

# **Mini Project:1**

**A case study on Cold Storage's plant temperature maintenance contracted to a professional company.**

**by**

**Akash Biswas**

**Great Lakes Institute of Management**

**PGP – Business Analytics & Business Intelligence**

# Table of Contents

1 Project Objective.....	4
2 The Company/Plant profile and her Associates.....	4
3 Issues related to this Analysis and Case Study.....	4
4 Exploratory Data Analysis – Step by step approach (#Problem-1)	5
4.1 Environment Set up and Data Import.....	7
4.1.1 Install necessary Packages and Invoke Libraries.....	7
4.1.2 Set up working Directory.....	7
4.1.3 Import and Read the Dataset.....	8
4.2 Variable Identification & Inferences.....	8
4.3 Summary of the dataset.....	9
4.4 Calculation of overall Average/Mean Temperature and Overall Standard Deviation at Cold Storage using the attach() function.....	9
4.5 Creation of subsets as per season in 2016 along with the Mean Temperature per season in 2016 at Cold Storage.....	10
4.6 Finding the probability of the overall temperature falling below 2 degree C.....	11
4.7 Finding the probability of the overall temperature rising above 4 degree C.....	12
4.8 Conclusion.....	13
5 Appendix A – Source Code Problem-1.....	14

6 Exploratory Data Analysis – Step by step approach (#Problem-2)	15
7.1 Environment Set up and Data Import.....	16
7.1.1 Install necessary Packages and Invoke Libraries.....	16
7.1.2 Set up working Directory.....	16
7.1.3 Import and Read the Dataset.....	16
8.1 Statistical Assumptions.....	17
8.2 Hypothesis Testing via Z-Test Method .....	17
8.3 Summary of the Dataset.....	18
8.4 Calculating Mean.....	18
8.5 Bell Curve Diagram Reference.....	20
8.6 Hypothesis Testing via t-test method.....	20
8.7 Conclusion and inferences.....	22
9 Appendix A – Source Code Problem-2.....	22

## **1. Project Overview:**

The objective of the report is to enact and ensure that an analytical approach has been taken to find out a solution to the two major problems regarding the systematic and adequately professional maintenance of Cold Storage's plant over the years. This report will be dedicated to performing a statistical analysis using analytical tools such as 'R' and MS Excel via installing packages, importing the provided datasets, taking insights from them and providing analytical solution with graphical exploration.

## **2. The Company/Plant profile and her Associates:**

'Cold Storage' a leading name in the business of storing and preserving dairy products have come into an Annual Maintenance Case (AMC) with a professional company handling their plant-maintenance. Cold Storage is associated with storing Pasteurized fresh whole or skimmed milk, Sweet Cream, Flavoured Milk Drinks etc.

## **3. Issues/Problems related to this Analysis and Case Study:**

In their varied day-to-day routine, they need to ensure that there are no changes of texture, body appearance and separation of fats for which they need to the optimal temperature to be maintained at between 2 deg C to 4 deg C. In the very first year of business i.e in 2016 they outsourced the plant maintenance work to a professional company with stiff penalty clauses. It was agreed between them that if it was statistically proven that probability of temperature going outside the 2 degrees C to 4 degrees C during that one-year of the contract was above 2.5% and less than 5% then the penalty would be 10% of AMC (annual maintenance case). In case it exceeded 5% then the penalty would be 25% of the AMC fee. We need to perform a detailed analytical study of the data provided to us for the first year of their operations and collaborations with the company. We need to

apply Descriptive as well as Inferential Statistics to know the requirement to inculcate changes if necessary, at Cold Storage. The AMC is an annual maintenance contract that the two parties decide upon for which Cold Storage pays the plant maintenance company an annual amount for repeated service and maintenance as and when required by them. Glitches and failures during and after the activities may lead to losses incurred by Cold Storage due to which they introduced a clause containing penalty. The average temperature data at date level is provided to us in the file “Cold\_Storage\_Temp\_Data.csv”

