# Predicting Infant Mortality using Binary Logistics Regression Model

#### Kuol Alier

#### 2023-12-17

#Importing Infant Mortality Data into R programming.

infant\_mortality <- read.csv("C:/Users/kuolm/Downloads/Infant.csv")

head(infant\_mortality)

## Age Tobacco\_use Smoke\_present Current\_wgt.kg. Height.cm. Birth\_wgt.g. Lbw

## 1 24 No yes 69 159 2500 low

## 2 28 No no 78 168 3300 normal

## 3 22 Yes yes NA NA 2700 low

## 4 29 No no NA NA 2900 normal

## 5 30 No no 52 NA 3000 normal

## 6 22 No yes 62 145 2100 low

## Marital\_status Died Sex2

## 1 Married still birth Female

## 2 Married live birth Female

## 3 Single/ separated still birth Female

## 4 Single/ separated live birth Female

## 5 Married live birth Male

## 6 Married still birth Male

View(infant\_mortality)

#Descriptive Statistics

summary(infant\_mortality)

## Age Tobacco\_use Smoke\_present Current\_wgt.kg.

## Min. :15.00 Length:318 Length:318 Min. : 43.00

## 1st Qu.:21.00 Class :character Class :character 1st Qu.: 55.00

## Median :24.00 Mode :character Mode :character Median : 60.00

## Mean :25.24 Mean : 61.74

## 3rd Qu.:28.00 3rd Qu.: 68.00

## Max. :45.00 Max. :101.00

## NA's :14

