COLD STORAGE CASE STUDY

SHALINI KRISHNAN

CONTENTS

1. PROJECT OBJECTIVE

The main objective of this report is to explore **COLD STORAGE** dataset in **R** and derive insights about the same.

The exploration report comprises the following:

- Importing the datasets in R(2 datasets are dealt with)
- Understanding the structure of the dataset
- Graphical exploration
- Descriptive Statistics
- Probability Distribution
- Hypothesis Testing(Statistical Inference)
- Reading insights from the dataset

2. ASSUMPTIONS (COLD STORAGE SAMPLE DISTRIBUTION)

- Unlike population distribution, the sample distribution always follows a normal distribution with increasing sample size (CENTRAL LIMIT THEOREM).

- Sample estimate will be the reflection of the reality.

- Low standard error (sigma/sqrt (n)) in turn leads to low sampling error (xbar-mu) which provides more accuracy in the measurements.
  sigma - sample standard deviation ,n - sample size
  xbar - sample mean, mu - population mean

3. Exploratory Data Analysis-STEP BY STEP APPROACH

A typical Data exploration consists of the following steps:

1. Environmental setup(working directory) and Data import
2. Variable Identification
3. Univariate analysis
4. Bi-variate analysis
5. Missing Values Treatment(not in scope for our project)
6. Outliers Treatment(not in scope for our project)
7. Variable Transformation/Feature Creation (not in scope for our project)
8. Feature Exploration(scope less for the project)

Follow the above mentioned steps for the given dataset.

3.1 ENVIRONMENTAL SETUP AND DATA IMPORT

3.1.1 INSTALL REQUIRED PACKAGES AND INVOKE LIBRARIES

In this section we install all of the necessary packages and the associated libraries are invoked. Having all the required packages at the same place makes the readability of the codes easier. Please refer Appendix A for the Source code.

3.1.2 SET UP THE WORKING DIRECTORY

Setting up the working directory in R helps to import and export data files and code files easy and effective. Basically, working directory is the location where the data and the codes to be worked on is stored.

3.1.3 IMPORT AND READ THE DATASET

The dataset to be worked on is in .csv format. Hence “read.csv” command is used to import the required dataset. Please refer Appendix A for the source code.

3.1.4 INVOKING NECCESARY LIBRARIES

All the required libraries are called to work on the given dataset. Refer Appendix A (source code).

3.2 VARIABLE IDENTIFICATION

The following functions are used to learn the details about the Variables of the given dataset.

names() – displays all the variable names

dim() - provides the dimension of the dataset (rows and columns)

head() – returns the first n rows of the dataset (default: 6 rows)

tail() – returns the last n rows of the dataset (default: 6 rows)

str() – displays the internal structure of the dataset

summary() – produces statistical summaries of the data

anyNA() – tests for any missing values

mean() – calculates the arithmetic mean of the variables

sd() – calculates the standard deviation of the variables

pnorm() – calculates the cumulative distribution function of a normal distribution

t.test() – calculates the t test value, p value for further evaluation of hypothesis testing

hist() – displays a histogram for any given numerical variable

boxplot() – visualizes the correlation between any numerical and categorical variable. It provides the five point summary of the distribution.

3.2.1 VARIABLE IDENTIFICATION INFERENCE

With employing the above functions in R we can simply derive the skeleton and characteristics of the given datasets. COLD STORAGE dataset is a data frame with 365 rows and 4 columns. The variables are of appropriate datatypes. Another dataset MARCH is a data frame with 35 rows and 4 columns.

