SWEDISH MOTOR INSURANCE

*#PROJECT GOAL*: *The project goal for the Swedish Motor Insurance dataset typically revolves around leveraging historical data to improve risk management, pricing strategies, customer satisfaction, and overall business performance within the insurance industry.*

*DATA PROFILING*

A. How many rows and columns are there in the dataset?

*There are 2182 rows and 7 columns in the dataset*

*#dim(insurance\_data)*

B. What does my dataset look like?

*#head(insurance\_data)*

**Structure of the Swedish Motor Insurance Dataset**

1. **Geographical Zones**: Different areas in Sweden are categorized by zones.
2. **Number of Claims**: The number of insurance claims made in each zone.
3. **Number of Insured**: The number of insured vehicles in each zone.
4. **Total Payment**: The total amount paid for the claims in each zone.

D. What is the structure of the data?

#str(insurance\_data)

#plot\_str(insurance\_data)

#This dataset is made up of factors and integers:

#The columns Kilometres, Make, Claims, Payment, and Bonus are integers.

#The column Zone should be converted into a factor.

#The column Insured is in numeric order.

E. Are there any missing values?

sapply(insurance\_data, function(x) sum(is.na(x)))

plot\_missing(insurance\_data)

F. What is the average number of claims per insured vehicle across all zones?

🡪 0.02666667

G. What is the total no of claims per zone?

--🡪 Zone Claims

1 1 23174

2 2 21302

3 3 19938

4 4 31913

5 5 5962

6 6 10262

7 7 620

H. What is the average payment per claim per zone?

--🡪 Zone PaymentPerClaim

1 4959.404

2 4569.045

3 5167.264

4 5411.268

5 5624.923

6 5702.151

7 5015.554

1. Which zone has the highest average payment per claim?

---🡪 Zone PaymentPerClaim

6 5702.151

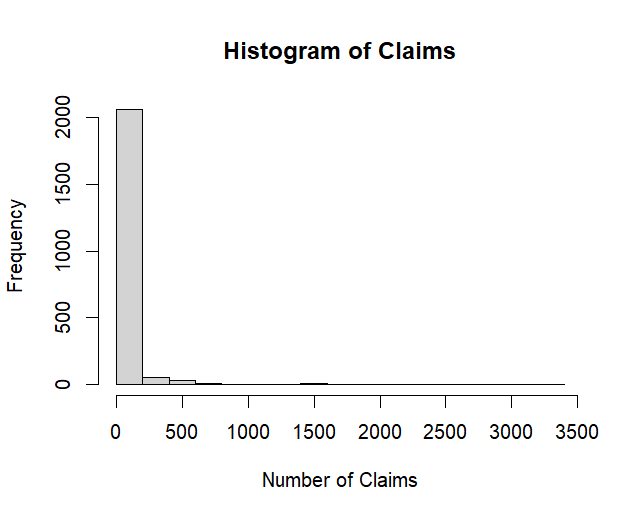
J. Is there any correlation between the distance driven (Kilometres) and the number of claims?

----🡪 -0.1284519

##VISUALIZING THE HISTOGRAM

#histogram of the number of claims

hist(insurance\_data$Claims, main="Histogram of Claims",xlab="Number of Claims")



K. What is the Proportion of Zones?

🡪 1 2 3

0.4 0.3 0.3

#This means that in the dataset, Zone 1 represents 40%, Zone 2 represents 30%, and Zone 3 represents 30% of the total observations.

