Project – Cold Storage Case Study

Feb 2020 Author: Snehal Gawand

Table of Contents

1. Project Objective

The objective of the report is to analyze the cold storage problems in R, generate insights about the problems via graphical explorations and derive solutions.

2. Assumptions

This temperature in the cold storage plant remains unaffected by the temperature and weather conditions outside the plant.

3 Problem 1 – Step by step approach

3.1 Environmental setup, dataset import and variable identification

- Import the dataset in R and set up working directories.
- Install packages required to read the csv file and performing graphical analysis.

The below listed functions are used to explore the variables used in the dataset

- dim → to check amount of data present in input dataset
- names → list down names of columns used in dataset
- str → view structure of the dataset
- summary → to check the summary of each column of dataset
- is.na → to check missing values in dataset

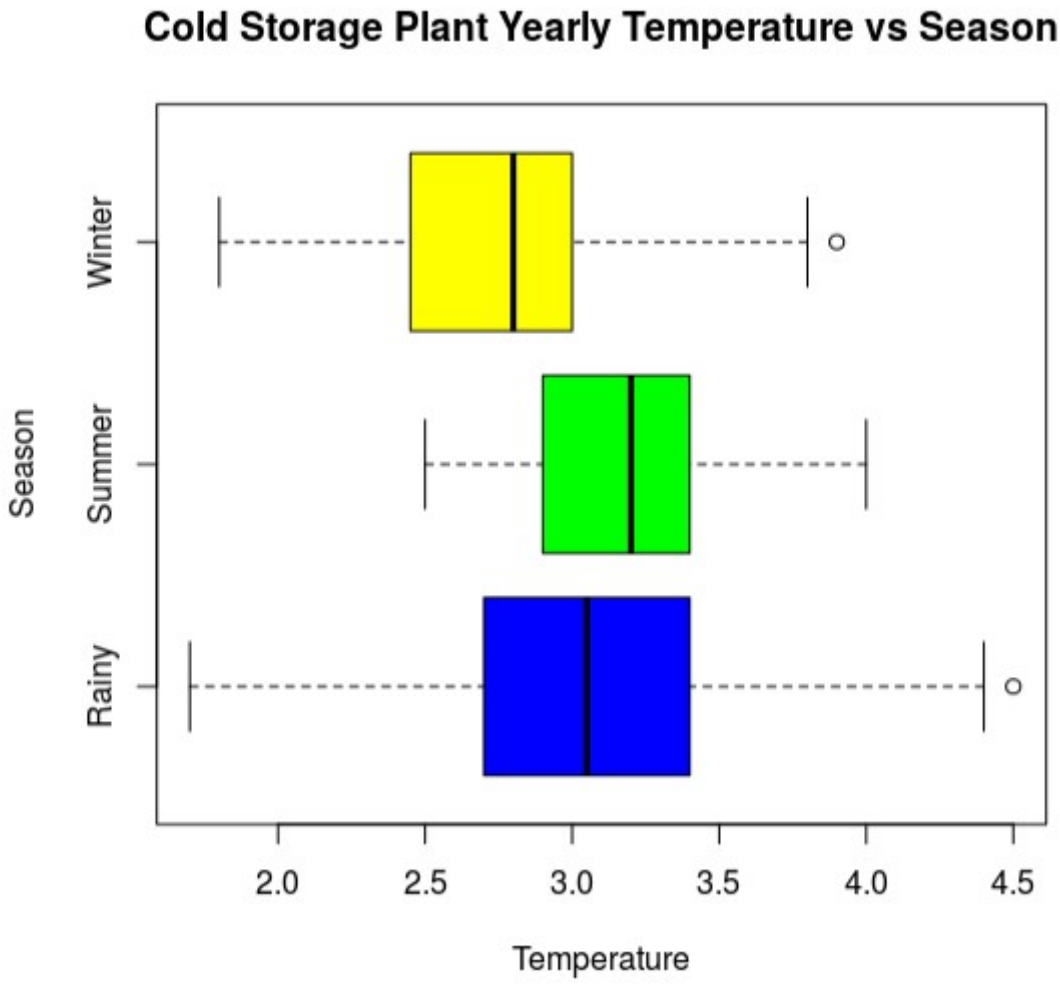
The result of these functions is described as follows:

- The dataset has the daily temperature of cold storage plant for entire year (365 days). The yearly data is divided into three categories as per season – Winter, Summer and Rainy
- The temperature recorded in dataset is in degree Celsius and consist of numeric value
- The minimum temperature observed in the cold storage plant during the year is 1 degree Celsius and maximum temperature is 4.5 degree Celsius
- There are no missing values in the dataset. All the columns have complete information about the temperature in cold storage plant.

Please refer Appendix A for Source Code.

