

**CSE 240 Data Science with R**

**STUDENT WORK BOOK**

|  |  |  |
| --- | --- | --- |
| **Name** | **:** | M SIVANT |
| **Unique ID** | **:** | E00119004 |
| **Year** | **:** | II |
| **Quarter** | **:** | Q6 |
| **Department** | **:** | B.Tech CSE (CyS & IoT or AI &ML) |
| **Faculty Name** | **:** | Prof.B.Nirmala or Prof.N.Chiranjeevi |
| **Academic Year** | **:** | 2020-2021 |

**Date: 07-11-2020**

**Questions:**

Perform the following operation:

1. T-Test ( score 1 and score 2) Manual calculation for the T-value Excel sheet R- Program

2. ANOVA Excel sheet R Program

3. Draw the Violin plot

4. Draw the Qplot for 5 Types

5. Demonstrate the skewness and kurtosis using built-in dataset or your own dataset to get the data summary.

**Program:**

# question 1

# T Test

a=c(3,3,3,12,15,16,17,19,23,24,32)

b=c(20,13,13,20,29,32,23,20,25,15,30)

print(t.test(a,b))

cat('\n')

# question 2

# Anova

d1 = c(0,2,3,5,8,10,12)

d2 = c(1,2,3,9,10,10,11)

d3 = c(1,4,5,5,8,9,10)

df = data.frame(d1,d2,d3)

stack\_group=stack(df)

a\_r = aov(values~ind,data=stack\_group)

print(a\_r)

cat('\n')

# question 3

library(ggplot2)

p <- ggplot(head(iris,130), aes(x=Petal.Length, y=Sepal.Width,fill=factor(Species))) +

  geom\_violin()

print(p)

# print(nrow(iris))

# question 4

# print(colnames(iris))

Species=factor(head(iris$Species,200))

Width=head(iris$Sepal.Width,200)

Length=head(iris$Sepal.Length,200)

# histogram

print(qplot(Width,geom="histogram", fill=Species))

# density

print(qplot(Width,geom="density", fill=Species))

# boxplot

print(qplot(Width,Length,geom="boxplot", fill=Species))

# dotplot

print(qplot(Width,Length,geom="dotplot", fill=Species))

# bar

print(qplot(Length,geom="bar", fill=Species))

# question 5

library(e1071)

library(ggplot2movies)

print(colnames(movies))

cat('\n')

# Right Skew

duration = head(movies$r1,20)

print(skewness(duration))

print(kurtosis(duration))

plot(density(duration),main="Right Skew")

cat('\n')

# Left Skew

duration = head(movies$r3,10)

print(skewness(duration))

print(kurtosis(duration))

plot(density(duration),main="Left Skew")

cat('\n')

# Normal Distribution

data=read.csv('D:\\cw\\2nd year\\Data Science with R\\CW\\Day 6\\income.csv')

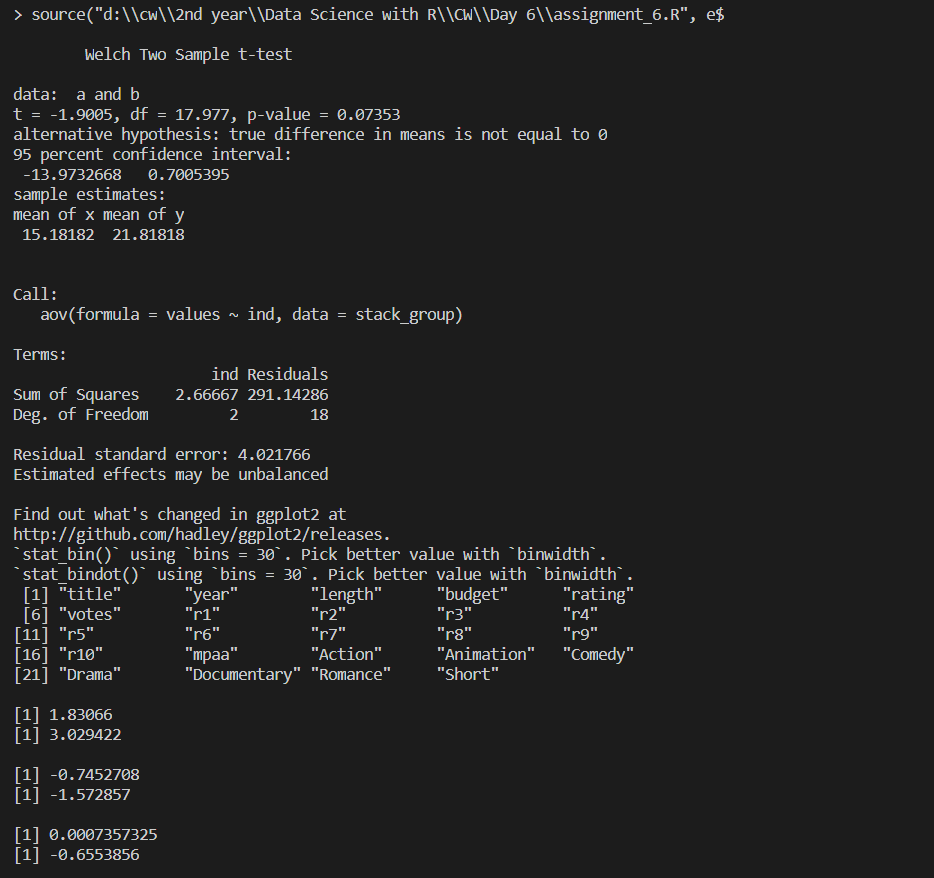
print(skewness(data$amount))