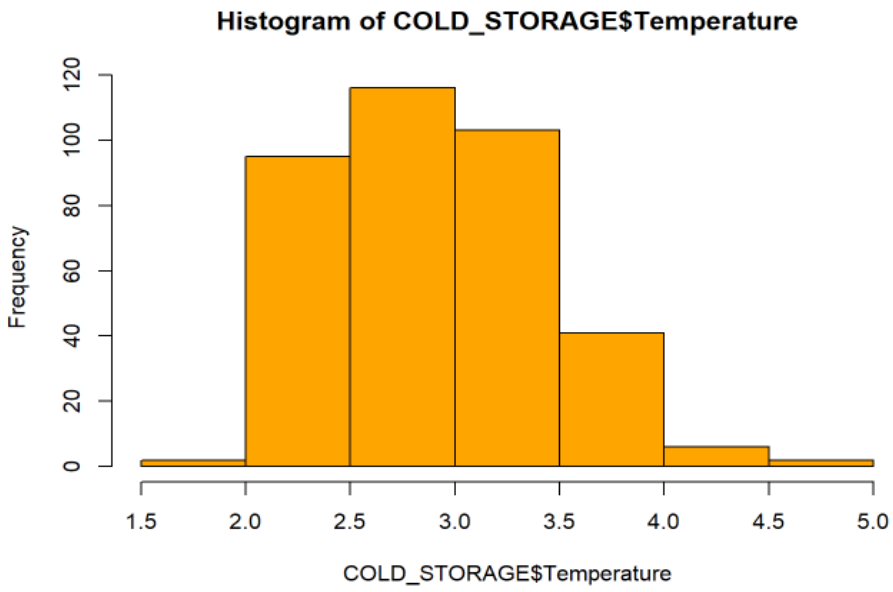
The variable names are Season, Month, Date and Temperature which are common for both the datasets.

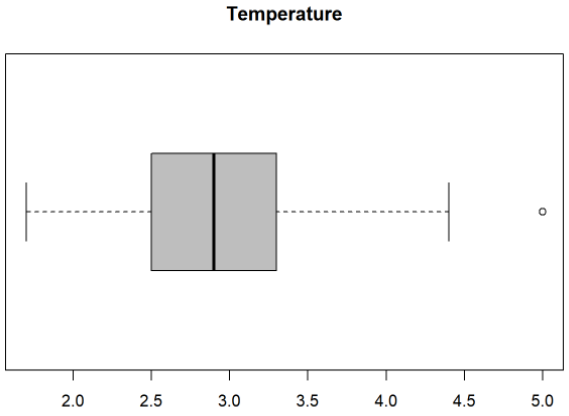
The 5 point summary is generated for the variables which provides the minimum value, maximum value, 1st quadrant, 3rd quadrant, mean and median (refer source code). To calculate the t.test static we need to input the confidence level (1-alpha). The dataset is a combination of factor, integer and numeric variables .All categorical variables are represented by factor datatype in R. Histogram shows the pattern distribution of the Temperature variable whereas the boxplot shows the centre, spread and shape of the variables.

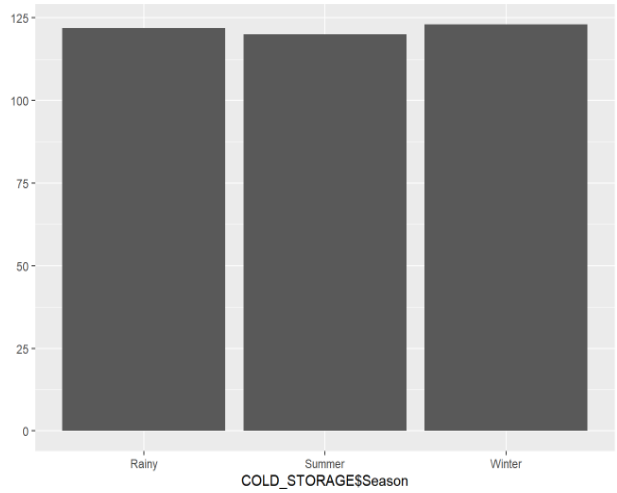
Since it doesn't have any deformities, directly we can pass on the functions to get the desired result.

3.3 UNIVARIATE ANALYSIS

The variable Temperature is the crucial measurement for this analysis. Season and Month are categorical variables. The data is recorded for a year and the same is given along with its months and dates.



