Assignment 1

Sharmin Akhter

2023-03-01

Contents

```
library(gridExtra)
library(reshape2)
library(ICSNP)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: ICS
```

```
library(datasets)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::combine() masks gridExtra::combine()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(dplyr)
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(qcc)
```

```
## Package 'qcc' version 2.7
## Type 'citation("qcc")' for citing this R package in publications.
```

```
library(ggplot2)
```

Problem I:

The following defects data collected from the last month's inspection reports for a particular type of tank. Construct a Pareto diagram and discuss the results.

Type of defects:

Parts damaged, Machine problem, supplied parts rusted,Masking insufficient, Misaligned weld, Processing out of order,unfinished fairing,incorrect dimension, Adhesive failure,Paint out of limits, improper test procedure

Frequency:

34, 29, 13, 17, 2, 4, 3, 36, 6, 10, 1

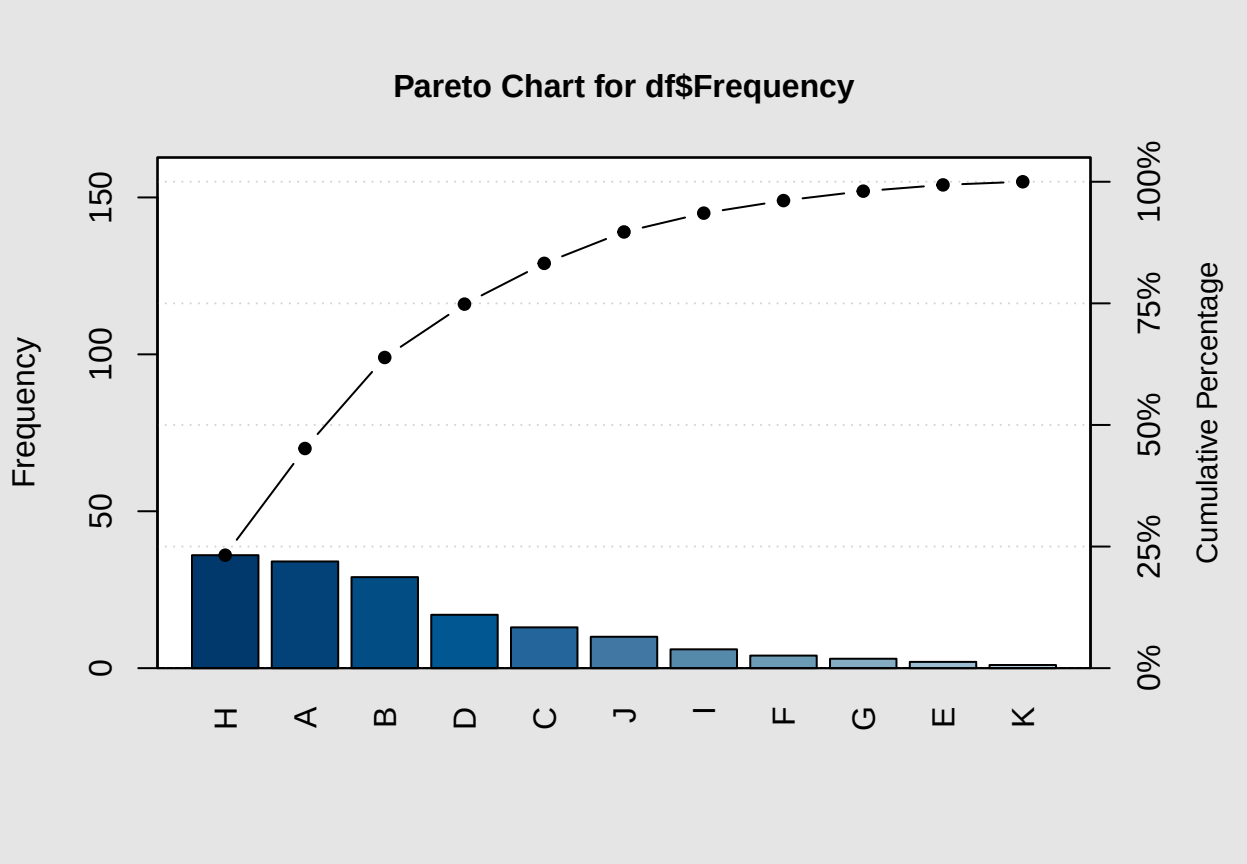
Create DataFrame with Type of defects and Frequency

```
df<- data.frame(Type_of_Defects = c('Parts damaged', 'Machine problem', 'Supplied parts rusted', 'Maski
df
```

```
##            Type_of_Defects Frequency
## 1            Parts damaged        34
## 2          Machine problem        29
## 3    Supplied parts rusted        13
## 4     Masking insufficient        17
## 5          Misaligned weld         2
## 6  Processing out of order         4
## 7       Unfinished fairing         3
## 8      Incorrect dimension        36
## 9         Adhesive failure         6
## 10     Paint out of limits        10
## 11 Improper test procedure         1
```

Pareto diagram

```
pareto.chart(df$Frequency)
```




```
##
## Pareto chart analysis for df$Frequency
##       Frequency   Cum.Freq.  Percentage Cum.Percent.
##   H  36.0000000  36.0000000  23.2258065   23.2258065
##   A  34.0000000  70.0000000  21.9354839   45.1612903
##   B  29.0000000  99.0000000  18.7096774   63.8709677
##   D  17.0000000 116.0000000  10.9677419   74.8387097
##   C  13.0000000 129.0000000   8.3870968   83.2258065
##   J  10.0000000 139.0000000   6.4516129   89.6774194
##   I   6.0000000 145.0000000   3.8709677   93.5483871
##   F   4.0000000 149.0000000   2.5806452   96.1290323
##   G   3.0000000 152.0000000   1.9354839   98.0645161
##   E   2.0000000 154.0000000   1.2903226   99.3548387
##   K   1.0000000 155.0000000   0.6451613  100.0000000
```

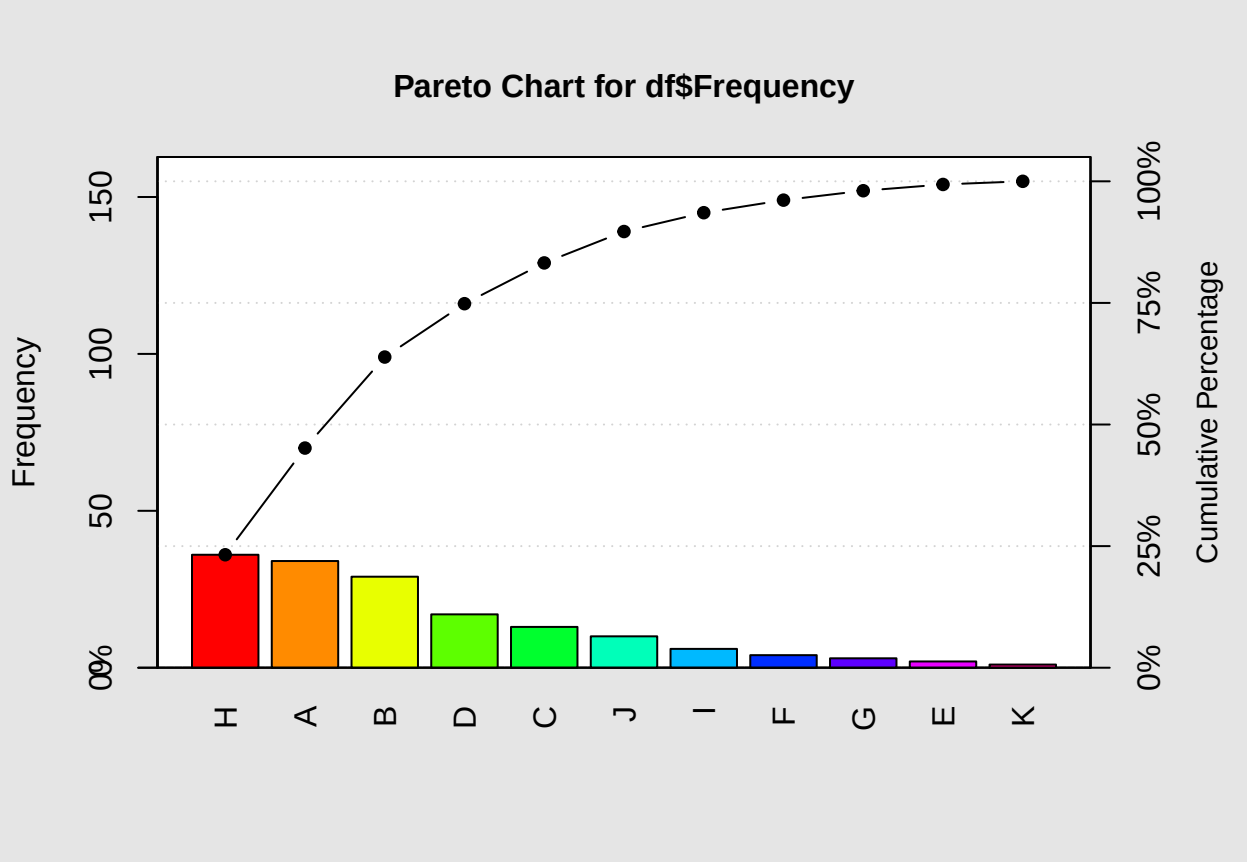
Modify Pareto diagram

```
bar_colors <- rainbow(length(df$Frequency))
pareto.chart(df$Frequency, col = bar_colors, ylab = "Frequency", yaxt = "n")
```

```
##
## Pareto chart analysis for df$Frequency
##       Frequency   Cum.Freq.  Percentage Cum.Percent.
##   H  36.0000000  36.0000000  23.2258065   23.2258065
##   A  34.0000000  70.0000000  21.9354839   45.1612903
##   B  29.0000000  99.0000000  18.7096774   63.8709677
```

```
##    D  17.0000000 116.0000000  10.9677419   74.8387097
##    C  13.0000000 129.0000000   8.3870968   83.2258065
##    J  10.0000000 139.0000000   6.4516129   89.6774194
##    I   6.0000000 145.0000000   3.8709677   93.5483871
##    F   4.0000000 149.0000000   2.5806452   96.1290323
##    G   3.0000000 152.0000000   1.9354839   98.0645161
##    E   2.0000000 154.0000000   1.2903226   99.3548387
##    K   1.0000000 155.0000000   0.6451613  100.0000000
```

```
axis(side = 2, at = seq(0,40,10), labels = paste0(seq(0,40,10), "%"))
```



