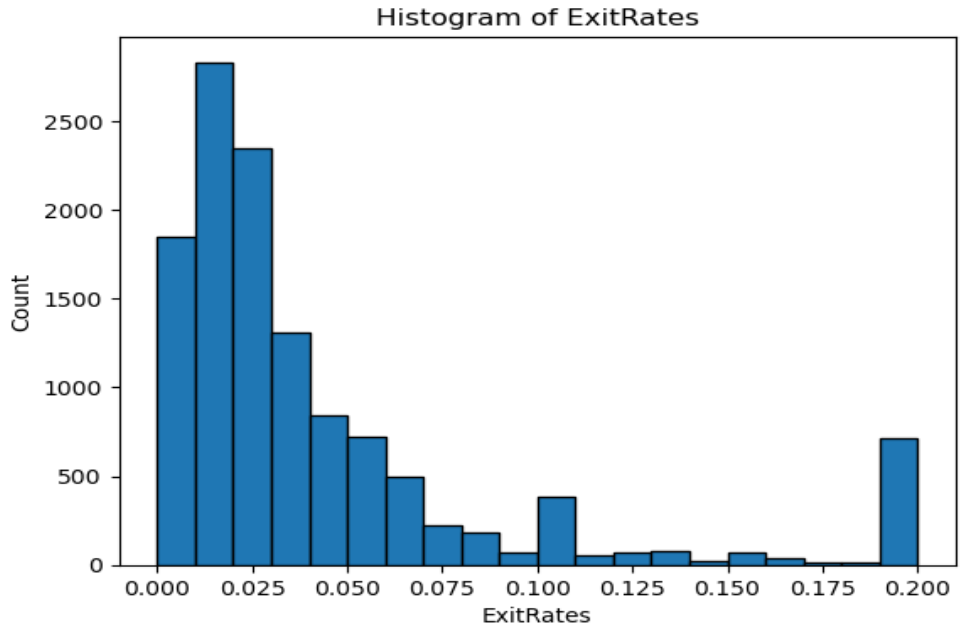
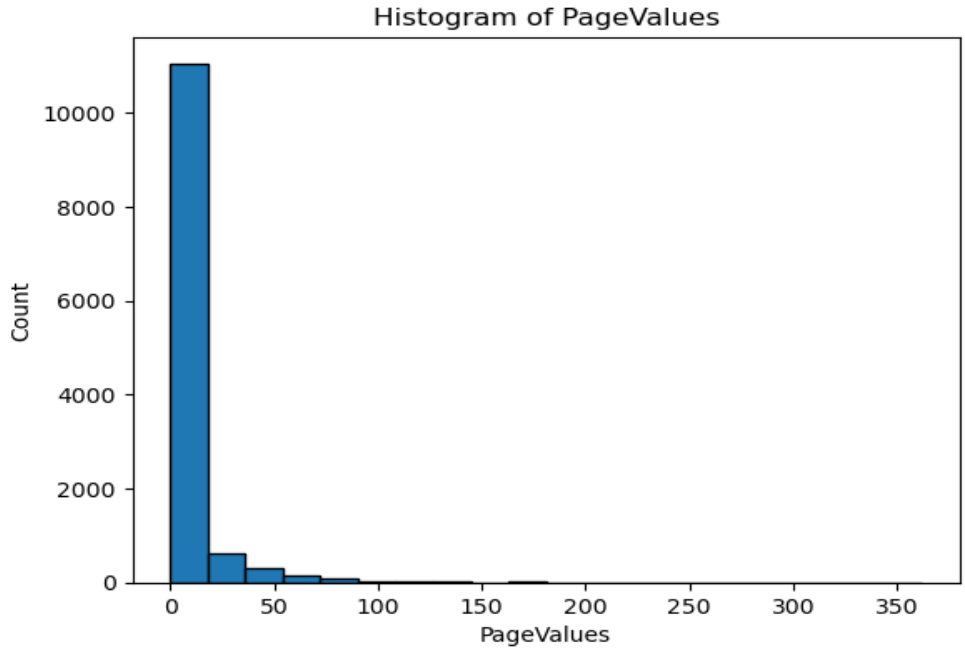
Question 1 :

