|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Distribution Analysis  Assignment-1 |
| |  |  |  | | --- | --- | --- | | Jay Panchal – 100960958 | 5/27/24 | Statistical and Predictive Modeling | |

**Question 1:**

**Scatterplot with smoothing for mpg vs wt.**

**A graph of a graph showing a line of fuel efficiency

Description automatically generated**

**Question 2:**

**2 (a): Identifiable Trends:**

**Direction:** There is a clear negative trend; as car weight (wt.) increases, fuel efficiency (mpg) decreases.

**Shape / form:** The smoothing line is slightly curved but mostly linear (using linear model – lm ), indicating a steady decline in mpg with increasing wt.

**Strength:** The points are close to the smoothing line, suggesting a strong negative correlation.

**Range:** The weight (wt.) ranges from 1.5 to 5.5 (in 1000 lbs.), and the miles per gallon (mpg) ranges from 10 to 35.

**Concentration of Data points:** Most data points are concentrated in the range of 2.5 to 3.5 (in 1000 lbs.) for wt. and 15 to 25 for mpg.

**2 (b): Anomalies:**

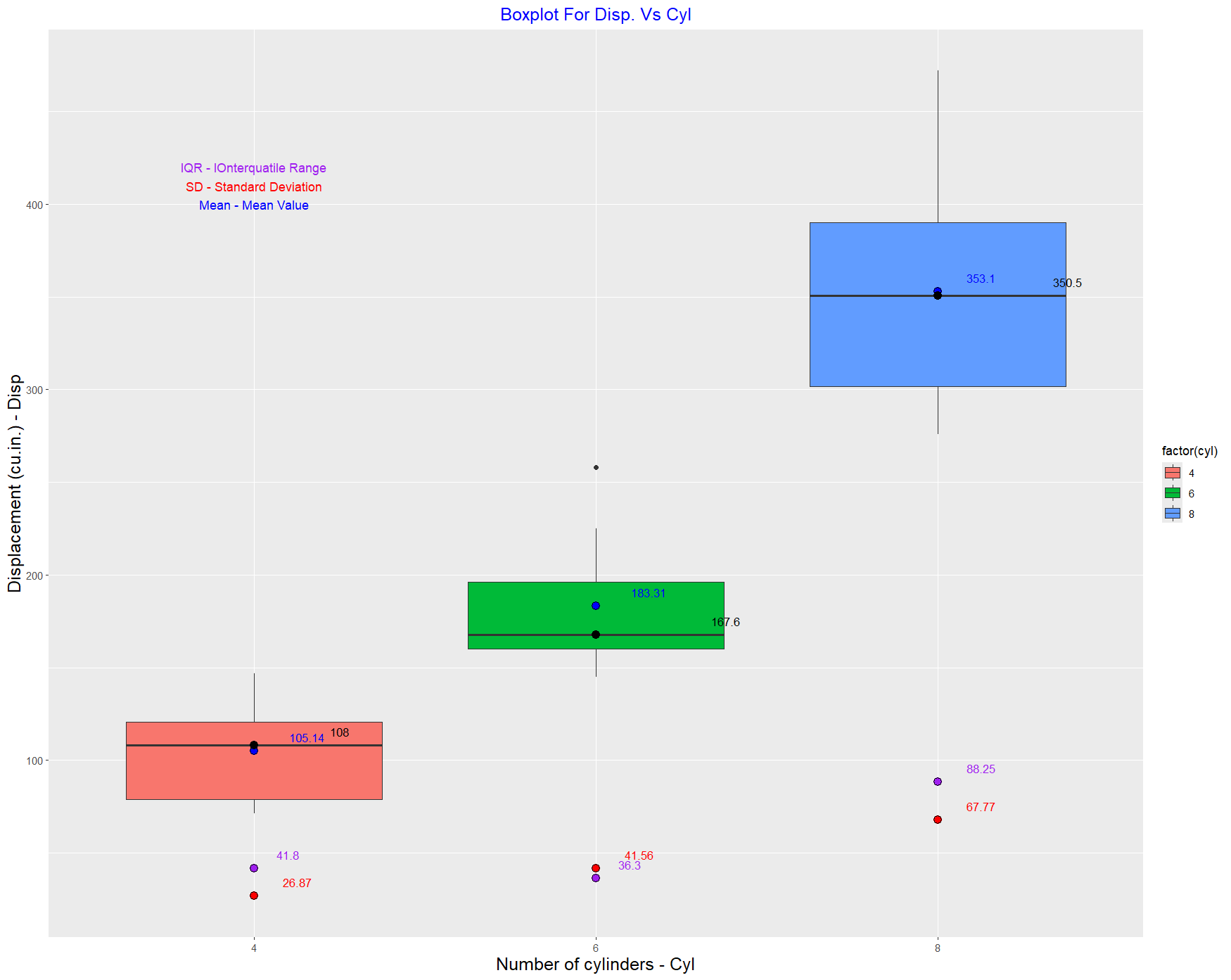
**Outliers:** There are few points with high and low wt. but relatively high mpg which are unusual and considered outliers.

