Project 1 - Cold Storage Case Study

Assignment Report

- By Samrat Mallik

Table of Contents

1. Project Objective

The objective of this assignment to explore the datasets “Cold_Storage_Temp_Data” and “Cold_Storage_Mar2018” in R and generate insights about the datasets. The exploration procedure will consist of the following:

- Importing the datasets in R
- Understanding the structure of the datasets
- Graphical exploration
- Statistical evaluation
- Generate meaningful insights from the datasets

We also want to the answer the following questions with regard to the datasets:

Problem 1(Cold_Storage_Temp_Data)

1. Find mean cold storage temperature for Summer, Winter and Rainy Season
2. Find overall mean for the full year
3. Find Standard Deviation for the full year
4. What is the probability of temperature having fallen below 2 degree C?
5. What is the probability of temperature having gone above 4 degree C?
6. What will be the penalty for the AMC Company?

Problem 2(Cold_Storage_Mar2018)

1. State the Hypothesis, do the calculation using z test
2. State the Hypothesis, do the calculation using t test
3. Give your inference after doing both the tests

2. Assumptions

We assume that both the datasets are normally distributed.

Problem 1

3. Exploratory Data Analysis for Problem 1

3.1. Environment Setup and Data Import

3.1.1. Installing Necessary Packages and Invoking Libraries

```
library(ggplot2)

## Registered S3 methods overwritten by 'ggplot2':
##   method         from 
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

3.1.2. Setting Up Working Directory

```
setwd("C:/Users/Sam/Documents/R/Presentations, Datasets and Codes")
getwd()

## [1] "C:/Users/Sam/Documents/R/Presentations, Datasets and Codes"
```

3.1.3. Importing and Reading the Dataset

```
data = read.csv("Cold_Storage_Temp_Data.csv", header = TRUE)
```

3.2. Variable Identification

```
dim(data)

## [1] 365   4

names(data)

## [1] "Season"      "Month"       "Date"        "Temperature"

summary(data)

##     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                                  

str(data)

## '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 ...
```

Inferences: the dataset contains 365 rows and 4 columns. The columns are named “Season”, “Month”, “Date” and “Temperature”. “Season” is a Factor of 3 levels namely “Rainy”, “Summer” and “Winter” having 122, 120 and 123 observations respectively. “Month” is a Factor of 12 levels corresponding to the 12 months in the year 2016 and “Date” signifies the day of each month. “Temperature” is a numeric vector which provides the cold storage temperature for each day of the year. It has a minimum value of 1.7, maximum value of 5.0 and median value of 2.9.

Question 1

3.3. Univariate Analysis

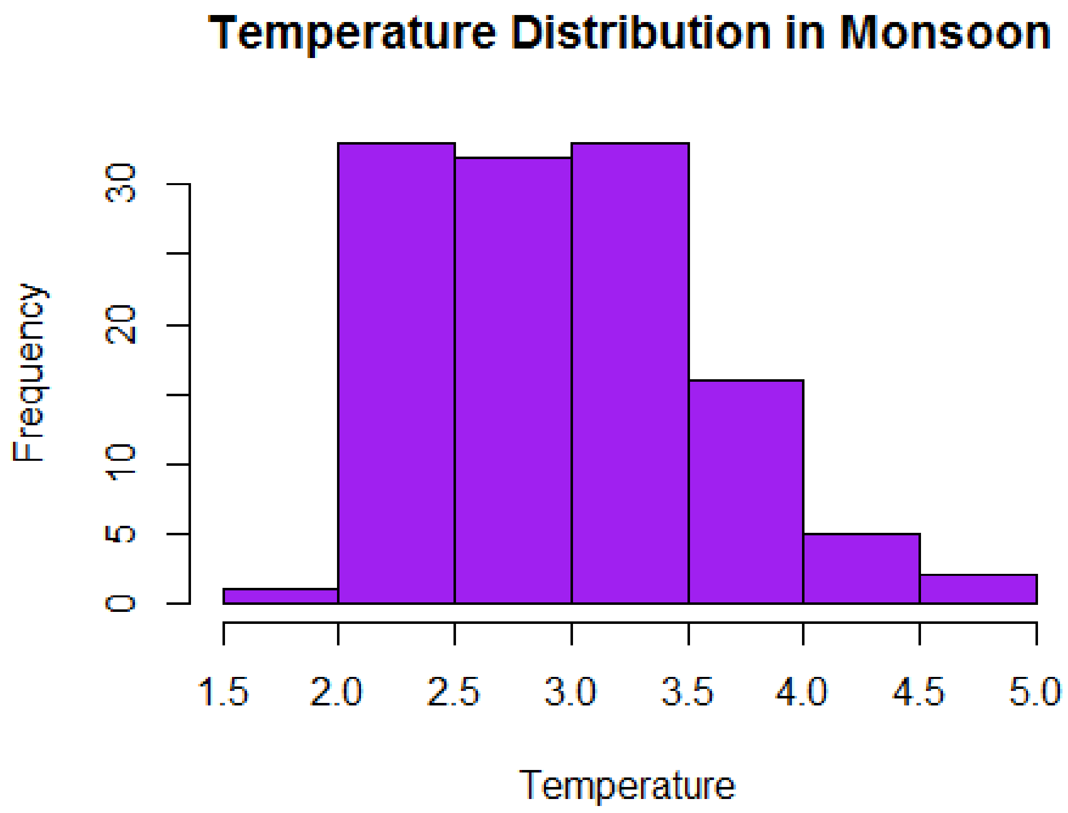
We perform analysis for each season separately to assist our project objectives.

3.3.1. Analysis for Rainy season