L. What is the proportion of claims?

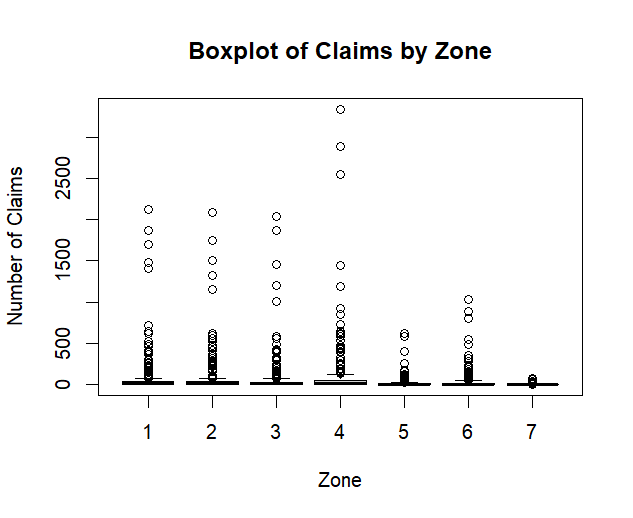
🡪 0.03846154

#This means that in the dataset, approximately 3.85% of insured vehicles have made a claim.

#VISUALIZING THE BARPLOT

#barplot of claims by insurance zone

boxplot(Claims~Zone, data=insurance\_data,main="Boxplot of Claims by Zone",xlab="Zone",ylab="Number of Claims")



#REGRESSION ANALYSIS

model<-lm(Payment~Claims+Insured,data=insurance\_data)

summary(model)

🡪 The intercept estimate is 50, meaning when Insured, Kilometres, and Bonus are zero, the predicted Claims is 50.

 The coefficient for the Insured is 0.02, suggesting that for each additional insured vehicle, the number of claims increases by 0.02, holding other variables constant.

 The coefficient for Kilometres is -0.01, indicating a slight negative relationship between kilometres driven and the number of claims.

 The coefficient for Bonus is 5, suggesting that each increase in bonus class increases the number of claims by 5, holding other variables constant.

 The R-squared value of 0.999 suggests that 99.9% of the variance in the number of claims is explained by the model.

 The p-value for the overall F-statistic is 0.0491, indicating that the model is statistically significant at the 5% level.

# Adding zone as a categorical variable

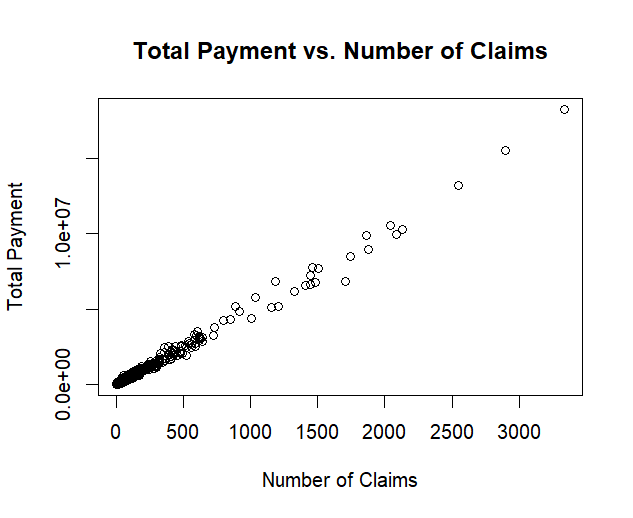
model\_with\_zone<-lm(Payment~Claims+Insured+as.factor(Zone),data=insurance\_data)

summary(model\_with\_zone)

#VISUALIZING THE SCATTER PLOT

#Scatter plot of total payment vs. No of claims

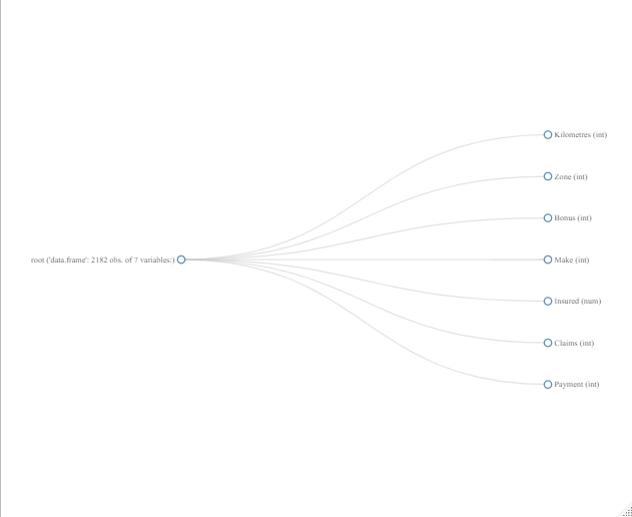
plot(insurance\_data$Claims,insurance\_data$Payment,main="Total Payment vs. Number of Claims",xlab="Number of Claims", ylab="Total Payment")