3.2 Graphical Exploratory Analysis

Temperature of Cold storage plant during the Seasons – Summer, Winter and Rainy



The graph between temperature and seasons of the cold storage plant indicates the following:

- Rainy and Winter season has few outliers in the temperature
- The mean and median values of temperature varies across all the seasons
- The Minimum temperature observed across all seasons is around 2.5 degree Celsius whereas the maximum temperature is 4.5 degree Celsius.

Please refer Appendix A for Source Code.

3.3 Answers to cold storage problem-1 questions:

3.3.1 Mean cold storage temperature for Summer, Winter and Rainy Season

```
> by(data=Cold_Storage_Temp_Data$Temperature,INDICES = Cold_Storage_Temp_Da
ta$Season,FUN = mean)
Cold_Storage_Temp_Data$Season: Rainy
[1] 3.087705
------------------------------------------------------------
Cold_Storage_Temp_Data$Season: Summer
[1] 3.1475
------------------------------------------------------------
Cold_Storage_Temp_Data$Season: Winter
[1] 2.776423
```

3.3.2 Overall Mean for the full year

```
> mean_year = mean(Cold_Storage_Temp_Data$Temperature)
[1] 3.002466
```

3.3.3 Standard deviation for the full year

```
> sd_year = sd(Cold_Storage_Temp_Data$Temperature)
[1] 0.4658319
```

3.3.4 Assuming normal distribution, probability of temperature having fallen below 2 degree Celsius

As we have to check the probability of normal distribution for temperature below 2 degree, the lower tail would be considered as true.

The mean and standard deviation values would be used to calculate the probability.

```
> temp_min = 2
> prob_less_than_min = pnorm(q=temp_min,mean = mean_year, sd=sd_year,lower.tail = T
RUE)
> prob_less_than_min
[1] 0.01569906
```

The probability of temperature having fallen below 2 degree Celsius is 0.01569906

3.3.5 Assuming normal distribution, probability of temperature having gone above 4 degree Celsius

As we have to check the probability of normal distribution for temperature above 2 degree, the lower tail would be considered as false.

The mean and standard deviation values would be used to calculate the probability.

```
> temp_max = 4
> prob_more_than_max = pnorm(q=temp_max,mean = mean_year, sd=sd_year,lower.tail = F
ALSE)
> prob_more_than_max
[1] 0.01612075
```

The probability of temperature having gone above 4 degree Celsius is 0.01612075

3.3.6 Penalty for AMC company

Penalty would be charged by Cold Storage Plant to the outsourced professional company if, it was statistically proven that the probability of temperature is going outside the 2 - 4 Celsius range during the one-year contract.

If the temperature was above 2.5% and less than 5% then the penalty would be 10% of AMC (annual maintenance contract).

In case it exceeded 5% then the penalty would be 25% of the AMC fee.

i.e. if total probability > 0.025 and total probability < 0.05

then penalty = 10%

else if total probability > 0.05

then penalty = 10%

calculate total probability

```
> #calculate total probability
> total_prob = prob_less_than_min + prob_more_than_max
> total_prob
[1] 0.03181981
```

```
> if ( total_prob > 0.025 && total_prob < 0.05) {
+   print("Penalty is 10%")
+ } else if ( total_prob > 0.05) {
+   print("Penalty is 25%")
+ } else {
+   print("No penalty")
+ }
[1] "Penalty is 10%"
```

Hence, as per the data in Cold Storage dataset, the outsourced professional company would be charged 10% penalty

3.3.7 One-Way ANOVA test to determine if there is a significant difference in Cold Storage temperature between rainy, summer and winter seasons

Consider null hypothesis Ho = The mean of temperature in rainy, summer and winter season is same Hence the alternative hypothesis Ha = mean of temperature in rainy season != mean of temperature in summer season != mean of temperature in winter season

```
> aov_coldstorage = aov(Temperature~Season,data = Cold_Storage_Temp_Data)
>
> summary(aov_coldstorage)
             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
```

The p-value is 5.08e-11 i.e. 0.0000000000508.

The p-value is very less than the standard significance level alpha (0.05).

Hence, it strongly leads to the conclusion to reject the null hypothesis that the mean of temperature in cold storage across rainy, winter and summer seasons is the same.

4 Problem 2 – Step by step approach

4.1 Environmental setup, dataset import and variable identification

- Import the dataset in R and set up working directories.
- Install packages required to read the csv file and performing graphical analysis.