```r
rainy_data = subset(data, Season == "Rainy")
mean(rainy_data$Temperature)

## [1] 3.039344

hist(rainy_data$Temperature, col = "purple",
     main = "Temperature Distribution in Monsoon",
     xlab = "Temperature")
```



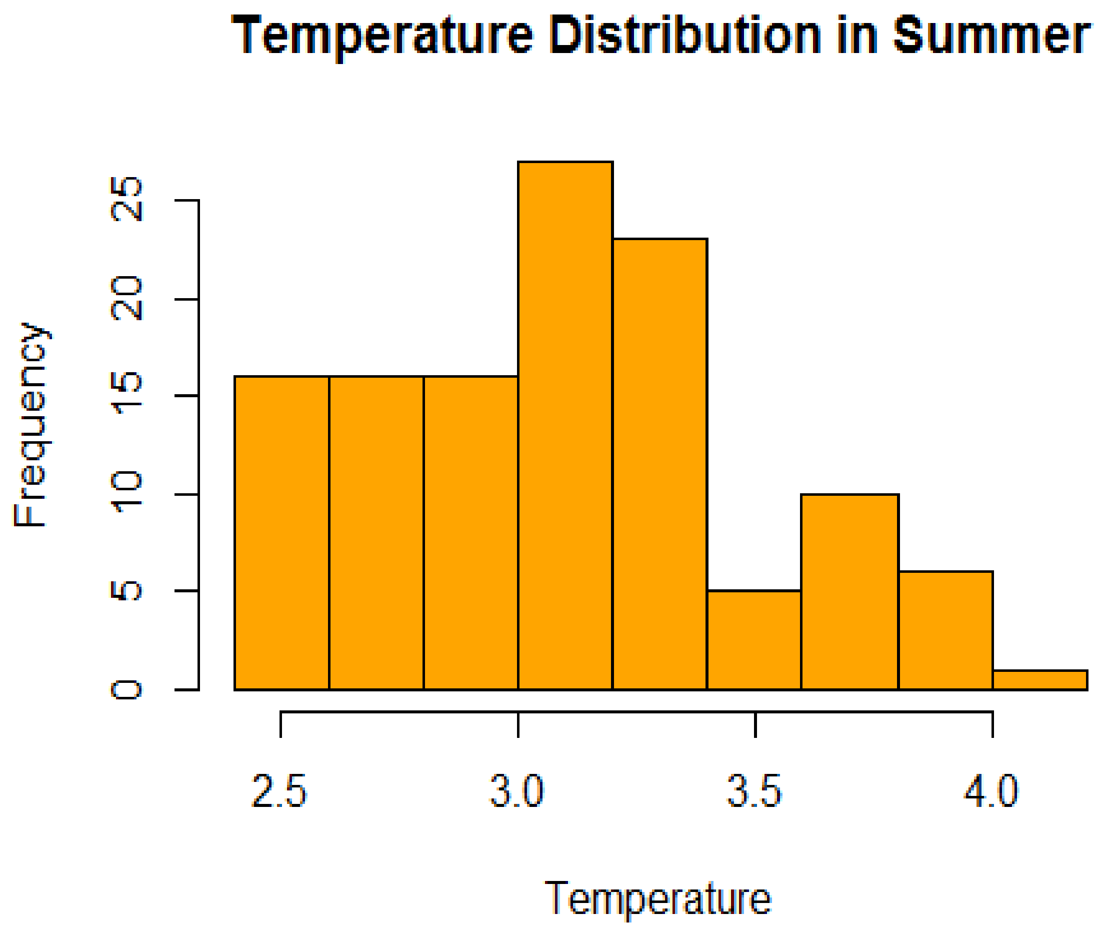
Upon Analysis we find that the mean temperature for rainy season in the given dataset is 3.039344. Graphical analysis reveals majority of the temperature distribution to be within 2 and 4 degree Celsius with highest temperature recorded at 5 degree Celsius and lowest at 1.7 degree C.

3.3.2. Analysis for Summer

