Mini Project – Cold Storage

Submission Date: August 30, 2019

Author: Ayush Jain
Mentor: Deepak Gupta

Table of Content

Contents

1 Project Objective 3

2 Assumptions 3

3 Exploratory Data Analysis – Step by step approach 3

3.1 Environment Set up and Data Import 4

3.1.1 Set up working Directory 4

3.1.2 Import and Read the Dataset 4

3.2 Variable Identification 4

4 Conclusion 5

5 Appendix A – Source Code 10

1 Project Objective

The objective of the report is to explore the Cold Storage data set (“Cold_Storage_Temp_Data.csv”) in R and identify whether the temperature of the Plant is maintained as per the Contract (between 2-degree C to 4-degree C). In case of Temperature going outside of 2-degree C and 4-degree C the Company must bear the penalty of 10% when it is within the range of 2.5% and 5%(inclusive) and 25% of the total Annual Maintenance Cost if it is more than 5%.

We will also be deriving the results on the below

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

In Mar 2018, Cold Storage started getting complaints from their Clients that the dairy products going sour and often smelling. Using the Hypothesis Testing (z-Test and t-Test) we need to Identify if any corrective measure is required in the Cold Storage Plant or the problem is from procurement side from where Cold Storage is getting the Dairy Products. For this purpose, we will take the sample data for last 35 days in the csv format (“Cold_Storage_Mar2018.csv”).

2 Assumptions

During the working, we will be assuming that the data is Normally Distributed achieving the Bell Curve.

3 Exploratory Data Analysis – Step by step approach

A Typical Data exploration activity consists of the following steps:

- Environment Set up and Data Import.
- Variable Identification.

We shall follow these steps in exploring the provided dataset.

3.1 Environment Set up and Data Import

3.1.1 Set up working Directory

Setting a working directory on starting of the R session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project.

```
### Environment Setup and data Import.

# Setting up working Directory.

setwd("C:/Users/290018451/Desktop/Personal")
```

Please refer Appendix A for Source Code.

3.1.2 Import and Read the Dataset

The given dataset is in .csv format. Hence, the command 'read.csv' is used for importing the file.

```
# Reading dataset in R.

ColdStorage_DS <- read.csv('Cold_Storage_Temp_Data.csv')
```

Please refer Appendix A for Source Code.

3.2 Variable Identification

This section holds the Variables/ Methods that are used during the Analysis of the problem. Below are the Functions that we have used for the Analysis.

- setwd(): setwd(dir) is used to set the working directory to dir.
- read.csv(): Reads a file in table format and creates a data frame from it.
- head(): Returns the first parts of a vector, matrix, table, data frame or function.
- str(): Compactly display the internal Structure of an R object.
- summary(): summary is a generic function used to produce result summaries of the results of various model fitting functions.
- mean(): Generic function for the (trimmed) arithmetic mean.
- sd(): This function computes the standard deviation of the values in x.
- pnorm(): Density, distribution function, quantile function and random generation for the normal distribution with mean equal to mean and standard deviation equal to sd.
- count(): Provides the count of the variable.
- t.test(): Performs one and two sample t-tests on vectors of data.

4 Conclusion

Problem 1

I. Find mean cold storage temperature for Summer, Winter and Rainy Season.

Solution:

We created 3 different Datasets namely ColdStorage_summer, ColdStorage_winter, ColdStorage_rainy that contains data for Summer, Winter and Rainy season from the ColdStorage_DS dataset. Using the function mean() we calculated the mean for each dataset.

```
#Subsetting the dataset for Summer Season and finding the mean.
ColdStorage_summer <- ColdStorage_DS[ColdStorage_DS$Season =="Summer",]
mean_summer <- mean(ColdStorage_summer$Temperature)
mean_summer

## [1] 3.153333

#Subsetting the dataset for Winter Season and finding the mean.
ColdStorage_winter <- ColdStorage_DS[ColdStorage_DS$Season =="Winter",]
mean_winter <- mean(ColdStorage_winter$Temperature)
mean_winter

## [1] 2.700813

#Subsetting the dataset for Rainy Season and finding the mean.
ColdStorage_rainy <- ColdStorage_DS[ColdStorage_DS$Season == "Rainy",]
mean_rainy <- mean(ColdStorage_rainy$Temperature)
mean_rainy

## [1] 3.039344
```