The below listed functions are used to explore the variables used in the dataset

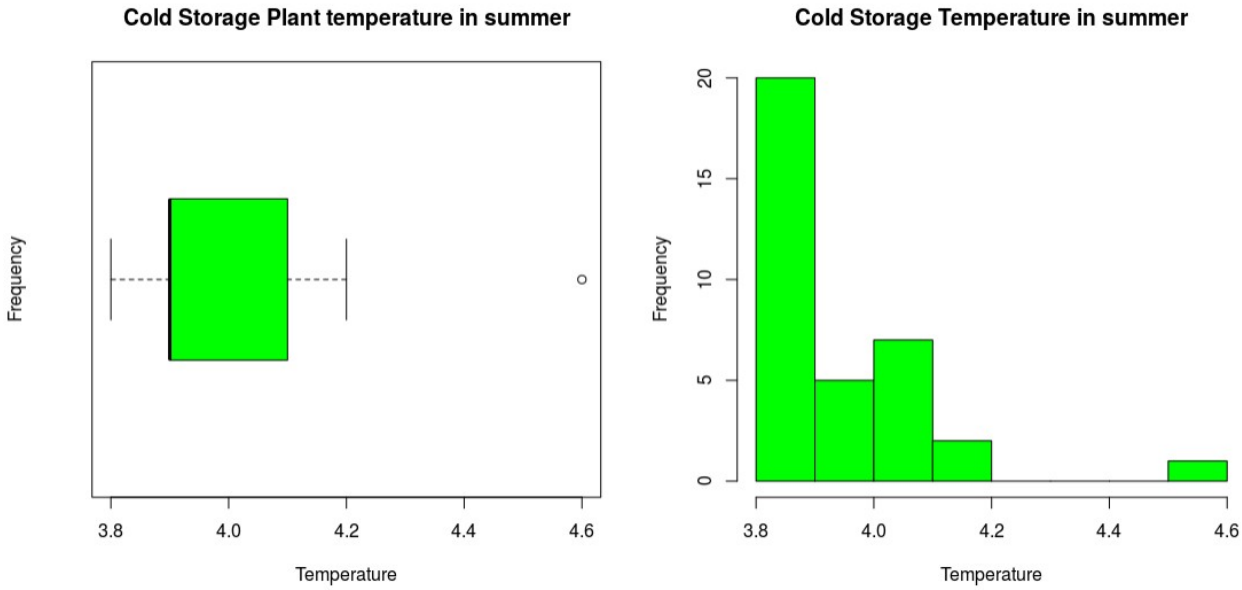
- dim → to check amount of data present in input dataset
- names → list down names of columns used in dataset
- str → view structure of the dataset
- summary → to check the summary of each column of dataset
- is.na → to check missing values in dataset

The result of these functions is described as follows:

- The dataset has the daily temperature of cold storage plant summer season, i.e. 2 months – Feb and March.
- The dataset contains total 35 observations and the minimum temperature observed is 3.8 degree Celsius and max temp is 4.6 degree Celsius.
- There are no missing values in the dataset. All the columns have complete information about the temperature in cold storage during summer.

4.2 Graphical Exploratory Analysis

Temperature of Cold Storage Plant during summer



The graph illustrates the following observations:

○ Summer season has few outliers in the temperature towards the right ○ The

temperature during summer has never been in the range of 4.2 to 4.5

4.3 Identify the type of hypothesis test to be performed to check if corrective action is needed at cold storage plant

Z-tests are statistical calculations that can be used to compare population means to a sample's. However, ztest requires population standard deviation and we do not have that value. So, we cannot perform Z-test.

T-tests are calculations used to test a hypothesis, but they are most useful when the population parameters (mean and SD are not known). Also t-test can be performed on small sample size, which fits the sample in Problem-2.

Hence, we can perform one sample t-test which will perform hypothesis of single group against unknown mean.

4.4 Hypothesis test and determine p-value

Hypothesis for t-test

Consider the null hypothesis Ho → Temperature in Cold Storage Plant is below or equal to 3.9 degree

Celsius, so no corrective action required i.e. Muo <= 3.9

Then alternative hypothesis would be Ha → Temperature in Cold Storage Plant exceeds 3.9 degree Celsius,

So corrective actions are required i.e. Muo > 3.9

As the alternative hypothesis is greater than null hypothesis value, we will perform the right-tailed t-test.

We assume 3. .9 C as the upper acceptable value for mean temperature and at alpha = 0.1.

Hence mean = 3.9

And alpha = 0.1

```
> t.test(Cold_Storage_Mar2018$Temperature,mu=3.9,alternative="greater")
        One Sample t-test
 data:
Cold_Storage_Mar2018$Temperature   t   =
2.7524, df = 34, p-value = 0.004711
alternative hypothesis: true mean is greater than 3.9 95
percent confidence interval:
 3.928648      Inf sample
estimates:
mean of x   3.974286
```

The P-value 0.004 is less than the alpha 0.1.

Hence, we reject the Null Hypothesis that temperature of the cold storage plant is less than or equal to 3.9 degree Celsius.

