SHC 798 Assignment 1, 2025

Richard Lubega

2025-07-14

SHC 798 Assignment 1, 2025

Part 1: Data Analysis with R

```
# Getting Started with the Dataset:
cat("\n=== Getting Started with the Dataset ===\n")
```

```
##
## === Getting Started with the Dataset ===
```

```
pacman::p_load(ggplot2) # checks if ggplot2 is installed;if it's not installed, it automatically instal
pacman::p_load(tidymodels) # tidymodels some useful packages and functionalities

cat("\n=== View first few rows of the dataset ===\n")
```

```
##
## === View first few rows of the dataset ===
```

```
head(mpg) # View first few rows of the dataset
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

```
cat("\n=== Get an overview of the dataset ===\n")
```

```
##
## === Get an overview of the dataset ===
```

```r
summary(mpg) # Get an overview of the dataset
```

```
##  manufacturer          model               displ            year
##  Length:234         Length:234         Min.   :1.600   Min.   :1999
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999
##  Mode  :character   Mode  :character   Median :3.300   Median :2004
##                                        Mean   :3.472   Mean   :2004
##                                        3rd Qu.:4.600   3rd Qu.:2008
##                                        Max.   :7.000   Max.   :2008
##       cyl           trans               drv                 cty
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00
##  Mean   :5.889                                         Mean   :16.86
##  3rd Qu.:8.000                                         3rd Qu.:19.00
##  Max.   :8.000                                         Max.   :35.00
##       hwy             fl                class
##  Min.   :12.00   Length:234         Length:234
##  1st Qu.:18.00   Class :character   Class :character
##  Median :24.00   Mode  :character   Mode  :character
##  Mean   :23.44
##  3rd Qu.:27.00
##  Max.   :44.00
```

```r
# --- Add ----
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

```r
# view(mpg)
# -------------

#Analyse the mpg dataset using descriptive methods

#(a)
# average city and highway fuel economy across all vehicle classes
cat("\n=== Average city and highway fuel economy, afe, across all vehicle classes ===\n")
```

```
##
## === Average city and highway fuel economy, afe, across all vehicle classes ===
```

```r
afe <- aggregate(cbind(cty, hwy) ~ class, data = mpg, FUN = mean)
afe
```

```
##        class      cty      hwy
## 1    2seater 15.40000 24.80000
## 2    compact 20.12766 28.29787
## 3    midsize 18.75610 27.29268
## 4    minivan 15.81818 22.36364
## 5     pickup 13.00000 16.87879
## 6 subcompact 20.37143 28.14286
## 7        suv 13.50000 18.12903
```