II. Find overall mean for the full year.

Solution:

Here we used the mean() for the dataset ColdStorage_DS on Temperature Variable and calculated the mean value for the population.

```
# Total mean for an Year.
mean_total <- mean(ColdStorage_DS$Temperature)
mean_total

## [1] 2.96274
```

III. Find Standard Deviation for the full year.

Solution:

Standard Deviation of the total Population is achieved using the method sd() on the ColdStorage_DS dataset.

```
# Finding the Standard Deviation.
stdv <- sd(ColdStorage_DS$Temperature)
stdv

## [1] 0.508589
```

IV. Assume Normal distribution, what is the probability of temperature having fallen below 2 deg C?

Solution:

The probability of temperature falling below 2 deg C is calculated using the method pnorm(). There are 2.9% chances of temperature falling below 2 deg C given that the mean of the population is 2.9627 with a Standard Deviation of 0.5085

```
# Probability of Temprature less then 2.
p_temp_less_2 <- pnorm(2,mean_total,stdv,lower.tail = TRUE)
p_temp_less_2

## [1] 0.02918146
```

V. Assume Normal distribution, what is the probability of temperature having gone above 4 deg C?

Solution:

The probability of temperature falling below 4 deg C is calculated using the method pnorm(). There are 2.07% chances of temperature going above 4 deg C given that the mean of the population is 2.9627 with a Standard Deviation of 0.5085

```
# Probability of Temprature greater then 4.
p_temp_gtr_4 <- pnorm(4,mean_total,stdv,lower.tail = FALSE)
p_temp_gtr_4

## [1] 0.02070077
```

VI. What will be the penalty for the AMC Company?

Solution:

Given, if the probability of temperature going outside of 2 deg C – 4 deg C is above 2.5% and less then 5% then the Company will be charged with the penalty of 10% of AMC(Annual Maintenance Cost). In case it exceeds 5% then the company would be charged as 25% of AMC.

To identify the penalty, we will calculate the probability of temperature outside of 2-4 deg C.

This is achieved by adding the Probability calculated for Temperature less than 2 and Temperature greater than 4 in the above section.

```
# Penalty of the Company
# Since the events are Mutually Exclusive we will sum the probability for
less then 2 and greater then 4
Total_Probability = p_temp_less_2 + p_temp_gtr_4
Total_Probability

## [1] 0.04988223
```

The calculated variation_percent is 4.98 which is within the range of 2.5 and 5, hence we can conclude that the Company will be charged with 10% of the Total AMC.

Problem 2

I. State the Hypothesis, do the calculation using z test.

Solution:

Here we are taking the sample data for last 35 days for the Analysis of the temperature maintained in the Plant to take the necessary actions.

We assume the Null Hypothesis(H0) is equal to 3.9 and Alternative Hypothesis(H1) greater then 3.9

It is given the Mean of the Total Population(mu) is 3.9 with Standard Deviation(sigma) 0.5085 and alpha as 0.1.

We calculated the mean of the Sample data using mean() which is 3.9742.

```
> X_Bar = mean(coldStorage_35D$Temperature)
> X_Bar
[1] 3.974286
```

We now calculate the z value for test by substituting the values in the below equation.

Z= (X_Bar - mu)/(sigma/sqrt(n))

The calculated z value is 0.8642.

Using norm.s.inv in Excel we get the z_critical as 1.2815 at a 90% confidence level.

From above we identified that, Z calculated is less then Z Critical hence Null Hypothesis is Accepted, rejecting the Alternative Hypothesis.

```
X_Bar = 3.974
mu = 3.9
sigma = 0.5085
n = 35
alpha = 0.1
z_critical = 1.2815

#====================== Z Test Hypothesis ==================================
=#

###H0 is Less then or Eq 3.9 and H1 is greater 3.9

z <- (X_Bar - mu)/(sigma/(n^0.5))
z

## [1] 0.8609438

x_bar_critical = z_critical*(sigma/(n^0.5)) + mu

# We calculated that z_critical = 1.2815 and got x_bar_critical = 4.0101, whi
ch is
# greater then z = 0.8609 hence Null Hypothesis (H0) is accepted and
# Alternative Hypothesis is rejected.
```

Please refer Appendix A for Source Code.