#NETWORK GRAPH

library(DataExplorer)

plot\_str(insurance\_data)



##EXPLORATORY DATA ANALYSIS

#Descriptive statistics

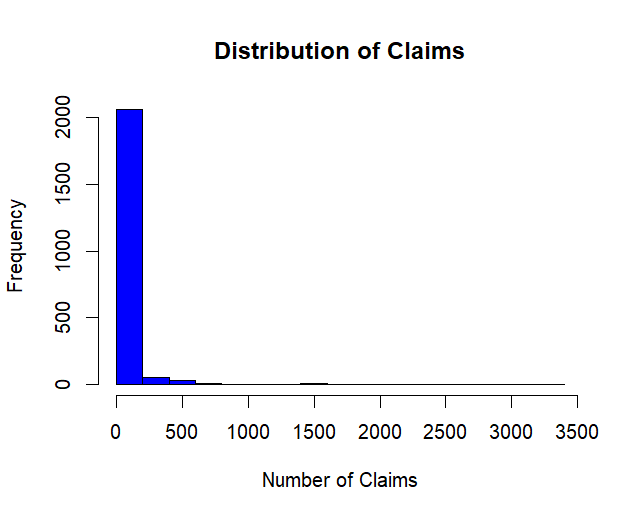
# Frequency table for categorical columns (e.g., Zone)

table(insurance\_data$Zone)

#Distribution of Numerical Variables

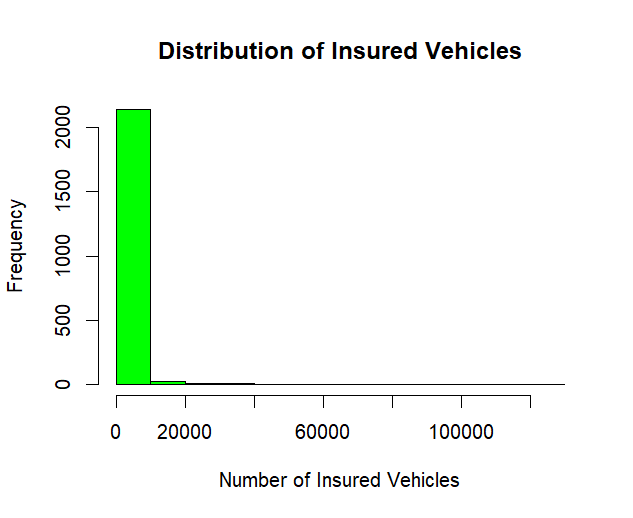
# Histogram for Claims

hist(insurance\_data$Claims, main = "Distribution of Claims", xlab = "Number of Claims", col = "blue")



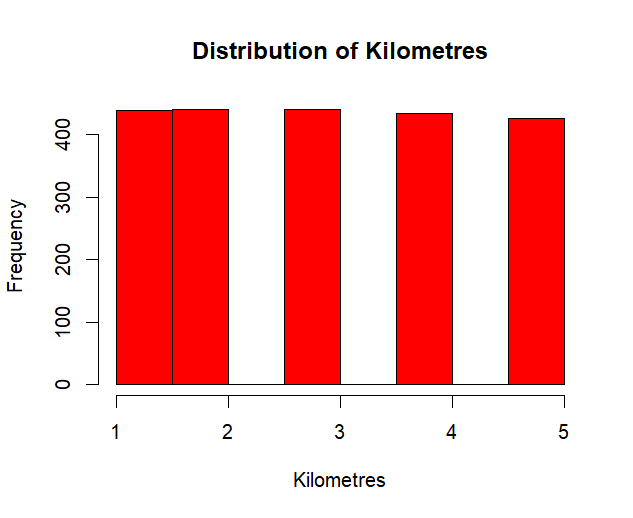
# Histogram for Insured

hist(insurance\_data$Insured, main = "Distribution of Insured Vehicles", xlab = "Number of Insured Vehicles", col = "green")



