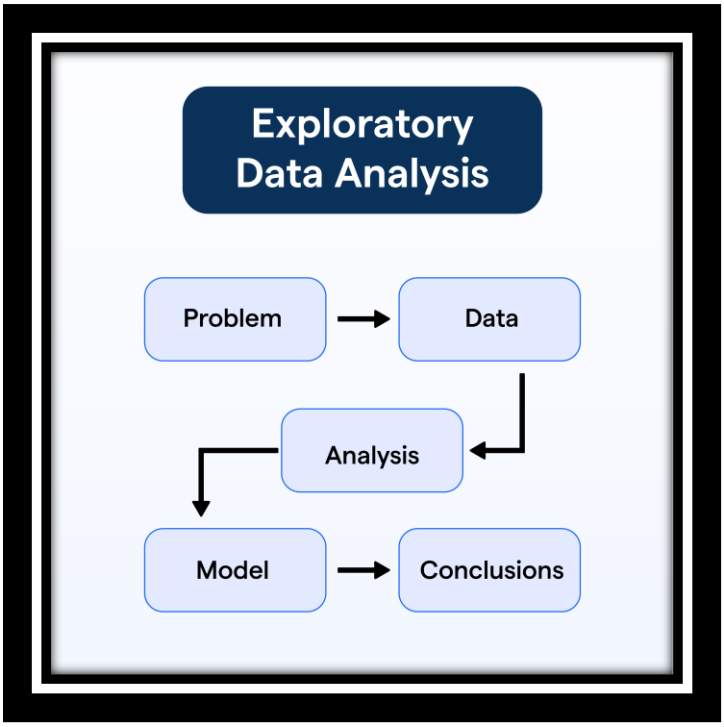
EXPLORATORY DATA ANALYSIS

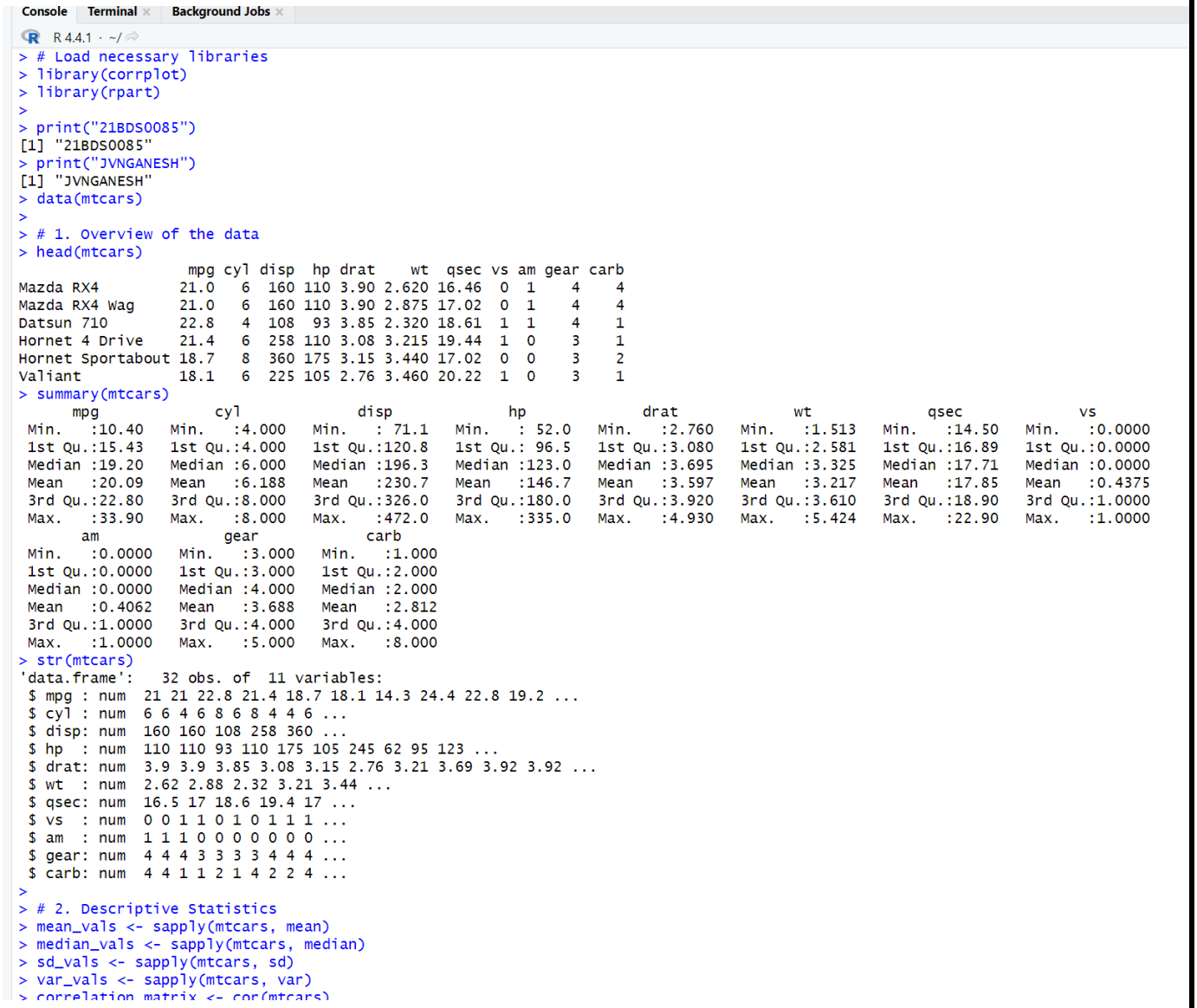
TASK -8



Name : Jvn Ganesh

Roll No : 21BDS0085

Screenshots:



```
Console  Terminal  Background Jobs
R 4.4.1 · ~/
> # Load necessary libraries
> library(corrplot)
> library(rpart)
>
> print("21BDS0085")
[1] "21BDS0085"
> print("JVNGANESH")
[1] "JVNGANESH"
> data(mtcars)
>
> # 1. Overview of the data
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
> summary(mtcars)
      mpg              cyl             disp             hp             drat             wt             qsec             vs
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695   Median :3.325   Median :17.71   Median :0.0000
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
       am              gear            carb
 Min.   :0.0000   Min.   :3.000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
 Median :0.0000   Median :4.000   Median :2.000
 Mean   :0.4062   Mean   :3.688   Mean   :2.812
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :1.0000   Max.   :5.000   Max.   :8.000
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
>
> # 2. Descriptive Statistics
> mean_vals <- sapply(mtcars, mean)
> median_vals <- sapply(mtcars, median)
> sd_vals <- sapply(mtcars, sd)
> var_vals <- sapply(mtcars, var)
> correlation_matrix <- cor(mtcars)
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Source

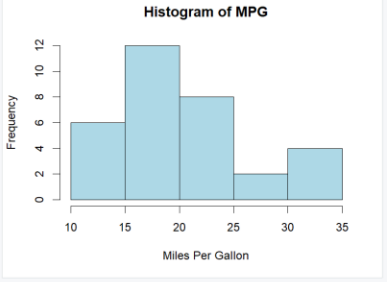
Console | Terminal | Background Jobs

R 4.4.1 · ~/

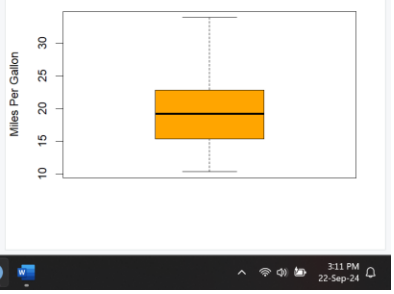
```
> # 2. Descriptive Statistics
> mean_vals <- sapply(mtcars, mean)
> median_vals <- sapply(mtcars, median)
> sd_vals <- sapply(mtcars, sd)
> var_vals <- sapply(mtcars, var)
> correlation_matrix <- cor(mtcars)
>
> # 3. Visualizations
> # Univariate
> hist(mtcars$mpg, main="Histogram of MPG", xlab="Miles Per Gallon", col="lightblue")
> boxplot(mtcars$mpg, main="Boxplot of MPG", ylab="Miles Per Gallon", col="orange")
>
> # Bivariate
> plot(mtcars$wt, mtcars$mpg, main="Weight vs MPG", xlab="Weight", ylab="MPG", pch=19, col="blue")
> boxplot(mpg ~ cyl, data=mtcars, main="MPG by Cylinder", xlab="Number of Cylinders", ylab="MPG", col=c("lightgreen", "lightblue", "lightcora
l"))
>
> # Multivariate
> pairs(mtcars, main="Pair Plot of mtcars Data", pch=21, bg=c("red", "green3", "blue")[unclass(mtcars$cyl)])
> corrplot(correlation_matrix, method="circle")
>
> # Advanced Visualization
> heatmap(correlation_matrix, main="Correlation Heatmap", col=topo.colors(10))
>
> # Principal Component Analysis (PCA)
> pca <- prcomp(mtcars, scale=TRUE)
> biplot(pca)
>
> # 4. Outlier Detection
> boxplot(mtcars$mpg, main="Boxplot for Outlier Detection")
> mtcars$mpg_zscore <- (mtcars$mpg - mean(mtcars$mpg)) / sd(mtcars$mpg)
> mtcars$mpg_outlier <- ifelse(abs(mtcars$mpg_zscore) > 3, "Outlier", "Not Outlier")
> table(mtcars$mpg_outlier)

Not Outlier
         32
>
> # 5. Handling Missing Data (Example)
> # Check for missing values
> colSums(is.na(mtcars))
        mpg         cyl        disp          hp        drat          wt        qsec          vs          am        gear        carb
          0           0           0           0           0           0           0           0           0           0           0
 mpg_zscore mpg_outlier
          0           0
>
> # 6. Feature Engineering
