exploratory

Jasmine Lu

2024-11-20

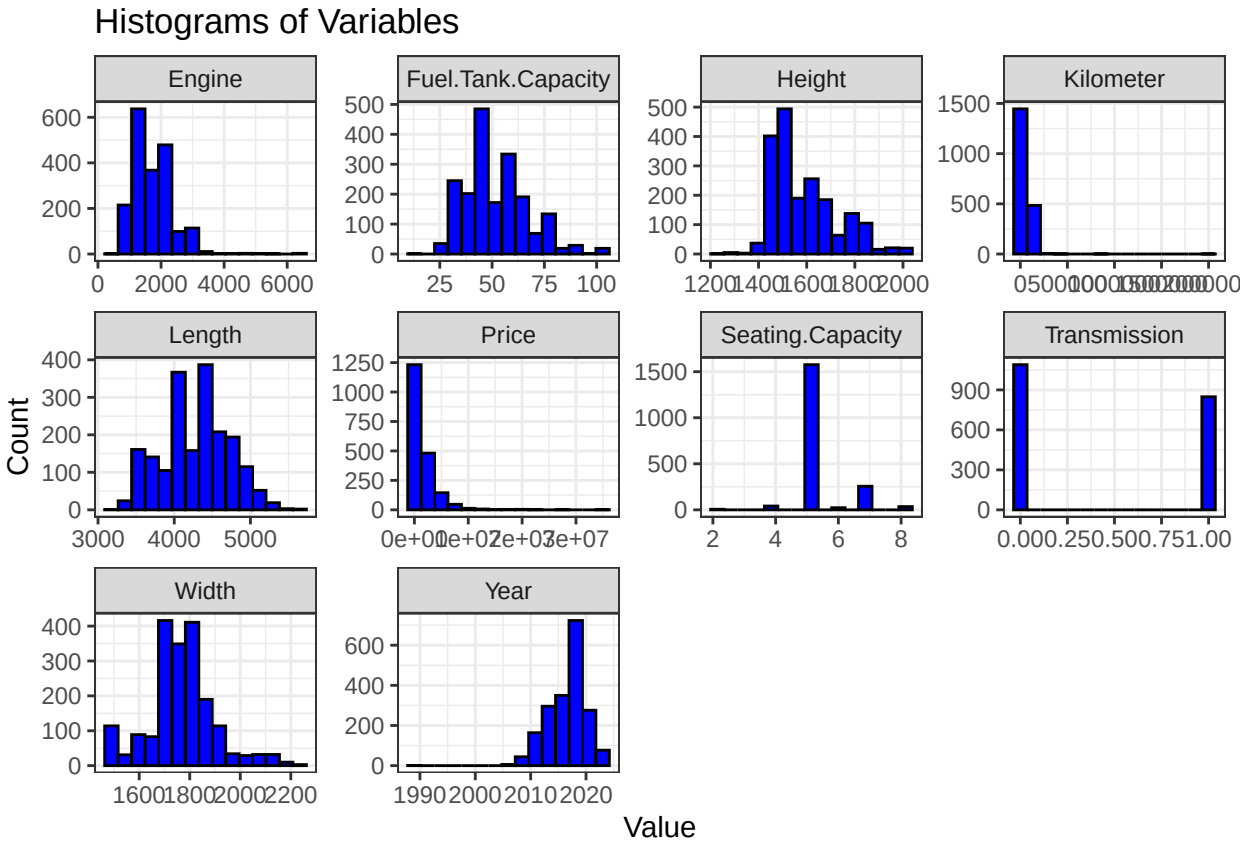
Libraries

Clean data

```
source("preprocess.R")
drop_cols <- c(
  "Model", "Location", "Color", "Owner", "Seller.Type", "Max.Power", "Max.Torque", "Drivetrain", "Make"
)
data <- read.csv("car details v4.csv")
df <- remove_cols(data, drop_cols)
df <- remove_null(df)
df$Transmission <- ifelse(df$Transmission == "Manual", 0, 1) #Manual = 0, Automatic = 1
df$Engine <- as.numeric(gsub("[^0-9]", "", df$Engine)) #Convert to numbers, ex. "123 cc" -> 123
```