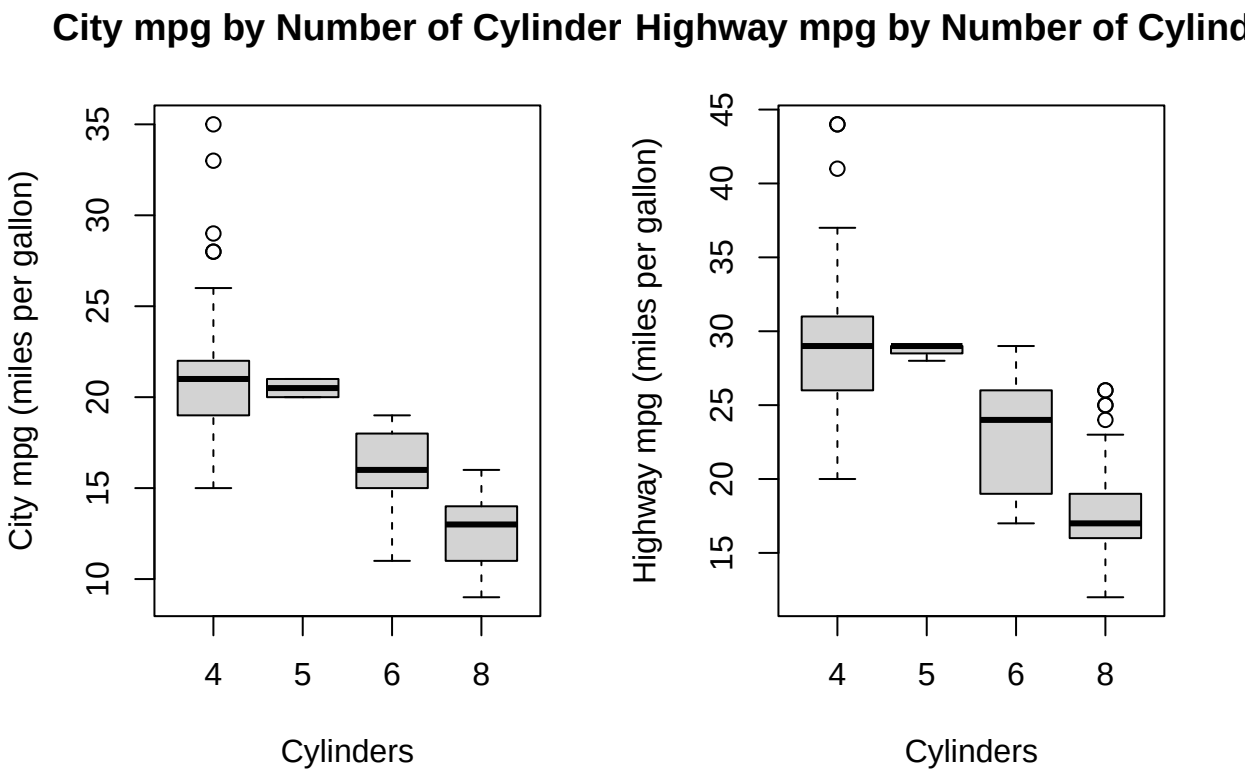
```r
#(b)
# Compare the fuel efficiency (cty and hwy)
cat("\n=== Comparing fuel efficiency for cty and hwy economies ===\n")
```

```
##
## === Comparing fuel efficiency for cty and hwy economies ===
```

```r
par(mfrow = c(1, 2)) # Set up a 1x2 plot layout for side-by-side boxplots

# Boxplot for city mpg by cylinders
boxplot(cty ~ cyl, data = mpg,
        main = "City mpg by Number of Cylinders",
        xlab = "Cylinders",
        ylab = "City mpg (miles per gallon)")

# Boxplot for highway mpg by cylinders
boxplot(hwy ~ cyl, data = mpg,
        main = "Highway mpg by Number of Cylinders",
        xlab = "Cylinders",
        ylab = "Highway mpg (miles per gallon)")
```



