COLD STORAGE PROJECT

Karthik Paranthaman

UNIVERSITY OF TEXAS - AUSTIN

PROJECT OVERVIEW

The main of the project is to analyze problems given based on the Cold Storage temperature maintenance.

The R-Studio Platform is used to analyze the dataset using Descriptive Statistical tools such as Mean, Median, IQR, Standard Deviation and Variances, Inferential Statistical tools like Probability Distribution Theorems, Hypothesis tools like Central Limit Theorem and Advanced Statistical Theorems like ANOVA tests.

The outputs are visualized using varies graphical methodologies to interpret with an aid of plots like Histogram, Bar Plot, Box Plot, Density Plot, and etc.

1. ASSUMPTIONS

It was assumed that all information provided by the customers are accurate. This will be confirmed using Sanity Check methodologies.

Normality and the equality of the variances of the dataset is assumed proved using various testing methodologies which will be discussed in the upcoming chapters.

In addition to that, it was assumed and eliminated the error rate of the measuring device in the temperature measuring device.

2. EXPLORATORY DATA ANALYSIS

In this Chapter we will be discussing about the data exploration approaches to find solutions to the given questions. The exploratory process consists of the following stages.

1. Environment Setup and Data Import
2. Variable Identification
3. Missing Value Identification
4. Business Solutions

3.1 Environment Setup and Data Import

3.1.1 WORKING DIRECTORY

The working directory has been set. The main dataset is called using the name

"ColdStorage_Data" and the sample population data set for problem – 2 is known as "ColdStorage_Data_Mar" in R-Studio (Refer to APPENDIX -1 – for R Code).

3.1.2 LIBRARY PACKAGES

The library packages used in this project are listed as follows:

1. readr
2. ggplot2
3. BSDA
4. RCMDR

3.2 Missing Value Identification

The "ColdStorage_Data" dataset contains 4 Variables and 365 observations.
Sanity Check output is to be FALSE and confirms to there are no NA entries.

Furthermore, output from "summary (ColdStorage_Data)" command also approves the Sanity Check from missing values and data classes.

In addition to that "head" & "tail" commands are used to call top & bottom 10 rows to identify for any formatting issue in the dataset given and there are none.

PROBLEM – 1

QUESTION – 01

The mean temperature maintained during all 3 seasons are tabulated in the table below.

SEASONS	**MEANS**
Rainy	3.088
Summer	3.147
Winter	2.776

QUESTION – 02

The average mean is stored under the variable name "Grand_Mean".

The average calculated temperature for the whole year is approximately **3.002°C.**

QUESTION – 03

The average mean is stored under the variable name "Grand_SD".

The average calculated temperature for the whole year is approximately **0.466°C.**

QUESTION – 04

The probability of temperature falling below 2.0°C is **1.57%.**

QUESTION – 05

The probability of temperature falling below 2.0°C is **1.612%.**

QUESTION – 06

The agreed terms on probability with the outsourced party is to maintain the temperature between 2°C-4°C to have probability range of 2.5 % to 5% and more than 5%.

In this scenario total probability of having the temperature outside the operation range is 1.57%+1.612% = **3.182%.**

Therefore, the 10% penalty will be fined from the outsourced parties from the AMC based on the initial agreed terms and conditions of the contract.