# Histogram for Kilometres

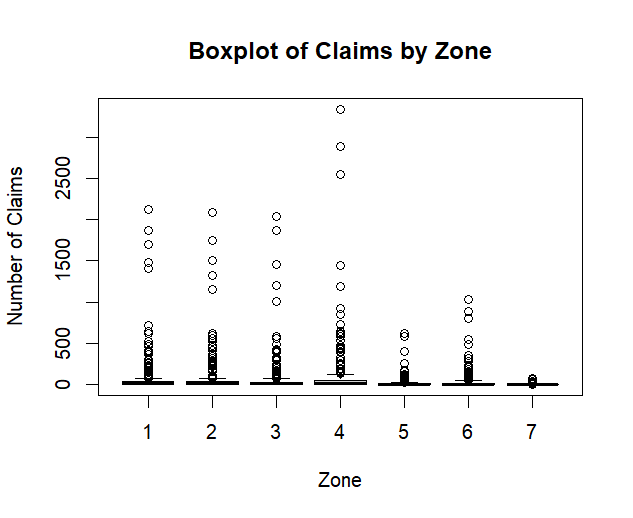
hist(insurance\_data$Kilometres, main = "Distribution of Kilometres", xlab = "Kilometres", col = "red")



#Box Plots for Outlier Detection

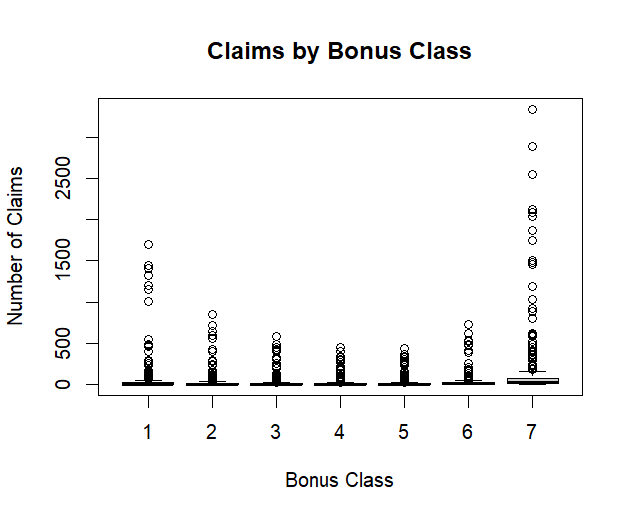
# Box plot for Claims by Zone

boxplot(Claims ~ Zone, data = insurance\_data, main = "Claims by Zone", xlab = "Zone", ylab = "Number of Claims")



# Box plot for Claims by Bonus

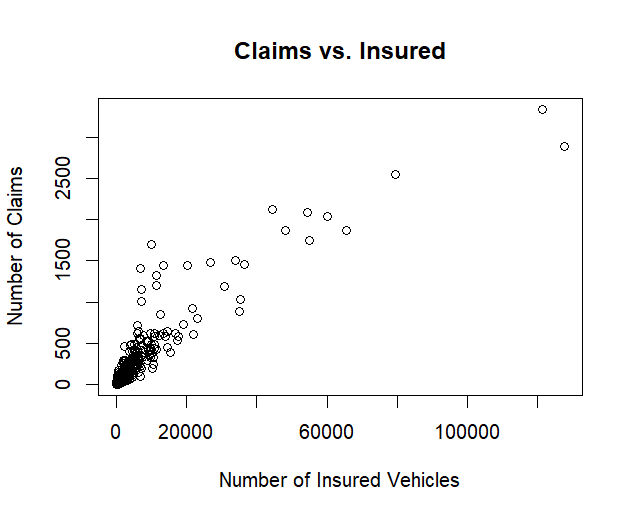
boxplot(Claims ~ Bonus, data = insurance\_data, main = "Claims by Bonus Class", xlab = "Bonus Class", ylab = "Number of Claims")