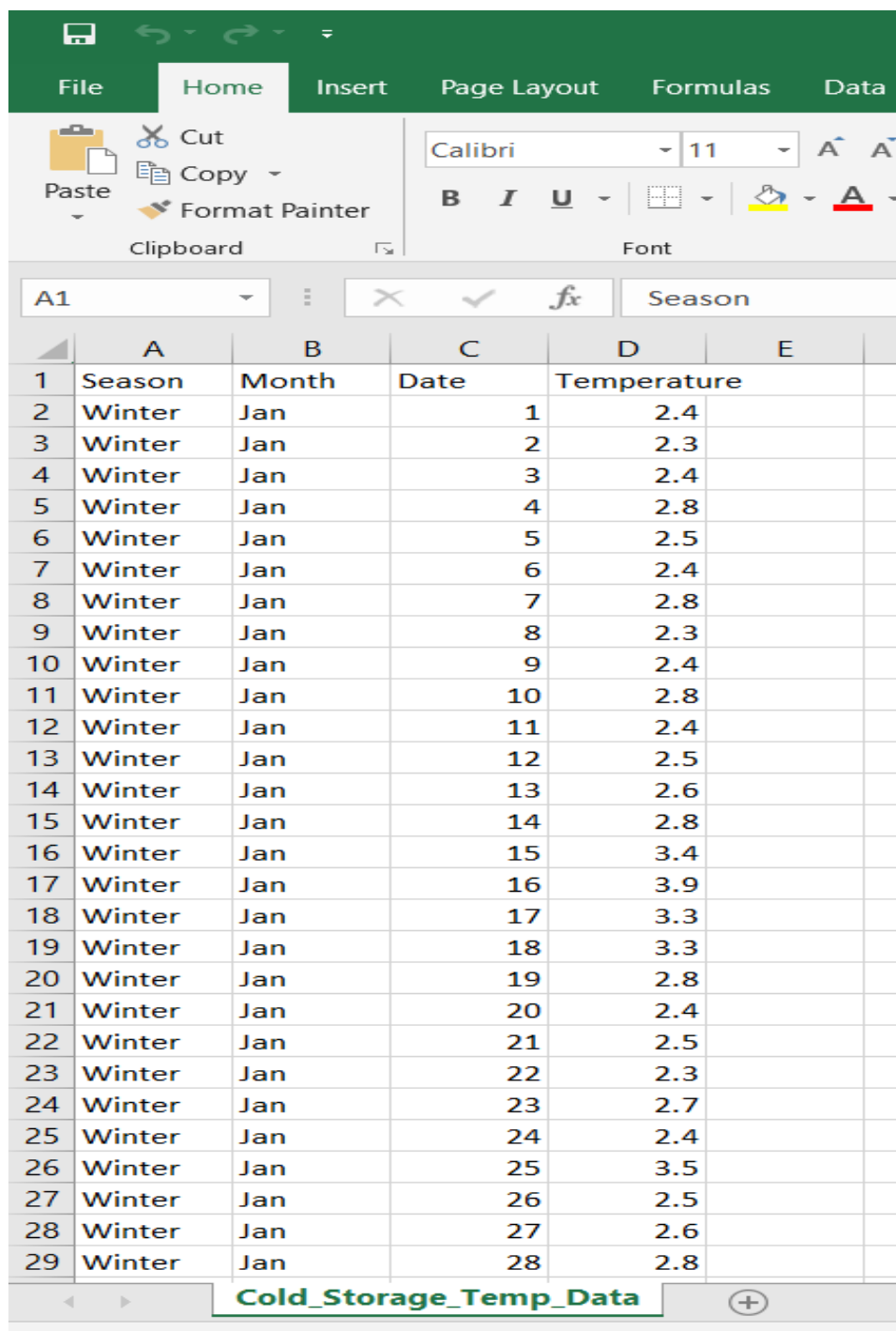
In March 2018, Cold Storage started getting complaints from their Clients that they have been getting complaints from end consumers of the dairy products going sour and often smelling. On getting these complaints, the supervisor pulled out data of last 35 days temperatures from the date of complaint. As a safety measure, the Supervisor has been vigilant to maintain the temperature below 3.9 deg C.

Assuming 3.9 deg C as the upper acceptable temperature range and at the Level of Significance ( $\alpha$ ) = 0.1 we need to hypothesize and find out if there is any need for some corrective action in the Cold Storage Plant or is it that the problem is from procurement side from where Cold Storage is getting their Dairy Products. Performing t-tests and z-tests will be core to solve the problem and rejection or acceptance of the null hypothesis may lead to an analytical decision making to perform at user-end. The data of the last 35 days has been provided in “Cold\_Storage\_Mar2018.csv”.

#### **4. Exploratory Data Analysis – Step by step approach(Problem-1)**

The first problem for the year 2016 leads us to a dataset “Cold\_Storage\_Temp\_Data.csv” which indicates a four variable .csv data with their elements collected in continuous observations from the 1<sup>st</sup> of January to the 31<sup>st</sup> of December 2016. The variables are c ← (Season, Month, Date & Temperature). The Seasons are separated into three categories namely Winter, Rainy & Summer. Similarly,

month contains all the months of 2016 (i.e from January to December) and their respective dates 1<sup>st</sup> to 28<sup>th</sup> /30<sup>th</sup> or 31<sup>st</sup> in the variable list Date. The vector within this dataset that we will analyse would be the Temperature Vector since we need to find a solution to the first problem. We will perform a few multivariate analysis to strengthen our claim to the solution. A illustrated example of the dataset is shown below for a formative reference:



	A	B	C	D	E
1	Season	Month	Date	Temperature	
2	Winter	Jan	1	2.4	
3	Winter	Jan	2	2.3	
4	Winter	Jan	3	2.4	
5	Winter	Jan	4	2.8	
6	Winter	Jan	5	2.5	
7	Winter	Jan	6	2.4	
8	Winter	Jan	7	2.8	
9	Winter	Jan	8	2.3	
10	Winter	Jan	9	2.4	
11	Winter	Jan	10	2.8	
12	Winter	Jan	11	2.4	
13	Winter	Jan	12	2.5	
14	Winter	Jan	13	2.6	
15	Winter	Jan	14	2.8	
16	Winter	Jan	15	3.4	
17	Winter	Jan	16	3.9	
18	Winter	Jan	17	3.3	
19	Winter	Jan	18	3.3	
20	Winter	Jan	19	2.8	
21	Winter	Jan	20	2.4	
22	Winter	Jan	21	2.5	
23	Winter	Jan	22	2.3	
24	Winter	Jan	23	2.7	
25	Winter	Jan	24	2.4	
26	Winter	Jan	25	3.5	
27	Winter	Jan	26	2.5	
28	Winter	Jan	27	2.6	
29	Winter	Jan	28	2.8	

We will use R console to find a solution to the first problem faced by Cold Storage and if we require to impose penalty to the maintenance company. The probability of the temperature having fallen below 2 deg C and above 4 deg C has to be retrieved for the year 2016 which will be presented using R.

## **4.1. Environment Set up and Data Import:**

### **4.1.1. Installation of necessary Packages and Invoking Libraries:**

A new R Script is chosen to perform the analysis named Project-1 Problem-1. Installation of various R packages which will assist us during our analysis is required to apply the functions we want to perform on different variables, vectors and lists of the dataset. We installed the below mentioned packages:

**install.packages("readr")** → for invoking libraries to read the .csv file into R.  
**install.packages("ggplot2")** → for invoking libraries to project graphical plots into R.