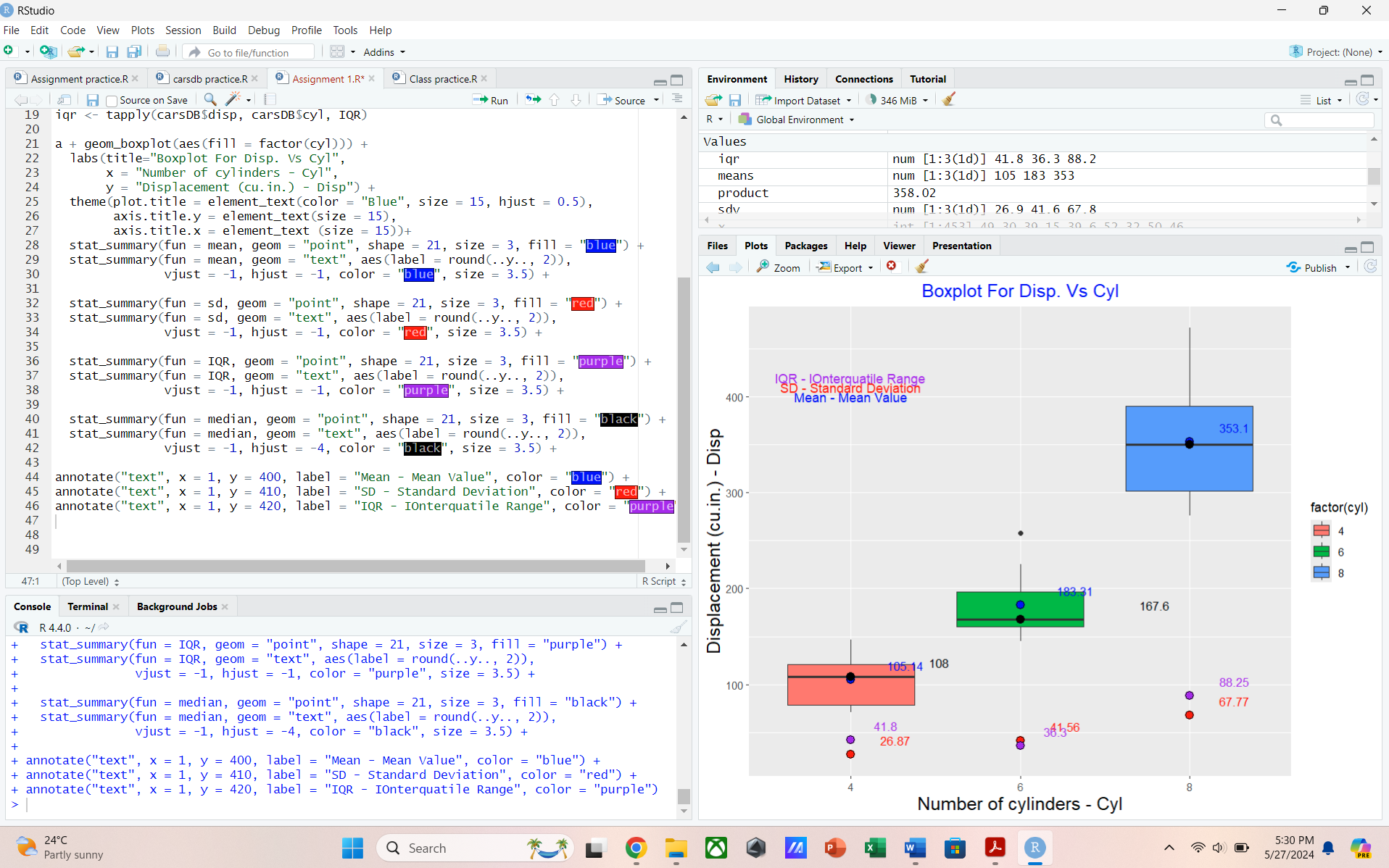
**Clusters:** There could be a cluster of data points around wt. of 3.0 to 3.5, indicating many cars have weights in this range.