plot(density(data$amount), main="Normal Distribution")

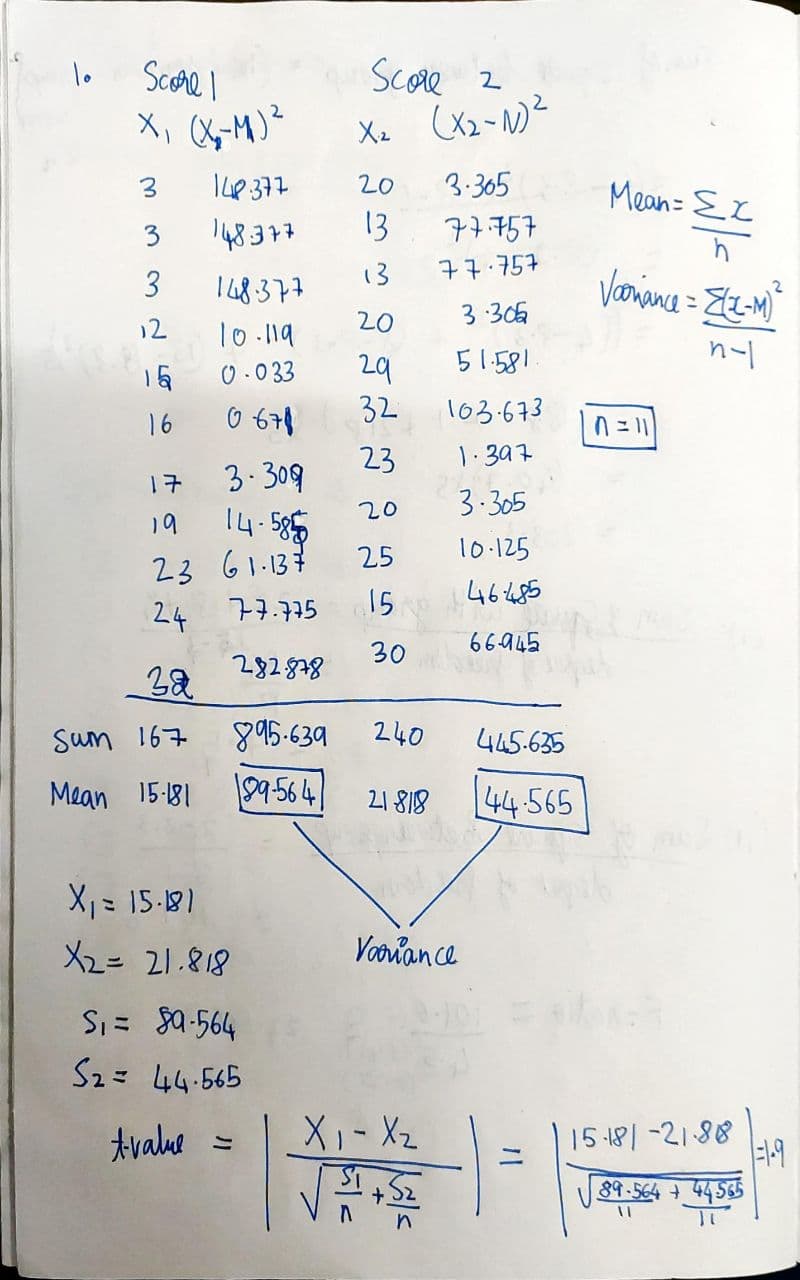
print(kurtosis(data$amount))