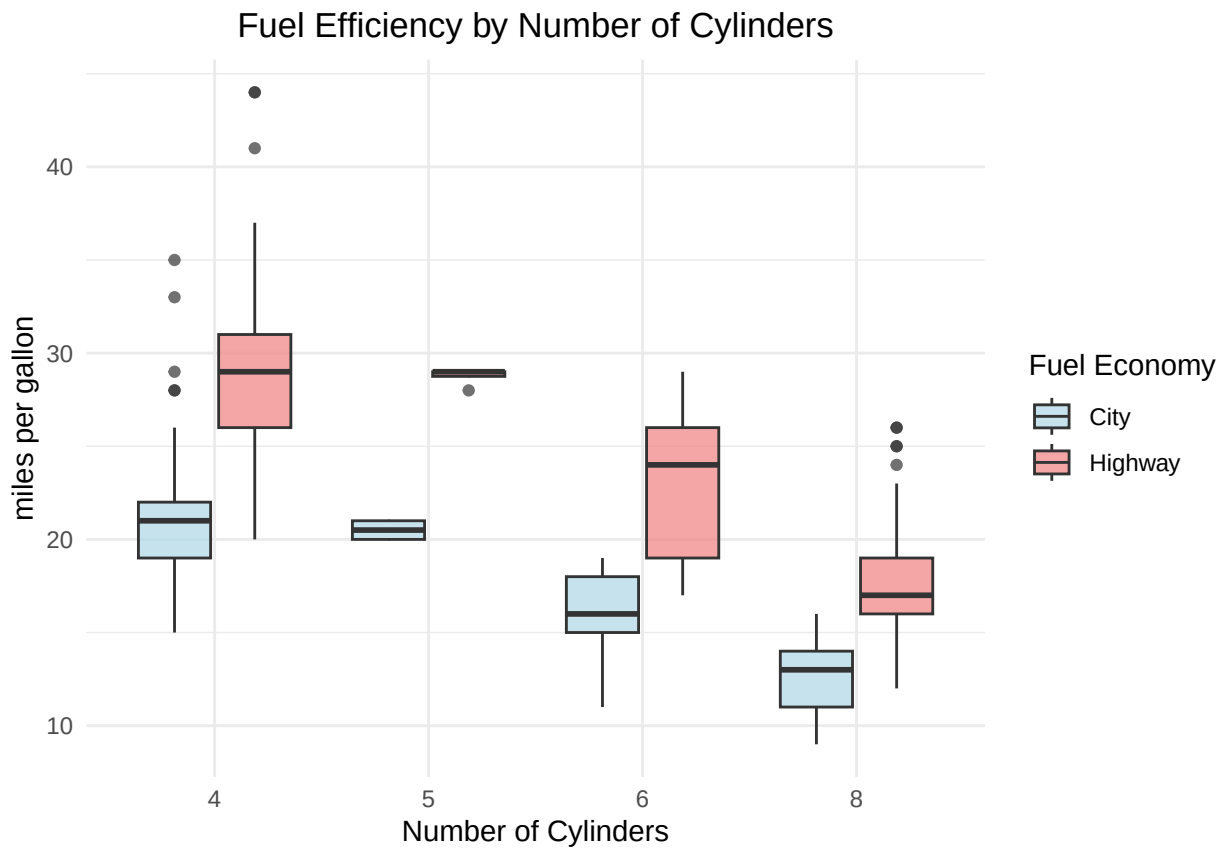

```
par(mfrow = c(1, 1)) # Reset plot layout to default

# Combine plots by faceting
cat("\n=== Combining the box plots for comparison ===\n")
```

```
##
## === Combining the box plots for comparison ===
```

```
mpg_comb <- mpg %>%
  select(cyl, cty, hwy) %>%
  pivot_longer(cols = c(cty, hwy), names_to = "fuel_econ", values_to = "mpg")

ggplot(mpg_comb, aes(x = factor(cyl), y = mpg, fill = fuel_econ)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Fuel Efficiency by Number of Cylinders",
       x = "Number of Cylinders",
       y = "miles per gallon",
       fill = "Fuel Economy") +
  scale_fill_manual(values = c("cty" = "lightblue", "hwy" = "lightcoral"),
                    labels = c("City", "Highway")) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
#---------  Median Values by cylinder count -------------
cat("\n=== Median Values by cylinder count ===\n")
```

```
##
## === Median Values by cylinder count ===
```

```
mpg %>%
  group_by(cyl) %>%
  summarise(
    median_cty = median(cty),
    median_hwy = median(hwy),
    .groups = 'drop'
  )
```

```
## # A tibble: 4 x 3
##     cyl median_cty median_hwy
##   <int>      <dbl>      <dbl>
## 1     4       21           29
## 2     5       20.5         29
## 3     6       16           24
## 4     8       13           17
```

```
cat("\n=== Trend Analysis ===\n")
```

```
##
## === Trend Analysis ===
```

Commenting on the Trend This analysis clearly demonstrates that engine size (cylinder count) is a major predictor of fuel efficiency, with smaller engines being visibly more fuel-efficient than larger ones. Some *outliers* exist (may be due to high-efficiency hybrids or low-efficiency compact cars).

- **Inverse relationship:** Based on the boxplots (where more cylinders = lower mpg), there's a clear *negative* correlation between the number of cylinders and fuel efficiency (mpg). As cylinder count increases, both city and highway mpg decrease.
- **Highway vs City efficiency:** Highway mpg is consistently higher than city mpg across all cylinder counts (as seen from the combined plot), which may be explained by the more efficient cruising speeds on highways. Generally, the **fuel efficiency difference** between city and highway driving becomes more pronounced in vehicles with fewer cylinders.
- **4-cylinder cars** are the most fuel-efficient, with median values of 21 mpg (for city) and 29 mpg (for highway). The rest in each category have lower values. **8-cylinder cars** are the least fuel-efficient, with median values of 13 mpg (for city) and 17 mpg (for highway).
- **5-cylinder cars** are the least common (narrower range) in both categories. This may be due to fewer models of these cars. **6-cylinder cars** have the most broad range compared to the others
- There is also **variability within cylinder** groups, and is most pronounced in **6-cylinder cars**, whch suggests that factors beyond cylinder count (including vehicle weight, engine technology, etc.) also influence fuel efficiency.

```
# (c)
# Correlation: Engine Displacement vs Highway Fuel Economy
cat("\nCorrelation: engine displacement (displ) and highway fuel economy (hwy) \n")
```

```
##
## Correlation: engine displacement (displ) and highway fuel economy (hwy)
```

```
# Calculate correlation coefficient
correlation_pearson <- cor(mpg$displ, mpg$hwy)
correlation_spearman <- cor(mpg$displ, mpg$hwy, method = "spearman")

cat("Pearson correlation coefficient:", round(correlation_pearson, 4), "\n")
```

```
## Pearson correlation coefficient: -0.766
```

```
cat("Spearman correlation coefficient:", round(correlation_spearman, 4), "\n")
```

```
## Spearman correlation coefficient: -0.8267
```

```
# Interpretation of correlation strength
interpret_correlation <- function(r) {
  abs_r <- abs(r)
  if (abs_r >= 0.7) return("Strong")
  else if (abs_r >= 0.3) return("Moderate")
  else return("Weak")
}

cat("Correlation strength:", interpret_correlation(correlation_pearson), "\n")
```

```
## Correlation strength: Strong
```

```r
cat("Direction:", ifelse(correlation_pearson > 0, "Positive", "Negative"), "\n")
```