> # Creating new variable based on weight
> mtcars$wt_category <- ifelse(mtcars$wt > 3, "Heavy", "Light")
> # Binning the mpg variable
> mtcars$mpg_bin <- cut(mtcars$mpg, breaks = c(10, 15, 20, 25, 30, 35), labels = c("10-15", "15-20", "20-25", "25-30", "30-35"))
>
> # 7. Hypothesis Testing
> # Corrected t-test for two levels (4 and 6 cylinders)
```

```
R 4.4.1 · ~/
>
> # 6. Feature Engineering
> # Creating new variable based on weight
> mtcars$wt_category <- ifelse(mtcars$wt > 3, "Heavy", "Light")
> # Binning the mpg variable
> mtcars$mpg_bin <- cut(mtcars$mpg, breaks = c(10, 15, 20, 25, 30, 35), labels = c("10-15", "15-20", "20-25", "25-30", "30-35"))
>
> # 7. Hypothesis Testing
> # Corrected t-test for two levels (4 and 6 cylinders)
> mtcars_4_6 <- subset(mtcars, cyl %in% c(4, 6))
> t_test_result <- t.test(mpg ~ cyl, data = mtcars_4_6)
> t_test_result

        Welch Two Sample t-test

data:  mpg by cyl
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

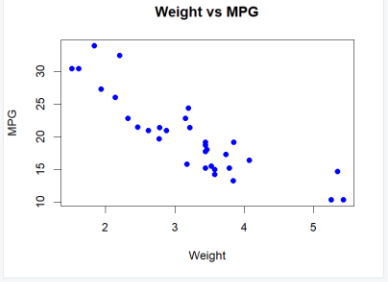
```
        mpg         cyl        disp          hp        drat          wt        qsec          vs          am        gear        carb
          0           0           0           0           0           0           0           0           0           0           0
 mpg_zscore mpg_outlier
          0           0

> # 6. Feature Engineering
> # Creating new variable based on weight
> mtcars$wt_category <- ifelse(mtcars$wt > 3, "Heavy", "Light")
> # Binning the mpg variable
> mtcars$mpg_bin <- cut(mtcars$mpg, breaks = c(10, 15, 20, 25, 30, 35), labels = c("10-15", "15-20", "20-25", "25-30", "30-35"))
>
> # 7. Hypothesis Testing
> # Corrected t-test for two levels (4 and 6 cylinders)
> mtcars_4_6 <- subset(mtcars, cyl %in% c(4, 6))
> t_test_result <- t.test(mpg ~ cyl, data = mtcars_4_6)
> t_test_result

        Welch Two Sample t-test

data:  mpg by cyl
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
```



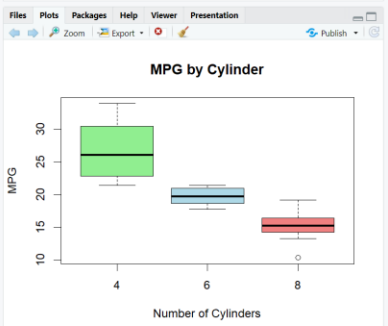

```
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

>
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
>
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
>
```



```
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

>
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
>
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
>
```

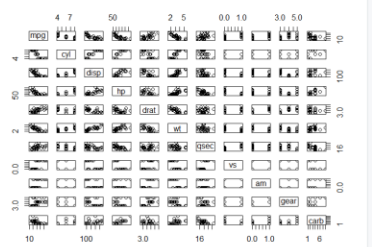


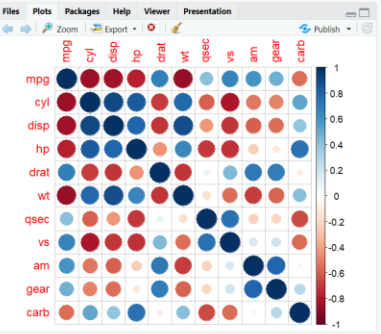

```
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

>
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
>
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
>
```

```
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

>
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
>
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
>
```

```
anova_result    List of 13
correlation_ma… num [1:11, 1:11] 1 -0.852 -0.848 -0.776…
kmeans_result   List of 9
lm_model        List of 12
mtcars          32 obs. of 15 variables
mtcars_4_6      18 obs. of 15 variables
pca             List of 5
t_test_result   List of 10
tree_model      List of 15
mean_vals       Named num [1:11] 20.09 6.19 230.72 146.69…
median_vals     Named num [1:11] 19.2 6 196.3 123 3.7 ...
sd_vals         Named num [1:11] 6.027 1.786 123.939 68.5…
var_vals        Named num [1:11] 3.63e+01 3.19 1.54e+04 4…
```