Code :

```
# Load the dataset
data <-  read.csv ("Desktop/R-final/online_shoppers_intention.csv")
View(data)
# Create separate data frames for ExitRates and PageValues
exitRates <- data$ExitRates
pageValues <- data$PageValues
# Plot histograms side by side
hist_exitRates <- ggplot(data, aes(x = ExitRates)) +
  geom_histogram(binwidth = 0.05, color = "black", fill = "lightblue") +
  labs(x = "ExitRates", y = "Frequency") +
  ggtitle("Histogram of ExitRates")

hist_pageValues <- ggplot(data, aes(x = PageValues)) +
  geom_histogram(binwidth = 0.05, color = "black", fill = "lightblue") +
  labs(x = "PageValues", y = "Frequency") +
  ggtitle("Histogram of PageValues")
# Arrange the histograms side by side
grid.arrange(hist_exitRates, hist_pageValues, ncol = 2)
```





Insights ;

Exit Rate Distribution: The histogram of Exit Rates exhibits a skewed binomial distribution, commonly with a right (positive) skew. This means that there could be a larger number of pages with lower exit rates and a smaller number of pages with higher exit rates.

Page Value Distribution: The histogram of Page Values might also display a skewed exponential distribution, but the skewness might vary depending on the website and the specific goals being measured. It could be either positively or negatively skewed.

Range and Spread:

- Exit Rate Distribution: The range of exit rates observed in the histogram might be relatively broad, ranging from very low exit rates to high exit rates. The spread of the distribution could be influenced by factors such as the quality and engagement of the website content, user experience, and the nature of the website's objectives.
- Page Value Distribution: The range of page values in the histogram might also be wide, encompassing a range of values representing different levels of effectiveness in driving conversions or goal completions. The spread of the distribution could be influenced by factors such as the conversion rate, user behavior patterns, and the value assigned to different goals.

Reasons for Differences:

- Metrics Focus: Exit Rates primarily measure abandonment behavior, while Page Values assess the contribution to conversions. The underlying factors and considerations for these metrics are distinct, leading to differences in their distributions.
- User Behavior: Exit Rates are influenced by factors such as the relevance and engagement of the content, navigation ease, and user intent to leave a particular page. Page Values, on the other hand, depend on user actions leading to conversions or goal completions, which can be influenced by the effectiveness of the page in driving those actions.
- Conversion Funnel: Exit Rates focus on the last page viewed in a session, while Page Values consider the entire conversion path. This difference in scope affects the way the metrics are distributed and the factors that contribute to their values.
- Weighting and Attribution: Page Values assign different weights or values to each page based on its contribution to conversions. This attribution process can introduce variations in the

distribution of Page Values compared to Exit Rates, which typically do not incorporate weighting.

Question 2 :

Code :

```
# Set seed for reproducibility
set.seed(123)
# Generate random indices for training set
trainIndex <- sample(1:nrow(data), size =floor(0.7 * nrow(data)))

# Create training and testing sets
trainData <- data[trainIndex, ]
testData <- data[-trainIndex, ]
# Logistic Regression
logit <- glm(Revenue ~ .,
             data = trainData, family="binomial")
summary(logit)

my_prediction <- predict(logit, testData, type="response")

# building a decision tree
library(rpart)
library(rpart.plot)
library(caret)
library(caTools)
library(pROC)

my_tree <- rpart(Revenue ~ .,
            data = trainData, method="class", cp=0.02)
rpart.plot(my_tree, type=1, extra=1)

predictions <- predict(my_tree, testData, type="prob")
```

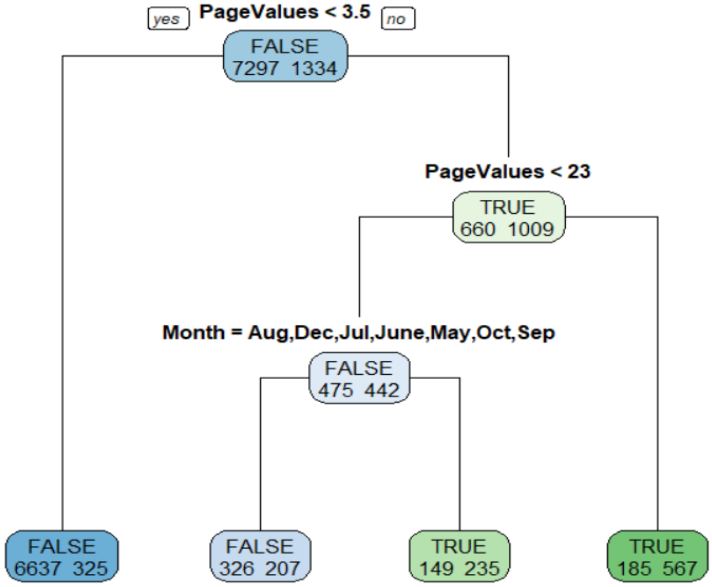
Insights :
Similarities between the logistic model and the tree:

- Both models are used for predictive modeling and can handle binary classification problems.
- Differences between the logistic model and the tree:
- Model Structure: The logistic regression model is a parametric model that estimates the regression coefficients to determine the relationship between the independent variables and the probability of the outcome. In contrast, the tree model is a non-parametric model that uses a hierarchical structure of decision rules to classify observations.
- Interpretability: Logistic regression provides estimates of regression coefficients, representing the magnitude and direction of the relationship between each independent variable and the log-odds of the outcome. These coefficients can be interpreted as the change in the log-odds for a one-unit increase in the corresponding independent variable, assuming all other variables are held constant. In contrast, tree models provide a graphical representation of decision rules and splits but do not directly provide coefficients for interpretation.

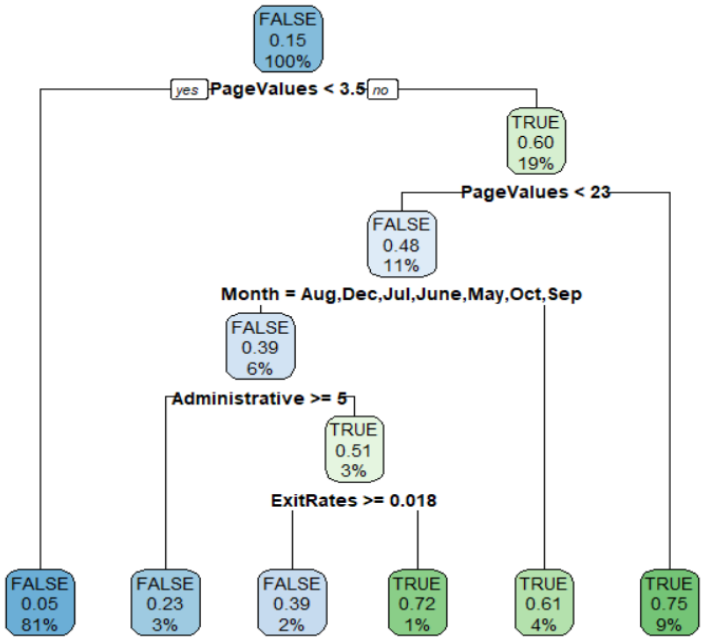
- Handling of Non-linear Relationships: Logistic regression assumes a linear relationship between the independent variables and the log-odds of the outcome. If there are non-linear relationships, additional transformations or interactions may be required. Tree models, on the other hand, can capture non-linear relationships without the need for explicit transformations.
- Handling of Categorical Variables: Logistic regression requires the conversion of categorical variables into dummy variables before fitting the model. Each level of the categorical variable is represented by a separate dummy variable. In tree models, categorical variables can be directly used without the need for dummy variable encoding.

Analyzing the regression coefficients from the logistic model, we can derive the following insights:

- Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, and ProductRelated_Duration have relatively small coefficients, suggesting a weak association with the log-odds of the outcome.
- BounceRates, ExitRates, and SpecialDay have negative coefficients, indicating that higher values of these variables are associated with lower log-odds of the outcome. For ExitRates, the coefficient estimate is -15.85. Exponentiating this coefficient gives us the odds ratio of exp(-15.85) = 3.39e-07. This indicates a significant decrease in the odds of the outcome for a one-unit increase in ExitRates, suggesting that higher ExitRates are associated with a significantly lower likelihood of the outcome.
- PageValues has a positive coefficient, indicating that higher values of PageValues are associated with higher log-odds of the outcome. The coefficient estimate is 0.08293. Exponentiating this coefficient gives us the odds ratio of exp(0.08293) = 1.086. This means that for a one-unit increase in PageValues, the odds of the outcome (e.g., a conversion) increase by approximately 8.6%.
- Month variables such as Dec, Feb, Mar, May, and Nov have negative coefficients, suggesting a lower likelihood of the outcome during these months compared to the reference month. Taking December (MonthDec) as an example, the coefficient estimate is -0.7638. Exponentiating this coefficient gives us the odds ratio of exp(-0.7638) = 0.465. This means that the odds of the outcome are approximately 46.5% lower in December compared to the reference month, while holding all other variables constant.
- OperatingSystems and VisitorTypeReturning_Visitor have negative coefficients, indicating a lower likelihood of the outcome for certain operating systems and returning visitors.
- Browser, Region, TrafficType, Weekend, and VisitorTypeOther have coefficients close to zero, suggesting a weak association with the log-odds of the outcome. For Weekend, The coefficient estimate is 0.1487. Exponentiating this coefficient gives us the odds ratio of exp(0.1487) = 1.16. This means that the odds of the outcome are 16% higher during weekends compared to weekdays, while holding all other variables constant.



Prunned tree :



Question 3 :

Code :

```
### Question 3 :
cm_log <- confusionMatrix(data= as.factor(as.numeric(my_prediction>0.5)),
            reference=as.factor(as.numeric(testData$Revenue)))
cm_tree <- confusionMatrix(data= as.factor(as.numeric(predictions[,2]>0.5)),
```

```
reference=as.factor(as.numeric(testData$Revenue)))
```

Insights:

Comparing the two confusion matrices, one from the logistic regression model and the other from the tree model, we can derive the following insights:

- Accuracy: The glm model has an accuracy of 0.8862, while the tree model has an accuracy of 0.8967. Both models perform well in terms of overall accuracy, with the tree model slightly outperforming the glm model.
- Sensitivity: Sensitivity refers to the ability of the model to correctly identify the positive class. The glm model has a sensitivity of 0.9766, indicating a high rate of correctly identifying the positive class. The tree model has a slightly lower sensitivity of 0.9390 but still performs well in this aspect.
- Specificity: Specificity measures the ability of the model to correctly identify the negative class. The glm model has a specificity of 0.4041, suggesting that it has a lower capability in correctly identifying the negative class. On the other hand, the tree model shows better specificity with a value of 0.6712.
- Balanced Accuracy: Balanced accuracy takes into account both sensitivity and specificity and provides an overall measure of model performance. The tree model has a higher balanced accuracy of 0.8051 compared to the glm model with a value of 0.6903.
- Kappa: Kappa is a statistical measure that assesses the agreement between the predicted and actual classes, considering the possibility of agreement by chance. The tree model has a higher kappa value of 0.6111, indicating a stronger agreement between predicted and actual classes compared to the glm model with a kappa value of 0.4707.

Overall, the tree model demonstrates better performance in terms of accuracy, specificity, balanced accuracy, and kappa compared to the glm model. However, the glm model shows a higher sensitivity, indicating a better ability to correctly identify the positive class. Depending on the specific objectives and requirements of the analysis, one model may be preferred over the other.

Question 4 :

Code :

```
 # Extract numerical variables

numerical_vars <- data[, sapply(data, is.numeric)]

df <- as.data.frame(numerical_vars)


# Create an empty vector to store the results

results <- vector("list", ncol(df))


# Loop through each numerical variable

for (i in 1:ncol(df)) {
```

```
  my_min <- try(min(df[, i], na.rm = TRUE))
  my_max <- try(max(df[, i], na.rm = TRUE))
  my_mean <- try(mean(df[, i], na.rm = TRUE))
  my_std <- try(sd(df[, i], na.rm = TRUE))
  my_median <- median(df[, i], na.rm = TRUE)
  results[[i]] <- c(my_min, my_mean, my_std, my_max, my_median)
}
# Print the results
for (i in 1:ncol(df)) {
  var <- colnames(df)[i]
  cat("Variable:", var, "\n")
  cat("Median:", results[[i]][5], "\n")
  cat("Standard Deviation:", results[[i]][3], "\n\n")
}
```

Insights :