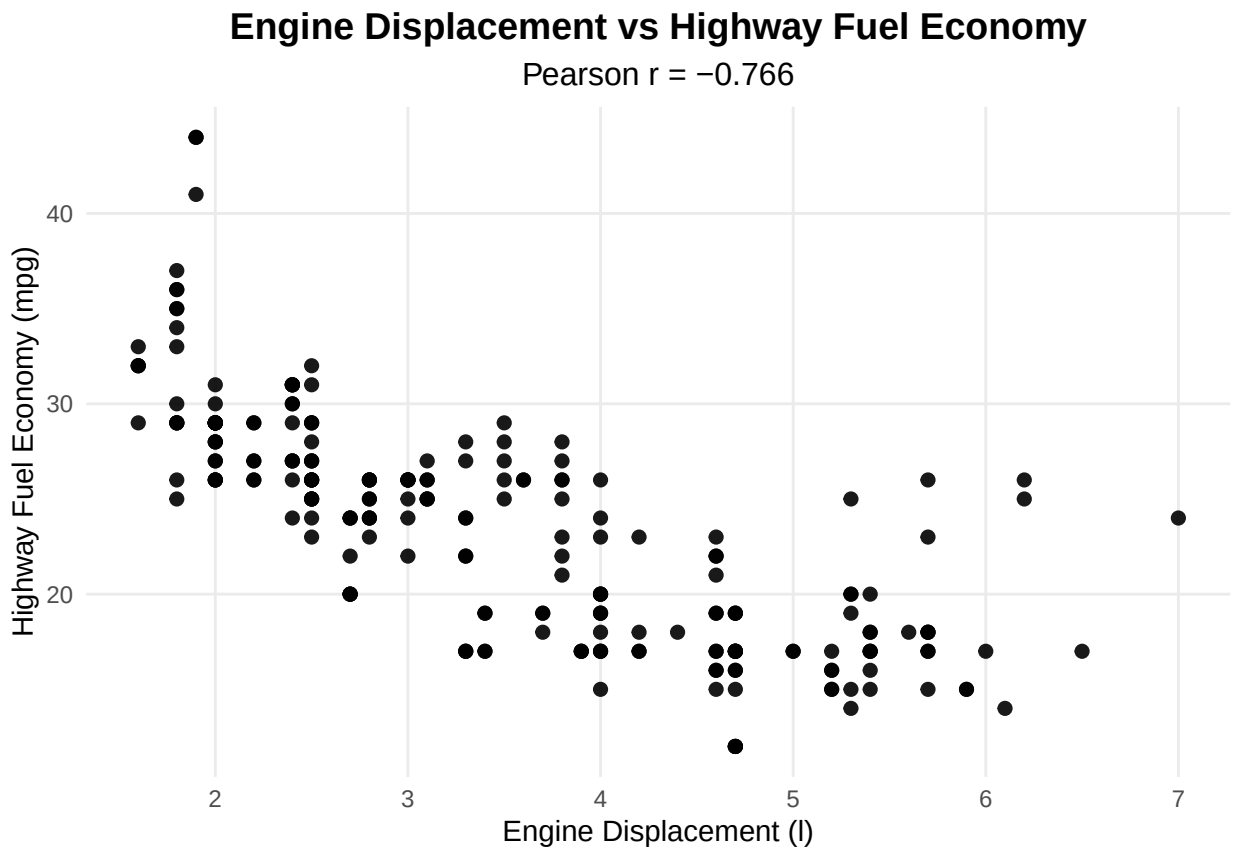
```
## Direction: Negative
```

```r
# Create basic scatter plot
cat("\n=== Creating a Basic Scatter Plot ===\n")
```

```
##
## === Creating a Basic Scatter Plot ===
```

```r
# Basic scatter plot
plot_dh <- ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point(alpha = 0.9, size = 2, color = "black") +
    labs(
    title = "Engine Displacement vs Highway Fuel Economy",
    subtitle = paste("Pearson r =", round(correlation_pearson, 3)),
    x = "Engine Displacement (l)",
    y = "Highway Fuel Economy (mpg)",
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5),
    axis.title = element_text(size = 11),
    panel.grid.minor = element_blank()
  )

print(plot_dh)
```




```
cat("\nTest the significance of the correlation \n")
```

```
##
## Test the significance of the correlation
```

```
cor_test <- cor.test(mpg$displ, mpg$hwy, method = "pearson")
cat("Pearson correlation test:\n")
```

```
## Pearson correlation test:
```

```
print(cor_test)
```

```
##
##  Pearson's product-moment correlation
##
## data:  mpg$displ and mpg$hwy
## t = -18.151, df = 232, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8142727 -0.7072539
## sample estimates:
##       cor
## -0.76602
```

```
cat("\nSignificance level: ", ifelse(cor_test$p.value < 0.001, "p < 0.001 (highly significant)",
                                     ifelse(cor_test$p.value < 0.01, "p < 0.01 (significant)",
                                            ifelse(cor_test$p.value < 0.05, "p < 0.05 (significant)", "no
```

```
##
## Significance level:  p < 0.001 (highly significant)
```