```
summer_data = subset(data, Season == "Summer")
mean(summer_data$Temperature)

## [1] 3.153333

hist(summer_data$Temperature, col = "orange",
     main = "Temperature Distribution in Summer",
     xlab = "Temperature")
```



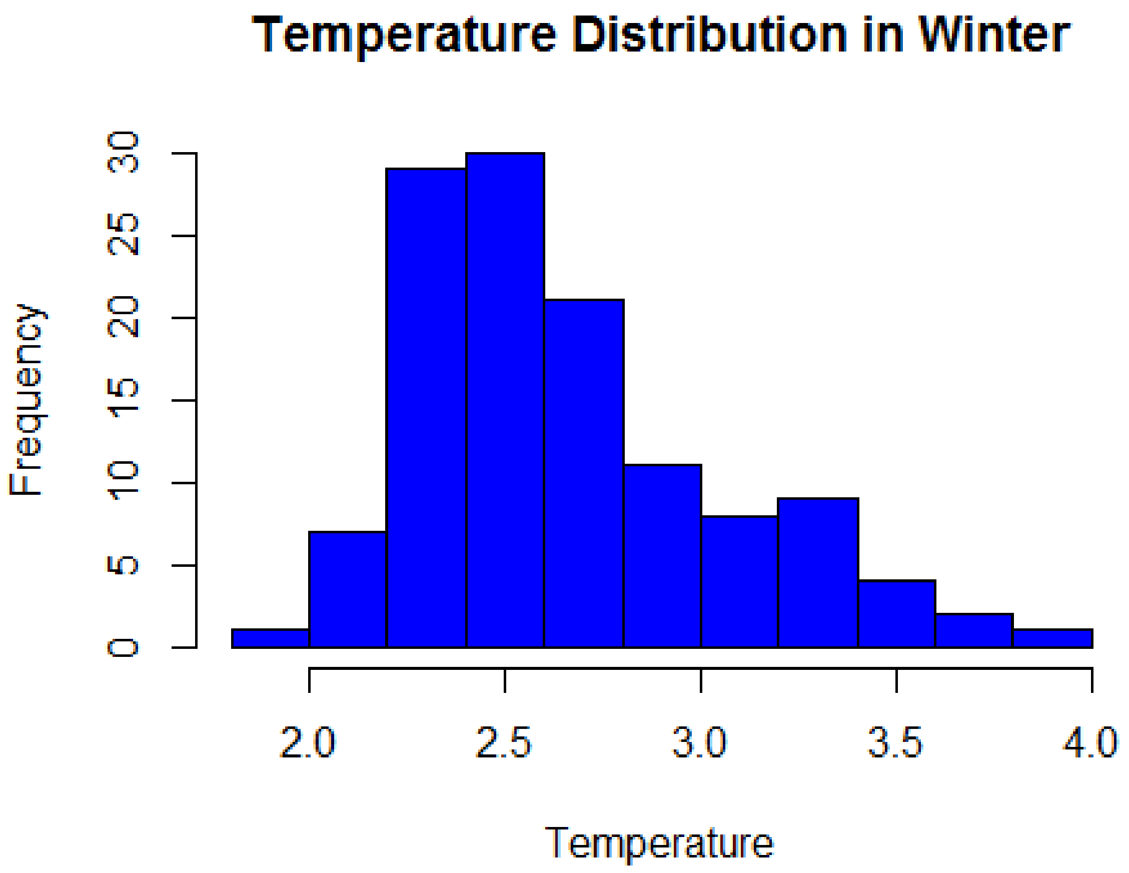
We find that the mean temperature for summer is 3.153333 which is higher than the rainy season mean. But upon graphical deduction we discover the temperature fluctuation to be lower than that of rainy season with highest at 4.1 degree C and lowest at 2.5 degree C.

3.3.3. Analysis for Winter

```r
winter_data = subset(data, Season == "Winter")
mean(winter_data$Temperature)