QUESTION – 07

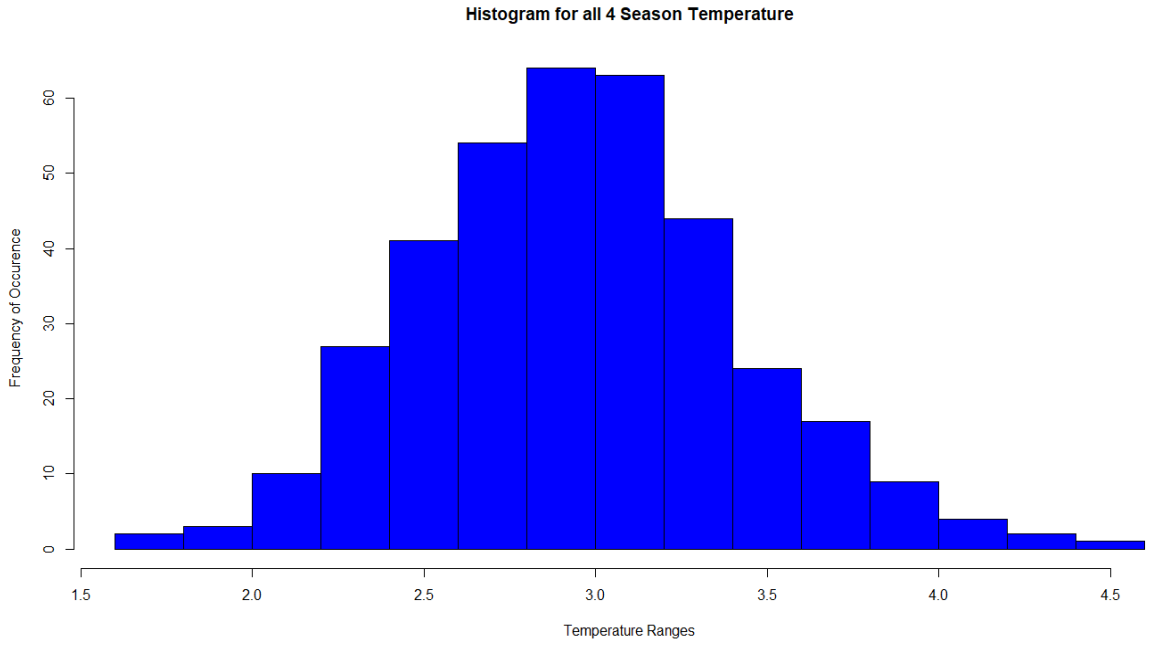
Before the ANOVA one way test a Descriptive Analysis must be done to understand the distribution and the nature of the data.

Summary of the distribution is shown below:

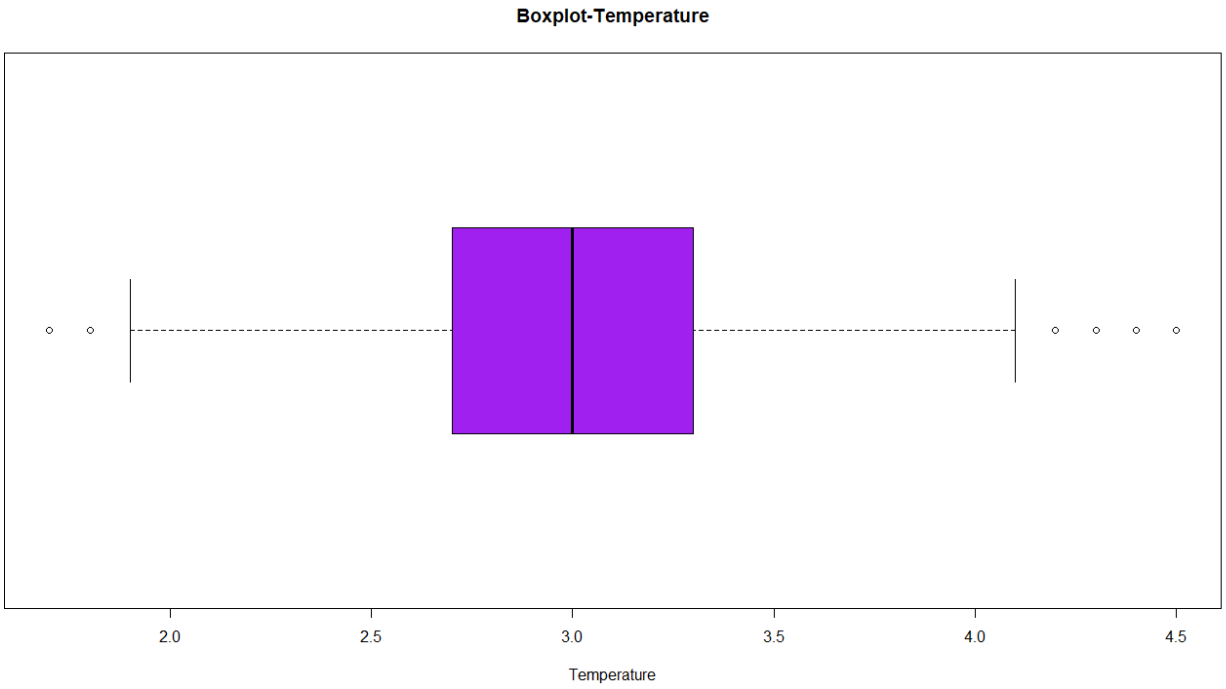
```
 Season        Month           Date       Temperature
 Rainy :122   Aug    : 31   1      : 12   Min.   :1.700
 Summer:120   Dec    : 31   2      : 12   1st Qu.:2.700
 Winter:123   Jan    : 31   3      : 12   Median :3.000
              Jul    : 31   4      : 12   Mean   :3.002
              Mar    : 31   5      : 12   3rd Qu.:3.300
              May    : 31   6      : 12   Max.   :4.500
              (Other):179   (Other):293

>
```