## Height.cm. Birth\_wgt.g. Lbw Marital\_status

## Min. : 3.0 Min. :1500 Length:318 Length:318

## 1st Qu.:154.0 1st Qu.:2700 Class :character Class :character

## Median :160.0 Median :3000 Mode :character Mode :character

## Mean :152.1 Mean :3015

## 3rd Qu.:162.0 3rd Qu.:3400

## Max. :175.0 Max. :4300

## NA's :25 NA's :4

## Died Sex2

## Length:318 Length:318

## Class :character Class :character

## Mode :character Mode :character

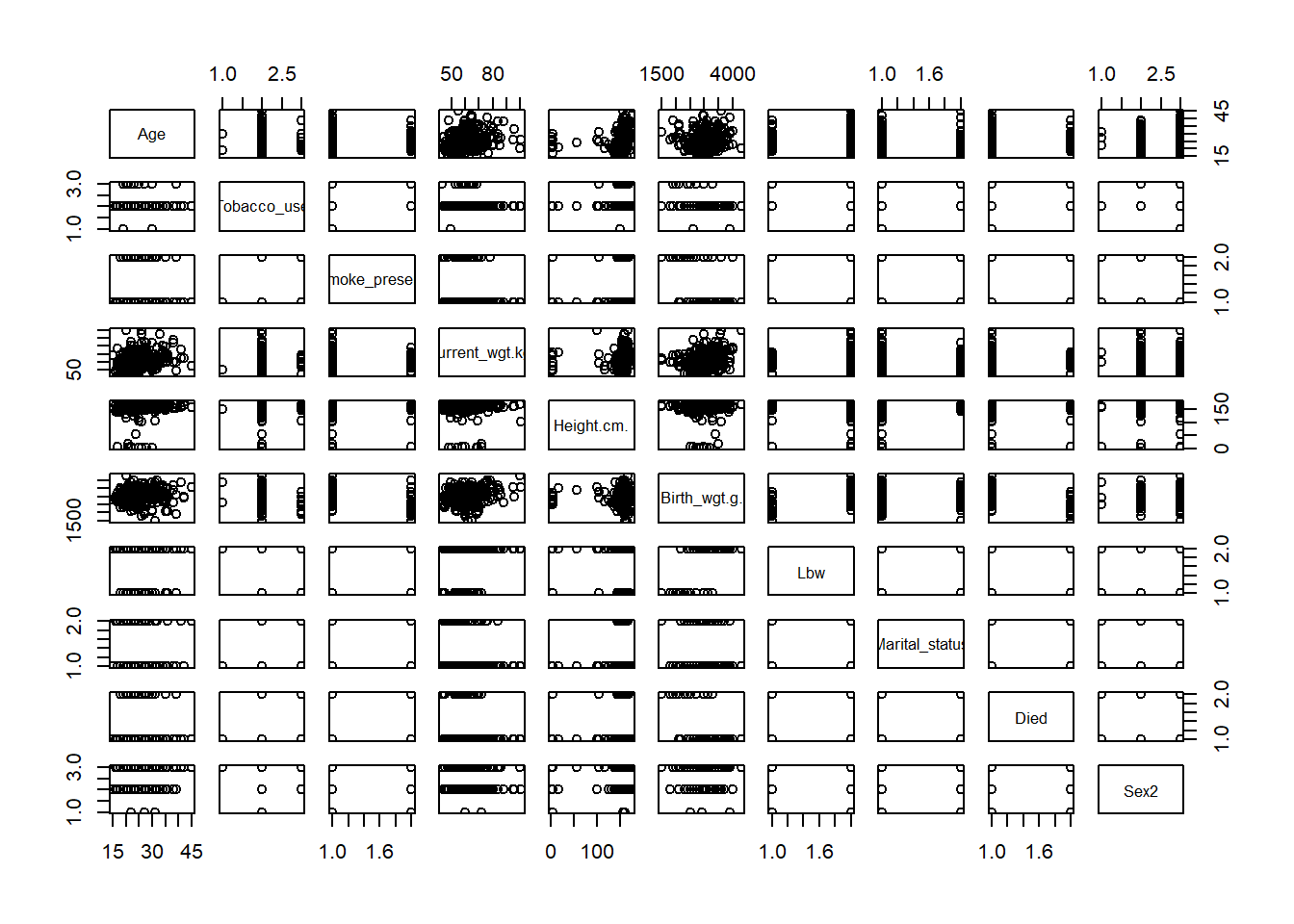
##

##

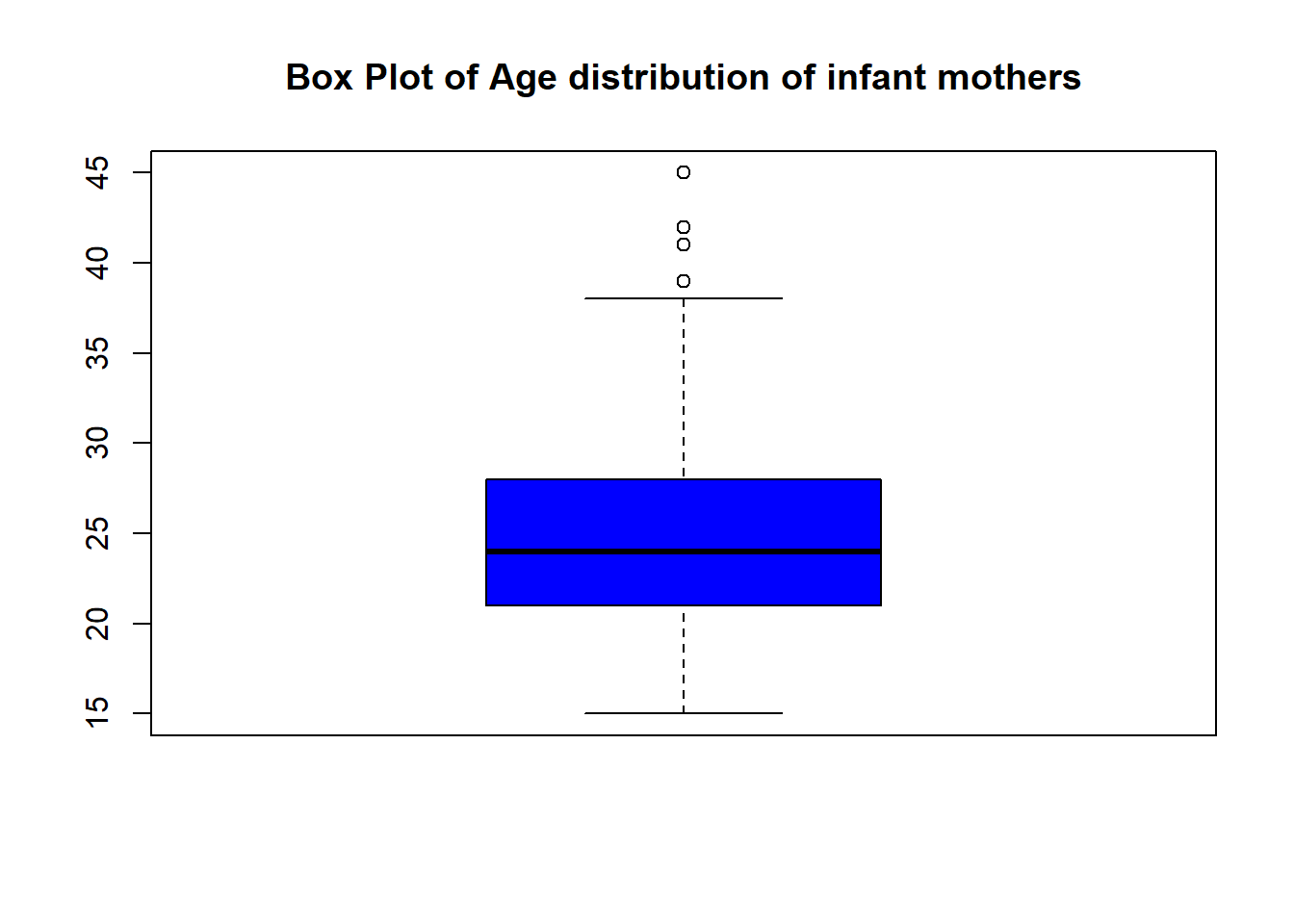
##

##

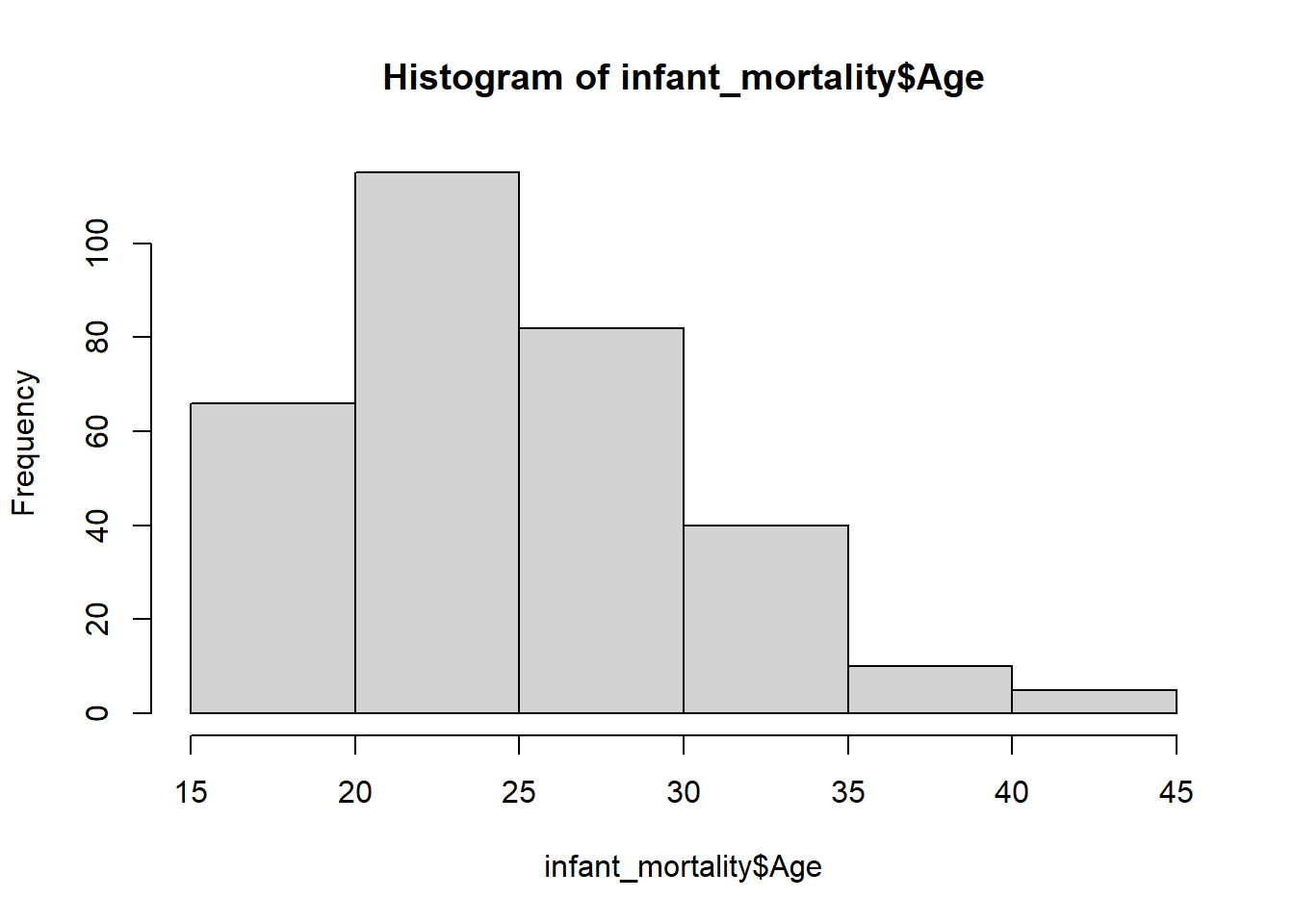
plot(infant\_mortality)