Therefore, based on the analysis, a **strong negative**, **highly** statistically **significant** correlation exists between engine displacement (displ) and highway fuel economy (hwy).

The scatter plot reinforces this because as displacement increases, highway mpg decreases.

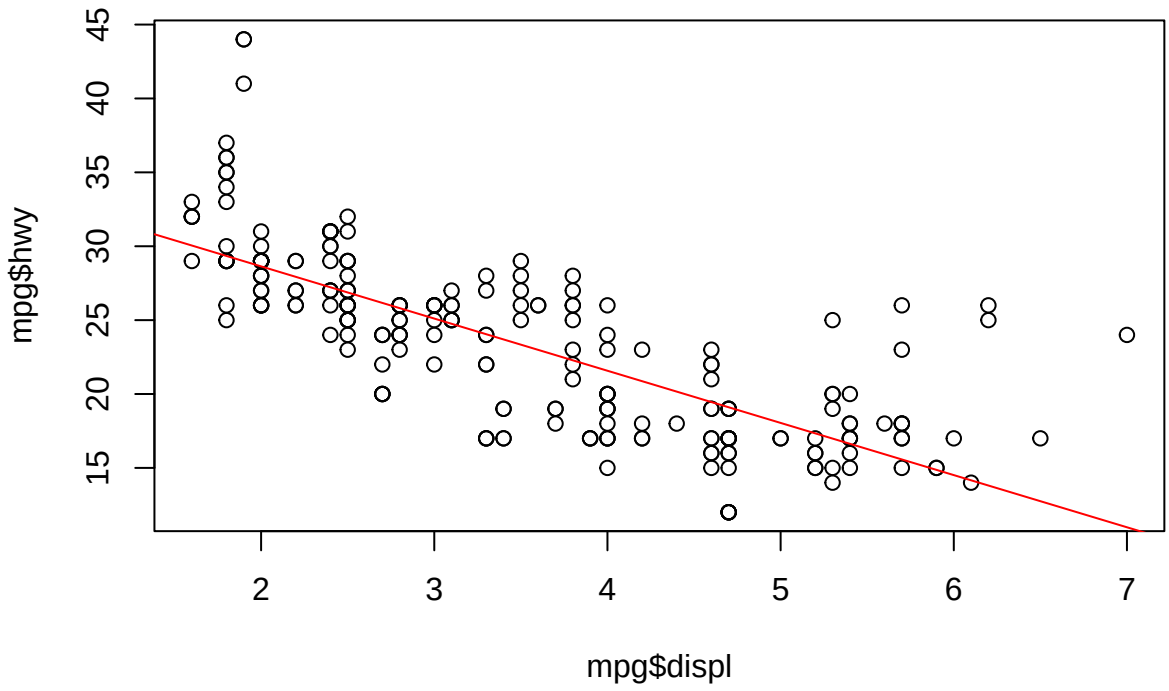
```
# Linear regression
cat("\n Linear Regression Model \n")
```

```
##
##  Linear Regression Model
```

```
lm_model <- lm(hwy ~ displ, data = mpg)
summary(lm_model)
```

```
##
## Call:
## lm(formula = hwy ~ displ, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1039 -2.1646 -0.2242  2.0589 15.0105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.6977     0.7204   49.55   <2e-16 ***
## displ        -3.5306     0.1945  -18.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.836 on 232 degrees of freedom
## Multiple R-squared:  0.5868, Adjusted R-squared:  0.585
## F-statistic: 329.5 on 1 and 232 DF,  p-value: < 2.2e-16
```

```
plot(mpg$displ, mpg$hwy)+
  abline(lm_model, col = "red")
```