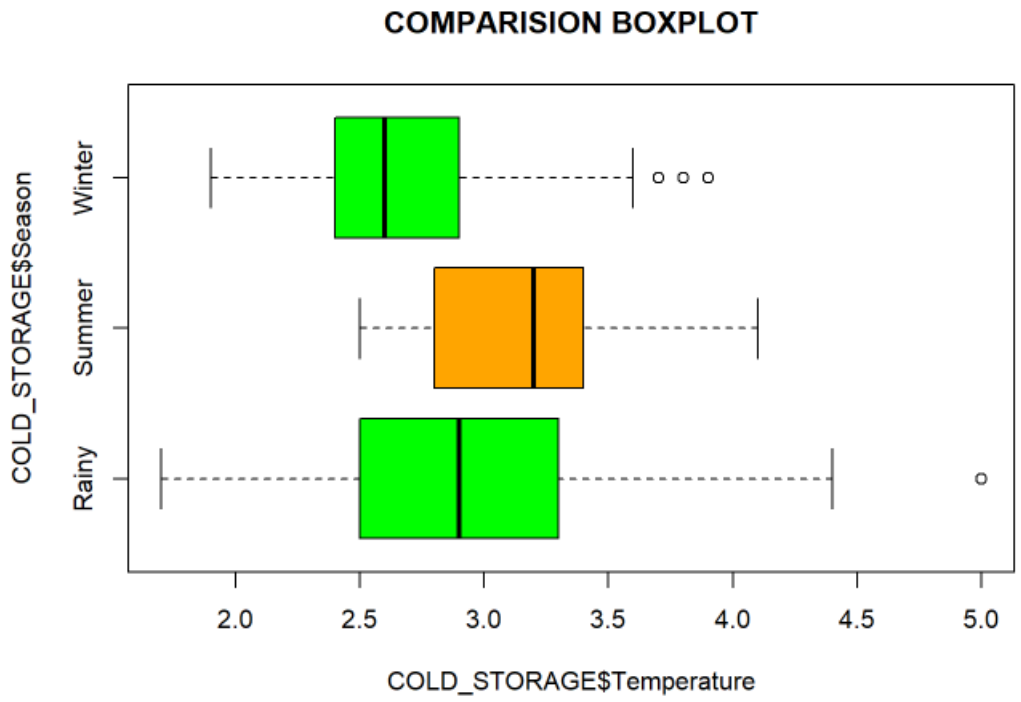




The temperature distribution seems to be slightly right skewed with minimum value 1.5C to maximum value 4.5C.The boxplot shows the presence of an outlier. The plot for the seasons shows that the temperatures are recorded for all the 3 seasons, where the readings of rainy and winter season ascends slightly than of the readings of summer season.

3.4 BI-VARIATE ANALYSIS

The correlation between Temperature and Season is plotted below.



The outlier however is in rainy season. The seasons have most of their temperature readings recorded between 2C - 4.5C.

4. MISSING VALUE IDENTIFICATION

The presence of missing values (NA) can be ascertained.

There are no missing values in the given dataset. Hence it doesn't require any of the missing value treatments.

5. OUTLIER IDENTIFICATION

Outliers are the values which bounds out extremely from the overall pattern of the distribution.

Here '5' is the outlier. However treatment of the outlier is not in scope for this project. It in no way alters our analysis.

6. DESCRIPTIVE STATISTICS

- MEAN-Measure of Central tendency

```
## THE OVERALL MEAN TEMPERATURE IS: 2.9627
```

- STANDARD DEVIATION-Measure of dispersion

```
## THE OVERALL STANDARD DEVIATION OF THE TEMPERATURE IS: 0.5086
```