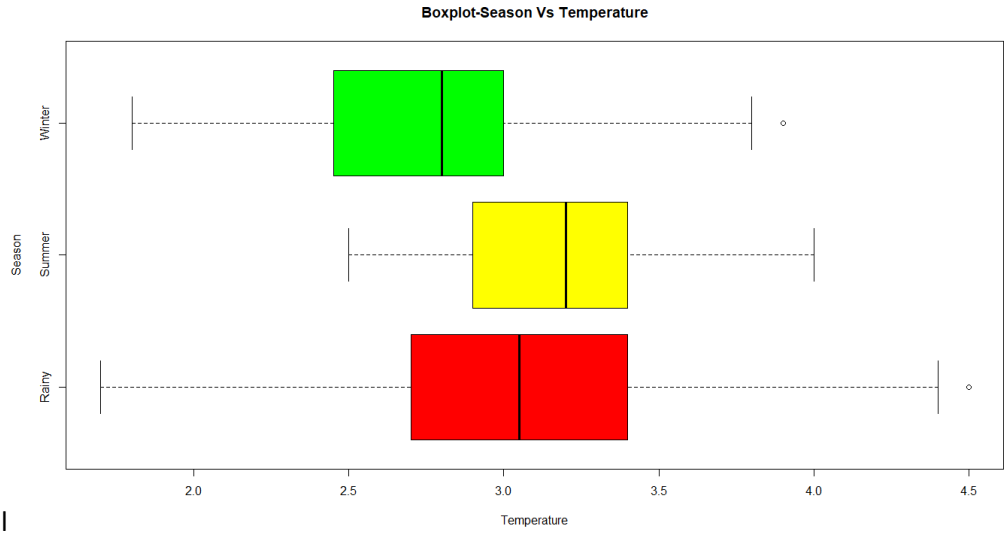
The minimum value of the temperature is 1.7°C while maximum value being 4.5°C. The following histogram figure illustrates the visual confirmation of the distribution.



The figure below illustrates the summary output in boxplot format.



The comparison between temperature range between each season is given below. According to the plot it can be said Summer and Rainy season has similar data while Winter season data to be further away. All IQR interfere between each other. To analyze further ANOVA one way test must be done.



The ANOVA one way test is sensitive to both normality and the equality of the of variance properties. Normality test on the overall population is done using **Shapiro-Wilk's** test.

ASSUMPTION - 1

The hypothesis for the Normality Test is given below.

H_0 = **Cold Storage temperature data follows the Normal Distribution**.

Against the alternative Hypothesis as,

H_a = **Cold Storage temperature data does not follow the Normal Distribution properties**.

```
        Shapiro-Wilk normality test

data:  ColdStorage_Data$Temperature
W = 0.99212, p-value = 0.05044
```

Since p-value of the test is around 5%, we fail to reject the null hypothesis that the response follows the Normal Distribution.

ASSUMPTION - 2

Next, we need to test the assumption on all four levels of the factor Seasons population variance is equal. In other words, the homogeneity of variance assumption must be satisfied.

```
Rainy Summer Winter
 0.278  0.124  0.172
```

As shown in the above table by running a variance calculation on the dataset it can be seen that variances are approximately equal. But this assumption must be further confirmed. By using a Hypothesis testing which could be formulated as,

$H_0: \delta_1 = \delta_2 = \delta_3$ =**Variance of Rainy, Summer & Winter**

Against the alterna1tive Hypothesis as,

H_a = **Atleast one of the variance is different from the other levels of factors**.

The hypothesis test for homogeneity using **"Levene Test"** is shown below.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   2   8.008 < 2.2e-16 ***
      362
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Number of samples for each season is displayed in the table:

SEASONS	NUMBER OF SAMPLES
Rainy	122
Summer	120
Winter	123

As we can see P value is very low which fails to reject the Null Hypothesis. But the number of samples at each level is almost equal this makes the F-test robust. Therefore, we can proceed with the One Way ANOVA test.

<u>One Way ANOVA Test</u>

Both Normality and Homogeneity level variances are being confirmed to proceed with the One Way ANOVA test with an aid of this R code. The test could be formulated as Hypothesis as given below:

$\boldsymbol{H_0}: \mu_1 = \mu_2 = \mu_3 =$ **Mean of all seasons are equal**.