1. Administrative and Administrative_Duration: These variables represent the number and duration of visits to administrative pages on the website. A low median value of 1 for Administrative suggests that most visitors do not frequently visit administrative pages. However, the high standard deviation of 3.32 indicates that there are some visitors who heavily engage with administrative pages, potentially indicating different user behaviors or preferences. The median duration of 7.5 for Administrative_Duration suggests that most administrative page visits are short in duration, but the high standard deviation of 176.78 indicates the presence of outliers with significantly longer durations. These outliers could represent users spending extended periods on administrative tasks or encountering difficulties.
2. Informational and Informational_Duration: These variables represent the number and duration of visits to informational pages on the website. The low median value of 0 for both variables suggests that the majority of visitors do not engage in informational page visits. The presence of outliers with non-zero values for both variables indicates that there are visitors who actively seek out informational content. The higher standard deviation for Informational_Duration (140.75) compared to Informational (1.27) suggests that the duration of informational visits is more variable, with some visitors spending longer periods exploring informational content.
3. ProductRelated and ProductRelated_Duration: These variables represent the number and duration of visits to product-related pages on the website. The moderate median value of 18 for ProductRelated suggests that visitors generally engage in a moderate number of product-related page visits. The high standard deviation of 44.48 for ProductRelated indicates a wide variation in the number of product-related visits, with some visitors exploring a large number of product-related pages. Similarly, the high standard deviation of 1913.67 for ProductRelated_Duration indicates a wide range of durations, with some visits lasting significantly longer. This suggests that there is diversity in user engagement with product-related content, potentially reflecting varying levels of interest, research, or purchase intent.

4. BounceRates and ExitRates: These variables represent the bounce rate and exit rate of visitors on the website. The low median values for both variables (0.0031 for BounceRates and 0.0252 for ExitRates) indicate that most visitors have relatively low bounce and exit rates. This implies that a significant portion of visitors tend to stay on the website and explore multiple pages before exiting. The standard deviations (0.0485 for BounceRates and 0.0486 for ExitRates) suggest some variability in these rates, with certain visitors having higher bounce or exit rates. This variability may be influenced by factors such as website design, content relevance, or user satisfaction.
5. PageValues: This variable represents the average value of the pages visited by a visitor, indicating the potential monetary value generated by the visitor's interactions. The median value of 0 suggests that a majority of visits do not result in immediate transactions or monetary value. However, the standard deviation of 18.57 indicates that there are visitors who generate higher page values, potentially indicating higher engagement or conversion rates. Analyzing the factors contributing to higher page values can provide insights into effective strategies for maximizing visitor value and optimizing conversion rates.
6. SpecialDay: This variable represents the proximity of the visit to a special day (e.g., holidays, festivals). The median value of 0 suggests that most visits occur on non-special days. The standard deviation of 0.1989 indicates some variability in the data, with some visits occurring on special days. Understanding visitor behavior and engagement patterns on special days can help businesses tailor their marketing and promotional activities to leverage these occasions effectively.
7. OperatingSystems: This variable indicates the operating system used by the visitors. The median value of 2 suggests that the majority of visitors use a particular operating system, which could correspond to a popular or widely adopted system. The standard deviation of 0.9113 indicates that there is some diversity in the operating systems used by visitors, with other less common systems also being present. Analyzing the distribution of operating systems can help identify the dominant platforms and optimize website compatibility and user experience accordingly.
8. Browser: This variable represents the web browser used by the visitors. The median value of 2 suggests that there might be a predominant browser choice among the visitors. The standard deviation of 1.7173 indicates some variability in browser preferences, with visitors using different browsers. Understanding the distribution of browsers can guide web development and testing efforts to ensure compatibility and optimize the website's performance across different browsers.
9. Region: This variable indicates the geographic region from which the visitors originate. The median value of 3 suggests that there might be a relatively balanced distribution of visitors across different regions. The standard deviation of 2.4016 indicates some variation in visitor distribution among different regions, with certain regions having more visitors than others. Analyzing the regional distribution can provide insights into target markets, customer segmentation, and localized marketing strategies.
10. TrafficType: This variable represents the type of traffic or referral source that led visitors to the website. The median value of 2 suggests that there might be a dominant traffic type or referral source. The standard deviation of 4.0252 indicates that there is a diverse range of traffic types, with multiple referral sources contributing to website visits. Understanding the distribution of traffic types can help identify effective marketing channels, optimize campaigns, and focus resources on the most valuable traffic sources.