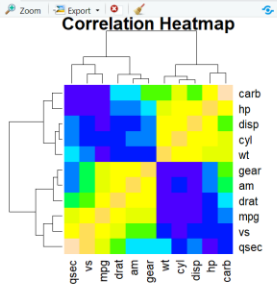


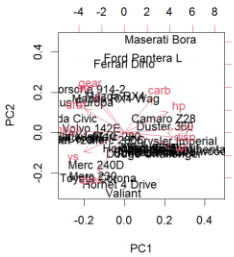



```
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

>
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
>
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
>
```

```
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

>
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
>
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
>
```

Environment	History	Connections	Tutorial
Data			
anova_result	List of 13		
correlation_ma…	num [1:11, 1:11] 1 -0.852 -0.848 -0.776…		
kmeans_result	List of 9		
lm_model	List of 12		
mtcars	32 obs. of 15 variables		
mtcars_4_6	18 obs. of 15 variables		
pca	List of 5		
t_test_result	List of 10		
tree_model	List of 15		
Values			
mean_vals	Named num [1:11] 20.09 6.19 230.72 146.69…		
median_vals	Named num [1:11] 19.2 6 196.3 123 3.7 ...		
sd_vals	Named num [1:11] 6.027 1.786 123.939 68.5…		
var_vals	Named num [1:11] 3.63e+01 3.19 1.54e+04 4…		

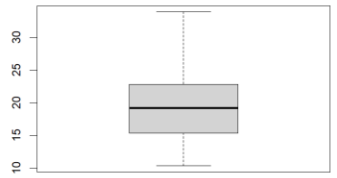




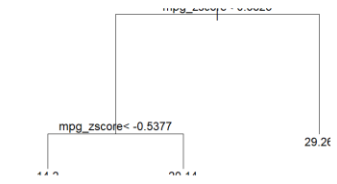

```
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

>
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
>
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
>
```

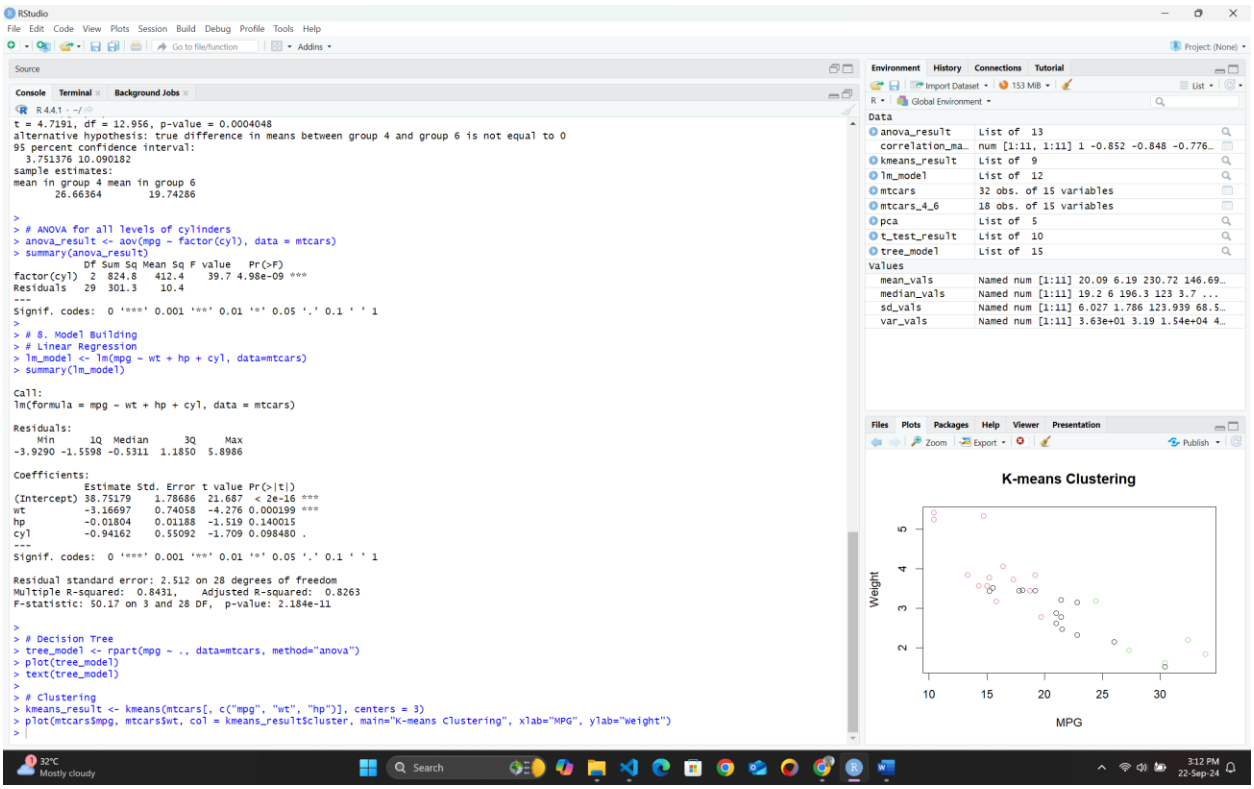
```
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

>
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
>
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
>
```