**library(readr)** → for invoking the library which contains the functions to read a .csv file into R.

**library(ggplot2)** → for invoking the library which contains the functions to plot graphical representation of the dataset into R.

### **4.1.2. Set up working Directory:**

The working directory has to be set from where the dataset has to be incorporated into R. The programme needs to know from where it needs to collect the .csv file saved at a random drive. Our .csv was saved in some folder in the E drive. So we assign the function:

**setwd("E:/from old hard disc 24.08.2019/Data Analytics and Data Science/Great Lakes PGP-BABI/2. Statistical Methods for Decision Making (SMDM)/PROJECT-1 (Cold Storage Case Study)")**

Note that all the back-slashes in the address has been converted to front slashes as R will not recognise the working directory otherwise.

We can also view the working directory by calling the function `getwd()`.

### 4.1.3. Import and Read the Dataset:

Now the dataset was imported and read by simultaneously storing it in a vector → `data_2016` using the function `read.csv`

The syntax used over here is:

```
data_2016 <- read.csv("Cold_Storage_Temp_Data.csv", header = TRUE)
```

Here, `Cold_Storage_Temp_Data.csv` is the .csv file and `header = TRUE` is to keep the header containing the variable names into the `data_2016` vector.

### 4.2. Variable Identification & Inferences:

Now we would like to explore the vector `data_2016` and identify its class and variables with some R functions. We use `class(data_2016)` function only to find out that it is a data.frame in the console. Data Frames are a kind of matrices which contains a host of different types of variables like character, numeric, logical etc.

Similarly we use this `class()` function to find the classes of the variables that we are working with and find out that the classes of Season, Month, Date & Temperature are Factor, Factor, Integer & Numeric (i.e with decimal places) respectively.

#### Console O/P: (an illustration for example)

```
(other):1/9
> class(data_2016)
[1] "data.frame"
> class(data_2016$Season)
[1] "factor"
> class(data_2016$Month)
[1] "factor"
> class(data_2016$Date)# summary of data_2016
[1] "integer"
> class(data_2016$Temperature)
[1] "numeric"
>
```

We also identify the structure of the vector `data_2016` by using `str(data_2016)` function which provides us with the number of observations, number of variables and the class of the variables as a whole.



## Console O/P: (an illustration for example)

```
> # structure of the vector
> str(data_2016)
'data.frame':   365 obs. of  4 variables:
 $ Season      : Factor w/ 3 levels "Rainy","Summer",...: 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 ...
 $ 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 ...
```

### 4.3. Summary of the dataset:

Summary of the dataset or the vector could be invoked by the function `summary(data_2016)`. In console:

```
> # summary of data_2016
> summary(data_2016)
  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
```

The summary shows us the number of observations in each season (i.e Rainy, Summer, Winter) taken. The month Aug, Dec, Jan, Jul, Mar and May has 31 observation and the rest in total are 179 observations. Since these two are having a class of factor, these prints out no. of observations each levels have. Similarly if we see the temperature summary we find the minimum value as 1.7 deg C, maximum as 5 deg C, 1<sup>st</sup> Quartile at 2.5 deg C, 3<sup>rd</sup> Quartile at 3.3 deg C, median at 2.9 deg C and Mean at 2.963 deg C. We can find the overall Mean of the temperature using another function stated later.

### 4.4. Calculation of overall Average/Mean Temperature and Standard Deviation at Cold Storage using the `attach()` function:

We can find the overall Mean of the temperature using another function as `mean()`. First of all we call the `attach(data_2016)` function which plays a significant part in this programme.

The `attach()` function is the database attached to the R search path. This means that the database is searched by R when evaluating the

variable, so objects in the database can be accessed by simply giving their names. This can be verified by `?attach()` function in R.

The overall mean of the temperature can be stated with a syntax:

```
Average_temp_2016 <- mean(Temperature)
```

**Average\_temp\_2016** prints out the mean of the overall temperature which is 2.96274 deg C. # Answer to question no 2.

We also find the Standard Deviation using the `sd()` function on the same variable Temperature. Syntax are as:

```
Standard_deviation <- sd(Temperature)
```

**Standard\_deviation** # standard deviation for the full year found = 0.508589 deg C (Answer to question no.3)

#### **4.5. Creation of subsets as per season in 2016 along with the Mean Temperature per season in 2016 at Cold Storage:**

Subsetting the dataset (`data_2016`) is required to make three smaller matrices or datasets which are variables and elements regarding winter, summer and rainy seasons. These subsets are required so that we could apply the `mean()` function to find out the temperature means of these three seasons. The syntax for subsetting `data_2016` is provided below:

```
winter_data_2016 <- data_2016[data_2016$Season=="Winter",]  
View(winter_data_2016)
```

```
summer_data_2016 <- data_2016[data_2016$Season=="Summer",]  
View(summer_data_2016)
```

```
rainy_data_2016 <- data_2016[data_2016$Season=="Rainy",]  
View(rainy_data_2016)
```

The subsetting data for the seasons summer, winter and rainy has been saved into new vectors namely, `winter_data_2016`,

summer\_data\_2016 and rainy\_data\_2016. The subsetting has been done using [] keeping data\_2016 in front of the brackets to indicate R which vector to subset. We extract the Season variable from the dataset by typing data\_2016\$Season and requesting Season to be only and only equal to (==) Summer, Winter and Rainy respectively. Now we put this within [] to clarify which rows and columns would we like to take into these new subsets. We want the rows containing the particular seasons but all columns pertaining to those seasons thus we write [data\_2016\$Season=="Rainy",] keeping the column blank after , which indicates that all columns should be taken. Now we use the same mean() function to calculate the means of the temperatures at Cold Storage for three separate seasons.