**Gaps:** There might be a gap in the wt. range from 4.5 to 5.0, indicating few cars have weights in this range.

**Question 3:**

**Box plot for Disp. Vs Cyl.**

****



**Question 4:**

**4 (a): Measure of Central Tendency:**

**Mean:** The mean is the average of the displacement (disp) values for each cylinder group, which is not directly visible on a boxplot. It can be calculated separately and can be shown in graph.

In the boxplot, the mean is represented by a blue point for each group.

Additionally, the mean value is labeled near the blue point to indicate its value.

**Values:**

**A screenshot of a computer

Description automatically generated**

**Median:** The median is the middle value of the displacement data when it is ordered from lowest to highest.

In the boxplot, the median is represented by a horizontal line inside the box.

The median is also marked by a black point and labeled near the point to indicate its value.

**Values:**

A screenshot of a computer

Description automatically generated

**4 (b): Measure of Dispersion:**

**Standard Deviation:** The standard deviation measures the spread of the data around the mean. While it isn’t directly shown on the boxplot, a wider box and longer whiskers generally indicate a higher standard deviation.

The sd is also marked by a red point and labeled near the point to indicate its value.

**Values:**

A screenshot of a computer

Description automatically generated

**Range:** The range is the difference between the maximum and minimum displacement values for cars with a specific number of cylinders.

For 4-cylinder cars, the range appears to be from around 70 to 150.

For 6-cylinder cars, the range appears to be from around 140 to 230.

For 8-cylinder cars, the range appears to be from around 270 to 475.

**Interquartile Range:** The range between the first quartile (25th percentile) and the third quartile (75th percentile). The IQR is represented by the height of the box in the boxplot.

**Values:**

A screenshot of a computer

Description automatically generated

**Code:**

**library(ggplot2)**

**#mpg vs wt Scatter plot**

g <- ggplot(data = carsDB, aes(x = wt,y = mpg))

g + geom\_point(color = "Red")+

labs(title="Fuel Efficiency by Weight",

x = "Weight - wt (1000 lbs)",

y = "Fuel Efficiency (mpg)")+

theme(plot.title = element\_text(color = "Blue", size = 15, hjust = 0.5),

axis.title.y = element\_text(size = 15),

axis.title.x = element\_text (size = 15))+

geom\_smooth(col = "Black", method = 'lm')

**#boxplot for disp. vs cyl**

a <- ggplot(data = carsDB,aes(x = factor(cyl), y = disp))

means <- tapply(carsDB$disp, carsDB$cyl, mean) **# Calculate mean value**

sdv <- tapply(carsDB$disp, carsDB$cyl, sd) **# Calculate median value**

iqr <- tapply(carsDB$disp, carsDB$cyl, IQR) **# Calculate IQR value**

a + geom\_boxplot(aes(fill = factor(cyl))) +

labs(title="Boxplot For Disp. Vs Cyl",

x = "Number of cylinders - Cyl",

y = "Displacement (cu.in.) - Disp") +

theme(plot.title = element\_text(color = "Blue", size = 15, hjust = 0.5),

axis.title.y = element\_text(size = 15),

axis.title.x = element\_text (size = 15))+

**# Calculate mean value in box plot**

stat\_summary(fun = mean, geom = "point", shape = 21, size = 3, fill = "blue") +

**# Write mean value in box plot**

stat\_summary(fun = mean, geom = "text", aes(label = round(..y.., 2)),

vjust = -1, hjust = -1, color = "blue", size = 3.5) +

**# Calculate sd value in box plot**

stat\_summary(fun = sd, geom = "point", shape = 21, size = 3, fill = "red") +

**# Write sd value in boxplot**

stat\_summary(fun = sd, geom = "text", aes(label = round(..y.., 2)),

vjust = -1, hjust = -1, color = "red", size = 3.5) +

**# Calculate IQR value in box plot**

stat\_summary(fun = IQR, geom = "point", shape = 21, size = 3, fill = "purple") +

**# Write IQR value in boxplot**

stat\_summary(fun = IQR, geom = "text", aes(label = round(..y.., 2)),

vjust = -1, hjust = -1, color = "purple", size = 3.5) +

**# Calculate median value in boxplot**

stat\_summary(fun = median, geom = "point", shape = 21, size = 3, fill = "black") +

**# Write median value in boxplot**

stat\_summary(fun = median, geom = "text", aes(label = round(..y.., 2)),

vjust = -1, hjust = -4, color = "black", size = 3.5) +

annotate("text", x = 1, y = 400, label = "Mean - Mean Value", color = "blue") +

annotate("text", x = 1, y = 410, label = "SD - Standard Deviation", color = "red") +

annotate("text", x = 1, y = 420, label = "IQR - IOnterquatile Range", color = "purple")