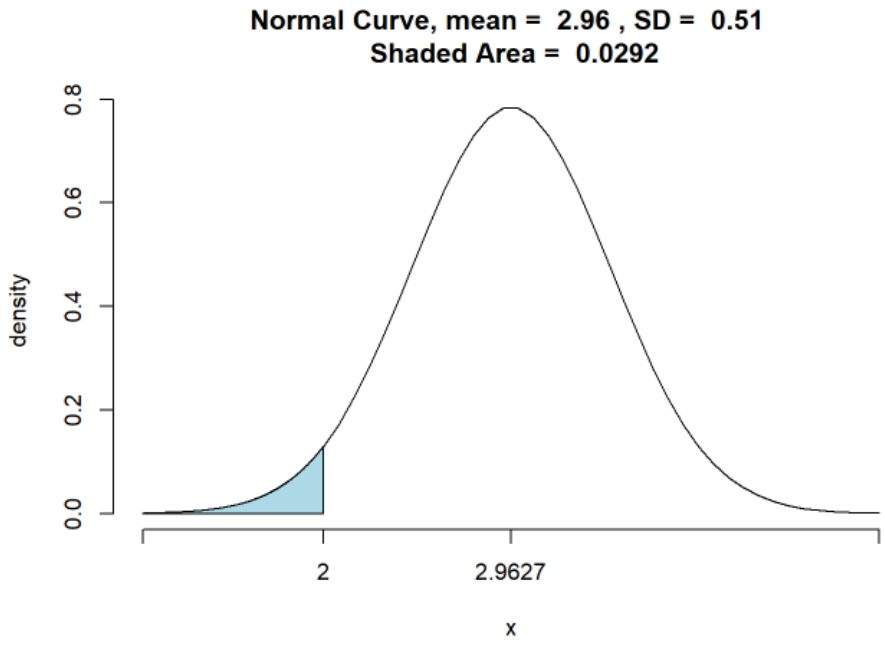
- The mean temperature for all the individual seasons

- ## THE MEAN TEMPERATURE OF SUMMER SEASON IS: 3.1533
- ## THE MEAN TEMPERATURE OF WINTER SEASON IS: 2.7008
- ## THE MEAN TEMPERATURE OF RAINY SEASON IS: 3.0393

7. INFERENTIAL STATISTICS

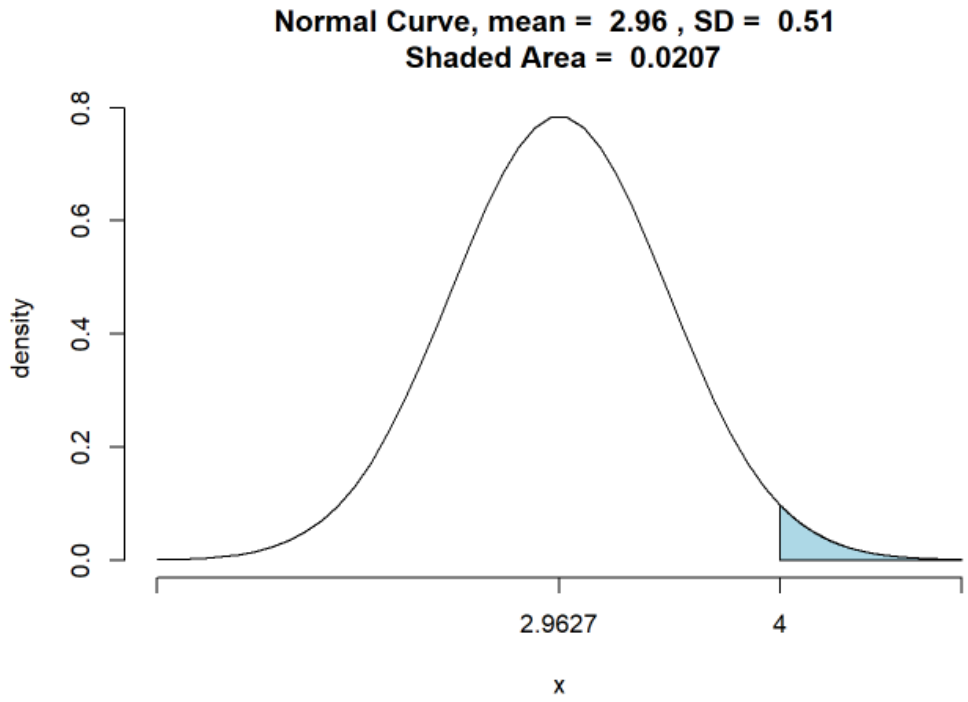
The probability distribution is calculated for values lesser than 2C and greater than 4C temperature. Assuming normal distribution, the probabilities of the sample distribution are calculated as shown below.