Problem II:

Fifty soft drink bottles of a specific brand are collected from one day production and measured its net weight, which are given below: The specification limits for this brand are (16+0:5oz) 15.8 16.3 16.2 16.1 16.6 16.3 15.9 15.9 16.2 16.4 16.1 16.2 16.5 16.4 16.3 16.3 16.2 15.9 16.4 16.2 16.1 16.1 16.4 16.5 16.0 16.1 15.8 16.7 16.6 16.4 16.1 16.3 16.5 16.1 16.5 16.2 16.1 16.2 16.1 16.3 16.3 16.2 16.4 16.3 16.5 16.6 16.3 16.4 16.1 16.5

1. Estimate the mean and standard deviation
2. Draw a histogram with superimposing the specification limits. Interpret the histogram focussing on how to improve the process.
3. Draw normal probability plot to justify your answer in (2)
4. Assuming the normality, estimate the sigma level of the process.
5. Estimate the percentage of soft drink bottle out side the lower specification and upper specification.

6. Since the soft drink bottles fell outside the upper specification are quite large, it is decided to lower the process mean setting by 0.2 units. If you are the production manager, how many bottles of drink you need to produce to get 10000 accepted bottles, with new process mean setting, but same standard deviation.

Problem II(1)

1. Estimate the mean and standard deviation

```
net_weights<- c(15.8, 16.3, 16.2, 16.1, 16.6, 16.3, 15.9, 15.9, 16.2, 16.4, 16.1, 16.2, 16.5, 16.4, 16.3
net_weights
```

```
##  [1] 15.8 16.3 16.2 16.1 16.6 16.3 15.9 15.9 16.2 16.4 16.1 16.2 16.5 16.4 16.3
## [16] 16.3 16.2 15.9 16.4 16.2 16.1 16.1 16.4 16.5 16.0 16.1 15.8 16.7 16.6 16.4
## [31] 16.1 16.3 16.5 16.1 16.5 16.2 16.1 16.2 16.1 16.3 16.3 16.2 16.4 16.3 16.5
## [46] 16.6 16.3 16.4 16.1 16.5
```

```
df<- data.frame(net_weights)
count(df)
```

```
##    n
## 1 50
```

Mean

```
sample_mean<- mean(net_weights)
cat("Sample mean:", round(sample_mean, 2), "\n")
```

```
## Sample mean: 16.26
```

Standard Deviation

```
sample_sd<- sd(net_weights)
cat("Sample Standard:", round(sample_sd, 2), "\n")
```

```
## Sample Standard: 0.21
```

Problem II(2)

2. Draw a histogram with superimposing the specification limits. Interpret the histogram focusing on how to improve the process.

Histogram

The specification limits for this brand are (16+0.5oz). So that the Upper limit is 16.5 and lower limit is 15.5

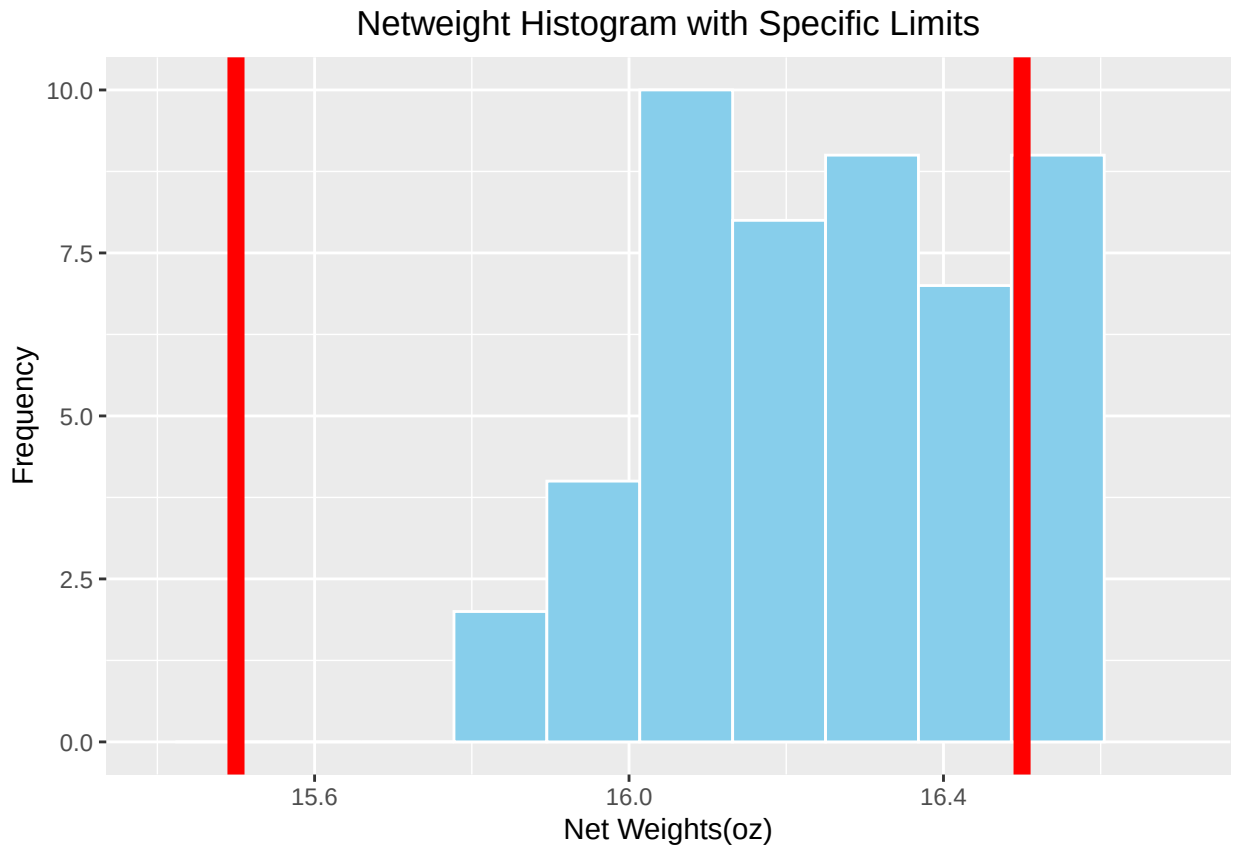
```
upper_limit<- 16.5
lower_limit<- 15.5
# Basic histogram
P<- ggplot(df, aes(x=net_weights), y = frequency) +
  geom_histogram(bins = 12, fill = "skyblue", color = "white") +
  ggtitle("Netweight Histogram with Specific Limits") +
  xlab("Net Weights(oz)") + ylab("Frequency") +
  xlim(c(15.4, 16.7)) +
  geom_vline(xintercept = upper_limit, col = "red", lwd = 3)+
```

```
  geom_vline(xintercept = lower_limit, col = "red", lwd = 3)+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
P
```

