

COLD STORAGE

CASE STUDY



August 29, 2019

SANKALP RAJ MATTA

sankalprajmatta@gmail.com

**TABLE OF CONTENT:**

1 Project Objective

2 Assumptions

3 Exploratory Data Analysis – Step by step approach

3.1 Environment Set up and Data Import

3.1.1 Install necessary Packages and Invoke Libraries

3.1.2 Set up working Directory

3.1.3 Import and Read the Dataset

3.2 Variable Identification

3.2.1 Variable Identification – Inferences

3.3 Univariate Analysis

3.4 Bi-Variate Analysis

3.5 Missing Value Identification

3.6 Outlier Identification

4 Solutions to the questions

5. Appendices

**OBJECTIVE**

To study the data sets of a COLD STORAGE and find out the solutions to the various problems and conduct hypothesis testing.

In Mar 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 pulls out data of last 35 days temperatures. As a safety measure, the Supervisor has been vigilant to maintain the temperature below 3.9 deg C.

Assume 3.9 deg C as upper acceptable temperature range and at alpha = 0.1 do you feel that there is 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 the Dairy Products.

**ASSUMPTIONS**

We are given the data sets of the Cold Storage and it is accurate.

**DESCRIPTIVE DATA ANALYSIS**

**RCODE**

summary(storage\_temp)

attach(storage\_temp)

describe(storage\_temp)

summary(storage\_temp)

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

Describe

|  |
| --- |
| 4 Variables 365 Observations  ---------------------------------------------------------------------------------------------------------------------  Season  n missing distinct  365 0 3    Value Rainy Summer Winter  Frequency 122 120 123  Proportion 0.334 0.329 0.337  ---------------------------------------------------------------------------------------------------------------------  Month  n missing distinct  365 0 12    Value Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep  Frequency 30 31 31 28 31 31 30 31 31 30 31 30  Proportion 0.082 0.085 0.085 0.077 0.085 0.085 0.082 0.085 0.085 0.082 0.085 0.082  ---------------------------------------------------------------------------------------------------------------------  Date  n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95  365 0 31 0.999 15.72 10.18 2 4 8 16 23 28 29  lowest : 1 2 3 4 5, highest: 27 28 29 30 31  ---------------------------------------------------------------------------------------------------------------------  Temperature  n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95  365 0 26 0.994 2.963 0.5665 2.3 2.4 2.5 2.9 3.3 3.7 3.9  lowest : 1.7 1.9 2.1 2.2 2.3, highest: 4.0 4.1 4.2 4.4 5.0  --------------------------------------------------------------------------------------------------------------------- |
|  |
| |  | | --- | | > | |

**EXPLORATORY DATA ANALYSIS**

**Rcode:**

hist(Temperature)

barchart()

?barchart

qplot(Month,fill=Season,data=storage\_temp, geom="bar",

main = "Month VS Season",

xlab = "Month",

ylab = "Season")

qplot(Season, Temperature, data = storage\_temp,

main = "season Vs Temperature",

xlab = "Season",

ylab = "Temperature")

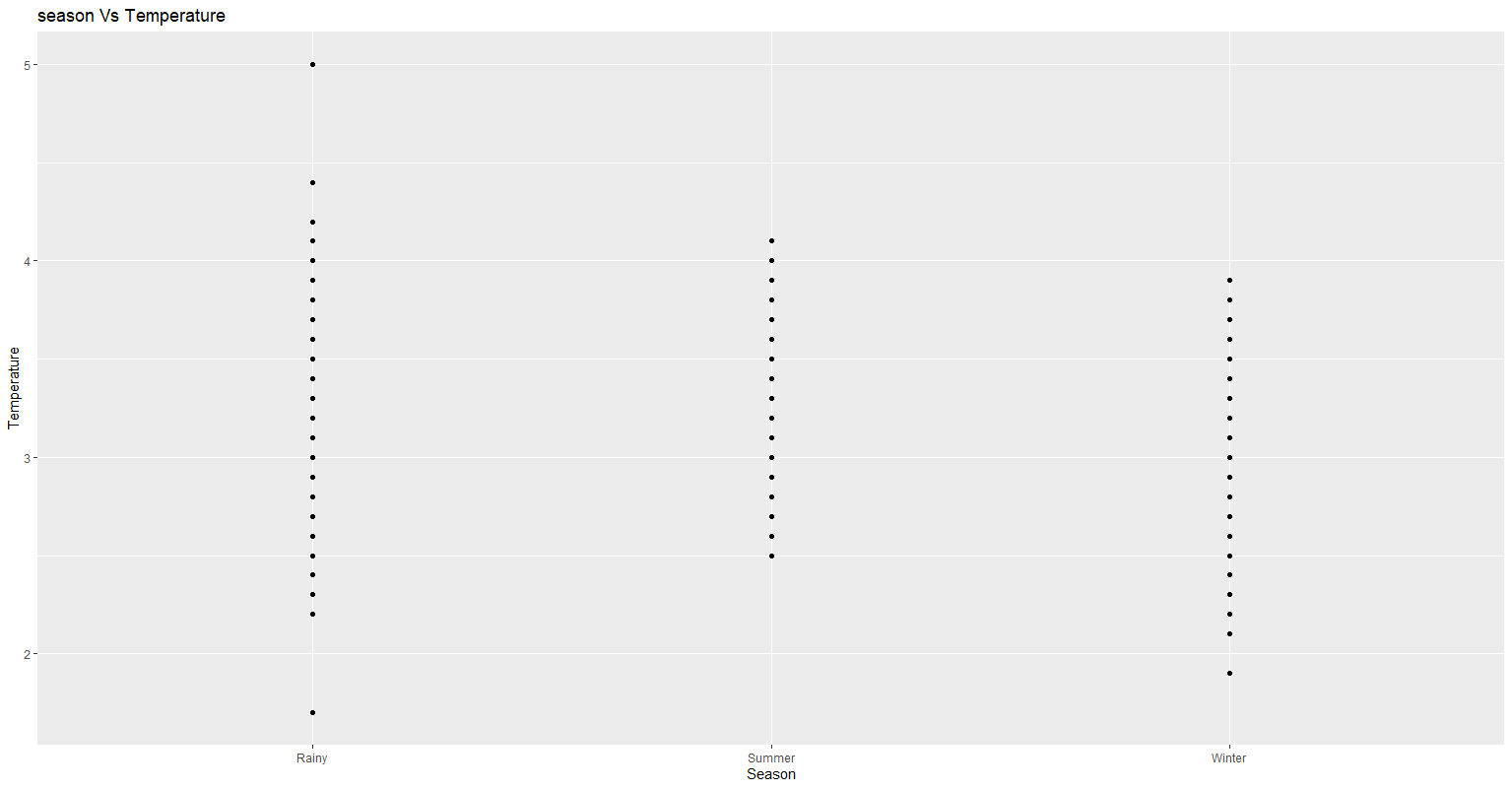
library(corrplot)