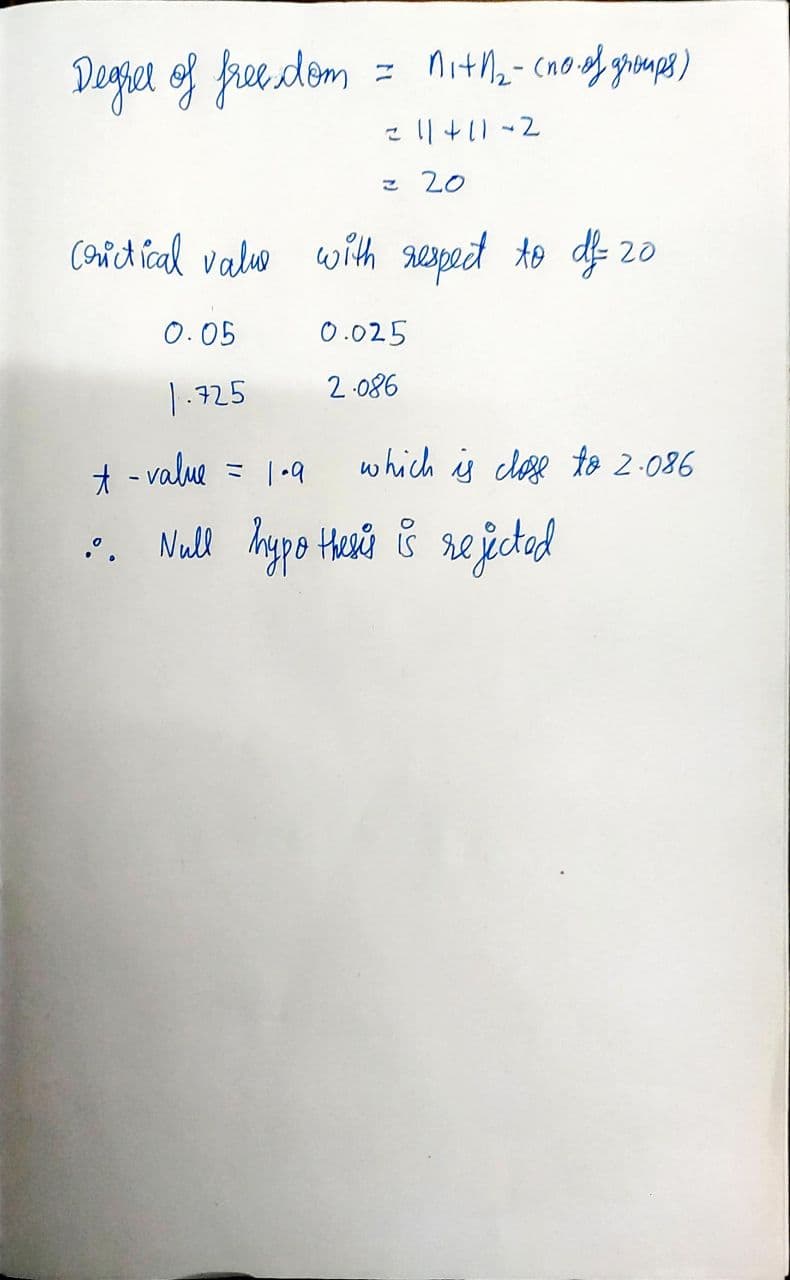
cat('\n')