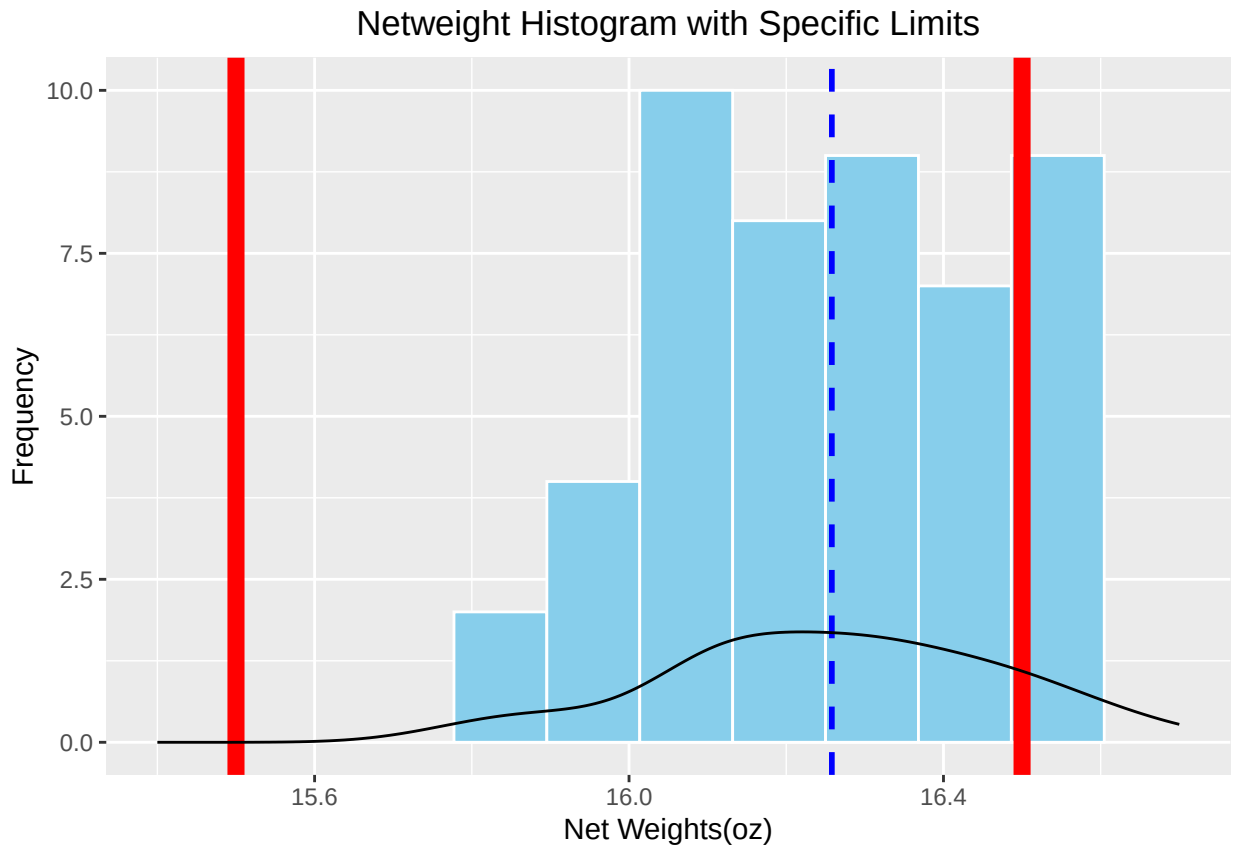
```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



Modified histogram with density curve and mean value line

```
P+geom_vline(aes(xintercept=mean(net_weights)),
             color="blue", linetype="dashed", size=1) + geom_density(color = "black")
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



Interpretation

From the above histogram we can see that there are several bottle drinks with net weights falling outside of the specified limits and somewhat normal with a mean value around 16.3oz. In fact, there are several bottles with net weights above the upper specification limit of 16.5 oz. This suggests that there is a problem with the process that is leading to some bottles being filled with more soda than desired.

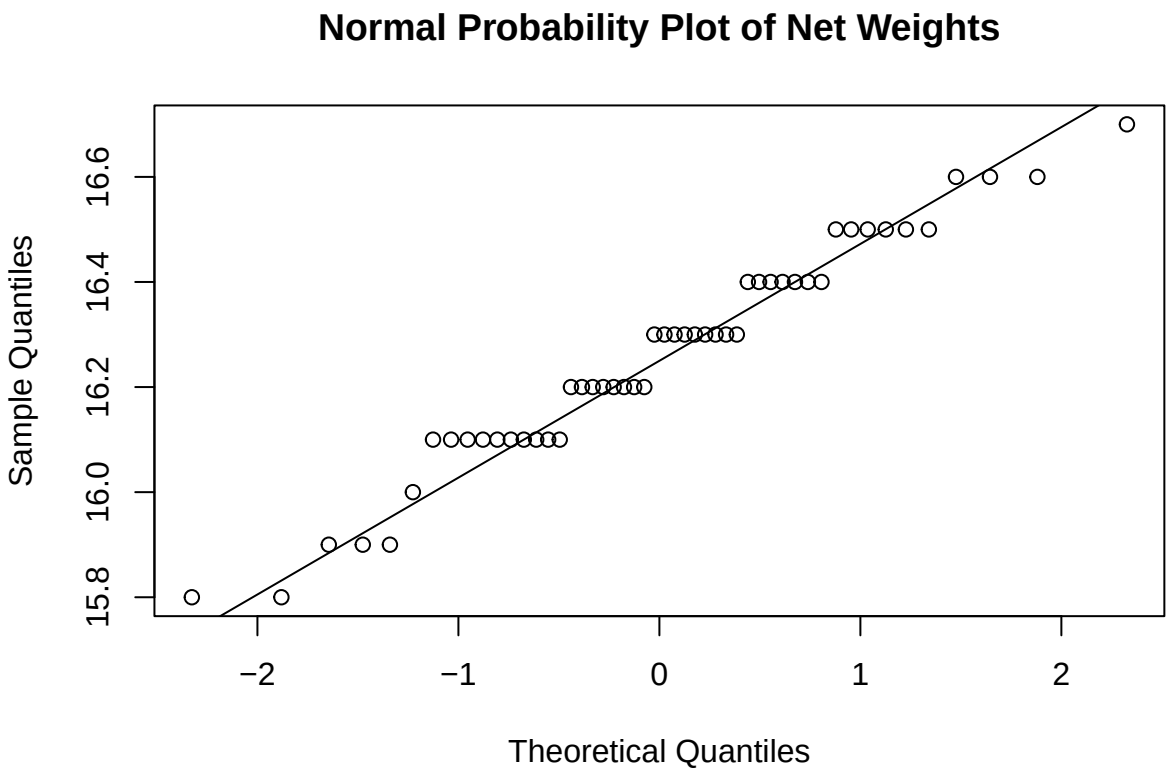
To improve the process, we could investigate the cause of the overfilling and make changes to reduce the variation in the filling process. This might involve making adjustments to the filling equipment or implementing additional quality control measures to ensure that the correct amount of soda is dispensed into each bottle.

Problem II(3)

3. Draw normal probability plot to justify your answer in (2)

Probability Plot

```
qqnorm(net_weights, main = "Normal Probability Plot of Net Weights")
qqline(net_weights)
```



Interpretation From the above normal probability plot we can see that the points form an approximately straight line. Hence the data is normally distributed.

Problem II(4)

4. Assuming the normality, estimate the sigma level of the process

Ans: Considering the normality, to estimate the sigma level of the process we can use

$$Z_{ST} = Z_{LT} + 1.5$$

and

Z_{LT} = min {(upper_limit- sample_Mean) / sample_sd, (sample_Mean – lower_limit) / sample_sd }

Here, upper_limit = 16.5, lower_limit = 15.5, sample_sd = 0.21, sample_mean = 16.26

Hence we get

```
Z_LT<- min((16.5-16.26)/0.21, (16.26-15.5)/0.21)
Z_LT
```

```
## [1] 1.142857
```

sigma level

```
Z_ST<- Z_LT+1.5
Z_ST
```

```
## [1] 2.642857
```

Therefore, the estimated sigma level of the process is 2.64, which suggests that the process is capable of producing products within 2.64 standard deviations of the mean on either side of the specification limits.