boxplot(infant\_mortality$Age, col = "Blue", main = "Box Plot of Age distribution of infant mothers")



hist(infant\_mortality$Age, border = "Black")



missing\_values <-colSums(is.na(infant\_mortality))

missing\_values

## Age Tobacco\_use Smoke\_present Current\_wgt.kg. Height.cm.

## 0 0 0 14 25

## Birth\_wgt.g. Lbw Marital\_status Died Sex2

## 4 0 0 0 0

#Imputation Technique- Data Preprocessing process.

infant\_mortality$Current\_wgt.kg.[is.na(infant\_mortality$Current\_wgt.kg.)]<-mean(infant\_mortality$Current\_wgt.kg., na.rm = TRUE)

infant\_mortality$Height.cm.[is.na(infant\_mortality$Height.cm.)] <- mean(infant\_mortality$Height.cm., na.rm = TRUE)

infant\_mortality$Birth\_wgt.g.[is.na(infant\_mortality$Birth\_wgt.g.)] <- median(infant\_mortality$Birth\_wgt.g., na.rm = TRUE)

missing\_values<-colSums(is.na(infant\_mortality))

missing\_values

## Age Tobacco\_use Smoke\_present Current\_wgt.kg. Height.cm.

## 0 0 0 0 0

## Birth\_wgt.g. Lbw Marital\_status Died Sex2

## 0 0 0 0 0

#Correlation between Associate variables

cor(infant\_mortality$Age, infant\_mortality$Birth\_wgt.g.)