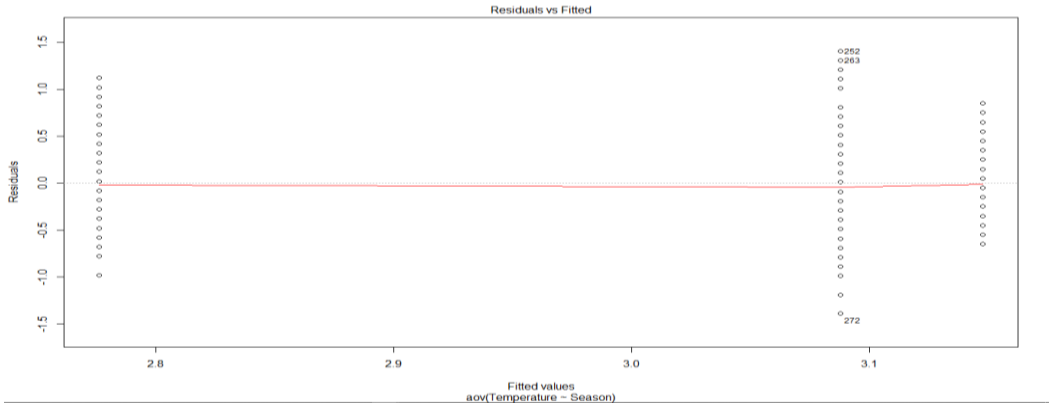
Against the alterna1tive Hypothesis as,

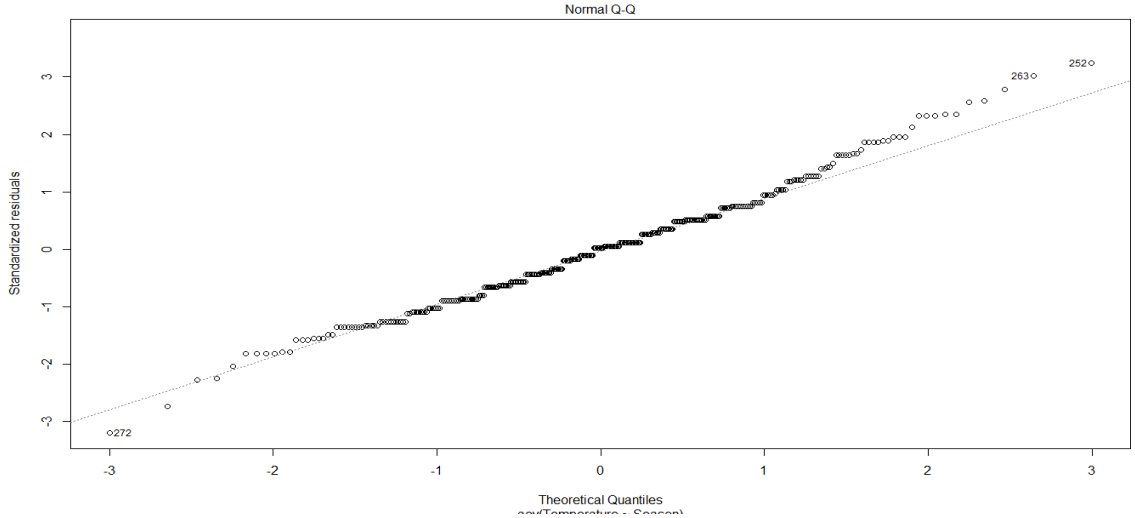
$\boldsymbol{H_a} =$ **Atleast one of the season has different mean from other two**.

```
> summary(AOV_Temp)
             Df Sum Sq Mean Sq F value   Pr(>F)
Season        2   9.70   4.848   25.32 5.08e-11 ***
Residuals   362  69.29   0.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Let us consider the summary output known as ANOVA Table. For the given problem sum of squares due to the factor Season (SSB) is 9.70 and the sum of squares due to error (SSW) is 69.29.The total sum of squares (SST) for the data is (9.70+69.29=78.99). Since the factor has 3 levels, DF corresponding to seasons is 3 – 1 = 2. Total DF for Residuals is 365 – 1 = 364. Hence DF due to error is 364 – 2 = 362.Mean sum of squares is obtained by dividing the sums of squares by corresponding DF. The value of the F-statistic is approximately 25 and the p-value is highly significant. (i.e: 5.08e-11). Based on the ANOVA test, therefore, reject the null hypothesis that the three population means are identical. At least one of the mean of the season sample population is different.