Problem II(5)

5. Estimate the percentage of soft drink bottle out side the lower specification and upper specification.

```
# calculate z-score for lower specification limit
z_lower <- (15.5 - 16.26) / 0.21

# calculate z-score for upper specification limit
z_upper <- (16.5 - 16.26) / 0.21

# percentage of bottles below the lower specification limit
p_lower <- pnorm(z_lower, lower.tail = TRUE) * 100

# percentage of bottles above the upper specification limit
p_upper <- (1 - pnorm(z_upper, lower.tail = TRUE)) * 100

# print the results
cat("Percentage of bottles below the lower specification limit:", round(p_lower, 2), "%\n")
```

```
## Percentage of bottles below the lower specification limit: 0.01 %
```

```
cat("Percentage of bottles above the upper specification limit:", round(p_upper, 2), "%\n")
```

```
## Percentage of bottles above the upper specification limit: 12.65 %
```

```
# calculate z-score for lower specification limit
z_lower <- ((16.26-15.5) / 0.21)

# calculate z-score for upper specification limit
z_upper <- ((16.5 - 16.26) / 0.21)

z_lower
```

```
## [1] 3.619048
```

```
z_upper
```

```
## [1] 1.142857
```

lower specification and upper specification

```
# percentage of bottles below the lower specification limit
p_lower <- pnorm(z_lower, lower.tail = TRUE)
p_lower * 100
```

```
## [1] 99.98522
```

```
# percentage of bottles above the upper specification limit
p_upper <- 1 - pnorm(z_upper, lower.tail = TRUE)
p_upper * 100
```

```
## [1] 12.6549
```

Interpretation

The percentage of bottles above the upper specification limit is 12.6549%. This means that approximately 12.66% of the produced bottles are above the upper specification limit of 16.5 oz, which indicates that the process may not be meeting the quality requirements and needs to be improved.

Problem II(6)

6. Since the soft drink bottles fell outside the upper specification are quite large, it is decided to lower the process mean setting by 0.2 units. If you are the production manager, how many bottles of drink you need to produce to get 10000 accepted bottles, with new process mean setting, but same standard deviation.

Ans: Lower the process mean setting by 0.2 units we get the new mean is 16.06 oz. Hence the new upper limit specification will be

```
new_upper_limit <- 16.06 + 3 * 0.21
new_upper_limit
```

```
## [1] 16.69
```

Calculate the z-score for the new upper specification limit:

```
z_new_upper <- (new_upper_limit - 16.06) / 0.21
z_new_upper
```

```
## [1] 3
```

Calculate the percentage of bottles that will be accepted with the new process mean setting:

```
p_new_upper <- pnorm(z_new_upper, lower.tail = TRUE)
p_new_upper * 100
```

```
## [1] 99.86501
```

Number of bottles

Calculate the number of bottles that need to be produced to get 10000 accepted bottles. Let N be the number of bottles produced to get 10000 accepted bottles.

```
N = 10000 / p_new_upper
N
```

```
## [1] 10013.52
```

Problem III:

The manager of the local hospital came to know about the SIX SIGMA methodology and overmuch interested to implement it in his hospital. You are hired to do a sample project as data scientist how to improve customer service quality. Since customer satisfaction is an important parameter for assessment, identify one or two quality characteristics related to the customer satisfaction. Explain how you can measure the quality characteristics you have identified.

#Ans:

As a data scientist, in order to improve customer service quality, we can focus on two important quality characteristics related to customer satisfaction:

Response Time:

It refers to the time taken by the hospital staff to respond to the queries, requests or complaints raised by the customers. A shorter response time indicates better customer service quality.

Communication:

It refers to the quality of communication between the hospital staff and customers. Clear and effective communication is essential to ensure that the customers' needs and expectations are understood and met.

To measure these quality characteristics, we can collect data through customer feedback surveys. The surveys can ask questions related to response time and communication quality, such as:

. How satisfied were you with the response time of hospital staff to your queries/requests/complaints?

. Did the hospital staff communicate clearly and effectively with you during your visit?

. How would you rate the overall communication experience with the hospital staff?

Based on the responses to these questions, we can calculate response time and communication scores for each customer, and then calculate the average scores for all customers. These scores can then be used as the basis for improvement initiatives.

For example, if the response time score is lower than expected, the hospital management can work on improving staff training and processes to reduce response time. Similarly, if the communication score is low, the hospital staff can be trained to improve their communication skills, and communication processes can be revisited and improved.

In summary, by focusing on important quality characteristics related to customer satisfaction, and by collecting and analyzing data through customer feedback surveys, we can identify areas for improvement and take corrective actions to improve customer service quality.

Problem IV

Perform an exploratory data analysis of the "IRIS Data". Give your interpretations / comments on each analysis you performed.

###Ans

EDA

```
iris <- datasets::iris   # Load the IRIS dataset
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
tail(iris)
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## 145          6.7         3.3          5.7         2.5 virginica
## 146          6.7         3.0          5.2         2.3 virginica
## 147          6.3         2.5          5.0         1.9 virginica
## 148          6.5         3.0          5.2         2.0 virginica
```

```
## 149          6.2         3.4          5.4         2.3 virginica
## 150          5.9         3.0          5.1         1.8 virginica
```

```
ncol(iris)
```

```
## [1] 5
```

```
nrow(iris)
```

```
## [1] 150
```

```
colnames(iris)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

```
str(iris)   # Check the structure of the dataset
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```
class(iris)
```

```
## [1] "data.frame"
```

#Comments The output of summary(iris) shows the summary statistics of each variable in the dataset. We can see that the sepal length ranges from 4.3 to 7.9 cm with a mean of 5.84 cm, the sepal width ranges from 2.0 to 4.4 cm with a mean of 3.06 cm, the petal length ranges from 1.0 to 6.9 cm with a mean of 3.76 cm, and the petal width ranges from 0.1 to 2.5 cm with a mean of 1.20 cm.

#Using apply() to calculate a particular statistic for multiple variables at the same time.

```
apply(iris[,1:4], 2, sd)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.8280661    0.4358663    1.7652982    0.7622377
```

#Summary by groups

```
# group mean
aggregate(.~Species, iris, mean)
```

```
##      Species Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     setosa        5.006       3.428        1.462       0.246
## 2 versicolor        5.936       2.770        4.260       1.326
## 3  virginica        6.588       2.974        5.552       2.026
```

```
# group sd
aggregate(.~Species, iris, sd)
```

```
##      Species Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     setosa    0.3524897   0.3790644    0.1736640   0.1053856
## 2 versicolor    0.5161711   0.3137983    0.4699110   0.1977527
## 3  virginica    0.6358796   0.3224966    0.5518947   0.2746501
```

#Table