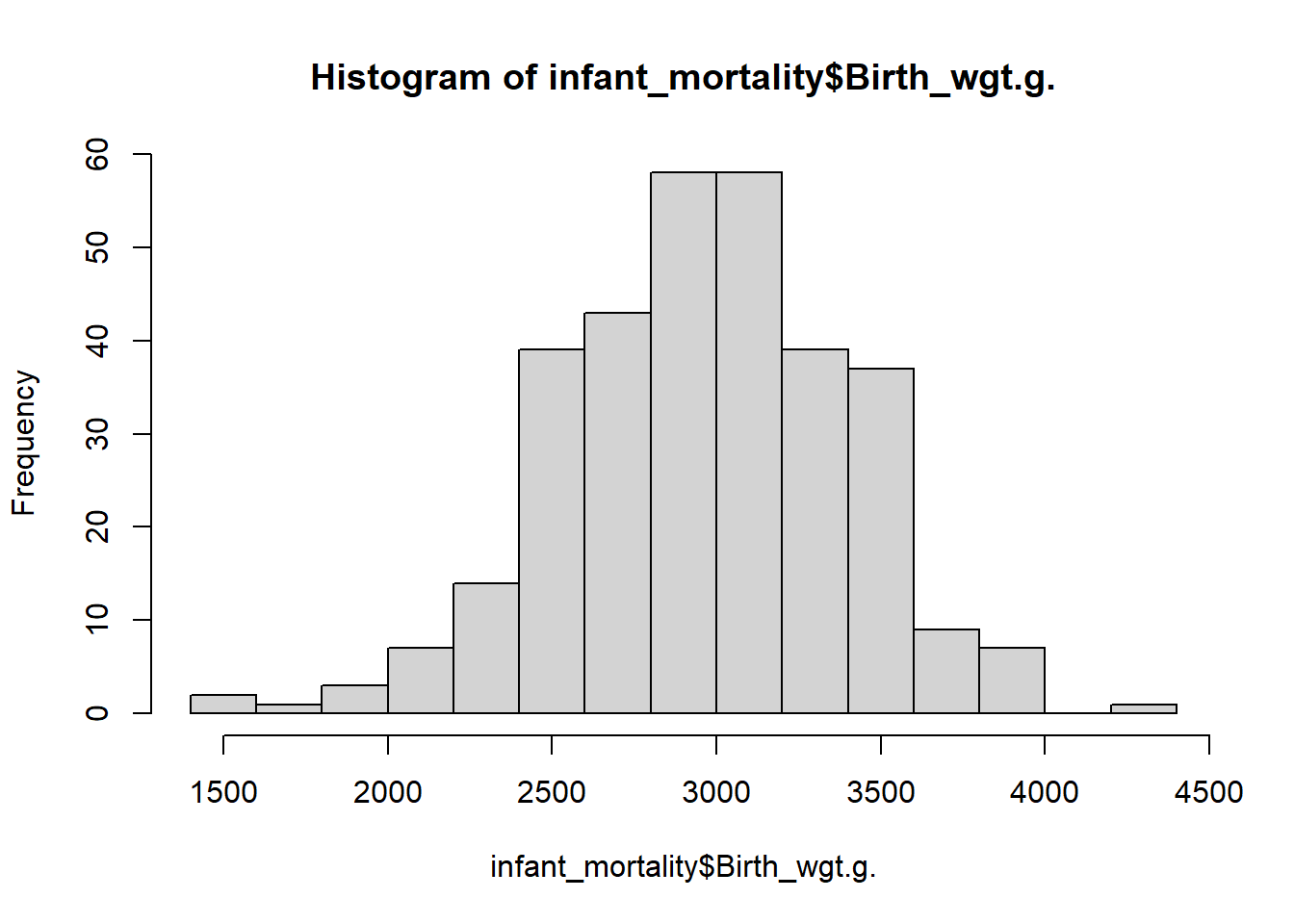
## [1] 0.09008078

cor(infant\_mortality$Age, infant\_mortality$Current\_wgt.kg.)