On the other hand the Residuals VsFitted Plot illustrates the homoscedasticity of variances properties and also the visually shows in depth understanding of how the two of the means are close together. Also the graph on Normal Q-Q plot Vs Season confirms the normality property of the distribution.





Now the next stage in the Oneway ANOVA test is to find out which group mean is away from the other two.

In order to identify for which Season type mean Temperature is different from other groups, the hypotheses may be stated as:

$\boldsymbol{H_0}: \mu_1 = \mu_2, \mu_1 = \mu_3, \mu_2 = \mu_3 =$ **All pairs of group means are equal**.

Against the alterna1tive Hypothesis as,

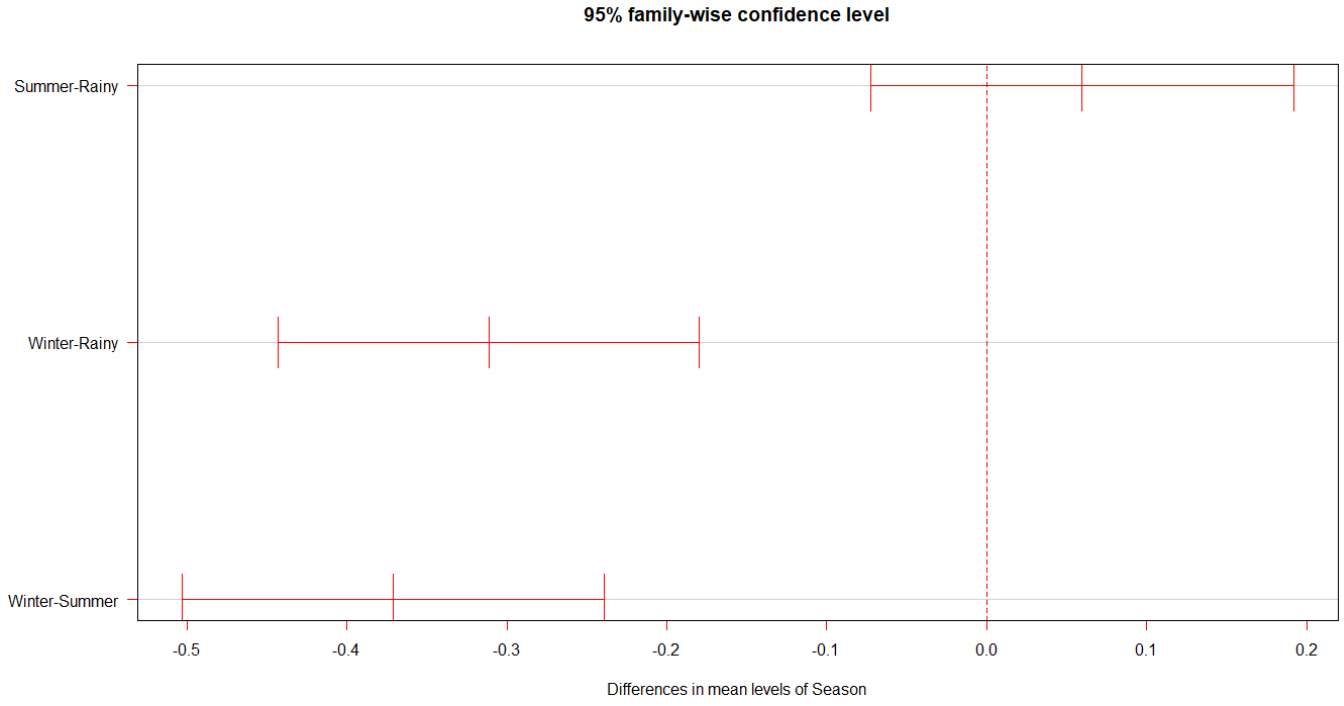
$\boldsymbol{H_a}: \mu_1 \neq \mu_2$ ***or*** $\mu_1 \neq \mu_3, \mu_2$ **or**
$\neq \mu_3$ **Atleast one of the season group has different mean from other two**.

where μ_1 represents mean Temperature of Winter, μ_2 is the mean temperature of the summer and μ_3 represents the Rainy season.

"TukeyHSD" function helps us to identify the mean which is away from other two.

