        Provides exceptionally clear and insightful explanations. |
| ·        Demonstrates a deep and comprehensive understanding of the report/project content. |
| ·        Can answer all questions, including complex ones, with depth and precision. |
| ·        Presentation is exceptionally structured, fluid, and engaging. |
| ·        Uses strong, relevant evidence consistently and skillfully to support statements. |