Histogram

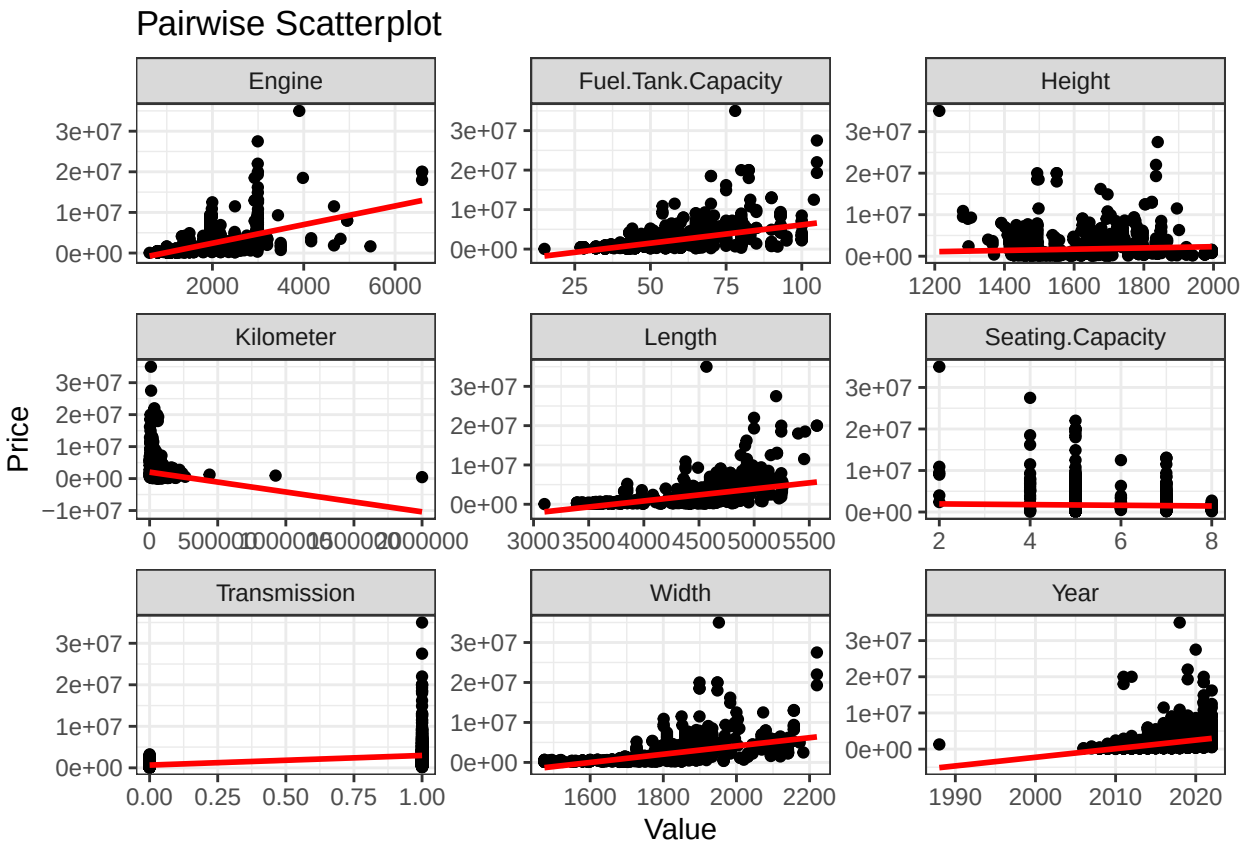
```
df %>%
  gather(key = "var", value = "value") %>%
  ggplot(aes(x = value)) +
    geom_histogram(bins = 15, fill = "blue", color = "black", stat="bin") +
    facet_wrap(~ var, scales = "free") +
    theme_bw() +
    labs(title = "Histograms of Variables", x = "Value", y = "Count")
```



Linearity

```
df %>%
  gather(-Price, key = "var", value = "value") %>%
  ggplot(aes(x = value, y = Price)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    facet_wrap(~ var, scales = "free") +
    theme_bw() +
  labs(title = "Pairwise Scatterplot", x = "Value", y = "Price")
```