```
table(iris$Species)
```

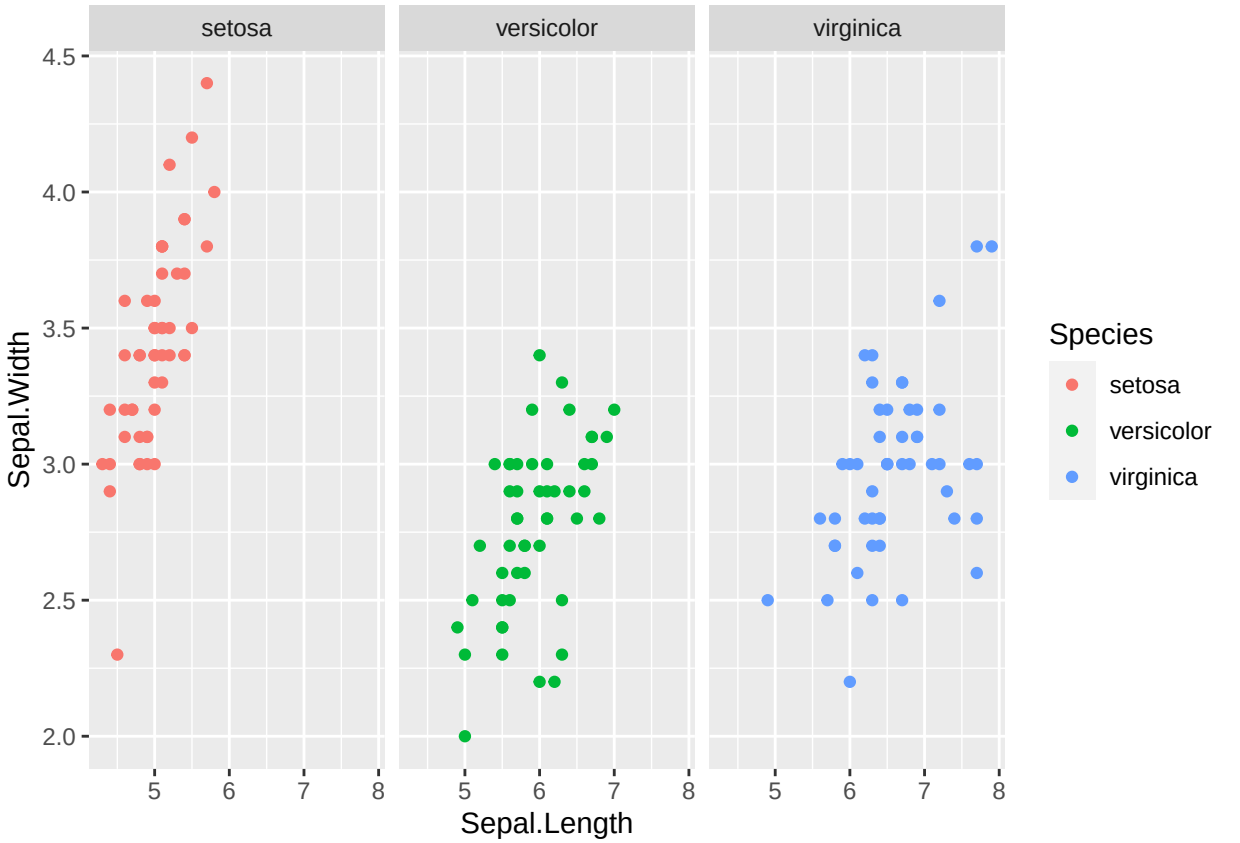
```
##
##     setosa versicolor  virginica
##         50         50         50
```

##Visualization

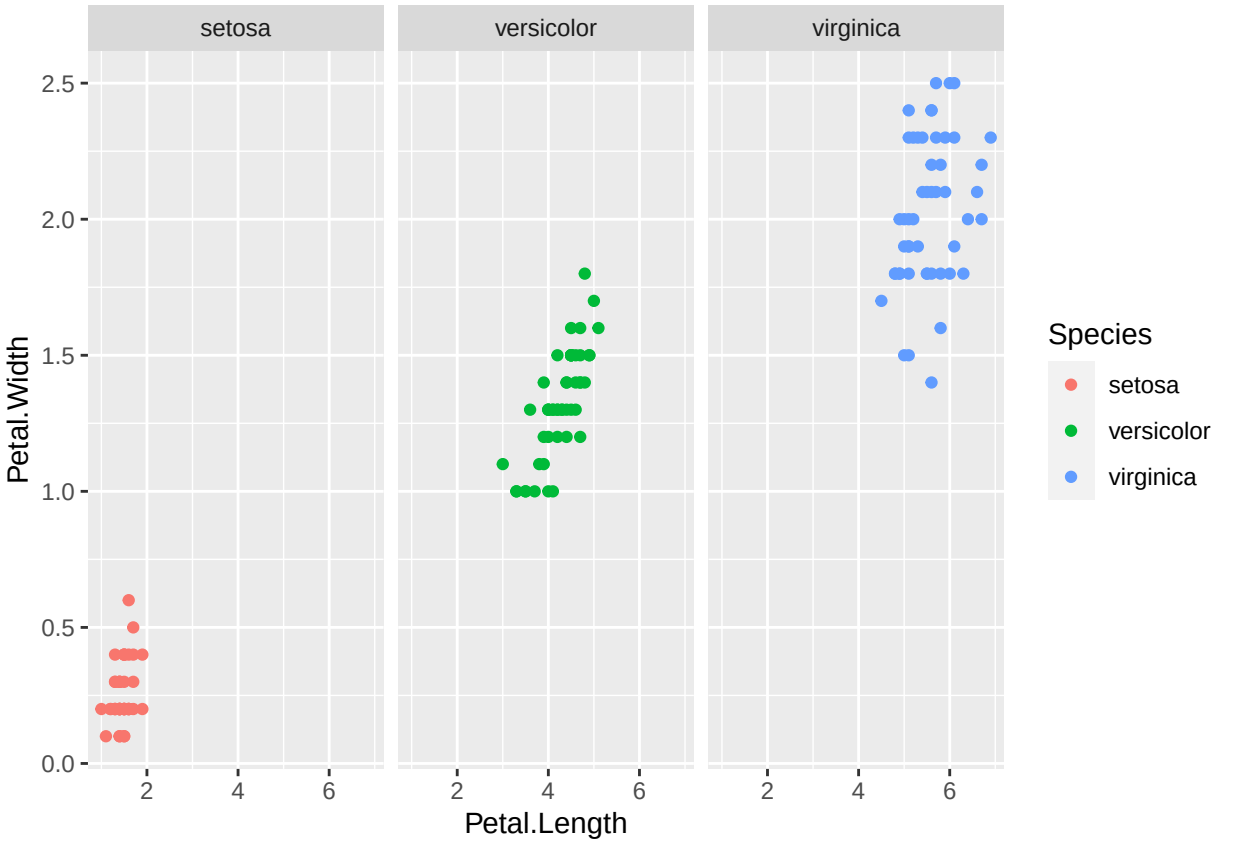
Scatter Plot

#Create a scatterplot matrix of the dataset

```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point() +
  facet_grid(. ~ Species)
```




```
ggplot(iris, aes(x = Petal.Length, y = Petal.Width, color = Species)) +
  geom_point() +
  facet_grid(. ~ Species)
```



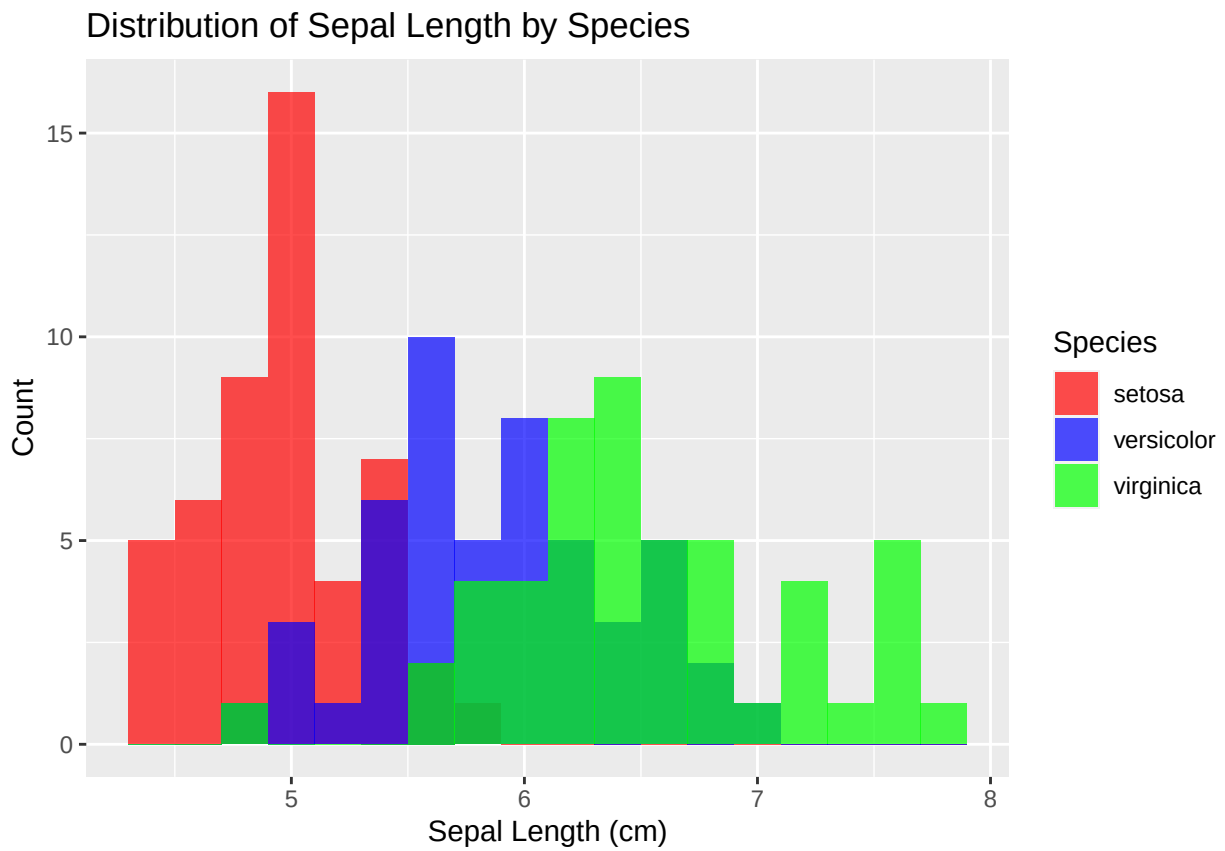
Comments

There seems to be a positive correlation between the length and width of all the species, however there is a distinguishing strong correlation and relationship between petal length and petal width.

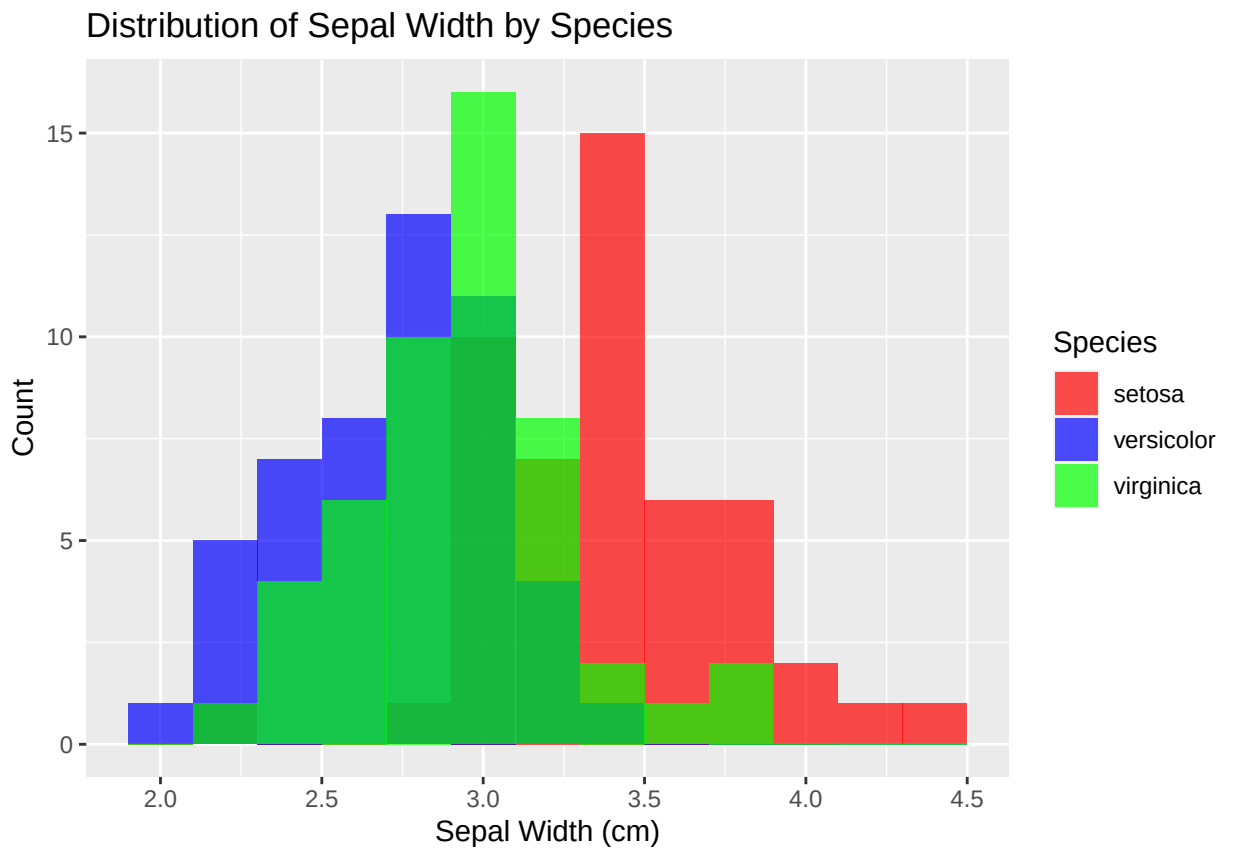
Histogram

#Create Histogram to see the distribution of species