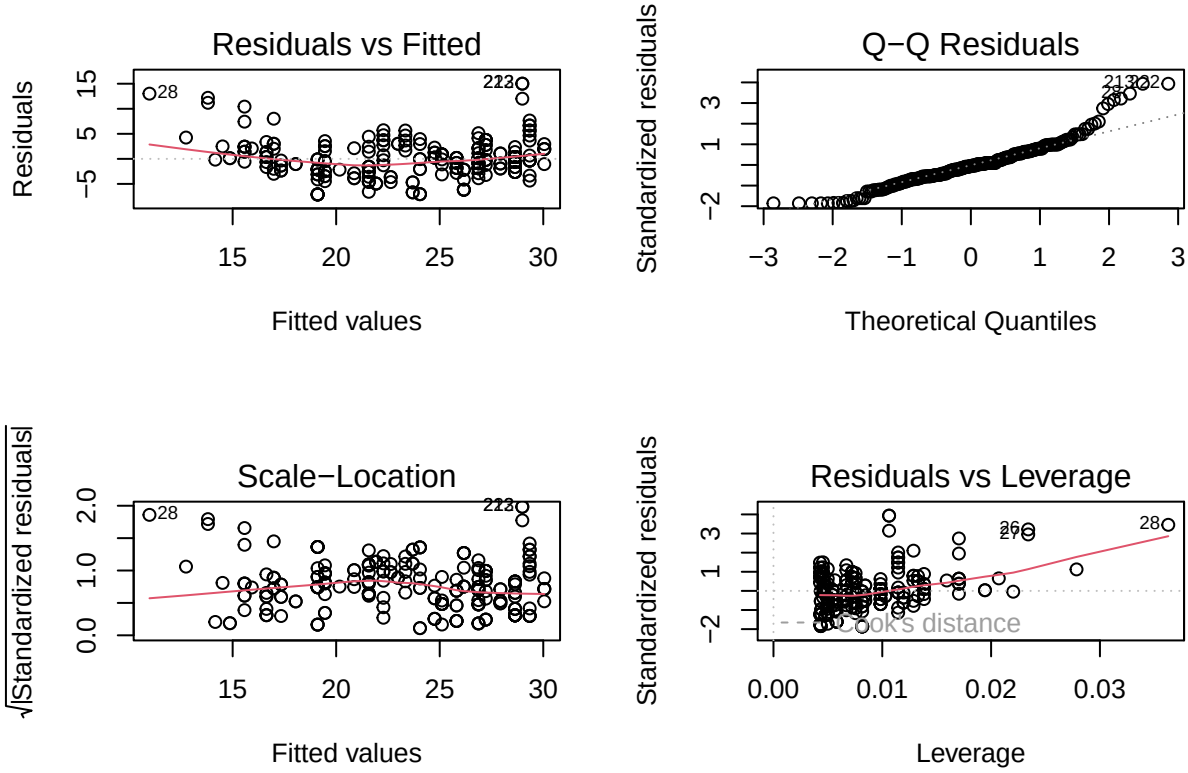



```
## integer(0)
```

```
cat("\n Model Diagnostics \n")
```