The syntaxes are:

```
# mean temperature for summer, rainy and winter
Average_temp_winter_2016 <- mean(winter_data_2016$Temperature)
Average_temp_winter_2016 # average found for winter 2016 = 2.700813 deg C (Answer to
question no.1)

Average_temp_rainy_2016 <- mean(rainy_data_2016$Temperature)
Average_temp_rainy_2016 # average found for rainy 2016 = 3.039344 deg C (Answer to
question no.1)

Average_temp_summer_2016 <- mean(summer_data_2016$Temperature)
Average_temp_summer_2016 # average found for summer 2016 = 3.039344 deg C (Answer to
question no.1)
```

#### 4.6. Finding the probability of the overall temperature falling below 2 degree C:

The Cold storage case has a clause with the maintenance company that if it was statistically proven that probability of temperature going outside the 2 degrees - 4 degrees C during the one-year contract was above 2.5% and less than 5% then the penalty would be 10% of AMC (annual maintenance case). In case it exceeded 5% then the penalty would be 25% of the AMC fee. So we need to find the probability distribution of temperature falling below 2 deg C using the pnorm() function the syntax suggests:

Therefore, statistically we need to find :

$X \sim N(2.96274, 0.508589)$  or  $X \sim N(\text{mean}, \text{standard deviation})$

$P(X < 2)$  #on the -ve side

**Prob\_value <- pnorm(lower limit, mean= , sd= )**

The pnorm() function is used since we assume the probability distribution is a Normal Distribution.

The probability we find of having the over all temperature below 2 deg C at Cold Storage is around 2.9%.

**# finding the probability of the temperature having fallen below 2 deg C**

**# (Assuming Normal Distribution)**

**Prob\_T\_less\_than\_2 <- pnorm(2, mean=Average\_temp\_2016 , sd=Standard\_deviation)**

**Prob\_T\_less\_than\_2 # probability found around 2.9% (Answer to question no. 4)**

#### **4.7. Finding the probability of the overall temperature rising above 4 degree C:**

Similarly,

$X \sim N(2.96274, 0.508589)$  or  $X \sim N(\text{mean}, \text{standard deviation})$

$P(X < 4)$  #on the +ve side

**Prob\_value <- 1-pnorm(upper limit, mean= , sd= )**

The probability we find of having the over all temperature above 4 deg C at Cold Storage is around 2.07%. The pnorm() function has been deducted from 1 since we require the probability of the area above 4 deg C and probability can never be more than 100% i.e 1.

#### **Syntax suggests:**

**# finding the probability of the temperature having gone above 4 deg C (Assuming Normal Distribution)**

**Prob\_T\_more\_than\_4 <- 1-pnorm(4, mean=Average\_temp\_2016 , sd=Standard\_deviation)**

**Prob\_T\_more\_than\_4 # probability found around 2.07% (Answer to question no. 5)**

#### 4.8. Conclusion:

As stated in the AMC that if it was statistically proven that probability of temperature going outside the range of 2 degrees C to 4 degrees C during the year 2016 of the contract was  $2.5\% < P < 5\%$  then the penalty would be 10% of AMC (annual maintenance case). In case it is  $> 5\%$  then the penalty would be 25% of the AMC fee. We, after a detailed statistical analysis and calculation of probability of the overall temperature of the year going above and below 2 and 4 deg C find out that they are 2.9% and 2.07% respectively. As per AMC having a probability within the range of  $2.5\% < P < 5\%$  for any one of the arguments (i.e  $< 2$  deg C and  $> 4$  deg C) will be chargeable upto 10% of the AMC paid to the plant maintenance company. Here we find out that the overall temperature being less than 2 deg C is 2.9% so, a 10% deduction on AMC amount should be considered. For a probability of having temperature greater than 4 deg C is less than 2.5% which is around 2.07%. Thus Cold Storage does not require to deduct here. We arrive to this conclusion for the first year of this deal. No percentage of probability is above 5% thus 25% of AMC as penalty isn't required.