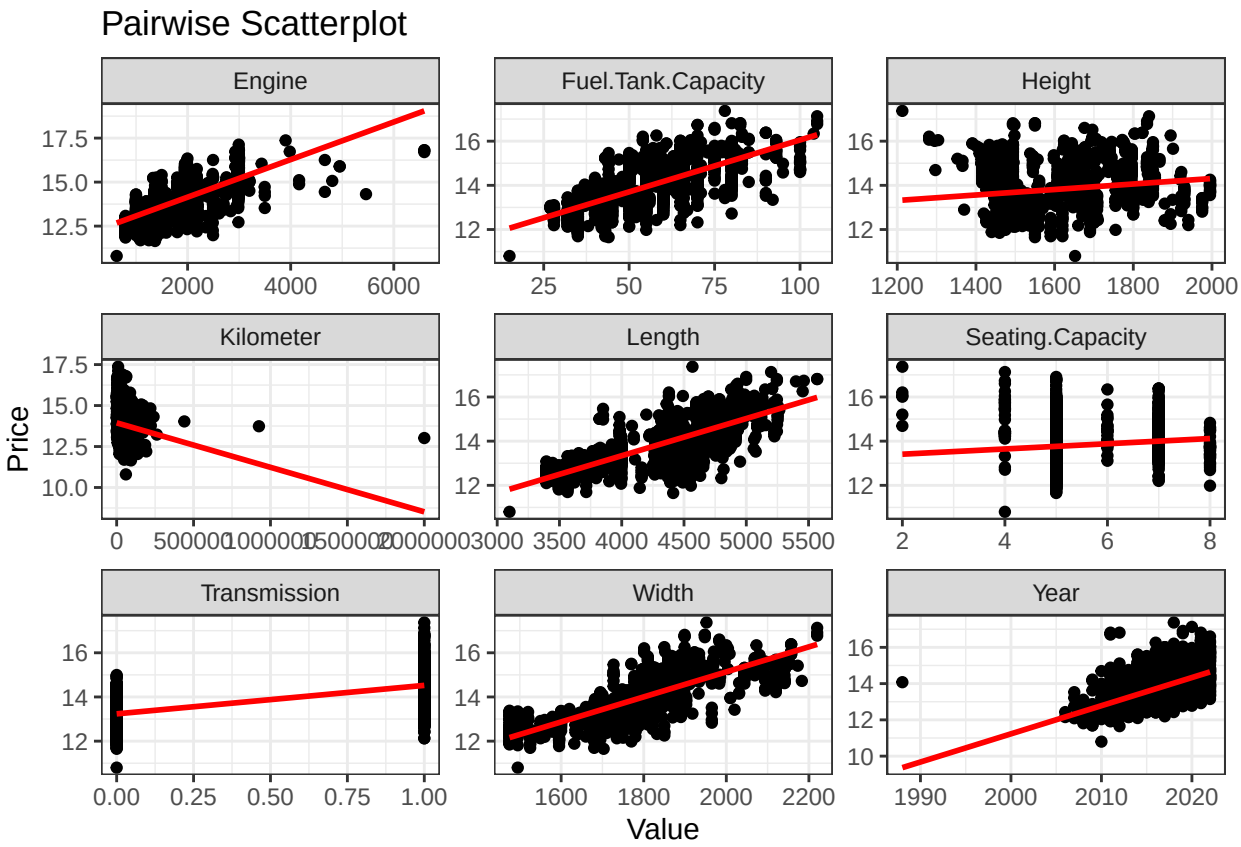
```
## `geom_smooth()` using formula = 'y ~ x'
```