Boxplot for Outlier Detection

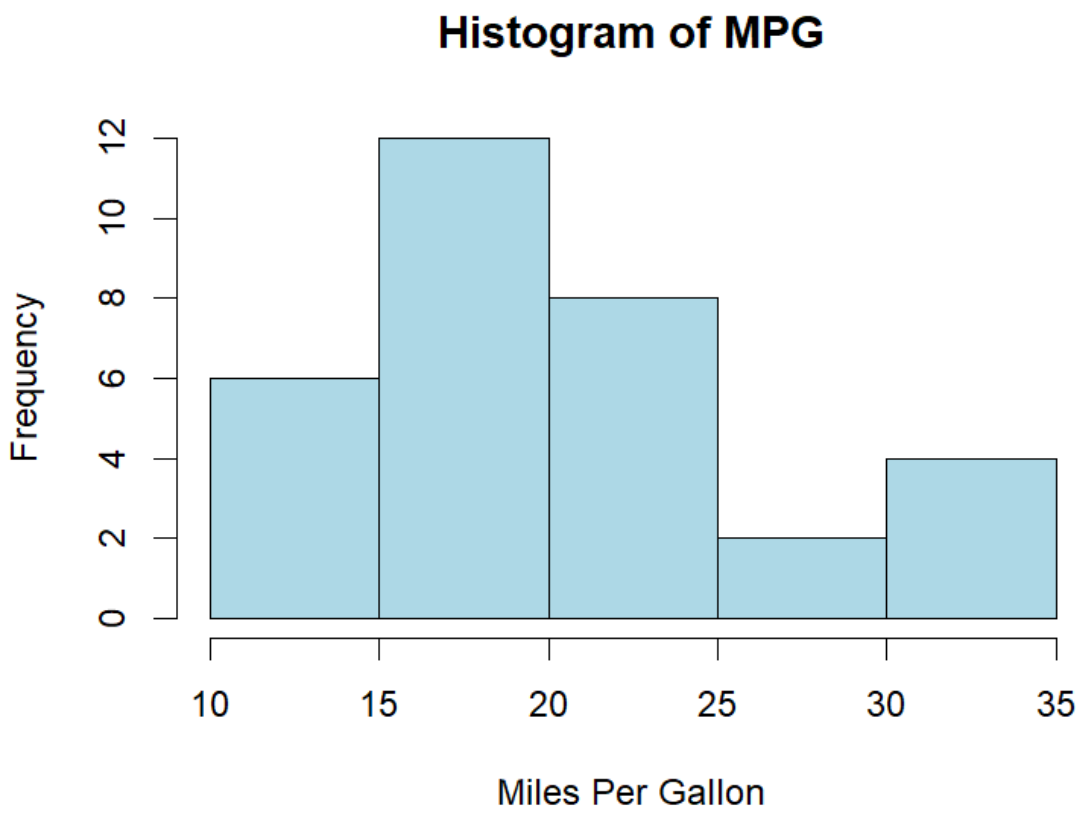


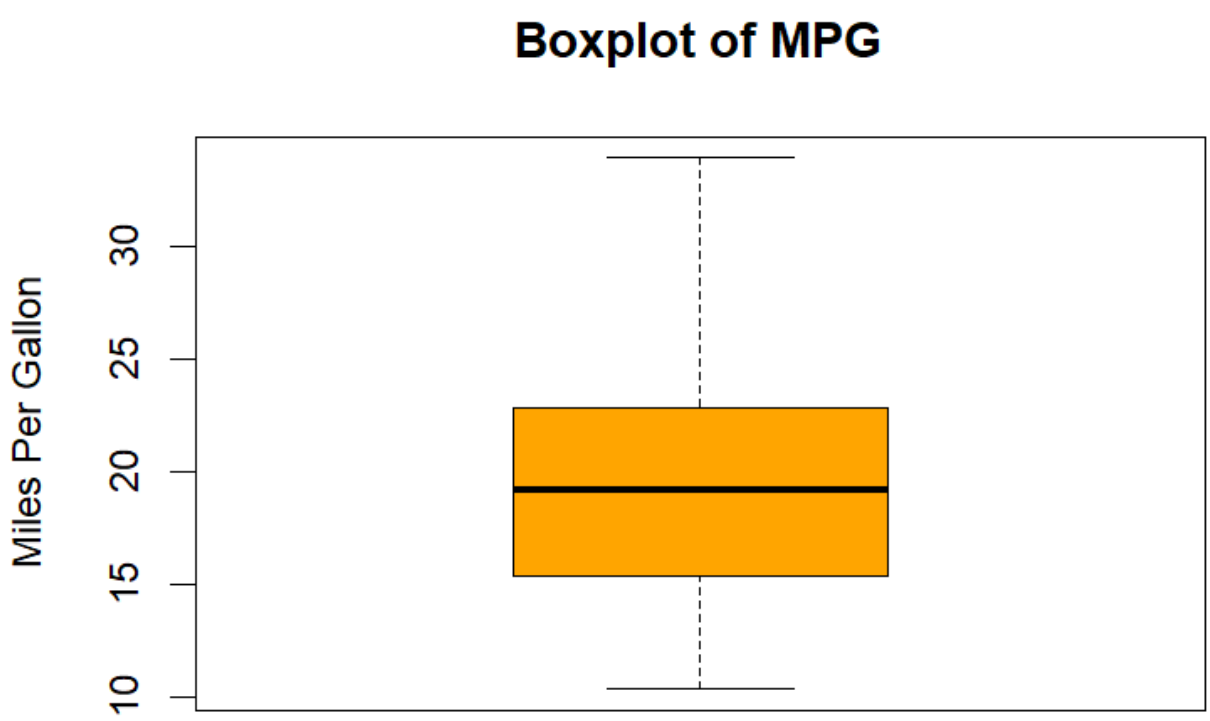

```
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
  3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286

>
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(cyl)  2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)

Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9290 -1.5598 -0.5311  1.1850  5.8986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
wt          -3.16697    0.74058  -4.276 0.000199 ***
hp          -0.01804    0.01188  -1.519 0.140015
cyl         -0.94162    0.55092  -1.709 0.098480 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11

>
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
>
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
>
```