# Please refer source code in APPENDIX-A for this problem.

## 5. Appendix A – Source Code # Problem – 1

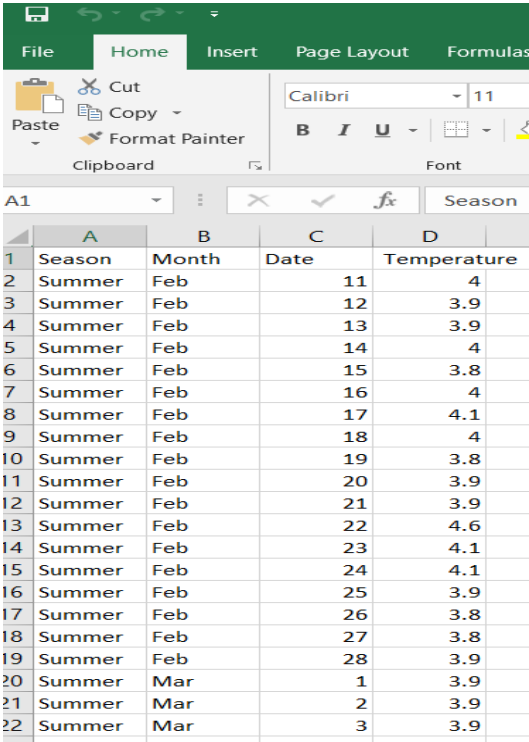
```
Project 1 R code.R | data_2016 | winter_data_2016 | summer_data_2016 | rainy_data_2016 | Project 1 Problem 2 R code.R | data_march_2018
Source on Save | Run | Source

1 # PROJECT-1 # PROBLEM-1
2
3 # readr package installed
4 install.packages("readr")
5
6 # readr package called
7 library(readr)
8
9 # working directory mentioned and called
10 setwd("E:/from old hard disc 24.08.2019/data Analytics and Data science/Great Lakes PGP-BABI/2. Statistical Methods for Decision Making (SMDM)/PROJECT-1 (Cold Storage Case Study)")
11
12 # working directory viewed OK.
13 getwd()
14
15 #The csv file 'cold_storage_Temp_Data.csv' is stored in the table named 'data_2016'.
16 data_2016 <- read.csv("cold_storage_Temp_Data.csv", header = TRUE)
17
18 # 'data_2016' viewed in a tabular form
19 view(data_2016)
20 class(data_2016) # Checking the class of the vector which is a data.frame
21 class(data_2016$Season)
22 class(data_2016$Month)
23 class(data_2016$Date)
24 class(data_2016$Temperature)
25
26 # structure of the vector
27 str(data_2016)
28
29 # summary of data_2016
30 summary(data_2016)
31
32 # we now find the overall mean of the full year in terms of temperature
33 # which is equal to 2.963 deg C that we get from the console.
34
35 # in another way we can find the overall mean by using the attach function and the mean function
36 ?attach()
37 attach(data_2016)
38 Average_temp_2016 <- mean(Temperature)
39 Average_temp_2016 # average found for the full year = 2.96274 deg C (Answer to question no.2)
40
41 # now we find the standard deviation of temperature maintained
42 Standard_deviation <- sd(Temperature)
43 Standard_deviation # standard deviation for the full year found = 0.508589 deg C (Answer to question no.3)
44
```

```
44
45 # now seperating the dataset into 3 different datasets for winter,rainy and summer and viewing the data frame
46
47 winter_data_2016 <- data_2016[data_2016$Season=="winter",]
48 view(winter_data_2016)
49
50 summer_data_2016 <- data_2016[data_2016$Season=="Summer",]
51 view(summer_data_2016)
52
53 rainy_data_2016 <- data_2016[data_2016$Season=="Rainy",]
54 view(rainy_data_2016)
55
56 # mean temperature for summer,rainy and winer
57 Average_temp_winter_2016 <- mean(winter_data_2016$Temperature)
58 Average_temp_winter_2016 # average found for winter 2016 = 2.700813 deg C (Answer to question no.1)
59
60 Average_temp_rainy_2016 <- mean(rainy_data_2016$Temperature)
61 Average_temp_rainy_2016 # average found for rainy 2016 = 3.039344 deg C (Answer to question no.1)
62
63 Average_temp_summer_2016 <- mean(summer_data_2016$Temperature)
64 Average_temp_summer_2016 # average found for summer 2016 = 3.039344 deg C (Answer to question no.1)
65
66 # finding the probability of the temperature having fallen below 2 deg C (Assuming Normal Distribution)
67 Prob_T_less_than_2 <- pnorm(2, mean=Average_temp_2016 , sd=Standard_deviation)
68 Prob_T_less_than_2 # probability found around 2.9% (Answer to question no. 4)
69
70 # finding the probability of the temperature having gone above 4 deg C (Assuming Normal Distribution)
71 Prob_T_more_than_4 <- 1-pnorm(4, mean=Average_temp_2016 , sd=Standard_deviation)
72 Prob_T_more_than_4 # probability found around 2.07% (Answer to question no. 5)
73
```

## 6. Exploratory Data Analysis – Step by step approach(Problem-2)

The second problem for the year 2018 leads us to a dataset “Cold\_Storage\_Mar2018.csv” which also indicates a four variable .csv data with their elements collected in continuous observations for the last 35 days by the supervisor. The supervisor has received complaints from their clients who moreover has received complaints from the end users that their dairy products were getting sour and often smelling. The supervisor put in an immediate action to pull out the data of temperature regulation for the last 35 days to evaluate statistically the probability of sustaining a malfunction or error at the plant. The Supervisor is vigilant to maintain a temperature below 3.9 deg C. Keeping the upper acceptable limit at 3.9 deg C we need to check that if some corrective action is needed or there is a problem from the side from where Cold Storage procures the products. The variables are  $c \leftarrow (\text{Season, Month, Date \& Temperature})$ . The vector within this dataset that we will analyse would be the Temperature Vector since we need to find a solution to the second problem. We will perform hypothesis testing to find out a probable solution to this problem. A illustrated example of the dataset is shown below for a formative reference:



	A	B	C	D
	Season	Month	Date	Temperature
1	Summer	Feb	11	4
2	Summer	Feb	12	3.9
3	Summer	Feb	13	3.9
4	Summer	Feb	14	4
5	Summer	Feb	15	3.8
6	Summer	Feb	16	4
7	Summer	Feb	17	4.1
8	Summer	Feb	18	4
9	Summer	Feb	19	3.8
10	Summer	Feb	20	3.9
11	Summer	Feb	21	3.9
12	Summer	Feb	22	4.6
13	Summer	Feb	23	4.1
14	Summer	Feb	24	4.1
15	Summer	Feb	25	3.9
16	Summer	Feb	26	3.8
17	Summer	Feb	27	3.8
18	Summer	Feb	28	3.9
19	Summer	Mar	1	3.9
20	Summer	Mar	2	3.9
21	Summer	Mar	3	3.9



## 7.1. Environment Set up and Data Import:

### 7.1.1. Installation of necessary Packages and Invoking Libraries:

A new R Script is chosen to perform the analysis named Project-1 Problem-2. Installation of various R packages which will assist us during our analysis is required to apply the functions we want to perform on different variables, vectors and lists of the dataset. We installed the below mentioned packages:

**install.packages("readr")** → for invoking libraries to read the .csv file into R.  
**install.packages("ggplot2")** → for invoking libraries to project graphical plots into R.

**library(readr)** → for invoking the library which contains the functions to read a .csv file into R.

**library(ggplot2)** → for invoking the library which contains the functions to plot graphical representation of the dataset into R.