Log Transformation

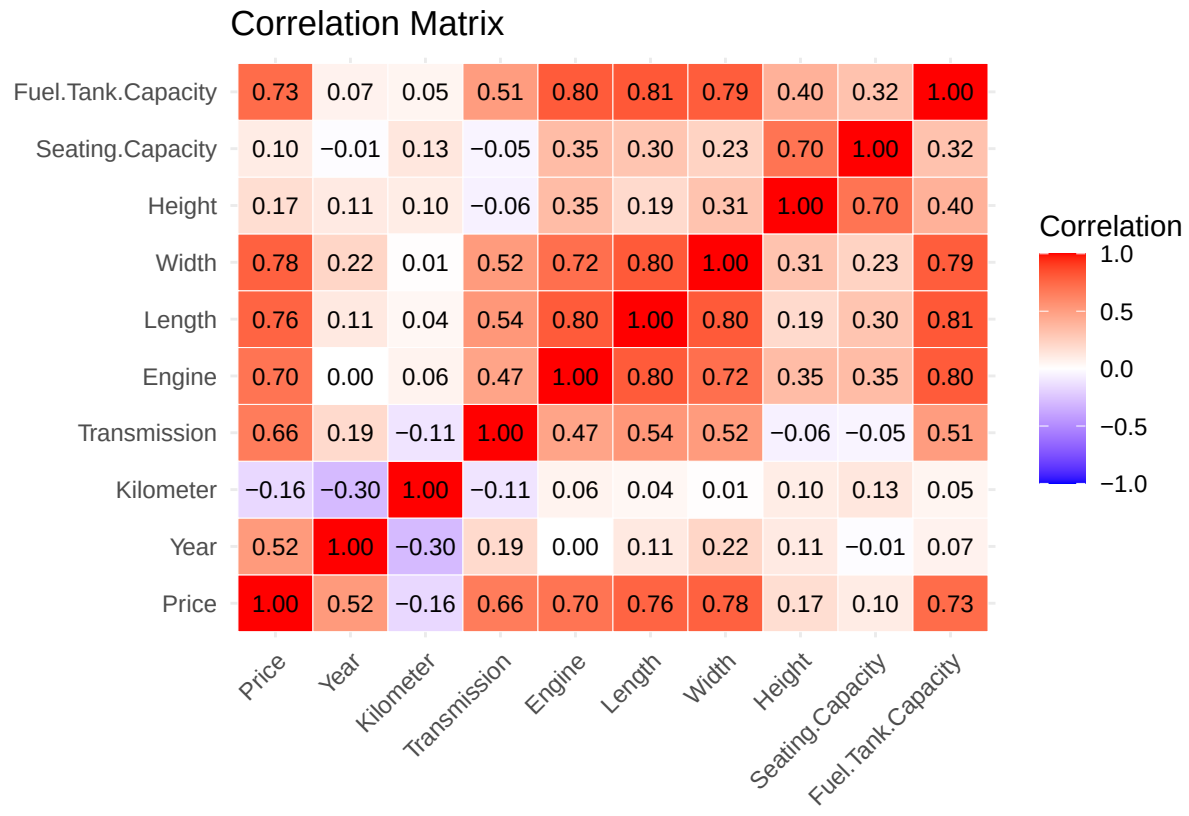
```
df_log <- df
df_log$Price <- log(df_log$Price)
df_log %>%
  gather(-Price, key = "var", value = "value") %>%
  ggplot(aes(x = value, y = Price)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    facet_wrap(~ var, scales = "free") +
    theme_bw() +
  labs(title = "Pairwise Scatterplot", x = "Value", y = "Price")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Correlation

```
cor_matrix <- cor(df_log[, sapply(df_log, is.numeric)])
cor_melted <- melt(cor_matrix)
ggplot(data = cor_melted, aes(x = Var1, y = Var2, fill = value)) +
    geom_tile(color = "white") +
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1, 1), space = "Lab",
                         name="Correlation") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
    labs(x = "", y = "", title = "Correlation Matrix") +
    geom_text(aes(label = sprintf("%.2f", value)), size = 3)
```



MLR Model

```
model <- lm(Price ~ ., data = df_log)
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ ., data = df_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93892 -0.19195 -0.00833  0.17540  2.66073
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.377e+02  5.026e+00 -47.292  < 2e-16 ***
## Year                 1.228e-01  2.522e-03  48.672  < 2e-16 ***
## Kilometer           -6.850e-07  1.306e-07  -5.244 1.75e-07 ***
## Transmission         3.419e-01  1.890e-02  18.086  < 2e-16 ***
## Engine               3.988e-04  2.188e-05  18.222  < 2e-16 ***
## Length               3.681e-04  3.869e-05   9.514  < 2e-16 ***
## Width                1.151e-03  1.057e-04  10.885  < 2e-16 ***
## Height              -3.217e-04  9.071e-05  -3.546    4e-04 ***
## Seating.Capacity    -1.098e-01  1.360e-02  -8.072 1.21e-15 ***
## Fuel.Tank.Capacity   1.258e-02  1.050e-03  11.983  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3162 on 1927 degrees of freedom
## Multiple R-squared:  0.8943, Adjusted R-squared:  0.8938
## F-statistic:  1811 on 9 and 1927 DF,  p-value: < 2.2e-16
```