1. P(X<2) =0.0292



The graph shows that approximately 2.9% of the values falls under the range of 2C temperature.

2. P(X>4) =0.0207



The graph shows that approximately 2% of the values falls above the range of 4C temperature.

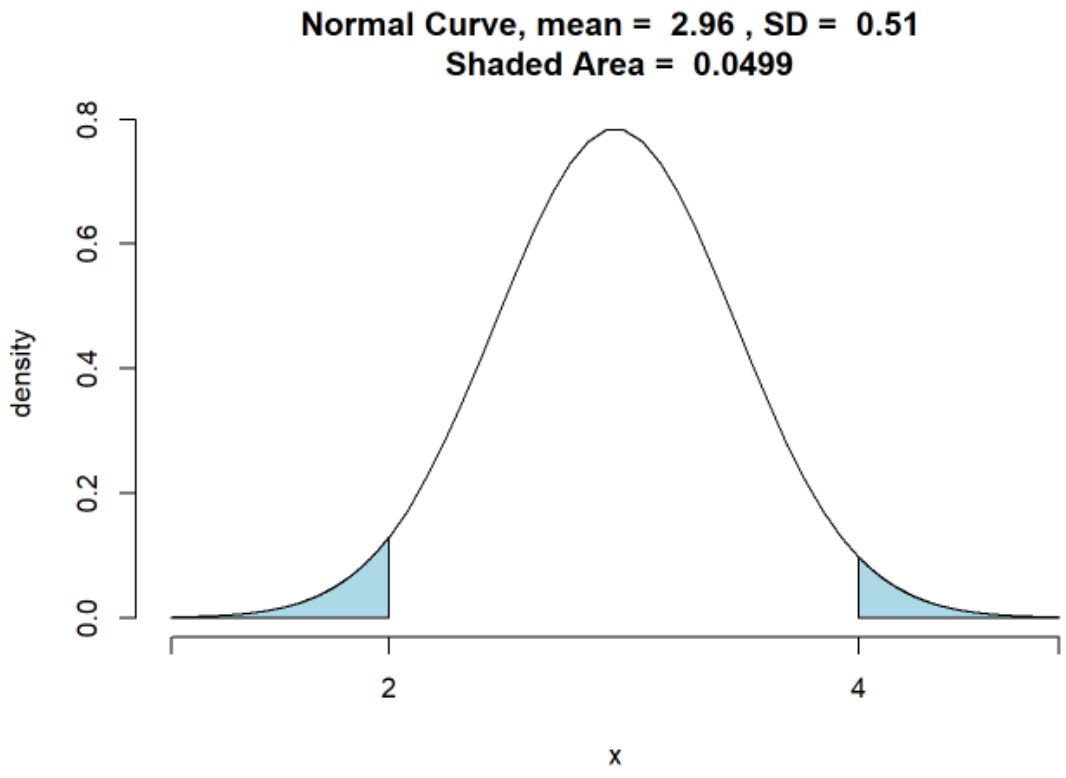
With the given conditions the penalty for the AMC Company could be easily proved. The temperature falling below 2C and above 4C are the values which are outbound and will further lead to penalty constraint.95% values falls within the range (2C-4C) and approximately 5% values fall out of the range.

7.1 PENALTY FOR THE AMC COMPANY

CLAUSE:

- If the temperature values falling under 2C and above 4C is between 2.5%-5% then the penalty for the AMC Company is 10% of the annual maintenance fee.
- If the temperature values are more than 5% then the penalty would be 25% of the annual maintenance fee.

Since the outbound values (4.9%) are lesser than 5%, there is only 10% penalty for the AMC Company.



With these evaluations we conclude with the COLD STORAGE dataset, with further analysis would be done with another dataset MARCH.