### 7.1.2. Set up working Directory:

The working directory has to be set from where the dataset has to be incorporated into R. The programme needs to know from where it needs to collect the .csv file saved at a random drive. Our .csv was saved in some folder in the E drive. So we assign the function:

**setwd("E:/from old hard disc 24.08.2019/Data Analytics and Data Science/Great Lakes PGP-BABI/2. Statistical Methods for Decision Making (SMDM)/PROJECT-1 (Cold Storage Case Study)")**

Note that all the back-slashes in the address has been converted to front slashes as R will not recognise the working directory otherwise.

We can also view the working directory by calling the function `getwd()`.

### 7.1.3. Import and Read the Dataset:

Now the dataset was imported and read by simultaneously storing it in a vector → `data_march_2018` using the function `read.csv`



The syntax used over here is:

```
data_march_2018 <- read.csv("Cold_Storage_Mar2018.csv", header  
= TRUE)
```

Here, `Cold_Storage_Mar2018.csv` is the .csv file and `header = TRUE` is to keep the header containing the variable names into the `data_march_2018` vector.

### 8.1. Statistical Assumptions:

Now from the provided data set we need to evaluate that do we need to introduce some corrective measures into Cold Storage's temperature maintenance activity or are they already procuring expired stock. We would like to believe that Cold Storage as a plant is maintaining the temperatures needed to keep their stock fresh. So we hypothesize that the temperature that is being maintained by the supervisor is TRUE (i.e  $< 3.9$  deg C) which in turn keeps the stock fresh. The level of significance has been observed as 0.1 or = 10%.

### 8.2. Hypothesis Testing via Z-Test Method:

Out of the total temperature data of 2018 from Cold Storage the supervisor had to collect the Temperature data from the date of immediate complaint. Thus we can compute and assign our hypothesis on the population mean =  $\mu$

Sample taken by the supervisor from the population =  $n$   
Where,  $n = 35$  (since data for 35 days taken)

$\alpha$  = Level of Significance = 0.1 or 10%

Now we have a clear picture of our hypothesis here:

Null Hypothesis which we want to reject =  $H_0: \mu = 3.9$  deg C

Alternative Hypothesis which we want to accept =  $H_a: \mu < 3.9$  deg C.

### 8.3. Summary of the dataset:

When we invoke the `summary()` function then we get the following observations:

```
> summary(data_march_2018)
   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
> |
```

We have 35 Observations for summer 18 in the month of February and 17 in the month of march. Mean of the temperature within this 35 day period is 3.97 deg C.

### 8.4. Calculating Mean:

We as per our given data store 35 into `n` which is defined as the sample size since the supervisor has decided to take the temperature data of 35 days. After using the `attach` function we store the average of the temperature in the past 35 days in `X_bar` using the `mean()` function. We find `mean = 3.974286` deg C. The syntax is provided below:

```
# we can find the mean of the last 35 days by using the attach function and the mean function
attach(data_march_2018)
X_bar <- mean(Temperature)
X_bar # average found for the last 35 days or the sample = 3.974286 deg C
```

We use the Standard Deviation `sd = 0.508589` deg C from the previous problem as stated.

```
sd <- 0.508589
```

Let us assume that the population mean or the point at which our hypothesis has to be checked be  $\mu = \text{Mu} = 3.9$  deg C.

```
Mu <- 3.9
```

We will compute the Z score for  $\mu$ :

$$Z_{\text{computed}} = (\bar{x} - \mu) / (\sigma / n^{0.5})$$

### Syntax used in R:

# now performing hypothesis testing (t-test) we consider  $H_0: \mu = 3.9 \text{ deg C}$ ,  $H_a: \mu \text{ not } < 3.9 \text{ deg C}$

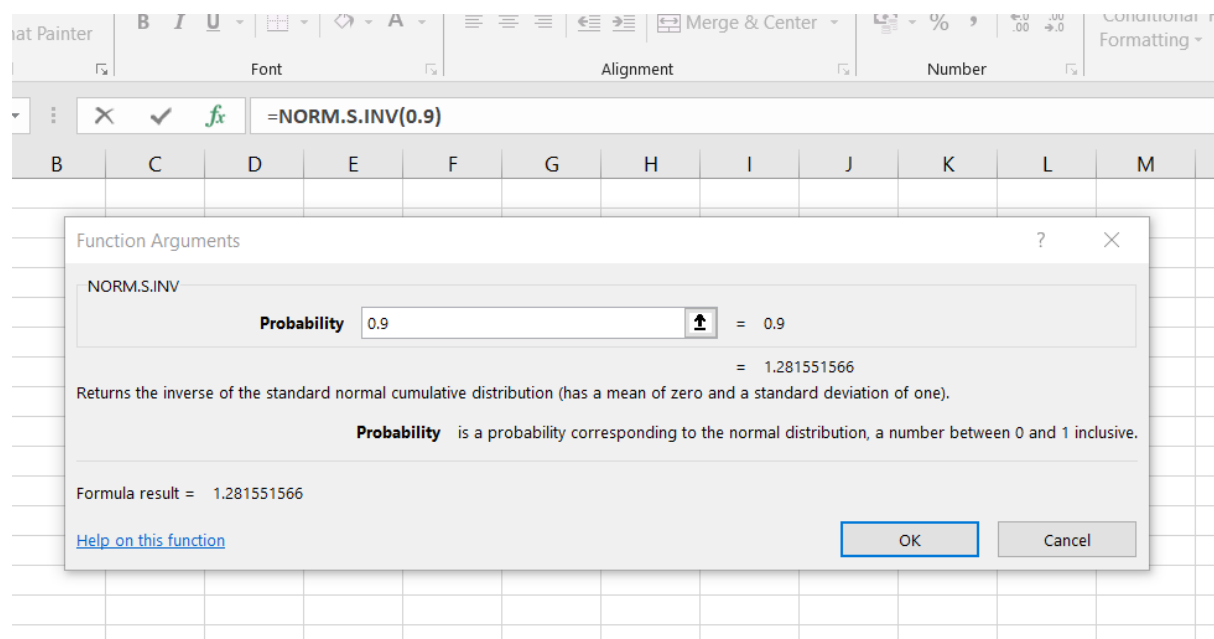
```
tstat <- (X_bar-Mu)/(sd/(n^0.5))  
tstat # t test statistics done and computed
```