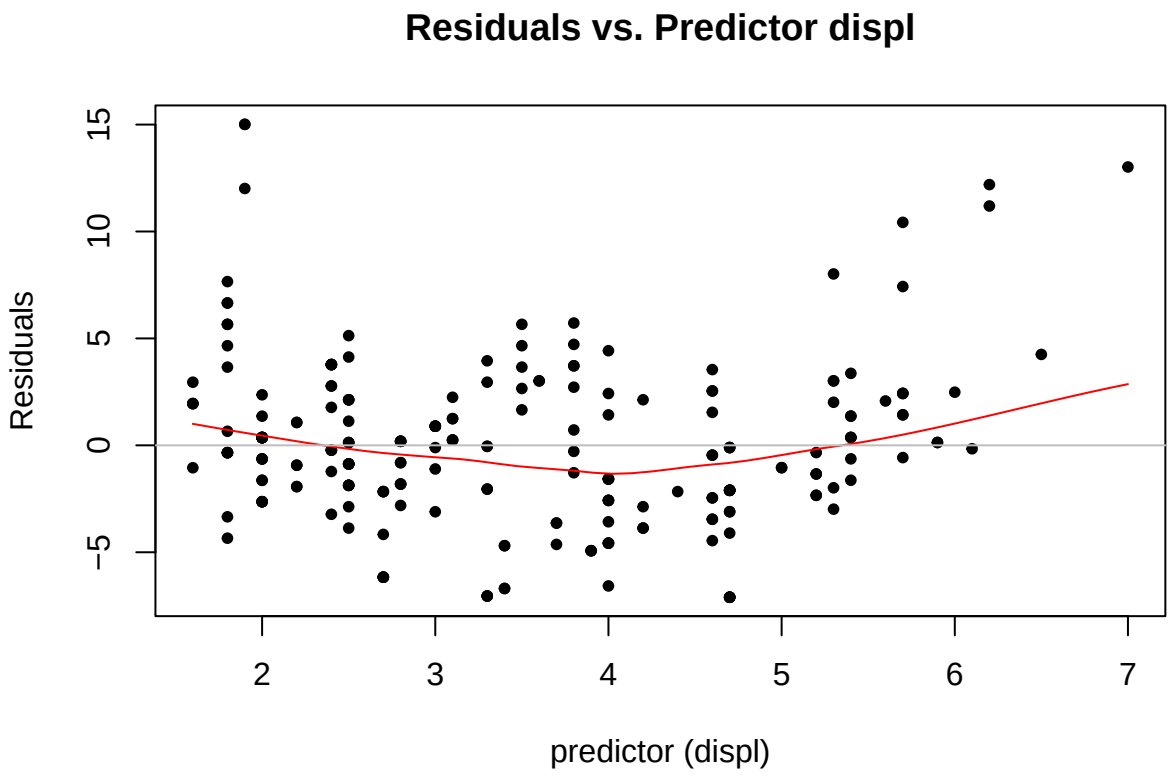
```
##
##  Model Diagnostics
```

```
# Diagnostics plots
par(mfrow = c(2,2))
plot(lm_model)
```



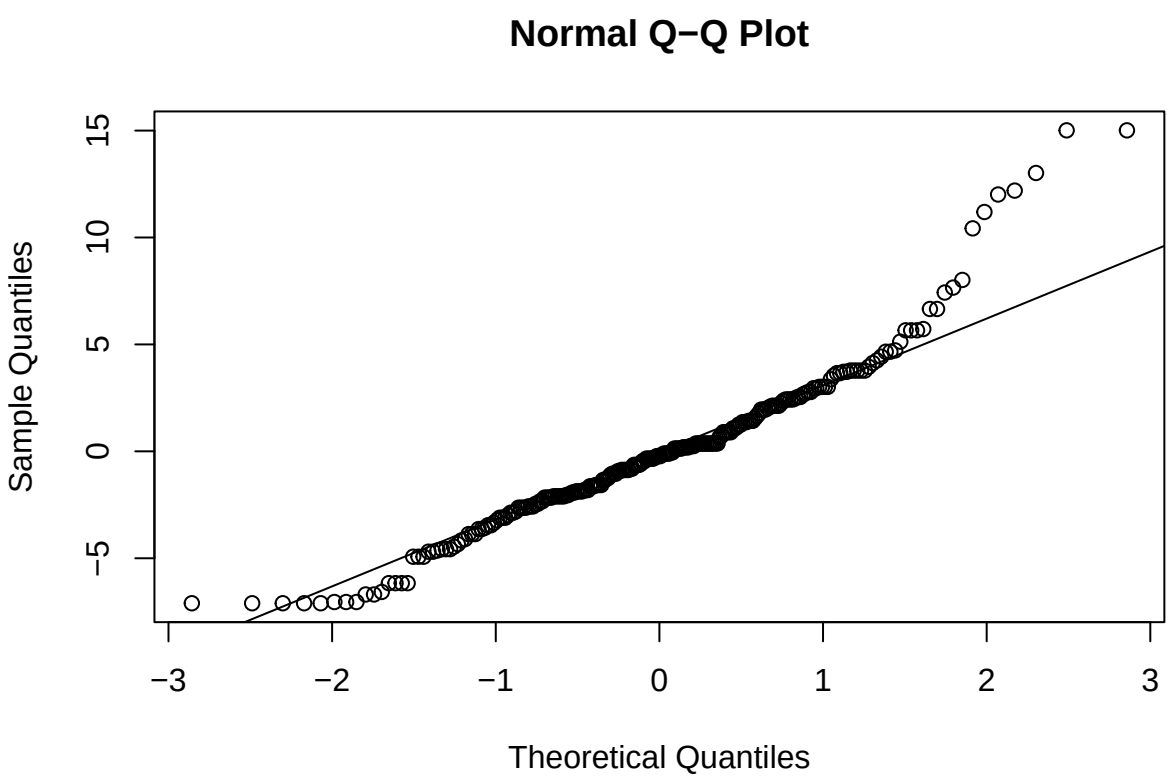

```r
par(mfrow = c(1,1))
# Tukey-Anscombe Plot
# plot(lm_model$fitted.values, lm_model$residuals, xlab="Fitted", ylab="Residuals", pch=20) +
#   title("Residuals vs. Fitted Values") +
#   lines(loess.smooth(lm_model$fitted.values, lm_model$residuals),col="red") +
#   abline(h=0, col="grey")

# Residuals vs. Predictor Plot
plot(mpg$displ, lm_model$residuals, xlab="predictor (displ)", ylab="Residuals", pch=20) +
  title("Residuals vs. Predictor displ") +
  lines(loess.smooth(mpg$displ, lm_model$residuals),col="red") +
  abline(h=0, col="grey")
```




```
## integer(0)
```

```
# Quantile-Quantile Plot
qqnorm(lm_model$residuals) #Quantile-Quantile Plot
qqline(lm_model$residuals) # adds the diagonal line
```