```
vif(model) #Looks like Length have a high Variance Inflation Factor so we can try removing it
```

```
##               Year          Kilometer       Transmission             Engine
##           1.299549           1.131932           1.703986           3.714295
##             Length              Width             Height   Seating.Capacity
##           5.622594           3.840612           2.906487           2.374768
## Fuel.Tank.Capacity
##           4.872971
```

```
model_without_length <- lm(Price ~ . - Length, data = df_log)
summary(model_without_length)
```

```
##
## Call:
## lm(formula = Price ~ . - Length, data = df_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18670 -0.19921 -0.00393  0.18232  2.83931
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.426e+02  5.114e+00 -47.451  < 2e-16 ***
## Year                1.257e-01  2.561e-03  49.065  < 2e-16 ***
## Kilometer          -6.244e-07  1.335e-07  -4.678  3.1e-06 ***
## Transmission        3.518e-01  1.931e-02  18.224  < 2e-16 ***
## Engine              4.763e-04  2.077e-05  22.929  < 2e-16 ***
## Width               1.563e-03  9.863e-05  15.847  < 2e-16 ***
## Height             -7.167e-04  8.250e-05  -8.687  < 2e-16 ***
## Seating.Capacity   -6.175e-02  1.292e-02  -4.779  1.9e-06 ***
## Fuel.Tank.Capacity  1.619e-02  1.001e-03  16.168  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3234 on 1928 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8888
## F-statistic:  1936 on 8 and 1928 DF,  p-value: < 2.2e-16
```

```
vif(model_without_length)
```

```
##               Year          Kilometer       Transmission             Engine
##           1.280539           1.129241           1.698742           3.198798
##              Width             Height   Seating.Capacity Fuel.Tank.Capacity
##           3.195116           2.297494           2.047352           4.236487
```

```r
cat("\nInterpretation of Model Parameter Estimates:\n")
```

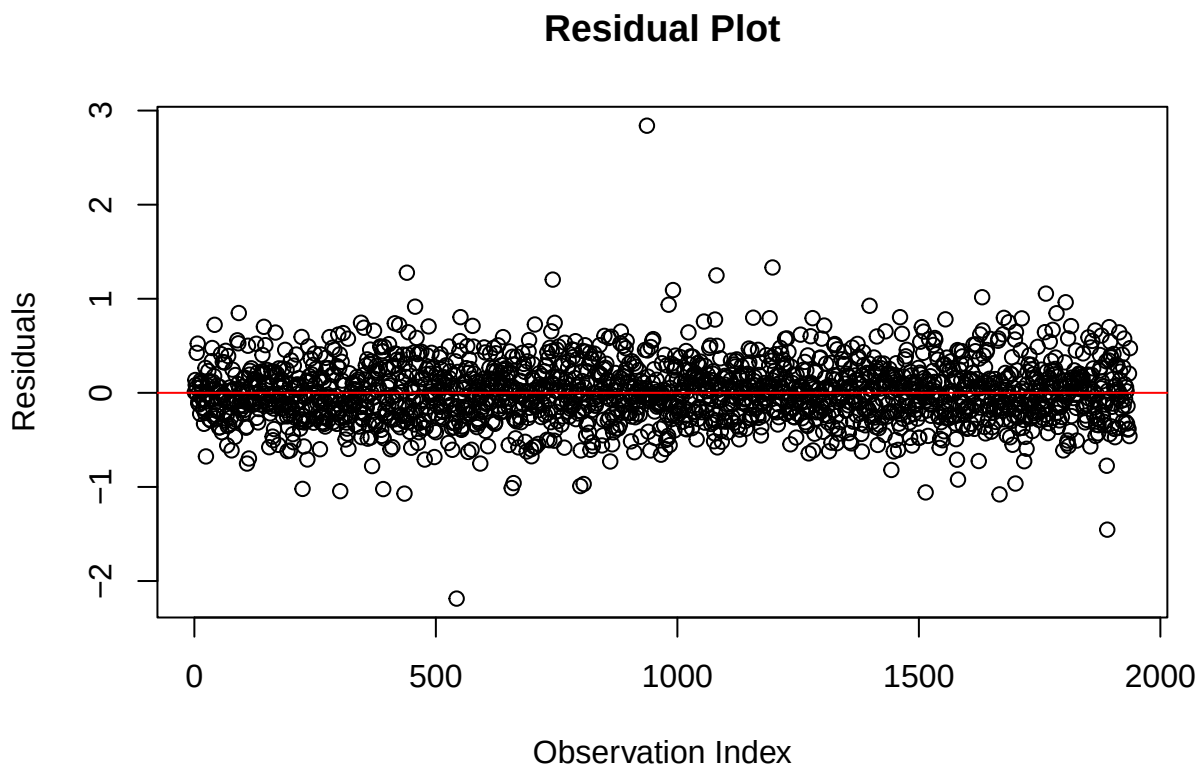
```
##
## Interpretation of Model Parameter Estimates:
```