We store the Z score in tstat as Z Computed which is = 0.8641166  
Now since our level of significance is = 0.1 therefore the probability of no type -I error will be  $1 - 0.1 = 0.9$  i.e 90% confidence level.

So now we compute Z Critical from excel putting probability = 0.9.  
We invoke the NORM.S.INV() function in excel to compute the Z Critical.

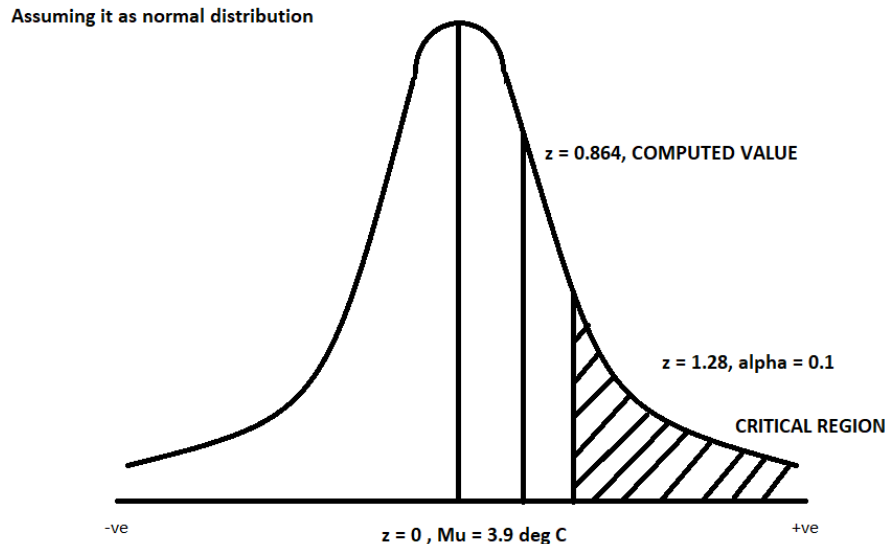
### NORM.S.INV(0.9)

Which is equal to 1.281552



Now we can clearly see that Z Critical > Z Computed

## 8.5. Bell Curve Diagram Reference:



Therefore, the Z Computed falls clearly outside the upper limit of the CRITICAL REGION. Therefore we have to accept the Null Hypothesis.

## 8.6. Hypothesis Testing via t-Test Method:

We can now construct a t-test model to confirm our study for this we need to find the P Value of the variable.

P -Value also known as the Actual Significance Level.

It signifies the actual risk or level of significance by which the null hypothesis is or might be rejected.

If  $P < \alpha$  then the Null Hypothesis is rejected.

### We compute the P-Value in R as:

```
# now we will prepare P-Value to check our hypothesis for upper acceptable temperature  
(one tailed) using pt() function  
P_value <- pt(tstat,34) # (n-1) = Degrees of freedom = 34 since, n = 35  
P_value
```

We use the pt() function to compute the P-Value  
`P_value <- pt(Z Computed,Degrees of freedom)`

Degrees of Freedom = (n-1)

Since it is a 1 tailed test we do not require to calculate for the other tail. Here P value = 0.8032103.

We find P value >  $\alpha$  thus accepting the Null Hypothesis.

```
#level of significance = alpha = 0.1 (already provided)
alpha <- 0.1
```

```
P_value > alpha
```

```
# Here we find that our P-value is > alpha or the level of significance (0.8032103 > 0.1)
# Thus we have to accept the Null Hypothesis (H0)
```

We can also check this with invoking the t.test() function on the temperature of 35 days.

### Syntax:

```
# Performing T-Test on Temperature
t.test(Temperature, mu = 3.9, alternative = "less", conf.level = 0.9)
```

Here Temperature is the variable in which we want to perform t-test, mu is the mean where our Null Hypothesis would be 0, its alternative will be the opposite i.e. less than 3.9 deg C and we have to calculate it on the confidence level of 90% since,  $\alpha = 0.1$ . We use the conf.level = to assign the confidence level.