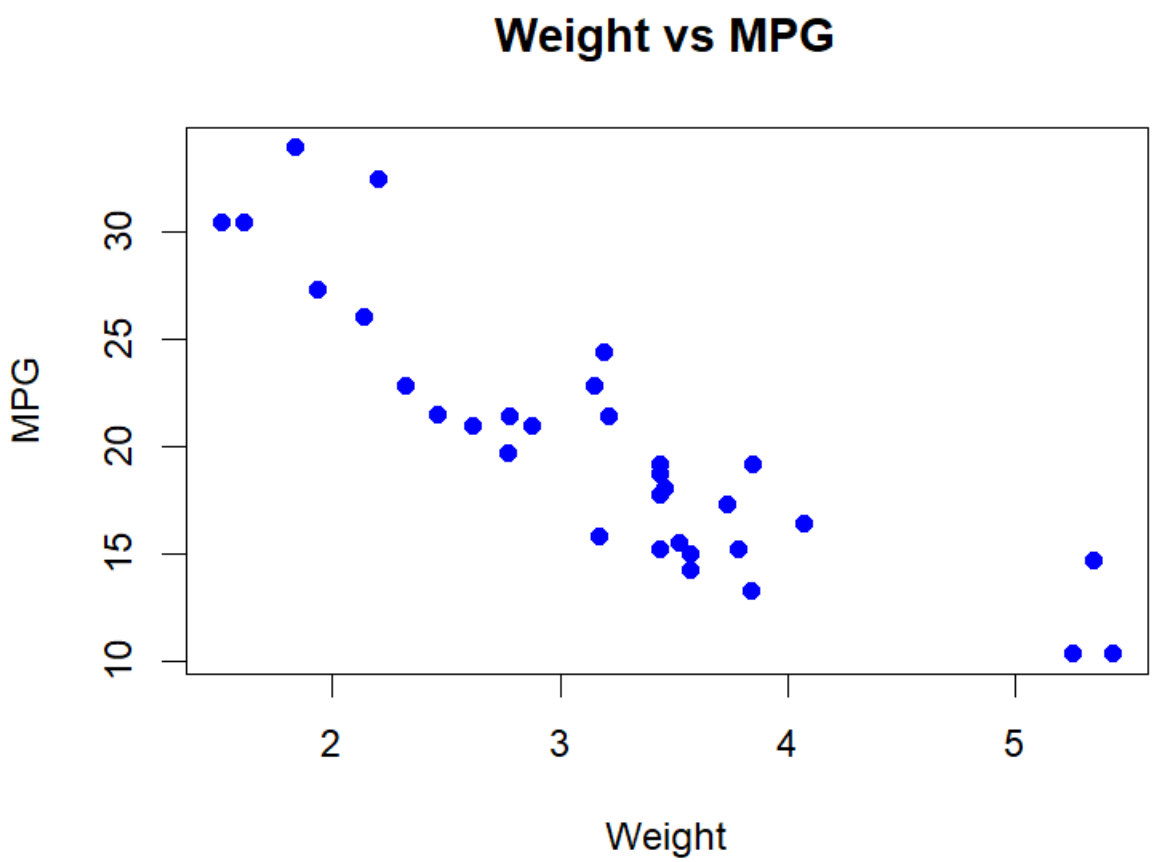



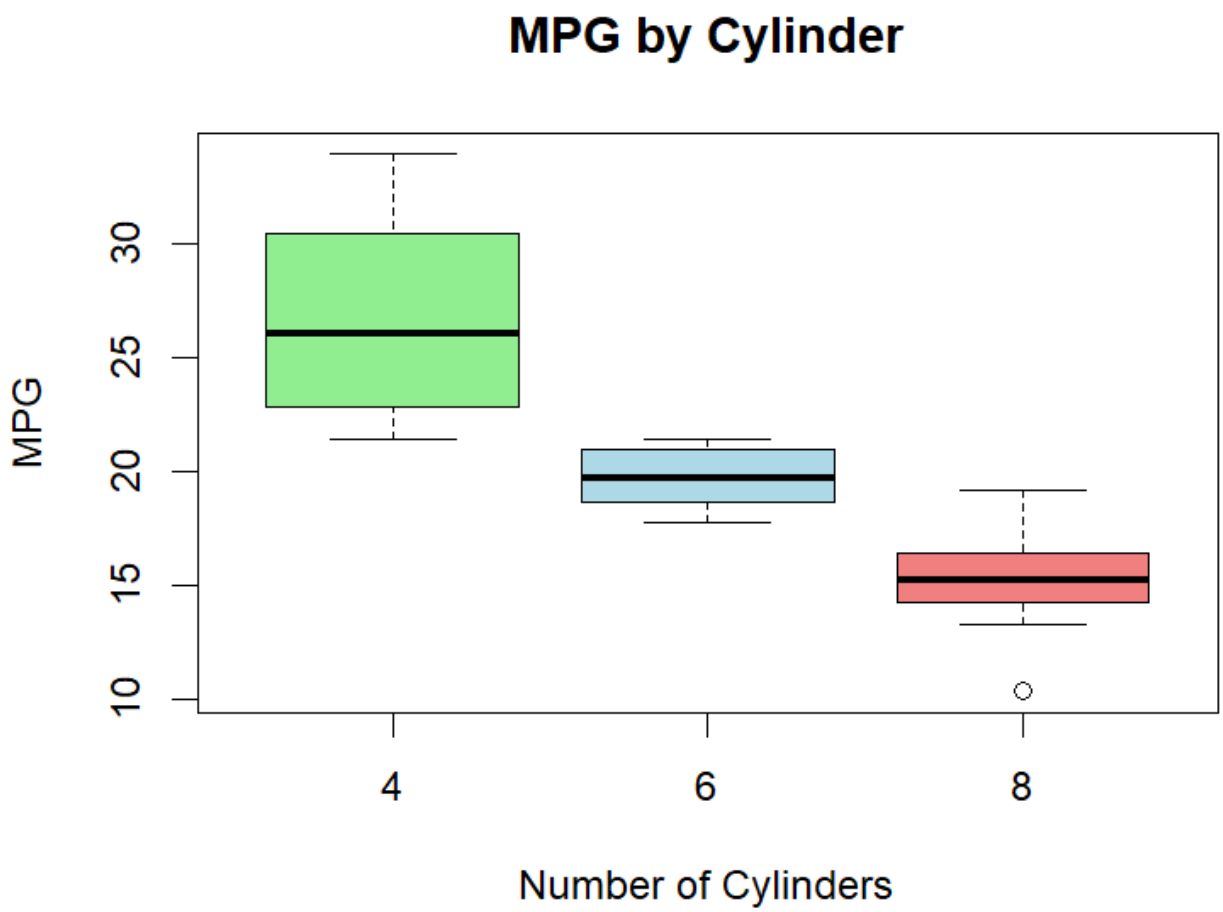

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R 4.4.1 · ~/
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means between group 4 and group 6 is not equal to 0
95 percent confidence interval:
3.751376 10.090182
sample estimates:
mean in group 4 mean in group 6
26.66364 19.74286
> # ANOVA for all levels of cylinders
> anova_result <- aov(mpg ~ factor(cyl), data = mtcars)
> summary(anova_result)
Df Sum Sq Mean Sq F value Pr(>F)
factor(cyl) 2 824.8 412.4 39.7 4.98e-09 ***
Residuals 29 301.3 10.4
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # 8. Model Building
> # Linear Regression
> lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
> summary(lm_model)
Call:
lm(formula = mpg ~ wt + hp + cyl, data = mtcars)
Residuals:
Min 1Q Median 3Q Max
-3.9290 -1.5598 -0.5311 1.1850 5.8986
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179 1.78686 21.687 < 2e-16 ***
wt -3.16697 0.74058 -4.276 0.000199 ***
hp -0.01804 0.01188 -1.519 0.140015
cyl -0.94162 0.55092 -1.709 0.098480 .
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared: 0.8431, Adjusted R-squared: 0.8263
F-statistic: 50.17 on 3 and 28 DF, p-value: 2.184e-11
> # Decision Tree
> tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
> plot(tree_model)
> text(tree_model)
> # Clustering
> kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
> plot(mtcars$mpg, mtcars$wt, col = kmeans_result$cluster, main="K-means Clustering", xlab="MPG", ylab="Weight")
Environment History Connections Tutorial
Import Dataset 153 MiB
Global Environment
Data
anova_result List of 13
correlation_ma... num [1:11, 1:11] 1 -0.852 -0.848 -0.776...
kmeans_result List of 9
lm_model List of 12
mtcars 32 obs. of 15 variables
mtcars_4_6 18 obs. of 15 variables
pca List of 5
t_test_result List of 10
tree_model List of 15
Values
mean_vals Named num [1:11] 20.09 6.19 230.72 146.69...
median_vals Named num [1:11] 19.2 6 196.3 123 3.7 ...
sd_vals Named num [1:11] 6.027 1.786 123.939 68.5...
var_vals Named num [1:11] 3.63e+01 3.19 1.54e+04 4...
Files Plots Packages Help Viewer Presentation
Zoom Export Publish
K-means Clustering
Weight
MPG
2
3
4
5
10
15
20
25
30
32°C
Mostly cloudy
Search
3:12 PM
22-Sep-24

Plots:

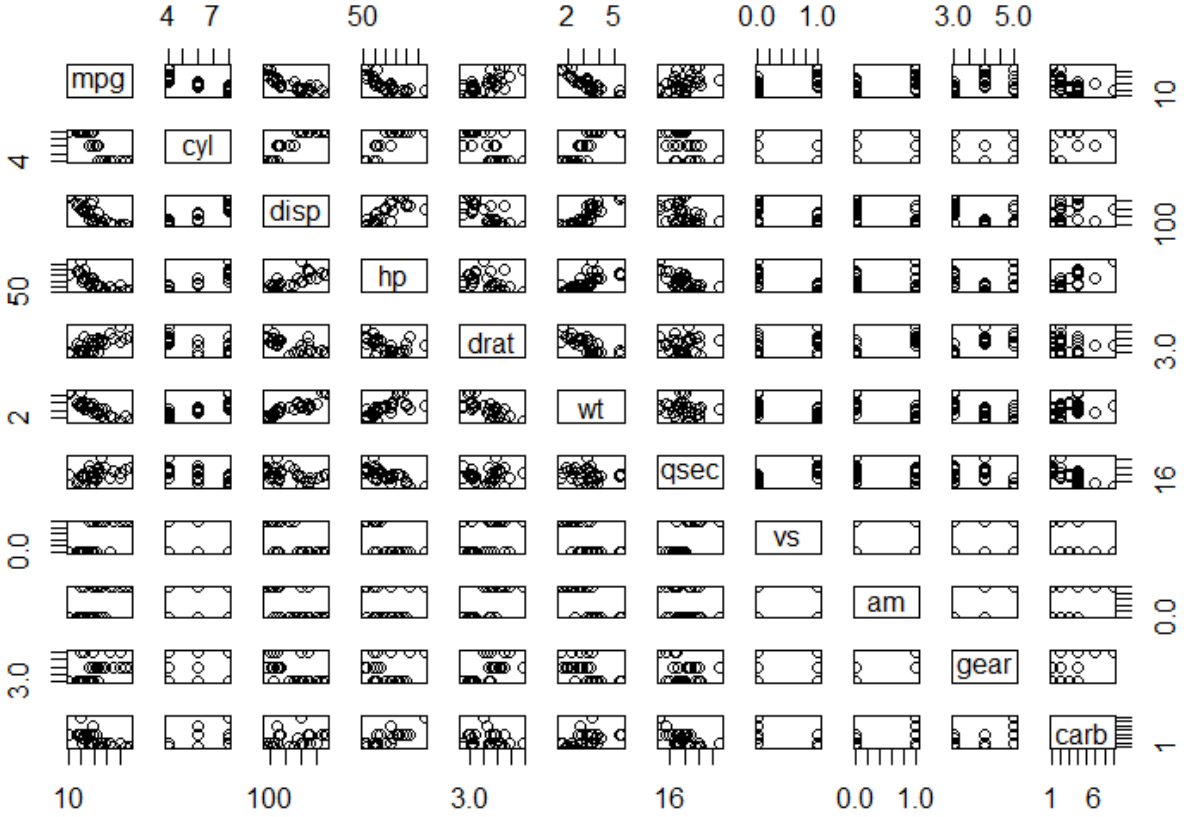

Histogram of MPG
Frequency
0
2
4
6
8
10
12
10
15
20
25
30
35
Miles Per Gallon

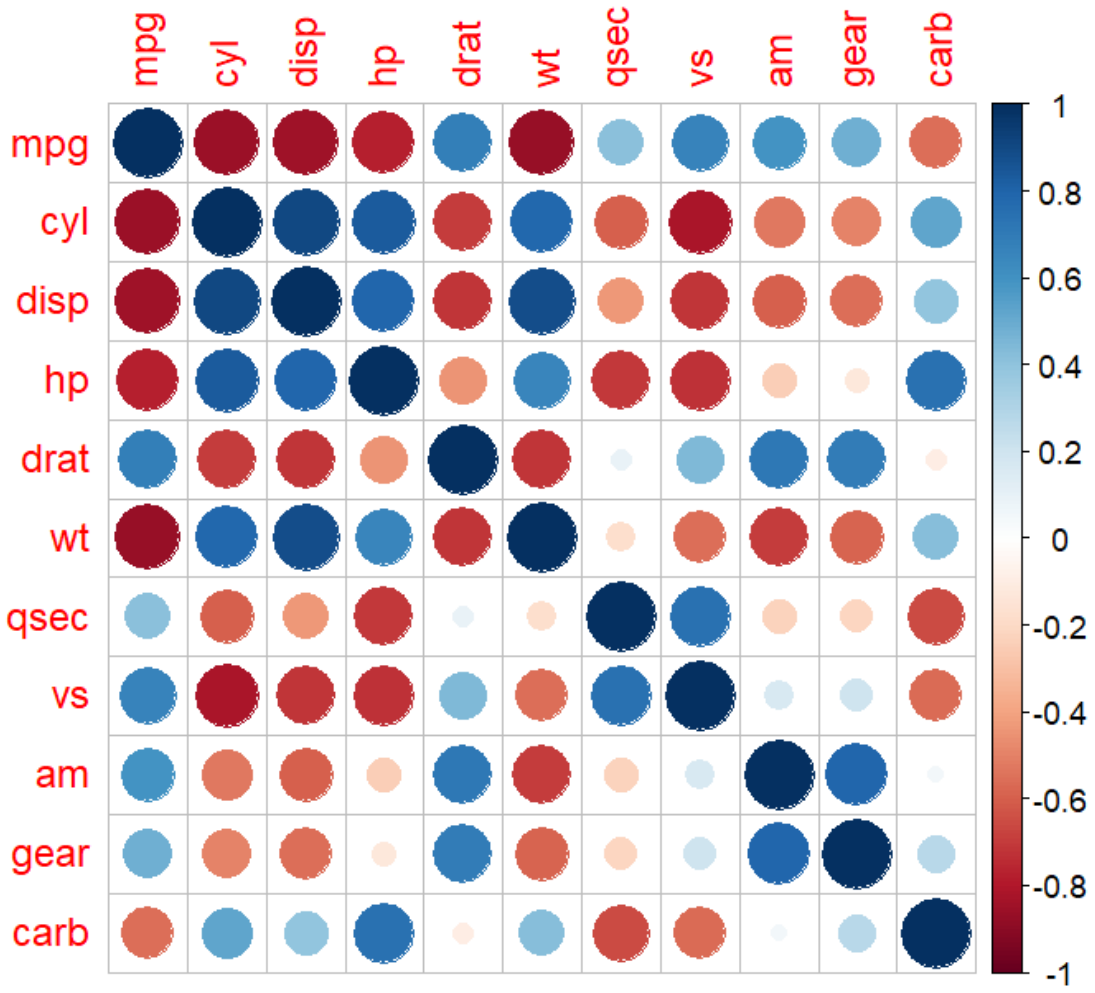

Boxplot of MPG
Miles Per Gallon
10
15
20
25
30

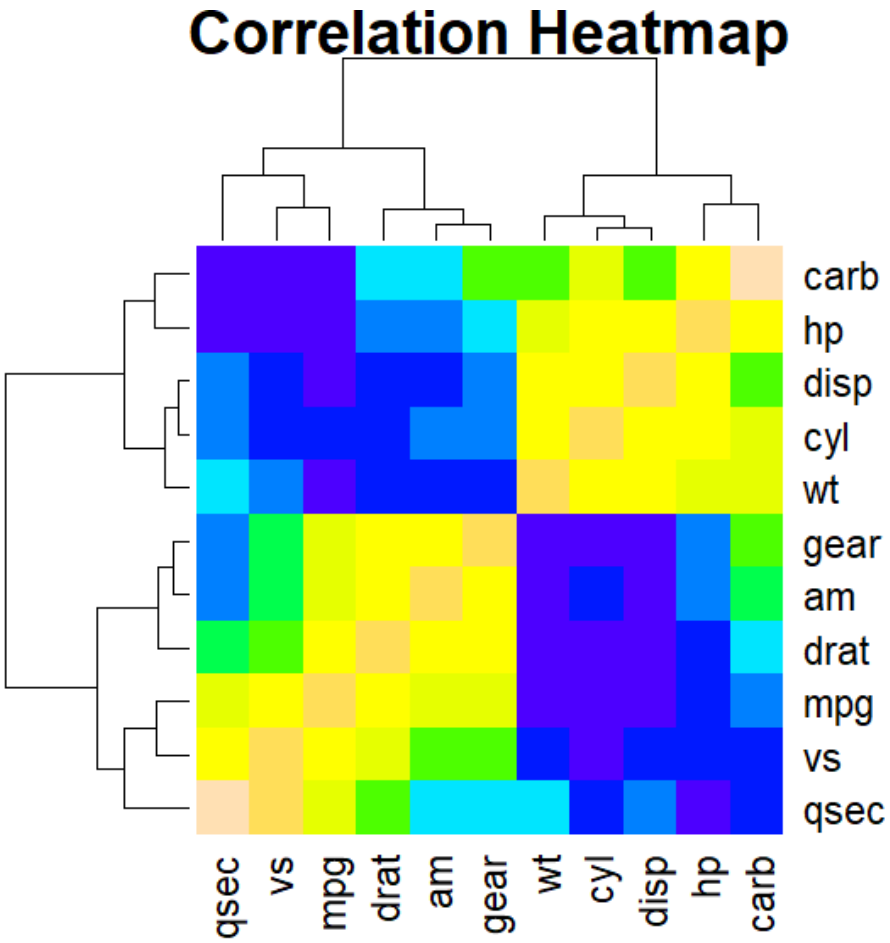

Weight vs MPG
MPG
10
15
20
25
30
2
3
4
5
Weight

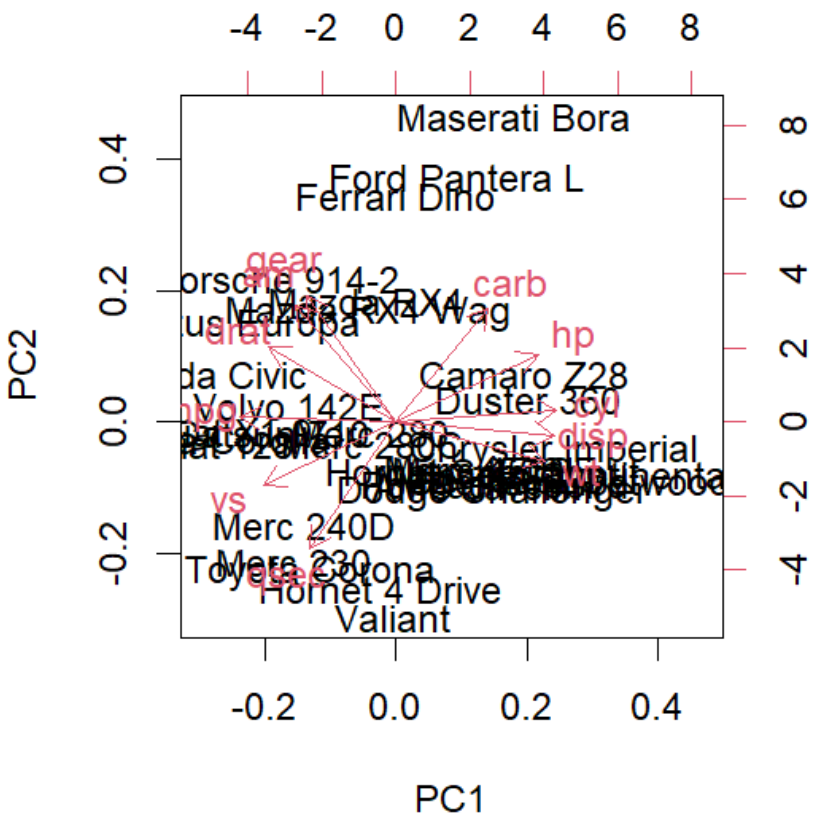

MPG by Cylinder
MPG
10
15
20
25
30
4
6
8
Number of Cylinders

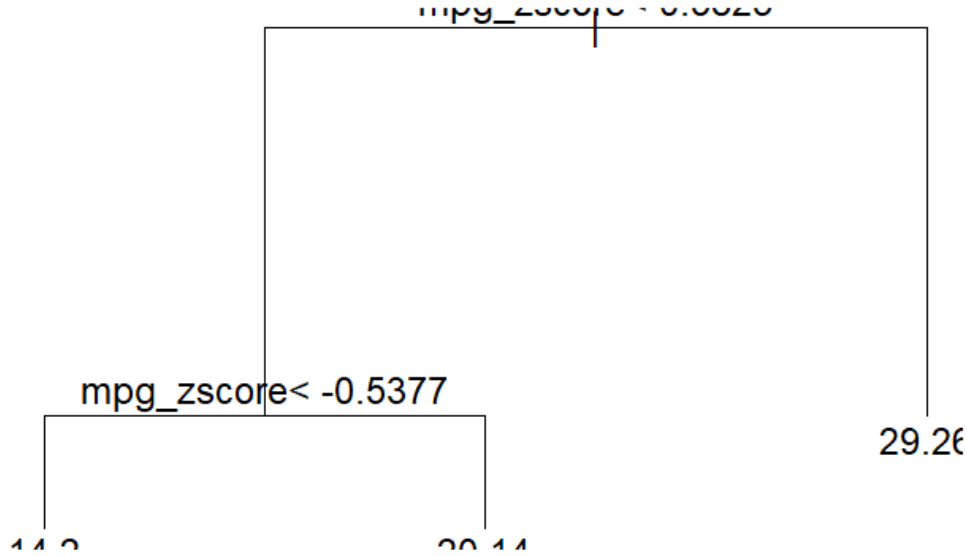
Pair Plot of mtcars Data

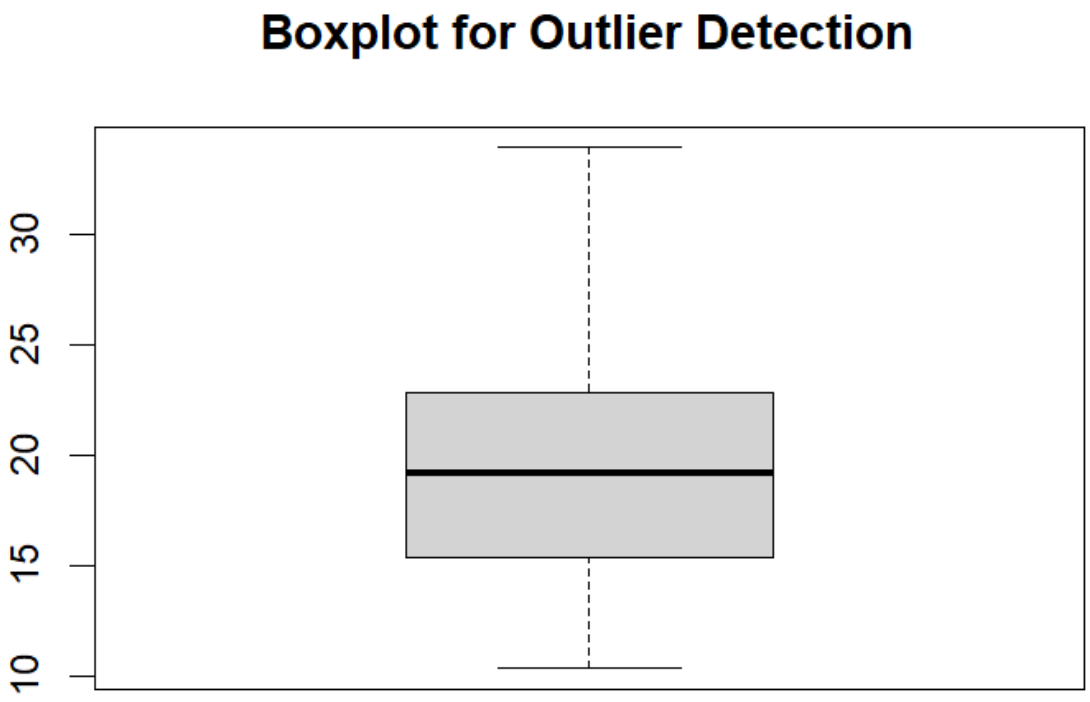


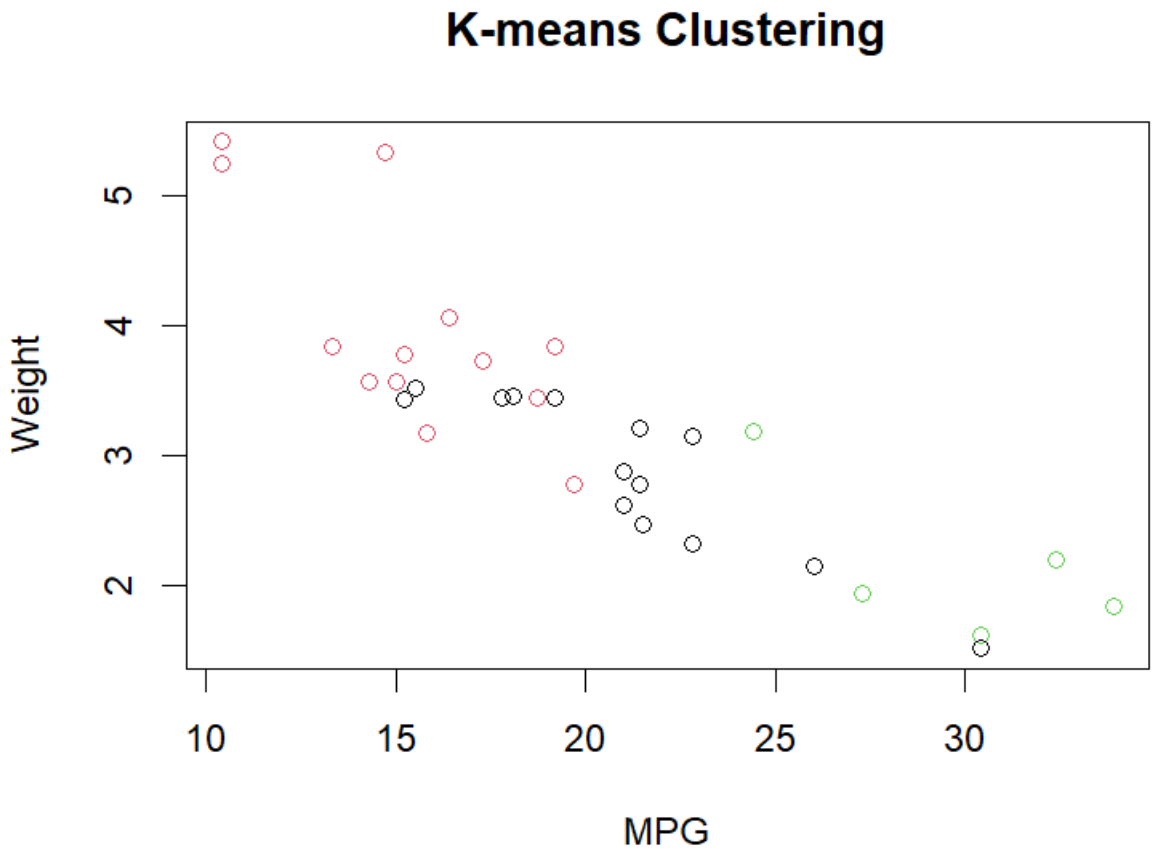

mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Correlation Heatmap
carb
hp
disp
cyl
wt
gear
am
drat
mpg
vs
qsec
qsec
vs
mpg
drat
am
gear
wt
cyl
disp
hp
carb


-4
-2
0
2
4
6
8
Maserati Bora
Ford Pantera L
Ferrari Dino
gear
carb
hp
Camaro Z28
Duster 360
disp
Chrysler Imperial
vs
Merc 240D
Hornet 4 Drive
Valiant
0.4
0.2
0.0
-0.2
PC2
PC1


mpg_zscore< -0.5377
29.26


Boxplot for Outlier Detection
10
15
20
25
30


K-means Clustering
Weight
2
3
4
5
10
15
20
25
30
MPG

Code ：

```
library(corrplot)
library(rpart)

# Load the dataset
data(mtcars)

# 1. Overview of the data
head(mtcars)
summary(mtcars)
str(mtcars)