```r
print(coef(summary(model_without_length)))
```

```
##                         Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)        -2.426416e+02 5.113501e+00 -47.451160  0.000000e+00
## Year                1.256547e-01 2.560979e-03  49.065086  0.000000e+00
## Kilometer          -6.243527e-07 1.334678e-07  -4.677928  3.099244e-06
## Transmission        3.518344e-01 1.930564e-02  18.224434  1.358330e-68
## Engine              4.763195e-04 2.077377e-05  22.928889 4.390701e-103
## Width               1.563021e-03 9.863386e-05  15.846701  2.933386e-53
## Height             -7.167373e-04 8.250395e-05  -8.687309  7.769957e-18
## Seating.Capacity   -6.174600e-02 1.292100e-02  -4.778732  1.897412e-06
## Fuel.Tank.Capacity  1.618922e-02 1.001303e-03  16.168141  3.070965e-55
```

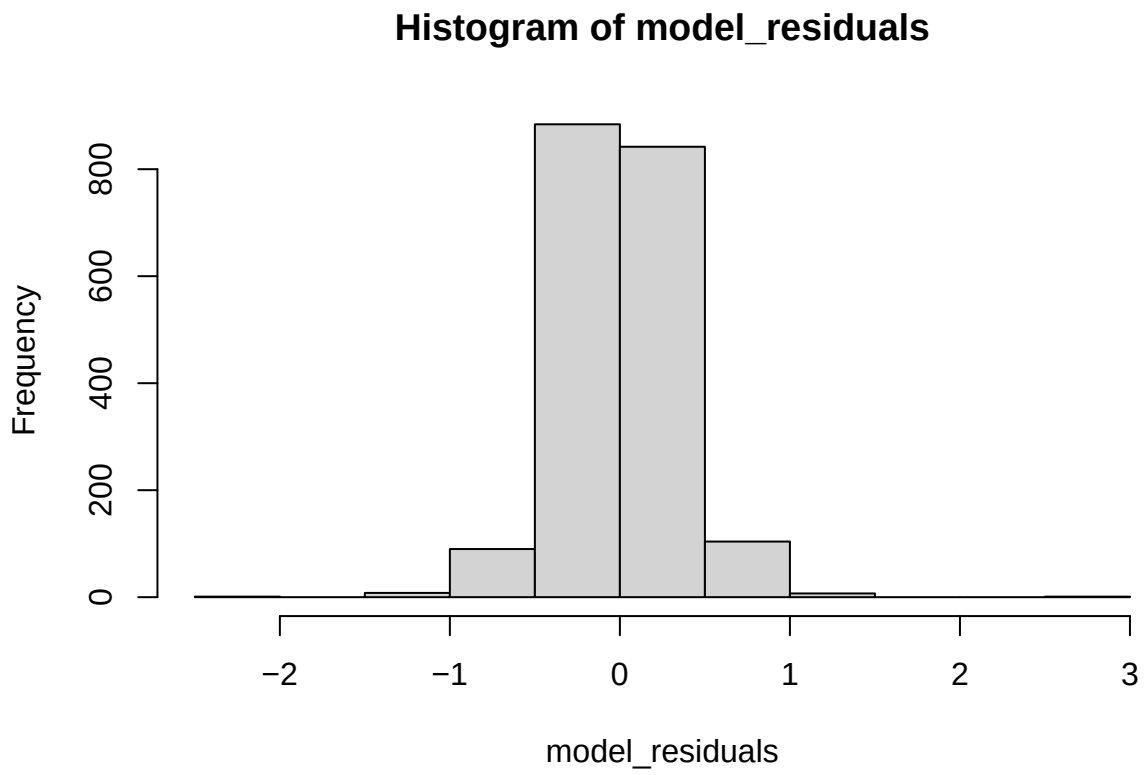
Residual plot