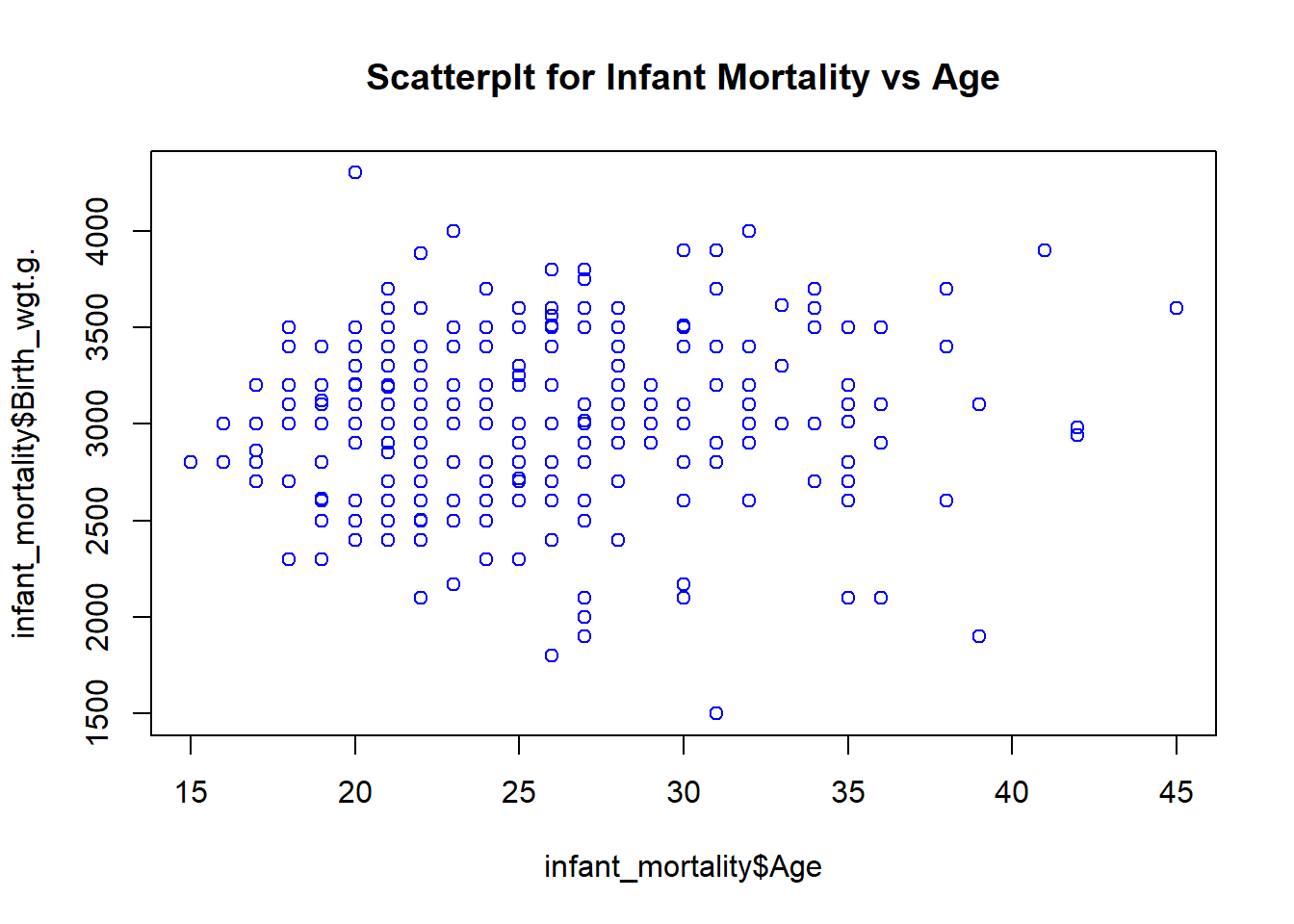
## [1] 0.2396127

#Data Visualization using Histogram, Scatter plot and Boxplot

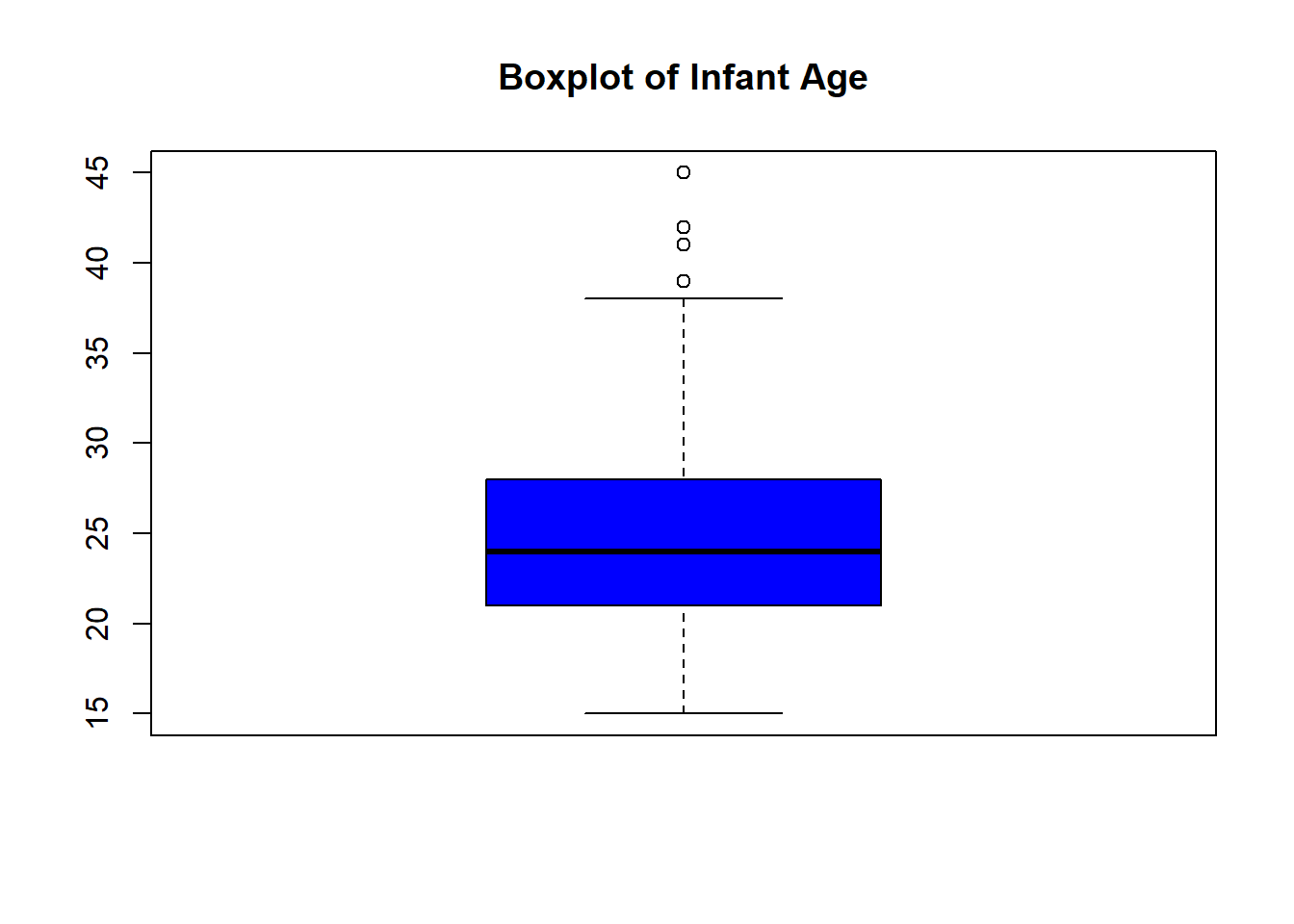
hist(infant\_mortality$Birth\_wgt.g.)