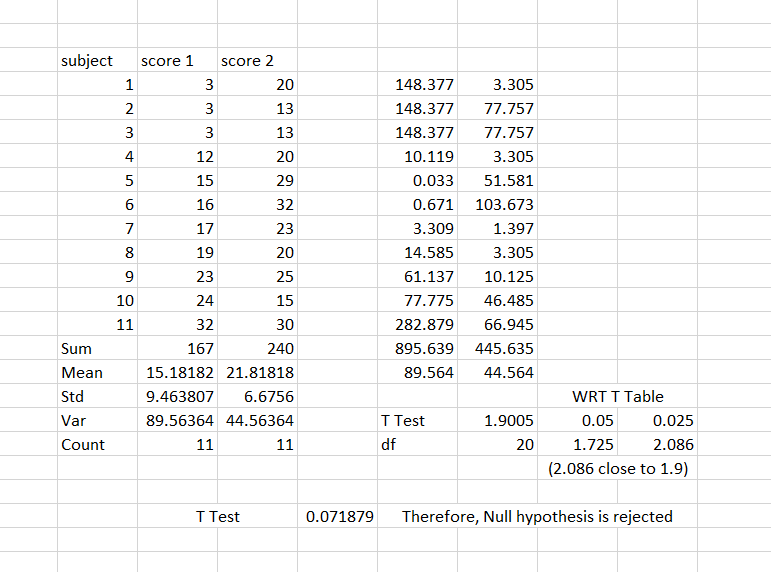
**Output:**



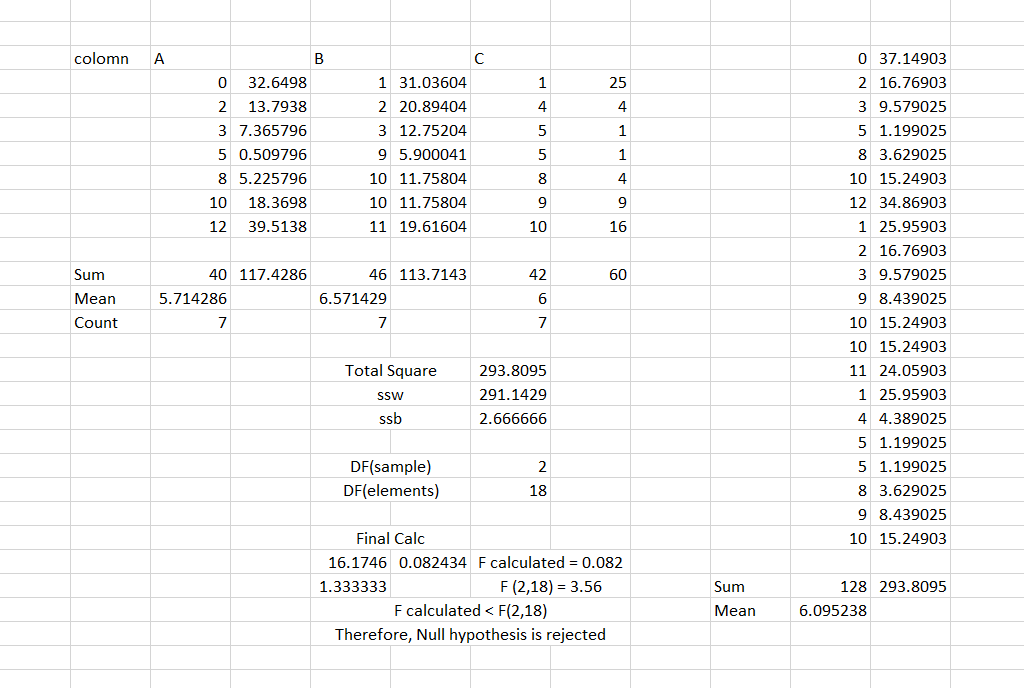
1. T Test:



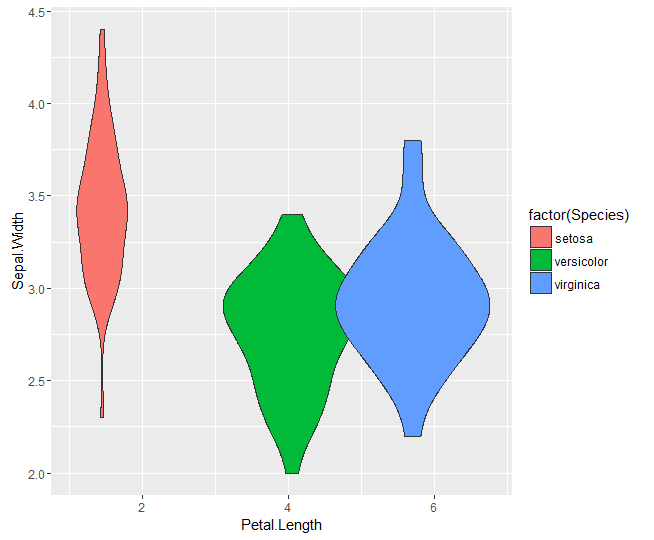




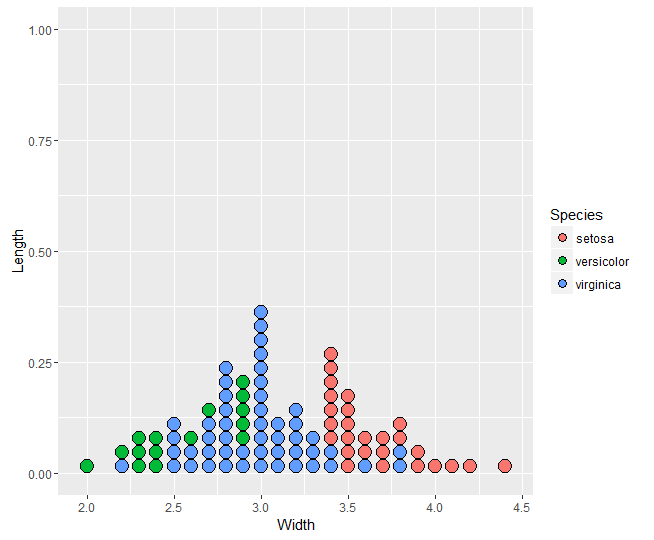
1. Anova



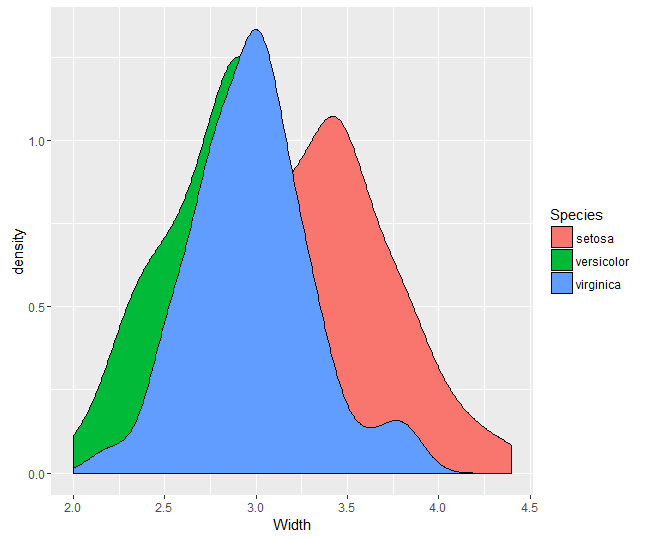
1. Violin Plot:



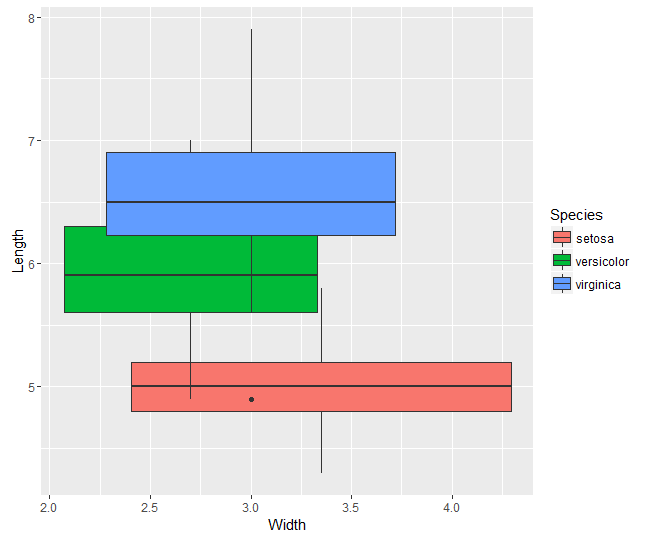
1. Q plot:
   1. Dot Plot



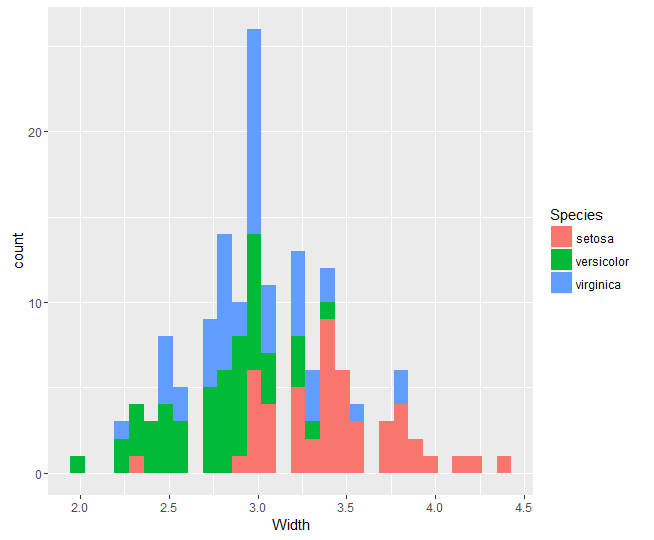
* 1. Density plot:



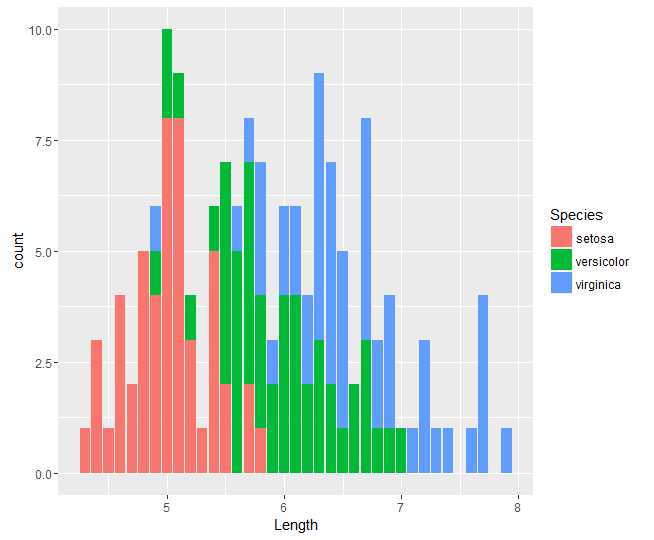
* 1. Bar plot:

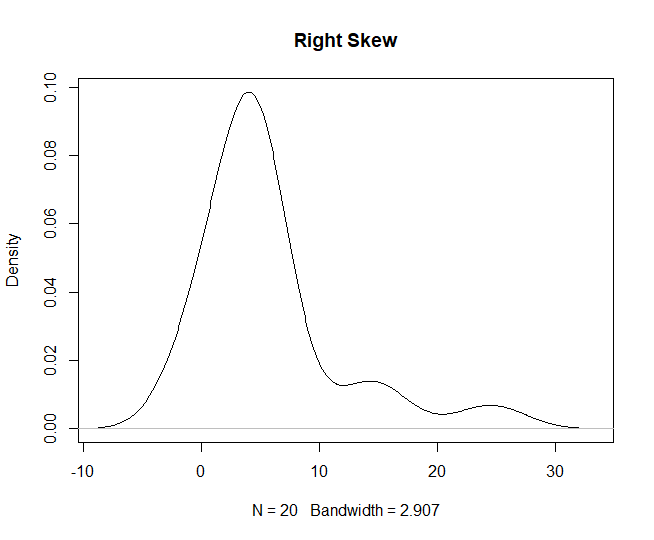
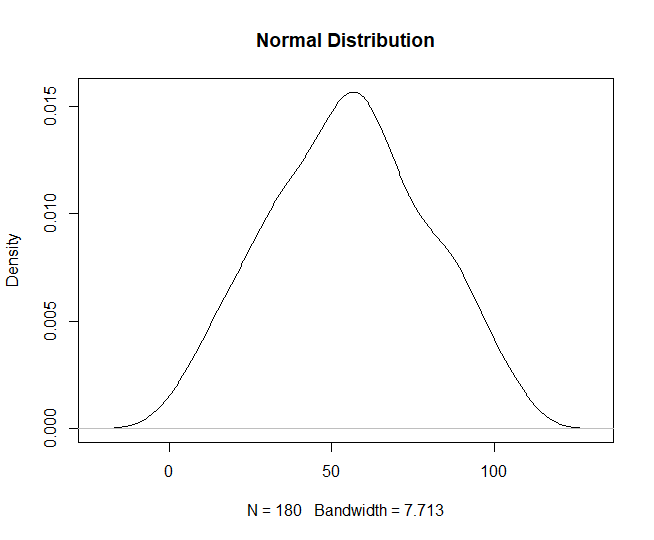