8. STATISTICAL HYPOTHESIS TESTING

A statistical hypothesis test is a method of statistical inference. It is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. The 6 step approach for the hypothesis testing are as follows:

1. Specify the Null hypothesis (H0)
2. Specify the alternate hypothesis (H1)
3. Set the significance level (alpha)
4. Check the sample size and decide which test to be performed
5. Calculate the test statistic and corresponding p value
6. Drawing a conclusion

For the given **MARCH** dataset, the hypothesis testing can be done as,

Null hypothesis H0: Mean temperature (mu) = 3.9C

Alternate hypothesis H1: Mean temperature (mu) < 3.9C

Alpha (level of significance): 0.10(specified)

Since it is single variable (Temperature) with unknown population mean and unknown population standard deviation, we perform single tailed t test. The alternative hypothesis is less than mu, so we term it has left tailed t test or t test on the negative side. The formulation for the left tailed t test is,

$$Tstatistic = (Xbar - mu) / (sd / sqrt(n))$$

Xbar-> sample mean=3.974

mu-> population mean (hypothesized) =3.9

sd->population standard deviation=0.1596

n->sample size=35

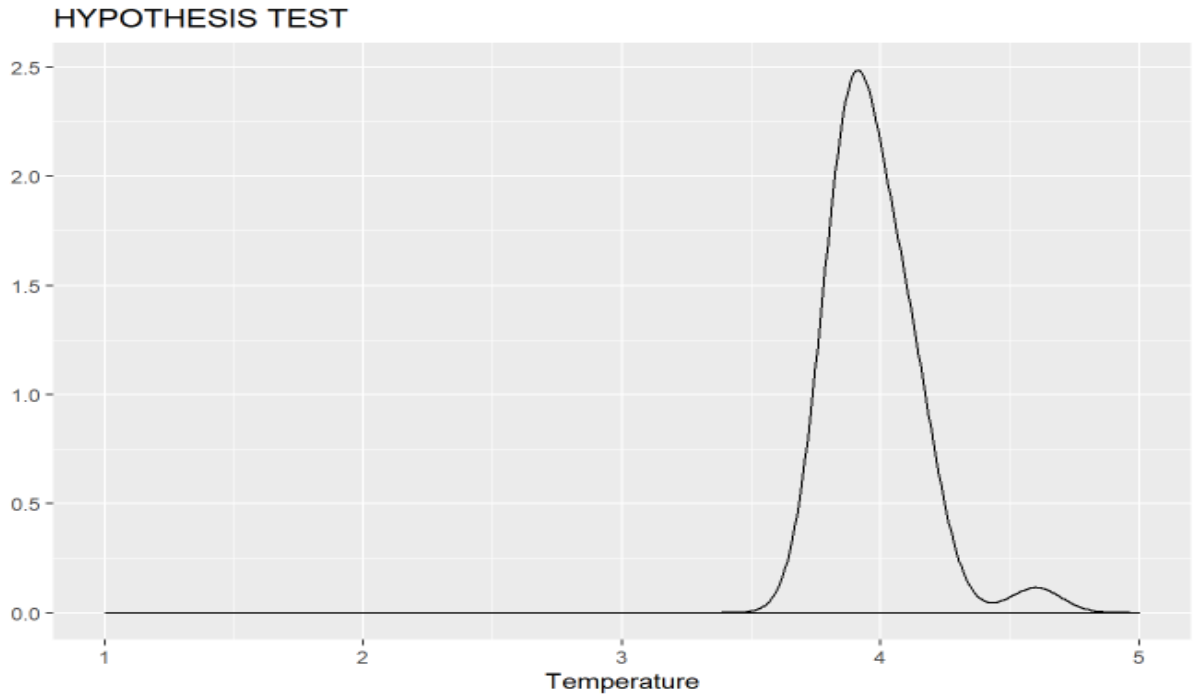
df->degrees of freedom (n-1) =34

Plugging in all the values the test statistic is calculated as

2.7524

p value = 0.9953

Since p value is greater than alpha (0.10), we fail to reject the null hypothesis. We tend to accept mu >= 3.9C. The readings shows that the temperature level maintained in the plant does not meet the expectancy level of the Supervisor (<3.9 C). Hence it is statistically proven that the corrective action is needed in the COLD STORAGE plant.