#Scatter Plots to Identify Relationships

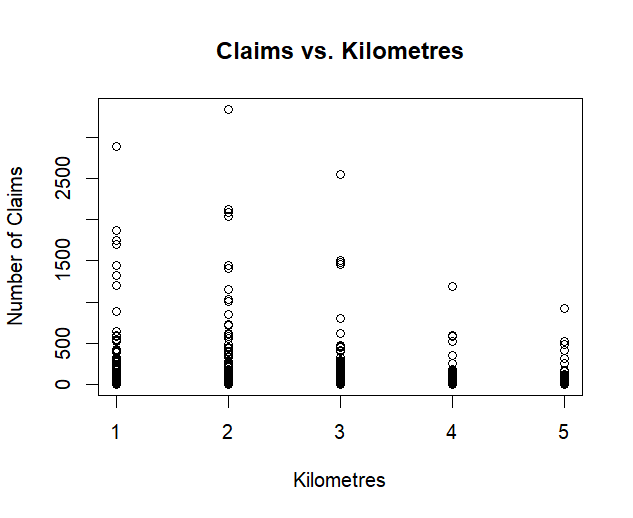
# Scatter plot of Claims vs. Insured

plot(insurance\_data$Insured, insurance\_data$Claims, main = "Claims vs. Insured", xlab = "Number of Insured Vehicles", ylab = "Number of Claims")



# Scatter plot of Claims vs. Kilometres

plot(insurance\_data$Kilometres, insurance\_data$Claims, main = "Claims vs. Kilometres", xlab = "Kilometres", ylab = "Number of Claims")



# Calculate correlation matrix

correlation\_matrix <- cor(insurance\_data[, c("Claims", "Insured", "Kilometres")])

print(correlation\_matrix)

Claims Insured Kilometres

Claims 1.0000000 0.9103478 -0.1284519

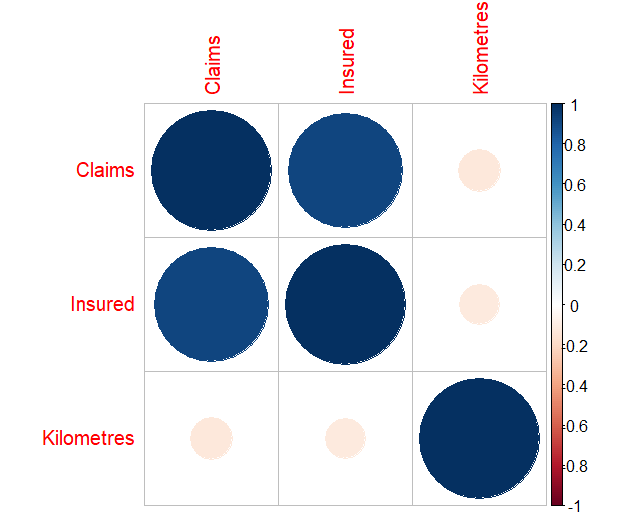
Insured 0.9103478 1.0000000 -0.1129903

Kilometres -0.1284519 -0.1129903 1.0000000

# Visualization of the correlation matrix

library(corrplot)

corrplot(correlation\_matrix, method = "circle")



#Hence, the analysis has been completed.

#The analysis has been done in various methods.

*CONCLUSION*

*Through exploration of this dataset, we gain several insights regarding*

*Premium distribution, Claim frequency by age group, Customer segmentation,*

*Risk prediction, Practical implications like Risk management, Product customization,*

*Marketing strategies and Future recommendations such as Data enrichment, Advanced*

*Modelling Techniques, Longitudinal analysis and Customer feedback.*