4.5 Inference

T-tests are calculations used to test a hypothesis, but they are most useful when the population parameters (mean and SD are not known). Also t-test can be performed on small sample size, which fits the sample in Problem-2.Hence, we performed one sample t-test.

According to the hypothesis test performed via t-test, we conclude that the temperature in cold storage plant has exceeded 3.9 degree Celsius. Hence, some corrective action is required in cold storage plant.

4. Appendix A – Source Code

```
#########################################################################
############                SOURCE CODE  FOR PROBLEM 1          ##################
#########################################################################

#install packages and invoke libraries
install.packages("readr") library("readr")

install.packages("ggplot2") library("ggplot2")

install.packages("corrplot") library("corrplot")

#setup working directory
setwd = ('/Users/snehal/Documents/Projects/Project 2') getwd()

#read input dataset
Cold_Storage_Temp_Data <- read_csv("Cold_Storage_Temp_Data.csv")

#variable Identification
#Check how much data is present in input dataset
> dim(Cold_Storage_Temp_Data)
[1] 365   4


#List down the variable names in dataset
> names(Cold_Storage_Temp_Data)
[1] "Season"      "Month"       "Date"        "Temperature"



#check  structure
> str(Cold_Storage_Temp_Data)
Classes ‘spec_tbl_df’, ‘tbl_df’, ‘tbl’ and 'data.frame':        365 obs. of  4 variables:
 $ Season     : chr  "Winter" "Winter" "Winter" "Winter" ...
 $ Month      : chr  "Jan" "Jan" "Jan" "Jan" ...
 $ Date       : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Temperature: num  2.3 2.2 2.4 2.8 2.5 2.4 2.8 3 2.4 2.9 ...
 - attr(*, "spec")=
  .. cols(
  ..   Season = col_character(),
  ..   Month = col_character(),
  ..   Date = col_double(),
  ..   Temperature = col_double()
  .. )
```

#summary statistics

```
> summary(Cold_Storage_Temp_Data)
    Season             Month                Date        Temperature
 Length:365         Length:365         Min.   : 1.00   Min.   :1.700
 Class :character   Class :character   1st Qu.: 8.00   1st Qu.:2.700
 Mode  :character   Mode  :character   Median :16.00   Median :3.000
                                       Mean   :15.72   Mean   :3.002
                                       3rd Qu.:23.00   3rd Qu.:3.300
                                       Max.   :31.00   Max.   :4.500
```

#check null values

```
> is.na(Cold_Storage_Temp_Data)
```

#Graphical analysis

#Convert the character variable 'Temperature' to factor for plotting graph

```
> Cold_Storage_Temp_Data$Season = as.factor(Cold_Storage_Temp_Data$Season)
```

#Create box plot between season and temperature

```
> plot(Cold_Storage_Temp_Data$Season, Cold_Storage_Temp_Data$Temperature, horizontal=TRUE,
col=c("Blue","Green","Yellow"), main='Cold Storage Plant Yearly Temperature vs Season')
```

##

############ SOURCE CODE FOR PROBLEM 2 ##################

##

#install packages and invoke libraries

#read input dataset

Cold_Storage_Mar2018<- read_csv("Cold_Storage_Mar2018.csv"

#variable Identification

#Check how much data is present in input dataset

```
> dim(Cold_Storage_Mar2018)
[1] 35  4
```

#List down the variable names in dataset

```
> names(Cold_Storage_Mar2018)
[1] "Season"      "Month"       "Date"        "Temperature"
```

#check structure

```
> str(Cold_Storage_Mar2018)
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':        35 obs. of  4 variables:
 $ Season     : chr  "Summer" "Summer" "Summer" "Summer" ...
 $ Month      : chr  "Feb" "Feb" "Feb" "Feb" ...
 $ Date       : num  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 ...
 - attr(*, "spec")=
  .. cols(
  ..   Season = col_character(),
  ..   Month = col_character(),
```

```
..   Date = col_double(),
..   Temperature = col_double()
.. )
```

#summary statistics

```
> summary(Cold_Storage_Mar2018)
    Season             Month                Date       Temperature
 Length:35          Length:35          Min.   : 1.0   Min.   :3.800
 Class :character   Class :character   1st Qu.: 9.5   1st Qu.:3.900
 Mode  :character   Mode  :character   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
```

#check null values

```
> is.na(Cold_Storage_Mar2018)
```

#Graphical analysis >

```
par(mfrow=c(1,2))
> boxplot(Cold_Storage_Mar2018$Temperature,main='Cold Storage Plant temperature in
summer', xlab = "Temperature", ylab = "Frequency",col = "green",horizontal = TRUE)
> hist(Cold_Storage_Mar2018$Temperature,main='Cold Storage Temperature in summer',xlab =
"T emperature", ylab = "Frequency",col = "green")

##########################################################################
############                    END OF CODE                  ##################
##########################################################################
```