## [1] 2.700813

hist(winter_data$Temperature, col = "blue",
     main = "Temperature Distribution in Winter",
     xlab = "Temperature")
```

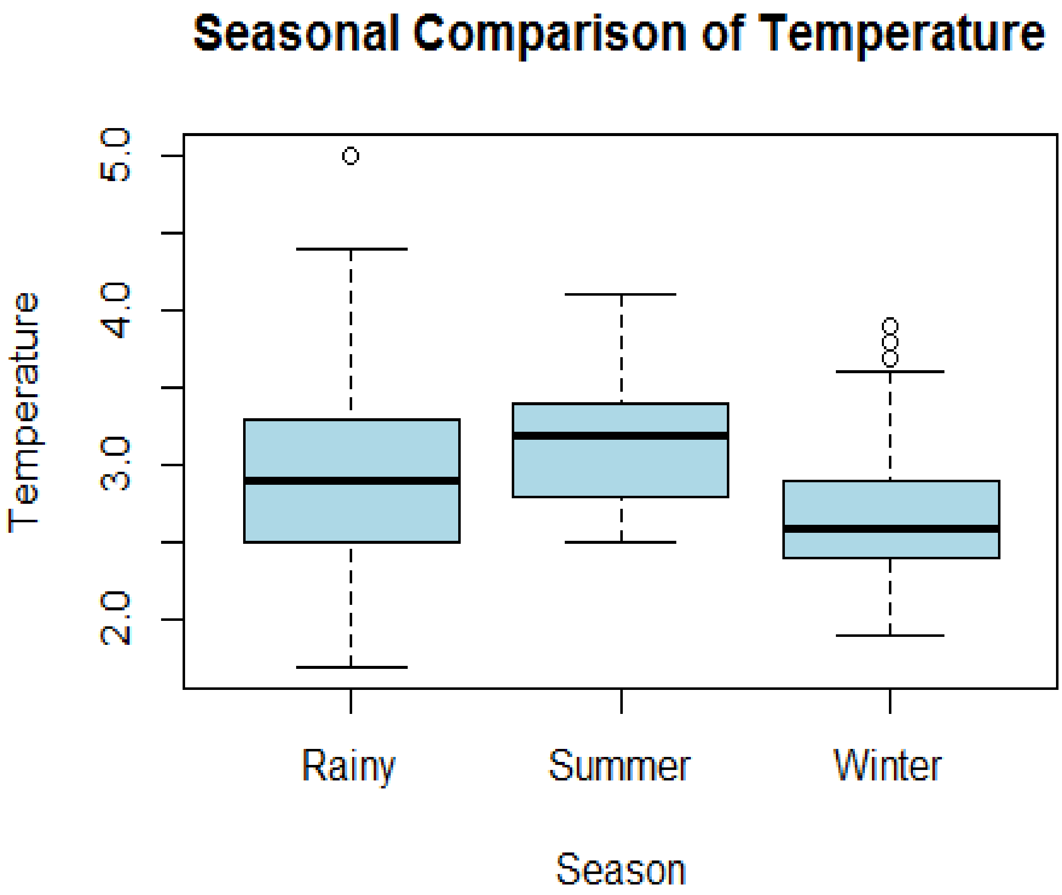


Winter has the lowest temperature mean at 2.700813. It ranges from 1.9 degree C to 3.9 degree C with majority distribution between 2.2 and 2.8 degree C.

3.4. Bi-Variate Analysis and Outlier Identification

We construct a comparative boxplot of the 3 seasons to better understand the correlation between them

```
plot(Temperature~Season, data = data, main = "Seasonal Comparison of
Temperature", col = "light blue")
```



We can easily see that rainy season has the highest temperature range with an outlier, followed by summer with a higher mean but no outlier signifying lesser fluctuations in temperature. Finally, winter has the lowest mean but also consists of multiple outliers.

Question 2

We find the overall mean for the entire year

```
data_mean = mean(data$Temperature)
print(data_mean)

## [1] 2.96274
```

Question 3

We calculate the Standard Deviation for the full year

```
data_sd = sd(data$Temperature)
print(data_sd)

## [1] 0.508589
```

Question 4

We calculate the probability of Temperature falling below 2 degree C

```
pnorm(2, mean = data_mean, sd = data_sd)

## [1] 0.02918146
```

We find that the probability of temperature having fallen below 2 degree C is 2.92 percent.

Question 5

We calculate the probability of Temperature going over 4 degree C

```
pnorm(4, mean = data_mean, sd = data_sd, lower.tail = FALSE)

## [1] 0.02070077
```

Since we need to find the area under the curve at a point right of the mean we take lower.tail = false. The calculated probability is 2.07 percent.

Question 6

For this question we need to find the probability of temperature going outside the 2- 4 degree C range. We do this by subtracting the probability of temperature lying within the specified range from 1.