8.1 INFERENCE

With subject to the above analysis, we come to a conclusion that the mean temperature level maintained in the COLD STORAGE in the month of March is greater than or equal to 3.9C.Though the supervisor was vigilant with the temperature maintained, he has to further notice that it is the month of summer where the atmospheric temperature is very high. So the maintained temperature level (>=3.9C) in the plant may not prove beneficial, which may lead to the products turning sour.

9. CONCLUSION

With the given data we conclude that,

- Statistically the maintained temperature (>=3.9C) may not prove beneficial in humid months like summer.

- If the dairy products turns sour even after maintaining the temperature level to <= 3.9C, then we strictly need to watch the temperature level in the plant for some more days with more number of samples.

- If everything works fine with the maintenance plant but if the customers are not satisfied with the products, the COLD STORAGE supplies, then necessary action can be taken against the vendor, from where all the dairy products are procured.

- The supervisor should be able to solve the problem immediately with the above cause - effect scenarios. If all these things taken into consideration, then no further problems may arise in the customer end.

10. APPENDIX A

SOURCE CODE

```
#############COLD STORAGE DATSET#################
###SET THE WORKING DIRECTORY###
setwd("C:/Training/BACP DECEMBER PROJECT")
###IMPORTING THE COLD STORAGE DATASET###
library(readr)
library(readxl)
library(dplyr)
library(ggplot2)
library(tigerstats)
read.csv("Cold_Storage_Temp_Data.csv")->COLD_STORAGE   ###READING THE DATASET
                                                                    INTO R
names(COLD_STORAGE)####GIVES ALL THE VARIABLE NAMES
## [1] "Season"      "Month"       "Date"        "Temperature"
dim(COLD_STORAGE)###GIVES THE DIMENSION OF THE DATASET
## [1] 365   4
str(COLD_STORAGE)###GIVES THE STRUCTURE OF THE DATASET
## 'data.frame':    365 obs. of  4 variables:
##  $ Season     : Factor w/ 3 levels "Rainy","Summer",..: 3 3 3 3 3 3 3 3 3
3 ...
##  $ Month      : Factor w/ 12 levels "Apr","Aug","Dec",..: 5 5 5 5 5 5 5 5
5 5 ...
##  $ Date       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Temperature: num  2.4 2.3 2.4 2.8 2.5 2.4 2.8 2.3 2.4 2.8 ...
summary(COLD_STORAGE)###GIVES THE 5 POINT SUMMARY
##     Season        Month          Date        Temperature
##  Rainy :122   Aug    : 31   Min.   : 1.00   Min.   :1.700
##  Summer:120   Dec    : 31   1st Qu.: 8.00   1st Qu.:2.500
##  Winter:123   Jan    : 31   Median :16.00   Median :2.900
##               Jul    : 31   Mean   :15.72   Mean   :2.963
##               Mar    : 31   3rd Qu.:23.00   3rd Qu.:3.300
##               May    : 31   Max.   :31.00   Max.   :5.000
```

```
##                (Other):179
```

```
anyNA(COLD_STORAGE)###CHECKS FOR ANY NA VALUES
```

```
## [1] FALSE
```

```
head(COLD_STORAGE)###GIVES FIRST 6 ROWS OF THE DATASET
```

```
##   Season Month Date Temperature
## 1 Winter   Jan    1         2.4
## 2 Winter   Jan    2         2.3
## 3 Winter   Jan    3         2.4
## 4 Winter   Jan    4         2.8
## 5 Winter   Jan    5         2.5
## 6 Winter   Jan    6         2.4
```

```
tail(COLD_STORAGE)###GIVES LAST 6 ROWS OF THE DATASET
```

```
##     Season Month Date Temperature
## 360 Winter   Dec   26         2.7
## 361 Winter   Dec   27         2.7
## 362 Winter   Dec   28         2.3
## 363 Winter   Dec   29         2.6
## 364 Winter   Dec   30         2.3
## 365 Winter   Dec   31         2.9
```

```
boxplot.stats(COLD_STORAGE$Temperature)$out###CHECKS FOR ANY OUTLIERS
```