```
ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
  geom_histogram(binwidth = 0.2, alpha = 0.7, position = "identity") +
  labs(title = "Distribution of Sepal Length by Species", x = "Sepal Length (cm)", y = "Count") +
  scale_fill_manual(values = c("red", "blue", "green"))
```




```
ggplot(iris, aes(x = Sepal.Width, fill = Species)) +
  geom_histogram(binwidth = 0.2, alpha = 0.7, position = "identity") +
  labs(title = "Distribution of Sepal Width by Species", x = "Sepal Width (cm)", y = "Count") +
  scale_fill_manual(values = c("red", "blue", "green"))
```




```
ggplot(iris, aes(x = Petal.Length, fill = Species)) +
  geom_histogram(binwidth = 0.2, alpha = 0.7, position = "identity") +
  labs(title = "Distribution of Petal Length by Species", x = "Petal Length (cm)", y = "Count") +
  scale_fill_manual(values = c("red", "blue", "green"))
```


```
ggplot(iris, aes(x = Petal.Width, fill = Species)) +
  geom_histogram(binwidth = 0.2, alpha = 0.7, position = "identity") +
  labs(title = "Distribution of Petal Width by Species", x = "Petal Width (cm)", y = "Count") +
  scale_fill_manual(values = c("red", "blue", "green"))
```

Comments

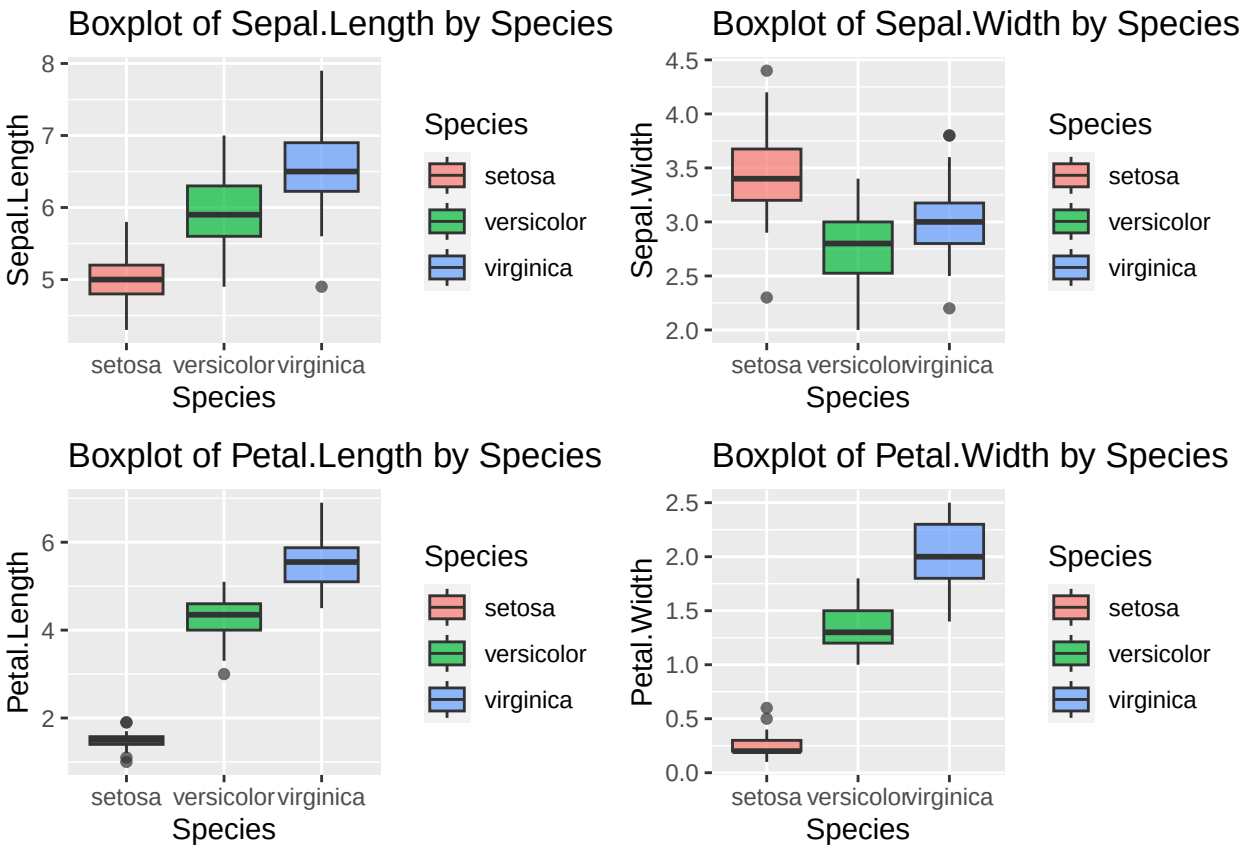
1. The distribution of iris-setosa petal is completely different from the other 2 species.
2. Using sepal length and sepal width, we can't separate one species from another as the distribution is overlapping.
3. iris-setosa is not normally distributed by sepal length and petal width.
4. Petal length can be used as a differentiating factor in terms of the distribution of the 3 flower species.

Boxplot

#Create a function to generate boxplots for each variable

```
graph <- function(y) {
  ggplot(iris, aes(x = Species, y = .data[[y]], fill = Species)) +
    geom_boxplot(alpha = 0.7) +
    labs(title = paste("Boxplot of", y, "by Species"), x = "Species", y = y)
}