# 2. Descriptive Statistics
mean_vals <- sapply(mtcars, mean)
median_vals <- sapply(mtcars, median)
sd_vals <- sapply(mtcars, sd)
var_vals <- sapply(mtcars, var)
correlation_matrix <- cor(mtcars)
```

```
# 3. Visualizations

# Univariate

hist(mtcars$mpg, main="Histogram of MPG",
xlab="Miles Per Gallon", col="lightblue")

boxplot(mtcars$mpg, main="Boxplot of MPG",
ylab="Miles Per Gallon", col="orange")


# Bivariate

plot(mtcars$wt, mtcars$mpg, main="Weight vs MPG",
xlab="Weight", ylab="MPG", pch=19, col="blue")

boxplot(mpg ~ cyl, data=mtcars, main="MPG by
Cylinder", xlab="Number of Cylinders", ylab="MPG",
col=c("lightgreen", "lightblue", "lightcoral"))


# Multivariate

pairs(mtcars, main="Pair Plot of mtcars Data", pch=21,
bg=c("red", "green3", "blue")[unclass(mtcars$cyl)])

corrplot(correlation_matrix, method="circle")
```

```
# Advanced Visualization
heatmap(correlation_matrix, main="Correlation Heatmap", col=topo.colors(10))

# Principal Component Analysis (PCA)
pca <- prcomp(mtcars, scale=TRUE)
biplot(pca)

# 4. Outlier Detection
boxplot(mtcars$mpg, main="Boxplot for Outlier Detection")
mtcars$mpg_zscore <- (mtcars$mpg - mean(mtcars$mpg)) / sd(mtcars$mpg)