II. State the Hypothesis, do the calculation using t-test.

Solution:

Assuming, Null Hypothesis(H0) equal to 3.9 and Alternative Hypothesis(H1) greater then 3.9 we now perform the t-Test for the above problem given the Population Mean(mu) = 3.9 at a confidence level of 90%

```r
#====================== T Test Hypothesis ==================================
==#

#H0 is Less then or Eq 3.9 and H1 is greater than 3.9

t.test(coldStorage_35D$Temperature,alternative = "greater",mu = mu,conf.level
= 0.90)

##
##  One Sample t-test
##
## data:  coldStorage_35D$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

# Since t = 2.7524 and p-value = 0.004711 which is smaller then alpha =0.1 he
nce the
# Null Hypothesis(H0) is rejected and Alternative Hypothesis(H1) is accepted.
```

We got t_stats = 2.7524 and p-value of 0.00471 at the confidence level of 90% which is smaller then the alpha, hence Accepting the Alternative Hypothesis (H1) and rejecting the Null Hypothesis(H0)

III. Give your inference after doing both the tests.

Solution:

Performing the Hypothesis testing using z-Test and t-Test we found that this is a contradictory statement as in the z-Test we accept the Null Hypothesis while in t-Test we are accepting the Alternative Hypothesis.

Here we choose z-test over t-test as the power of the t-Test is 61.1% with n as 35 which is too low to be considered. To attain the power as 90% we at least need n as 80. Hence rejecting the t-Test due to insufficient number of observations.

```
sdv = sd(coldStorage_35D$Temperature)
delta <- X_Bar - 3.9

power.t.test(n=35,delta = delta, sd = sdv, alternative = "one.sided")

##
##      Two-sample t test power calculation
##
##               n = 35
##           delta = 0.07428571
##              sd = 0.159674
##       sig.level = 0.05
##           power = 0.6110049
##     alternative = one.sided
##
## NOTE: n is number in *each* group

# Power to the t-test with n= 35, delta = 0.0742, sd = 0.1596 is 61.1%

power.t.test(power = .90,delta = delta, sd= sdv, alternative = "one.sided")

##
##      Two-sample t test power calculation
##
##               n = 79.81848
##           delta = 0.07428571
##              sd = 0.159674
##       sig.level = 0.05
##           power = 0.9
##     alternative = one.sided
##
## NOTE: n is number in *each* group

# To achieve the power as 90%, it is required to have the sample size of 80.
```

We can conclude that no Corrective Measures are needed in the Cold Storage Plant, it is the problem from the Procurement side from where Cold Storage is getting the Dairy Products.