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

```
###UNIVARIATE ANALYSIS

hist(COLD_STORAGE$Temperature,col="orange")

boxplot(COLD_STORAGE$Temperature,horizontal = TRUE,col="gray",main="Temperatu
re")

qplot(COLD_STORAGE$Season)

####BIVARIATE ANALYSIS

boxplot(COLD_STORAGE$Temperature~COLD_STORAGE$Season,horizontal = TRUE,col=c(
"green","orange"),main="COMPARISION BOXPLOT")

###OVERALL MEAN OF TEMPERATURE

mean(COLD_STORAGE$Temperature)->MEAN

cat("THE OVERALL MEAN TEMPERATURE IS:",round(MEAN,4))

## THE OVERALL MEAN TEMPERATURE IS: 2.9627
```

```
###OVERALL STANDARD DEVIATION OF TEMPERATURE
sd(COLD_STORAGE$Temperature)->S.D.
cat("THE OVERALL STANDARD DEVIATION OF THE TEMPERATURE IS:",round(S.D.,4))
## THE OVERALL STANDARD DEVIATION OF THE TEMPERATURE IS: 0.5086
by(COLD_STORAGE,INDICES = Season,FUN =summary)
## Season: Rainy
##     Season        Month         Date        Temperature
##  Rainy :122   Aug    :31   Min.   : 1.00   Min.   :1.700
##  Summer:  0   Jul    :31   1st Qu.: 8.00   1st Qu.:2.500
##  Winter:  0   Jun    :30   Median :16.00   Median :2.900
##               Sep    :30   Mean   :15.75   Mean   :3.039
##               Apr    : 0   3rd Qu.:23.00   3rd Qu.:3.300
##               Dec    : 0   Max.   :31.00   Max.   :5.000
##               (Other): 0
## ------------------------------------------------------------
## Season: Summer
##     Season        Month         Date        Temperature
##  Rainy :  0   Mar    :31   Min.   : 1.00   Min.   :2.500
##  Summer:120   May    :31   1st Qu.: 8.00   1st Qu.:2.800
##  Winter:  0   Apr    :30   Median :15.50   Median :3.200
##               Feb    :28   Mean   :15.53   Mean   :3.153
##               Aug    : 0   3rd Qu.:23.00   3rd Qu.:3.400
##               Dec    : 0   Max.   :31.00   Max.   :4.100
##               (Other): 0
## ------------------------------------------------------------
## Season: Winter
##     Season        Month         Date        Temperature
##  Rainy :  0   Dec    :31   Min.   : 1.00   Min.   :1.900
##  Summer:  0   Jan    :31   1st Qu.: 8.00   1st Qu.:2.400
##  Winter:123   Oct    :31   Median :16.00   Median :2.600
##               Nov    :30   Mean   :15.88   Mean   :2.701
##               Apr    : 0   3rd Qu.:23.50   3rd Qu.:2.900
##               Aug    : 0   Max.   :31.00   Max.   :3.900
##               (Other): 0
```

```
###IMPORTING THE MARCH DATASET
read.csv("Cold_Storage_Mar2018.csv")->TEMPORARY_SET
summary(TEMPORARY_SET)
##     Season   Month          Date       Temperature
##  Summer:35   Feb:18   Min.   : 1.0   Min.   :3.800
##              Mar:17   1st Qu.: 9.5   1st Qu.:3.900
##                       Median :14.0   Median :3.900
##                       Mean   :14.4   Mean   :3.974
##                       3rd Qu.:19.5   3rd Qu.:4.100
##                       Max.   :28.0   Max.   :4.600
####HYPOTHESIS TESTING
###H0=3.9,H1<3.9
sd(TEMPORARY_SET$Temperature)
## [1] 0.159674
t.test(TEMPORARY_SET$Temperature,mu=3.9,alternative = "less",conf.level = 0.9
)
##
##  One Sample t-test
##
## data:  TEMPORARY_SET$Temperature
## t = 2.7524, df = 34, p-value = 0.9953
## alternative hypothesis: true mean is less than 3.9
## 90 percent confidence interval:
##     -Inf 4.00956
## sample estimates:
## mean of x
##  3.974286

#############END################################################
```