```
prob =round(1 -(pnorm(4, mean = data_mean, sd = data_sd) - pnorm(2, mean =
data_mean, sd = data_sd)),4)
print(prob)

## [1] 0.0499
```

Thus, the probability of temperature exceeding the 2 - 4 degree C range is found to be 4.99 percent. According to the terms of the contract, a penalty of 10% of annual maintenance cost should be imposed on the AMC Company.

Problem 2

4. Exploratory Data Analysis for Problem 2

4.1. Importing and Reading the Dataset

```
new_data = read.csv("Cold_Storage_Mar2018.csv", header = TRUE)
```

4.2. Analysis of the Dataset

```
dim(new_data)

## [1] 35  4

summary(new_data)

##     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

str(new_data)

## 'data.frame':    35 obs. of  4 variables:
##  $ Season     : Factor w/ 1 level "Summer": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Month      : Factor w/ 2 levels "Feb","Mar": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Date       : int  11 12 13 14 15 16 17 18 19 20 ...
##  $ Temperature: num  4 3.9 3.9 4 3.8 4 4.1 4 3.8 3.9 ...

head(new_data)

##   Season Month Date Temperature
## 1 Summer   Feb   11         4.0
## 2 Summer   Feb   12         3.9
## 3 Summer   Feb   13         3.9
## 4 Summer   Feb   14         4.0
## 5 Summer   Feb   15         3.8
## 6 Summer   Feb   16         4.0
```

March 2018 dataset has 35 observations during the “Summer” season for the months of “February” and “March”. Since sample size > 30 it falls under the Central Limit theorem. We are required to perform z-test and t-test on this sample to ascertain whether corrective action is required at the Cold Storage Plant or not. We notice “Temperature” in this dataset ranges from 3.8 to 4.6 degree C.

Question 1

5. Statistical Analysis of Dataset

5.1. z-test

Here we take the Null Hypothesis (Ho): Mu = 3.9 (corrective action is not required)
and the Alternative Hypothesis (Ha): Mu > 3.9 (corrective action is required)

We begin by calculating the mean and standard deviation of the sample population. Given, Mu = 3.9 and Alpha = 0.1.

```r
attach(new_data)

new_data_mean = mean(Temperature)
new_data_sd = sd(Temperature)

Mu = 3.9

zstat = (new_data_mean - Mu)/(new_data_sd/sqrt(length(Temperature)))

P.Ho = pnorm(zstat, lower.tail = FALSE)
print(P.Ho)

## [1] 0.002958384
```

We calculate the zstat by dividing the sample error by the standard error. We find out the P-value by using the pnorm function. Since the zstat is positive, lower.tail in pnorm is taken as false. We find the P-value to be 0.29 percent.

Question 2

5.2. t-test

Next we carry out a one sample Student's t-test on the dataset. We perform the test at 90% confidence level since Alpha is 0.1. It is a right-tailed test.

Here also, Null Hypothesis (Ho): Mu = 3.9 (corrective action is not required)
and Alternative Hypothesis (Ha): Mu > 3.9 (corrective action is required)

```r
t.test(Temperature, alternative = c("greater"), mu = Mu, conf.level = 0.9)

##
##  One Sample t-test
##
## data:  Temperature
## t = 2.7524, df = 34, p-value = 0.004711
## alternative hypothesis: true mean is greater than 3.9
## 90 percent confidence interval:
##  3.939011      Inf
## sample estimates:
```

```
## mean of x
##  3.974286
```

We find that the tstat is 2.7524, degrees of freedom is 34 and the P-value is 0.47 percent.

Question 3

6. Conclusion

On performing the z-test we find the P-value to be 0.29 percent which is lesser than Alpha (10%). Hence, we reject the null hypothesis at 90% confidence level. The t-test gives us a P-value of 0.47 percent which is also less than Alpha. Moreover, we see that the true mean at 90 percent confidence level is 3.94 which is greater than Mu(3.9). Hence, in this case also the null hypothesis is rejected. Since in both the tests P-value < Alpha, the Null Hypothesis is rejected and the Alternative Hypothesis is accepted.

Hence, we may conclude by confirming that corrective action is required at the Cold Storage Plant with 90% confidence.