# create a 2x2 grid of subplots for each variable
grid.arrange(
  graph("Sepal.Length"),
  graph("Sepal.Width"),
  graph("Petal.Length"),
  graph("Petal.Width"),
  ncol = 2
)
```



Comments

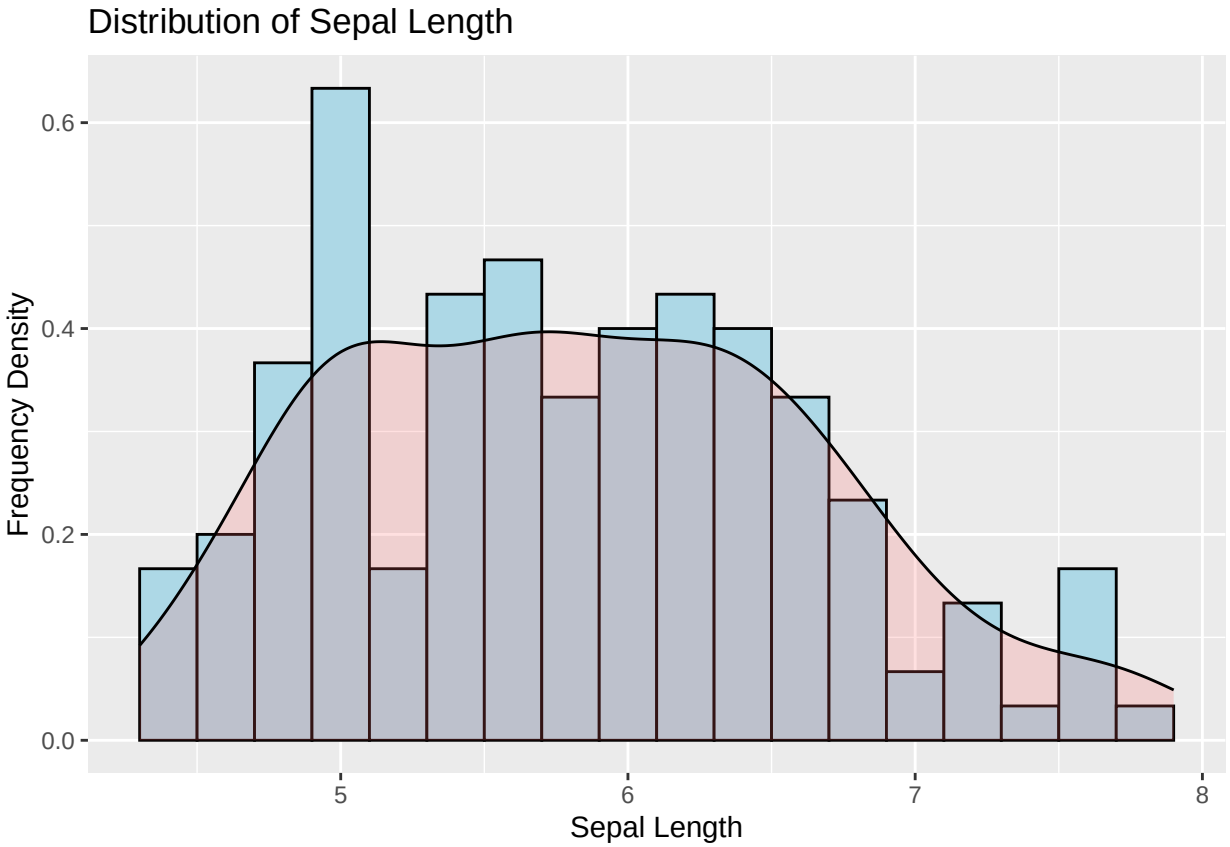
From the above box plots it's clear sighted that the Sepal Length for virginica and Sepal Width of Setosa both have outliers (They are the dots that out run the whiskers). While all the other boxplots looked perfectly balanced, we can see that that petal width for both setosa and versicolor are positively skewed as the median lie at the lower end of the boxplot.

Histogram with Density

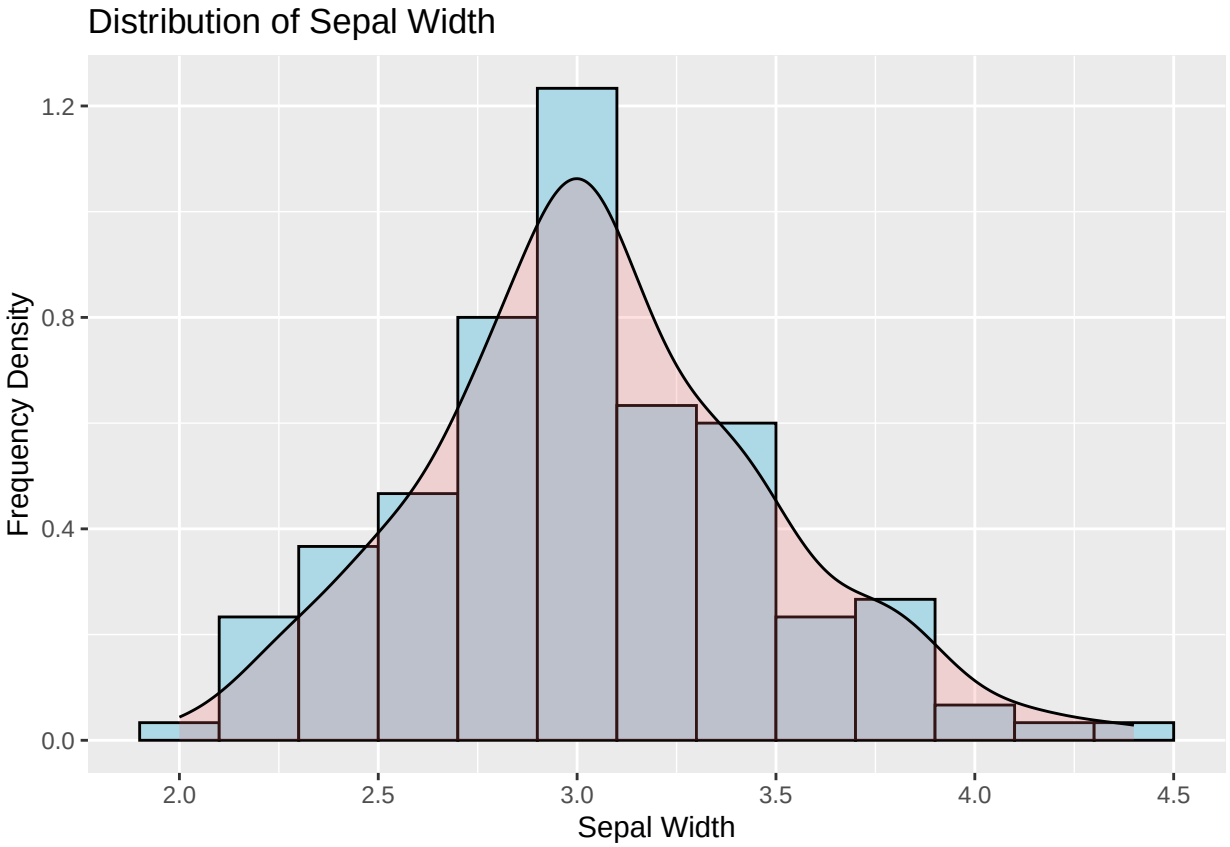
#Create a histogram of variable with density curve in the IRIS dataset