plot(infant\_mortality$Age, infant\_mortality$Birth\_wgt.g., col= "Blue", main = "Scatterplt for Infant Mortality vs Age")



boxplot(infant\_mortality$Age, col = "Blue", main = "Boxplot of Infant Age")



selected\_variable <- infant\_mortality[, c("Age", "Current\_wgt.kg.", "Height.cm.", "Birth\_wgt.g.")]

cor\_matrix <- cor(selected\_variable)

cor\_matrix

## Age Current\_wgt.kg. Height.cm. Birth\_wgt.g.

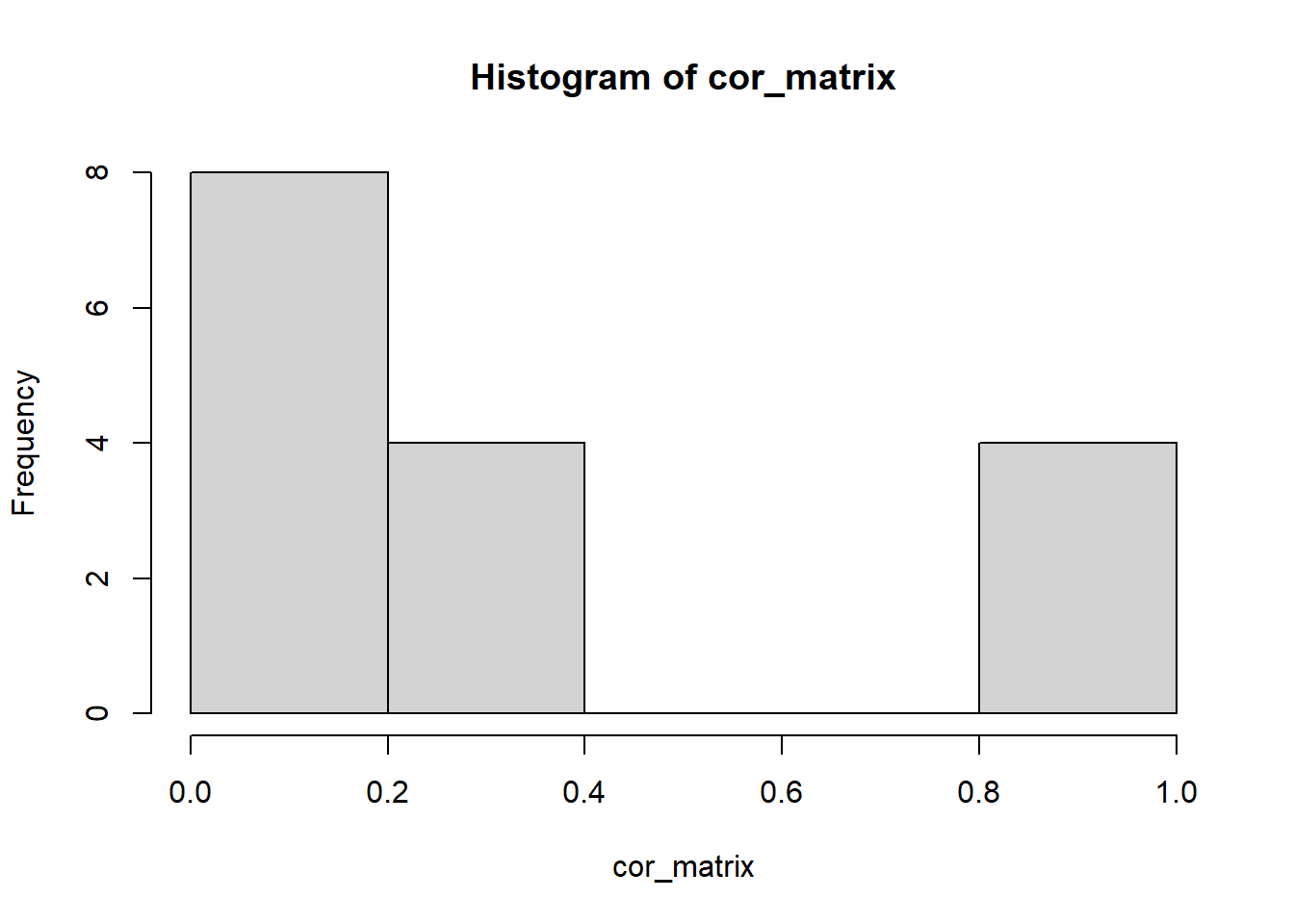
## Age 1.00000000 0.23961267 0.11615234 0.09008078