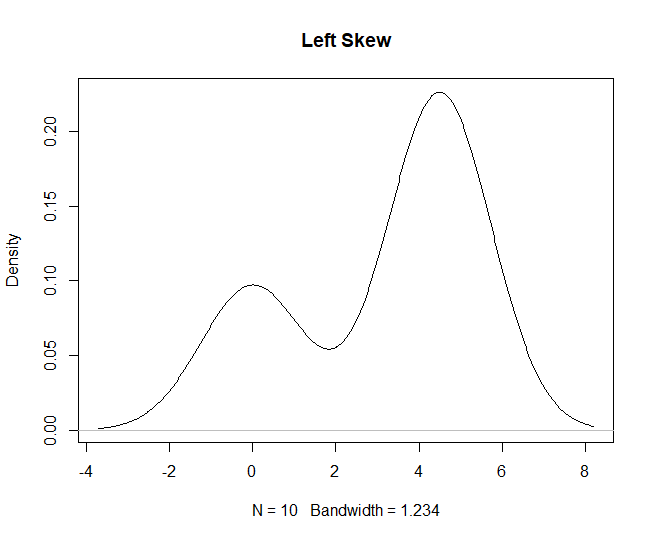
* 1. Density plot:



* 1. Histogram:



1. Skewness and Kurtosis:



**Explanation:**

1. T Test:

A t-test is commonly used to determine whether the mean of a population significantly differs from a specific value (called the hypothesized mean) or from the mean of another population.

1. P Value:

The p-value or probability value is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.

1. Critical Value:

The critical values of a statistical test are the boundaries of the acceptance region of the test.

The p-value is the variable that allows us to reject the null hypothesis (H₀: µ₁=µ₂) or, in other words, to establish that the two groups are different. However, since the p-value is just a value, we need to compare it with the critical value (⍺):

p\_value > ⍺ (Critical value): Fail to reject the null hypothesis of the statistical test.

p\_value ≤ ⍺ (Critical value): Reject the null hypothesis of the statistical test.

The critical value that most statisticians choose is ⍺ = 0.05. This 0.05 means that, if we run the experiment 100 times, 95% of the times we will be able to reject the null hypothesis and 5% we will not.

1. ANOVA:

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.

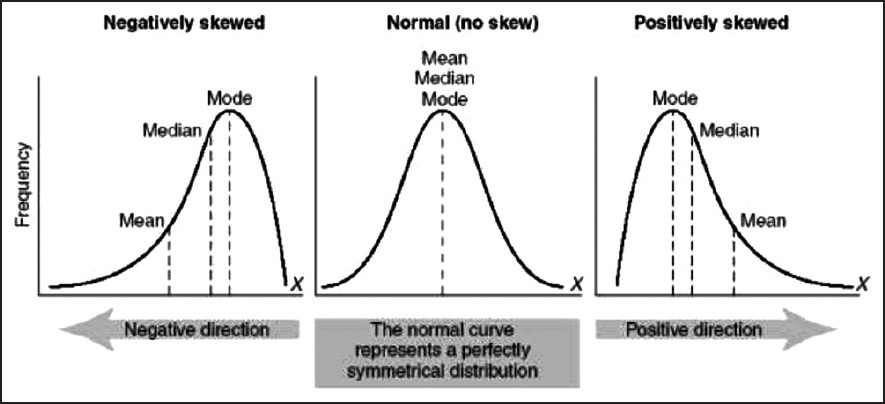
1. ANOVA vs T Test:

A Student’s t-test will tell you if there is a significant variation between groups. A t-test compares means, while the ANOVA compares variances between populations.

You could technically perform a series of t-tests on your data. However, as the groups grow in number, you may end up with a lot of pair comparisons that you need to run. ANOVA will give you a single number (the f-statistic) and one p-value to help you support or reject the null hypothesis.

1. Skewness:

Skewness refers to distortion or asymmetry in a symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. Negatively-skewed distributions are also known as left-skewed distributions.



1. Kurtosis:

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

