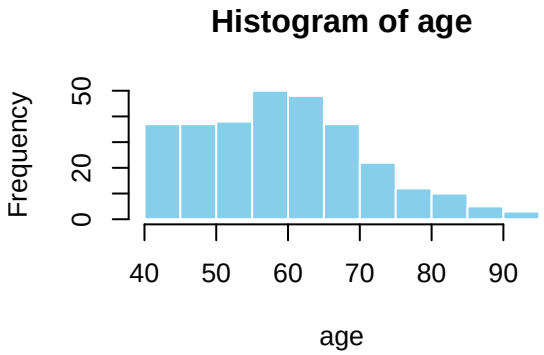
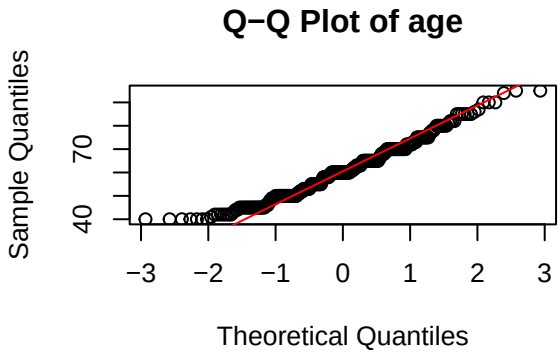
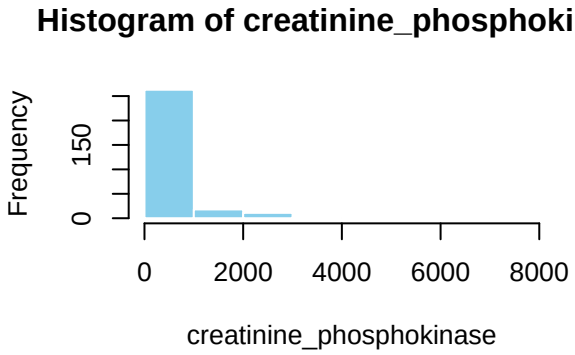
Linearity_Normality

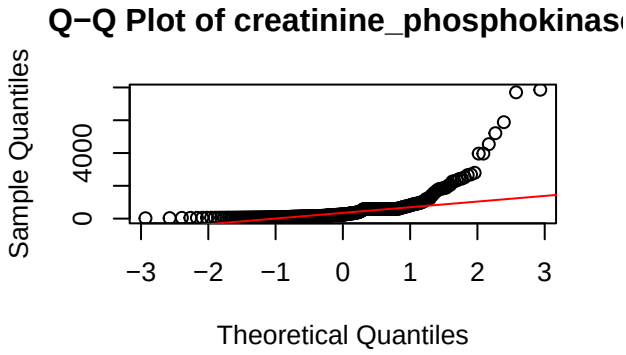
2024-11-20

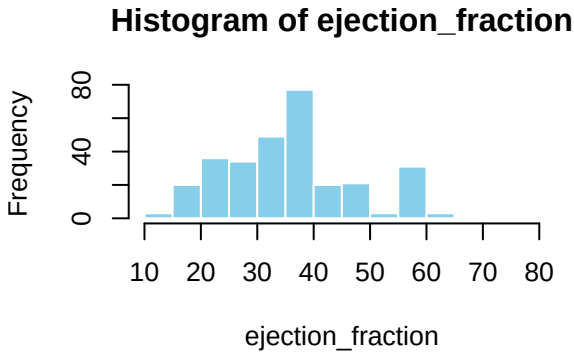
```
data <- read.csv("heart_failure_clinical_records_dataset.csv")
continuous_vars <- c("age", "creatinine_phosphokinase", "ejection_fraction",
                     "platelets", "serum_creatinine", "serum_sodium")

par(mfrow = c(2, 2))

for (i in 1:4) {
  var <- continuous_vars[i]

  hist(data[[var]], main = paste("Histogram of", var), xlab = var, col = "skyblue", border = "white")

  qqnorm(data[[var]], main = paste("Q-Q Plot of", var))
  qqline(data[[var]], col = "red")
}
```

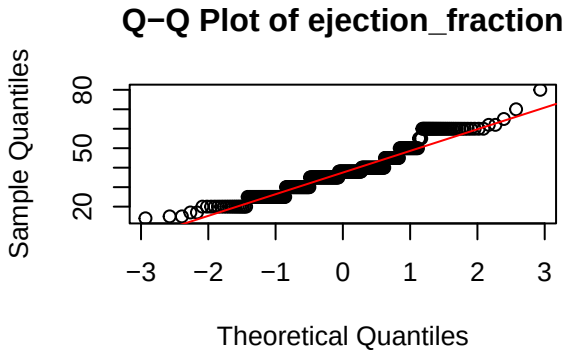


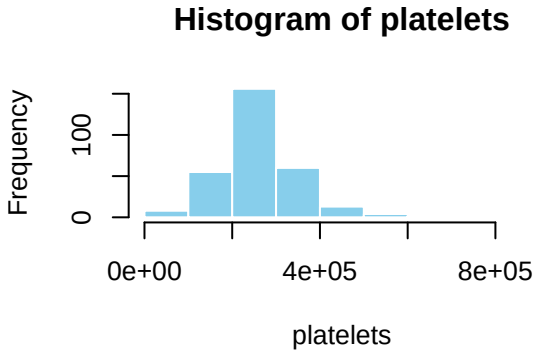


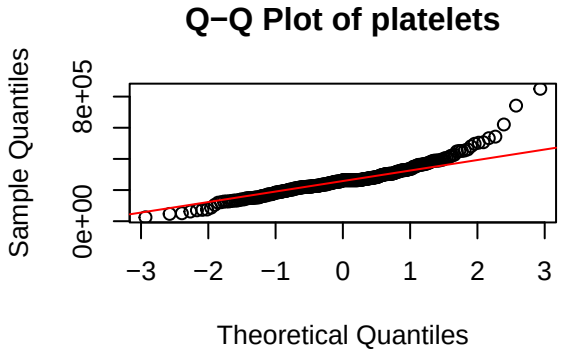












For all variables tested, the Shapiro-Wilk test produced p-values below 0.05, meaning none of these variables follow a normal distribution

almost all of the histograms were either skewed right or left which indicates a low level of normality. the platlets and serum sodium were the ones the were the most normally dsitributed.

using the qq plots i can see that age, ejection, serum sodium, and platlets seem to be the most normal. the other variables were not as normal.

The logistic regression results show that the predictor variables (like age or serum creatinine) are significantly linked to the likelihood of a DEATH_EVENT. The model seems to fit well, as shown by the lower residual deviance.

adding the squared term didn't help the model much. This means the relationship between the variables (like age or creatinine levels) and the risk of death looks pretty straight-line, not curvy. So, keeping it simple with a linear model makes the most sense here.