```
ggplot(iris, aes(x = Sepal.Length)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, color = "black", fill = "lightblue") +
  geom_density(alpha = .2, fill = "#FF6666") +
  labs(title = "Distribution of Sepal Length", x = "Sepal Length", y = "Frequency Density")
```

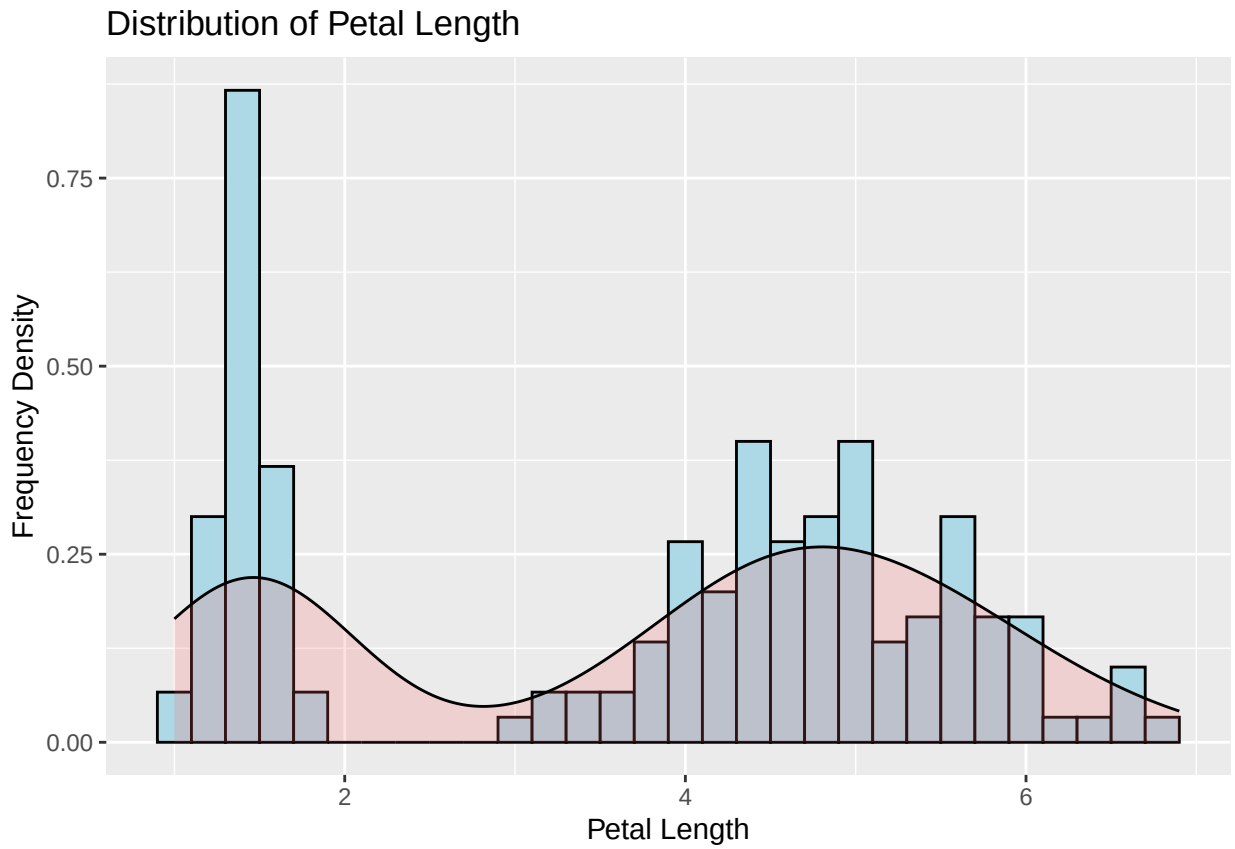
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```



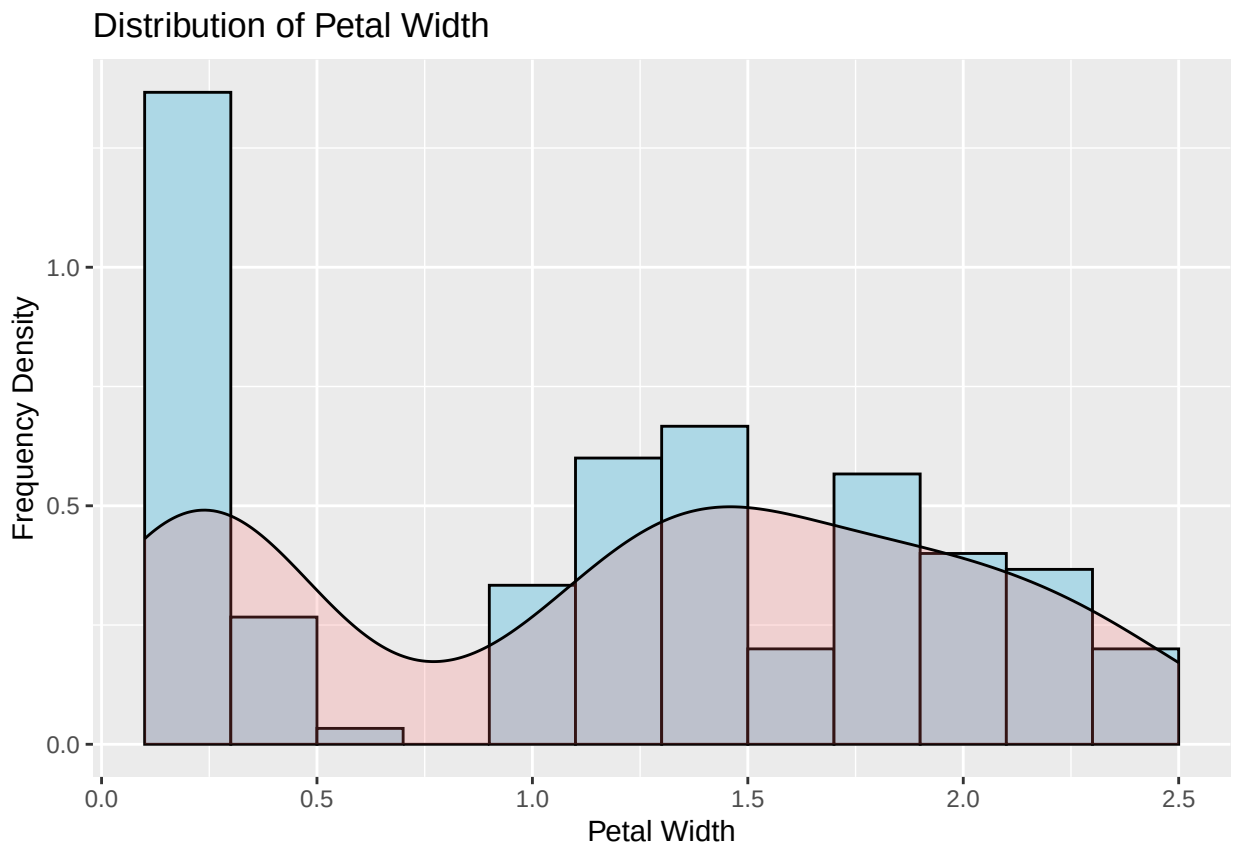

```
ggplot(iris, aes(x = Sepal.Width)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, color = "black", fill = "lightblue") +
  geom_density(alpha = .2, fill = "#FF6666") +
  labs(title = "Distribution of Sepal Width", x = "Sepal Width", y = "Frequency Density")
```




```
ggplot(iris, aes(x = Petal.Length)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, color = "black", fill = "lightblue") +
  geom_density(alpha = .2, fill = "#FF6666") +
  labs(title = "Distribution of Petal Length", x = "Petal Length", y = "Frequency Density")
```




```
ggplot(iris, aes(x = Petal.Width)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, color = "black", fill = "lightblue") +
  geom_density(alpha = .2, fill = "#FF6666") +
  labs(title = "Distribution of Petal Width", x = "Petal Width", y = "Frequency Density")
```



Comment

Looking at the overall distribution, petal length and petal width does not have a normal distribution, whereas sepal length and sepal width are uniformly distributed.

#One way ANOVA test

ANOVA

```
# perform one-way ANOVA test
anova_result <- aov(Sepal.Length ~ Species, data = iris)

# summary of the ANOVA test
summary(anova_result)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment

The test resulted in a very small p-value (less than 2e-16), which indicates strong evidence against the null hypothesis. Therefore, we can conclude that there is a significant difference in Sepal Lengths among the three species in the Iris dataset.

```r
# Split the iris dataset into two samples: first 50 observations for set 1, and next 50 observations fo

set1 <- iris[1:50, 1:4]
set2 <- iris[51:100, 1:4]

# Calculate the T-squared test statistic and associated p-value
result <- HotellingsT2(set1, set2)
result
```

```
##
##  Hotelling's two sample T2-test
##
## data:  set1 and set2
## T.2 = 625.46, df1 = 4, df2 = 95, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0,0)
```

```r
# Print the results
cat("Hotelling's T-squared test statistic:", result$T2, "\n")
```

```
## Hotelling's T-squared test statistic:
```

```r
cat("p-value:", result$p.value, "\n")
```

```
## p-value: 0
```

Comment

The null hypothesis is that the mean vectors of the two samples are equal. If the p-value is less than the significance level (e.g., 0.05), we reject the null hypothesis and conclude that the mean vectors are significantly different. p-value 0 indicates that we reject the null hypothesis and conclude that there is strong evidence of a difference between the means of the two groups being compared. However, it is important to consider the specific context and assumptions of the test before drawing conclusions.

#Multivariate version of ANOVA - MANOVA

MANOVA

The main goal of MANOVA is to determine whether there are significant differences between the means of two or more groups on a combination of dependent variables.

```r
# select two variables and categorical variable from iris dataset
Y <- cbind(iris$Sepal.Length, iris$Sepal.Width)
cate <- iris$Species

# perform MANOVA
manova1 <- manova(Y ~ as.factor(cate))

# print summary with Pillai test
summary(manova1, test = "Pillai")
```

```
##                  Df  Pillai approx F num Df den Df    Pr(>F)
## as.factor(cate)   2 0.94531   65.878      4    294 < 2.2e-16 ***
## Residuals       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment

We reject the null hypothesis of equality of means.

Summary

```
summary.aov(manova1)
```

```
##  Response 1 :
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(cate)   2 63.212  31.606  119.26 < 2.2e-16 ***
## Residuals       147 38.956   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response 2 :
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(cate)   2 11.345  5.6725   49.16 < 2.2e-16 ***
## Residuals       147 16.962  0.1154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```