mtcars$mpg_outlier <- ifelse(abs(mtcars$mpg_zscore) > 3, "Outlier", "Not Outlier")
table(mtcars$mpg_outlier)
```

```
# 5. Handling Missing Data (Example)

# Check for missing values

colSums(is.na(mtcars))

# Impute missing values with mean (if any)

# mtcars$mpg[is.na(mtcars$mpg)] <-
mean(mtcars$mpg, na.rm = TRUE)


# 6. Feature Engineering

# Creating new variable based on weight

mtcars$wt_category <- ifelse(mtcars$wt > 3, "Heavy",
"Light")

# Binning the mpg variable

mtcars$mpg_bin <- cut(mtcars$mpg, breaks = c(10,
15, 20, 25, 30, 35), labels = c("10-15", "15-20", "20-25",
"25-30", "30-35"))


# 7. Hypothesis Testing

# t-test

t_test_result <- t.test(mpg ~ cyl, data = mtcars)
```

```
# ANOVA
anova_result <- anova(lm(mpg ~ cyl + wt + hp, data = mtcars))

# 8. Model Building
# Linear Regression
lm_model <- lm(mpg ~ wt + hp + cyl, data=mtcars)
summary(lm_model)

# Decision Tree
tree_model <- rpart(mpg ~ ., data=mtcars, method="anova")
plot(tree_model)
text(tree_model)

# Clustering
kmeans_result <- kmeans(mtcars[, c("mpg", "wt", "hp")], centers = 3)
```

```
plot(mtcars$mpg, mtcars$wt, col =
kmeans_result$cluster, main="K-means Clustering",
xlab="MPG", ylab="Weight")
```