From the **model diagnostics (Tuskey-Anscombe plot)**, the red LOESS line is slightly curved which indicates non-linearity. Nonetheless, the expectation of the residuals can be considered zero. The variance of the errors increases with fitted values and homoskedasticity is violated.

The **Q-Q plot** indicates that the bulk of the residuals (in the central region) are approximately Gaussian distributed. The data exhibits heavy tails (skewness) and has outliers at the extremes. The noticeable presence of extreme positive residuals suggests a right-skewed distribution (departure from normality), the assumption of Gaussian errors is violated by the model.

To improve the model, variable transformation is required (to stabilize the spread and ensure error normality).

Comment on Model Outputs: The regression model, **lm_model** predicts highway fuel economy (hwy, in mpg) as a function of engine displacement (displ, in litres).

- **Regression Coefficients**:
  - **Intercept** (35.6977) implies that when engine displacement is theoretically 0 litres, the predicted highway fuel economy is approximately 35.7 mpg. It's p-value (< 2e-16) is very small and indicates that it is statistically significant.
  - **Slope** (-3.5306): For each 1-litre increase in engine displacement, highway fuel economy decreases by approximately 3.53 mpg, on average. The t-value (-15.07) and p-value (< 2e-16) indicate this coefficient is highly significant, confirming a strong negative relationship.
- **Statistical Significance**:

The p-value for displ is very small (< 2.2e-16), meaning the relationship is statistically significant. Engine size is a strong/meaningful predictor of fuel efficiency.

- **Model Fit**:
  - The **multiple R-squared** (0.5868) and **adjusted R-squared** (0.585) indicate that approximately 58.68% of the variability in highway fuel economy is explained by engine displacement. This suggests a moderately strong negative relationship, but other factors (e.g., vehicle weight, transmission type) may also play a role.

Implication of Model Outputs on the Relationship

- The **negative coefficient** for *displ* (-3.5306) supports the belief that cars with smaller engines have better fuel efficiency. As engine size increases, highway fuel economy decreases significantly, with a 1-liter increase in displacement leading to a 3.53 mpg reduction in fuel efficiency, on average. The highly significant **p-values** for both the displ coefficient and the overall model ($<$ 2e-16) confirm that the negative relationship between engine size and fuel efficiency is *not due to random chance*. This strengthens the conclusion that engine size is a reliable predictor of fuel efficiency.
- The **R-squared value** (0.5868) indicates that engine size alone doesn't explain all the variability in fuel efficiency, so the remaining 41.32% of variability implies other factors (like, vehicle weight, car transmission type, or car drive type) also influence fuel efficiency.