```
                diff   lwr   upr p adj
Summer-Rainy    0.06 -0.07  0.19  0.54
Winter-Rainy   -0.31 -0.44 -0.18  0.00
Winter-Summer -0.37 -0.50 -0.24  0.00
```

We can observe that Winter-Rainy & Winter-Summer means are equal and Summer-Rainy is not equal. Therefore, all null hypothesis will be rejected. It can be said that Summer -Rainy season mean temperatures are similar. So we can say that mean temperature of Winter Season is significantly different from other two seasons. Evidently pairwise different plot also indicates a similar output.

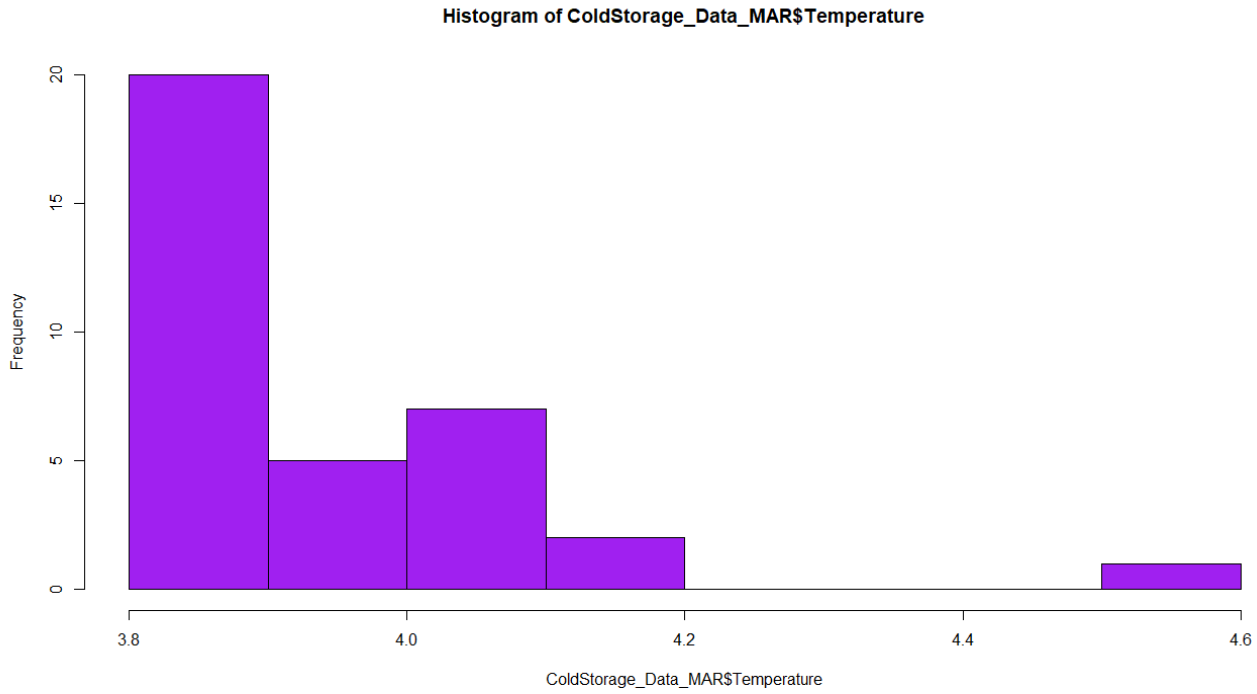


PROBLEM – 2

QUESTION – 01

Most suitable hypothesis test for this scenario would be t.test.

Given Cold_Storage_Mar2018.csv has been extracted from the Cold_Storage_Temp_Data.csv. True Mean of the population is 3.0025°C and the Standard Deviation is 0.4618. The number of rule of thumb to carry on Z.test is at least 30 observations. However from the histogram plot below it can be said that data is not normally distributed and this occurs due to the sample size. Therefore, to overcome this I have decided to use the t-statistics test check the Hypothesis Testing.



QUESTION – 02

H_0: Cold Storage Doesnt Require Correction

Against the alterna1tive Hy6ttt6pothesis as,

H_a: Cold Storage Does Require Correction

The ttest output is given as follows:

```
One Sample t-test

data:  ColdStorage_Data_MAR$Temperature
t = 147.25, df = 34, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
99 percent confidence interval:
 3.908399      Inf
sample estimates:
mean of x
 3.974286
```

The P-value is approximately 2.2e-16.

QUESTION – 03

From the ttest output P-value is less than the 1% significance level. Therefore, it can be said that Null hypothesis is failed to accept. Hence, the the Cold Storage does require a correction in maintaining the desired temperature.