5 Appendix A – Source Code

```
#==========================================================================#
#
 #
#
 #
# Exploratory Data Analysis - Cold Storage Problem
 #
#
 #
#
 #
#==========================================================================#

### Environment Setup and data Import.

# Setting up working Directory.
```

```r
setwd("C:/Users/290018451/Desktop/Personal")

# Reading dataset in R.

ColdStorage_DS <- read.csv('Cold_Storage_Temp_Data.csv')

# Looking at the top 10 observation.
head(ColdStorage_DS, n=10)

##    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
## 7  Winter   Jan    7         2.8
## 8  Winter   Jan    8         2.3
## 9  Winter   Jan    9         2.4
## 10 Winter   Jan   10         2.8

# Looking at the Structure of the dataset.
str(ColdStorage_DS)

## '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 of the dataset.
summary(ColdStorage_DS)

##     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

#Subsetting the dataset for Summer Season and finding the mean.
ColdStorage_summer <- ColdStorage_DS[ColdStorage_DS$Season =="Summer",]
mean_summer <- mean(ColdStorage_summer$Temperature)
mean_summer

## [1] 3.153333

#Subsetting the dataset for Winter Season and finding the mean.
ColdStorage_winter <- ColdStorage_DS[ColdStorage_DS$Season =="Winter",]
mean_winter <- mean(ColdStorage_winter$Temperature)
mean_winter
```

```
## [1] 2.700813

#Subsetting the dataset for Rainy Season and finding the mean.
ColdStorage_rainy <- ColdStorage_DS[ColdStorage_DS$Season == "Rainy",]
mean_rainy <- mean(ColdStorage_rainy$Temperature)
mean_rainy

## [1] 3.039344

# Total mean for an Year.
mean_total <- mean(ColdStorage_DS$Temperature)
mean_total

## [1] 2.96274

# Finding the Standard Deviation.
stdv <- sd(ColdStorage_DS$Temperature)
stdv

## [1] 0.508589

# Probability of Temprature less then 2.
p_temp_less_2 <- pnorm(2,mean_total,stdv,lower.tail = TRUE)
p_temp_less_2

## [1] 0.02918146

# Probability of Temprature greater than 4.
p_temp_gtr_4 <- pnorm(4,mean_total,stdv,lower.tail = FALSE)
p_temp_gtr_4

## [1] 0.02070077

# Penalty of the Company
# Since the events are Mutually Exclusive we will sum the probability for l
ess then 2 and greater than 4
Total_Probability = p_temp_less_2 + p_temp_gtr_4


#======================= END OF PROBLEM ===============================#

##====================== Start of Problem ==============================#

coldStorage_35D <- read.csv("Cold_Storage_Mar2018.csv")

head(coldStorage_35D,n =10)

##    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
## 7  Summer   Feb   17         4.1
## 8  Summer   Feb   18         4.0
```

```
## 9  Summer   Feb   19          3.8
## 10 Summer   Feb   20          3.9
```

```r
str(coldStorage_35D)
```

```
## '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 ...
```

```r
summary(coldStorage_35D)
```

```
##     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
```

```r
X_Bar = mean(coldStorage_35D$Temperature)
mu = 3.9
sigma = 0.5085
n = 35
alpha = 0.1
z_critical = 1.2815

#======================= Z Test Hypothesis ===============================#

###H0 is Less then or Eq 3.9 and H1 is greater 3.9

z <- (X_Bar - mu)/(sigma/(n^0.5))
z
```

```
## [1] 0.8642679
```

```r
x_bar_critical = z_critical*(sigma/(n^0.5)) + mu

# We calculated that z_critical = 1.2815 and got x_bar_critical = 4.0101, w
hich is
# greater then z = 0.8609 hence Null Hypothesis (H0) is accepted and
# Alternative Hypothesis is rejected.


#======================= T Test Hypothesis ===============================#

#H0 is Less then or Eq 3.9 and H1 is greater than 3.9

t.test(coldStorage_35D$Temperature,alternative = "greater",mu = mu,conf.lev
el = 0.90)
```

```
##
##  One Sample t-test
##
## data:  coldStorage_35D$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
```

```
# Since t = 2.7524 and p-value = 0.004711 which is smaller then alpha =0.1
hence the
# Null Hypothesis(H0) is rejected and Alternative Hypothesis(H1) is accepte
d.

sdv = sd(coldStorage_35D$Temperature)
delta <- X_Bar - 3.9

power.t.test(n=35,delta = delta, sd = sdv, alternative = "one.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 35
##           delta = 0.07428571
##              sd = 0.159674
##       sig.level = 0.05
##           power = 0.6110049
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

```
# Power to the t-test with n= 35, delta = 0.0742, sd = 0.1596 is 61.1%

power.t.test(power = .90,delta = delta, sd= sdv, alternative = "one.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 79.81848
##           delta = 0.07428571
##              sd = 0.159674
##       sig.level = 0.05
##           power = 0.9
##     alternative = one.sided
##
## NOTE: n is number in *each* group
```

```
# To achieve the power as 90%, it is required to have the sample size of 80
.

# Since the power of t-test with the sample size of 35 is 0.6110
# we consider z-Test over t-Test

#======================== END OF PROBLEM 2 ==============================#
```