## Current\_wgt.kg. 0.23961267 1.00000000 0.09725095 0.22145053

## Height.cm. 0.11615234 0.09725095 1.00000000 0.04657013

## Birth\_wgt.g. 0.09008078 0.22145053 0.04657013 1.00000000

hist(cor\_matrix)

#Fitting Linear Regression Model for Cuurrent Wgt (Depnt\_Var) and Height cm (Indept\_Var)

model<-lm(infant\_mortality$Current\_wgt.kg. ~ infant\_mortality$Height.cm., data = infant\_mortality)

summary(model)

##

## Call:

## lm(formula = infant\_mortality$Current\_wgt.kg. ~ infant\_mortality$Height.cm.,

## data = infant\_mortality)

##

## Residuals:

## Min 1Q Median 3Q Max

## -18.990 -6.628 -0.624 4.248 40.911

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 56.83477 2.87161 19.792 <2e-16 \*\*\*

## infant\_mortality$Height.cm. 0.03222 0.01855 1.737 0.0834 .

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 9.464 on 316 degrees of freedom

## Multiple R-squared: 0.009458, Adjusted R-squared: 0.006323

## F-statistic: 3.017 on 1 and 316 DF, p-value: 0.08336

#Converting categorical variables into Numerical factors

infant\_mortality$Marital\_status <- ifelse(infant\_mortality$Marital\_status=="Married", 1, 0)

infant\_mortality$Sex2 <- ifelse(infant\_mortality$Sex2=="Female", 1, 0)

infant\_mortality$Tobacco\_use <- ifelse(infant\_mortality$Tobacco\_use=="Yes", 1, 0)

infant\_mortality$Lbw <- ifelse(infant\_mortality$Lbw=="normal", 1, 0)

infant\_mortality$Died <-ifelse(infant\_mortality$Died=="live birth",1, 0)

#Fitting Logistics Regression Model to predict infant mortality using maternal risk factors

model2 <-glm(infant\_mortality$Died ~ infant\_mortality$Smoke\_present + infant\_mortality$Marital\_status, data = infant\_mortality, family = binomial)

summary(model2)

##

## Call:

## glm(formula = infant\_mortality$Died ~ infant\_mortality$Smoke\_present +

## infant\_mortality$Marital\_status, family = binomial, data = infant\_mortality)

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 3.871 1.010 3.831 0.000127 \*\*\*

## infant\_mortality$Smoke\_presentyes -23.746 1926.639 -0.012 0.990166

## infant\_mortality$Marital\_status 17.821 1926.639 0.009 0.992620

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 266.40 on 317 degrees of freedom

## Residual deviance: 40.92 on 315 degrees of freedom

## AIC: 46.92

##

## Number of Fisher Scoring iterations: 20

model3 <- glm(infant\_mortality$Died ~ infant\_mortality$Tobacco\_use + infant\_mortality$Smoke\_present, data = infant\_mortality, family = binomial)

summary(model3)

##

## Call:

## glm(formula = infant\_mortality$Died ~ infant\_mortality$Tobacco\_use +

## infant\_mortality$Smoke\_present, family = binomial, data = infant\_mortality)

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 22.51 2873.57 0.008 0.994

## infant\_mortality$Tobacco\_use -22.51 2873.57 -0.008 0.994

## infant\_mortality$Smoke\_presentyes -24.40 2873.57 -0.008 0.993

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 266.402 on 317 degrees of freedom

## Residual deviance: 32.365 on 315 degrees of freedom

## AIC: 38.365

##

## Number of Fisher Scoring iterations: 21