```r
residuals_values <- residuals(model_without_length)
plot(residuals_values, type = 'p', main = "Residual Plot", xlab = "Observation Index", ylab = "Residuals
abline(h = 0, col = "red")
```

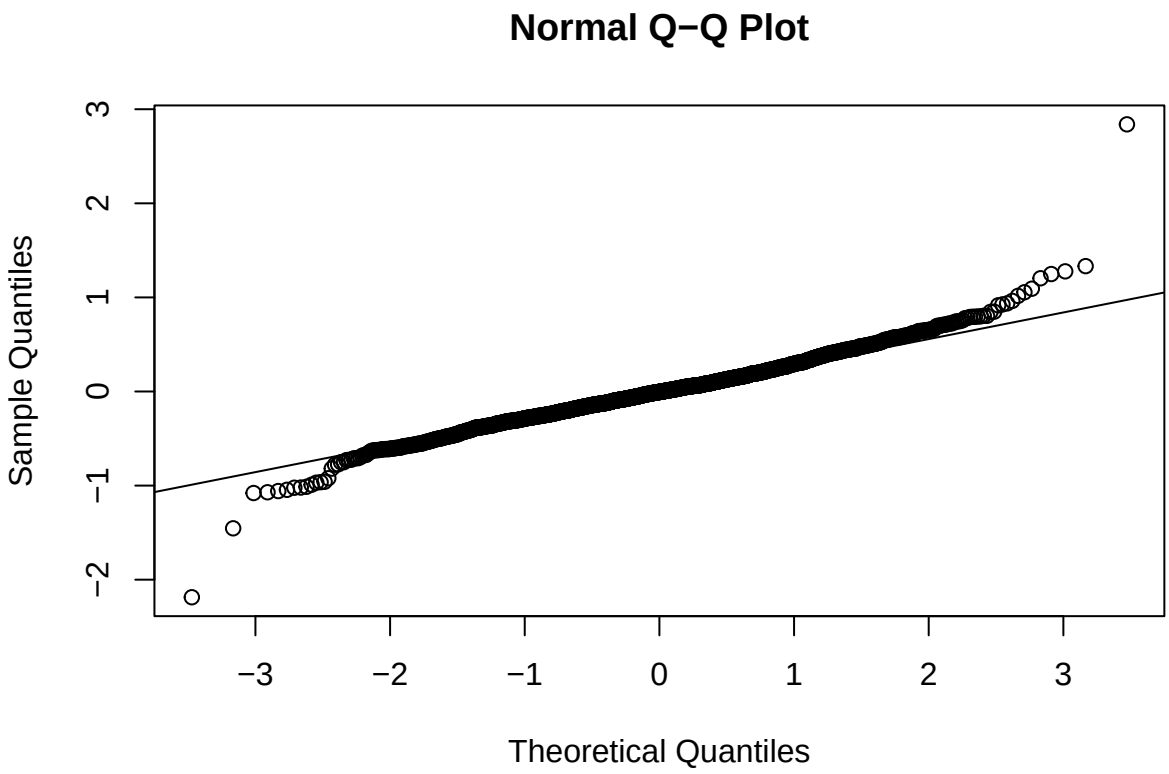


Histogram of Residuals

```r
model_residuals = model_without_length$residuals
hist(model_residuals)
```




```r
qqnorm(model_residuals)
qqline(model_residuals)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3