### In console:

```
> # Performing T-Test on Temperature
> t.test(Temperature, mu = 3.9, alternative = "less", conf.level = 0.9)

One Sample t-test

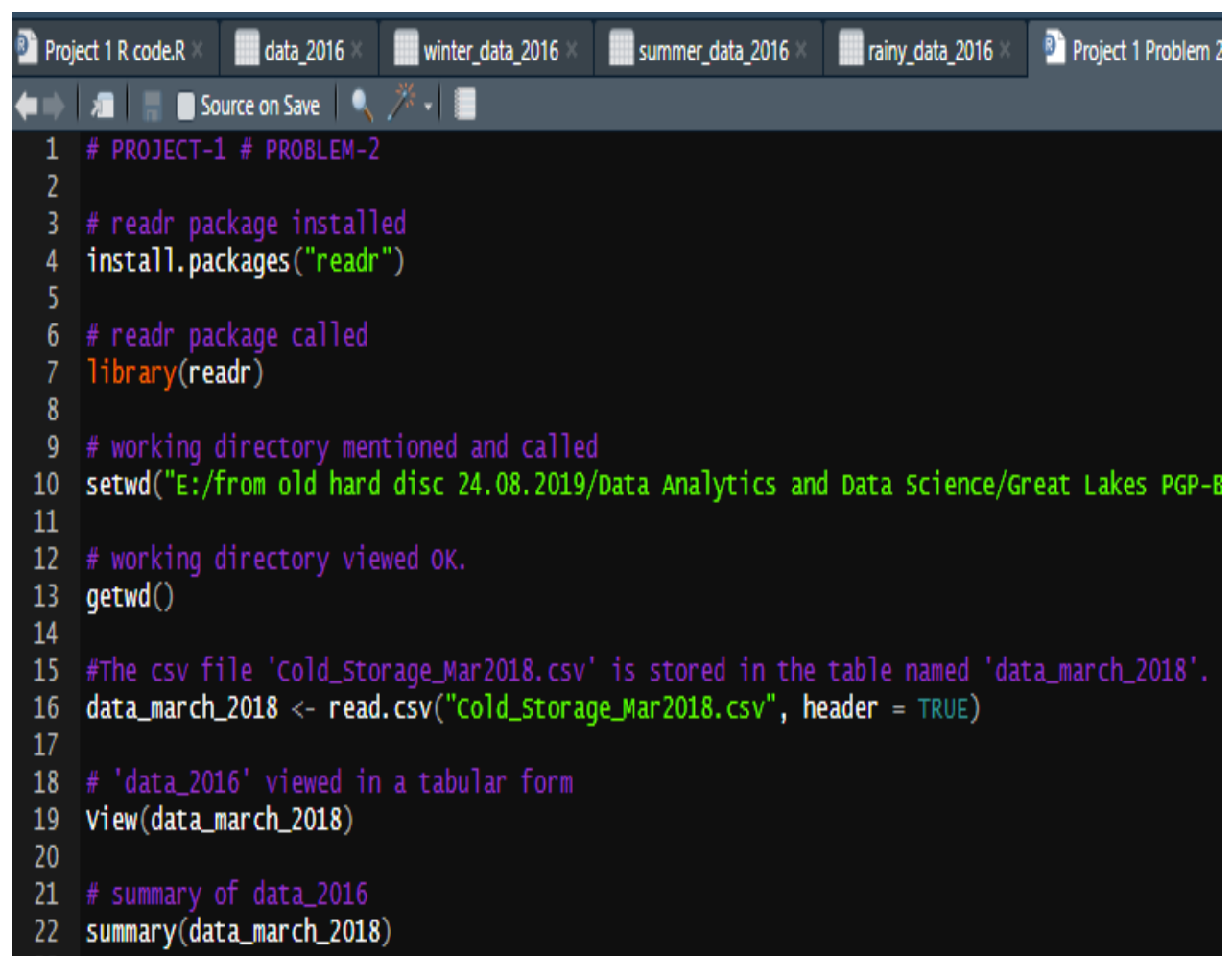
data:  Temperature
t = 2.7524, df = 34, p-value = 0.9953
alternative hypothesis: true mean is less than 3.9
90 percent confidence interval:
 -Inf 4.00956
sample estimates:
mean of x
 3.974286
```

## 8.7. Conclusion and inferences:

Therefore, from evidences collected from assumption of the null and alternative hypothesis the test is significant and Cold Storage needs to take necessary corrective actions at their plant for temperature regulation and maintenance. Our Null hypothesis has been accepted after applying both Z-test and t-test on the data of the past 35 days and has a P value greater than the level of significance. These implies that Cold Storage has not been checking their temperature maintenance module which has lead to the supply of rotten dairy products to the end-user ultimately.

# Please refer source code in APPENDIX-B for this problem.

## 9. Appendix B – Source Code:



```
1 # PROJECT-1 # PROBLEM-2
2
3 # readr package installed
4 install.packages("readr")
5
6 # readr package called
7 library(readr)
8
9 # working directory mentioned and called
10 setwd("E:/from old hard disc 24.08.2019/Data Analytics and Data Science/Great Lakes PGP-B
11
12 # working directory viewed OK.
13 getwd()
14
15 #The csv file 'Cold_Storage_Mar2018.csv' is stored in the table named 'data_march_2018'.
16 data_march_2018 <- read.csv("Cold_Storage_Mar2018.csv", header = TRUE)
17
18 # 'data_2016' viewed in a tabular form
19 view(data_march_2018)
20
21 # summary of data_2016
22 summary(data_march_2018)
23
```

```

#now sample size is 35 since the data is of the last 35 days, therefore:
n <- 35

# we can find the mean of the last 35 days by using the attach function and the mean function
attach(data_march_2018)
X_bar <- mean(Temperature)
X_bar # average found for the last 35 days or the sample = 3.974286 deg C

#population standard deviation is to be considered = 0.508589 deg C from the first problem
sd <- 0.508589

# Taking Mu = 3.9 deg C
Mu <- 3.9

# now performing hypothesis testing (t-test) we consider H0:Mu = 3.9 deg C, Ha:Mu not < 3.9 deg C
tstat <- (X_bar-Mu)/(sd/(n^0.5))
tstat # t test statistics done and computed

# now we will prepare P-Value to check our hypothesis for upper acceptable temperature (one tailed) using pt() function
P_value <- pt(tstat,34) # (n-1) = Degrees of freedom = 34 since, n = 35
P_value

#level of significance = alpha = 0.1 (already provided)
alpha <- 0.1

P_value > alpha

# Here we find that our P-value is > alpha or the level of significance (0.8032103 > 0.1)
# Thus we have to accept the Null Hypothesis (H0)

# Performing T-Test on Temperature
t.test(Temperature, mu = 3.9, alternative = "less", conf.level = 0.9)

```