library(GGally)

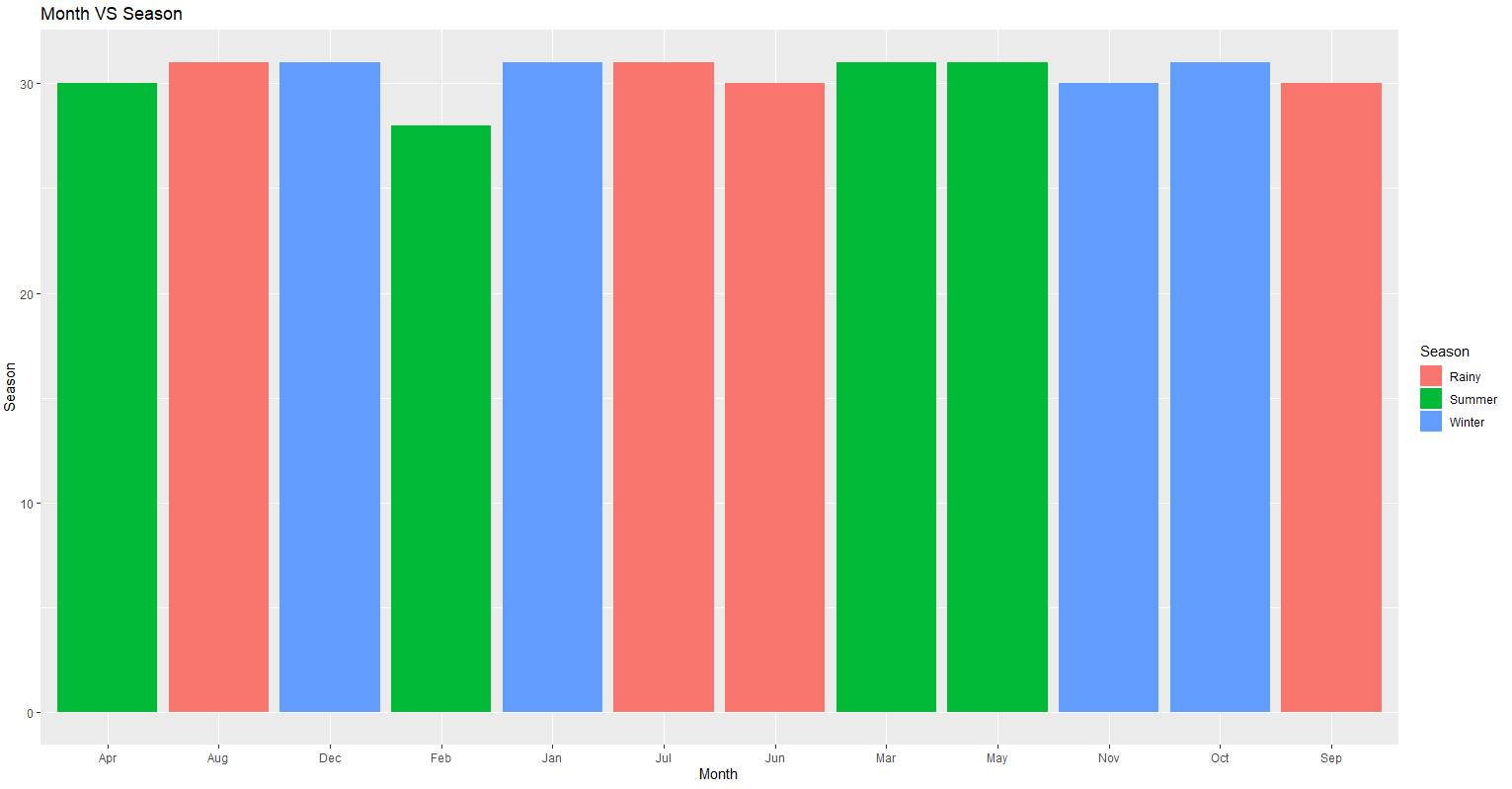
ggcorr(storage\_temp)

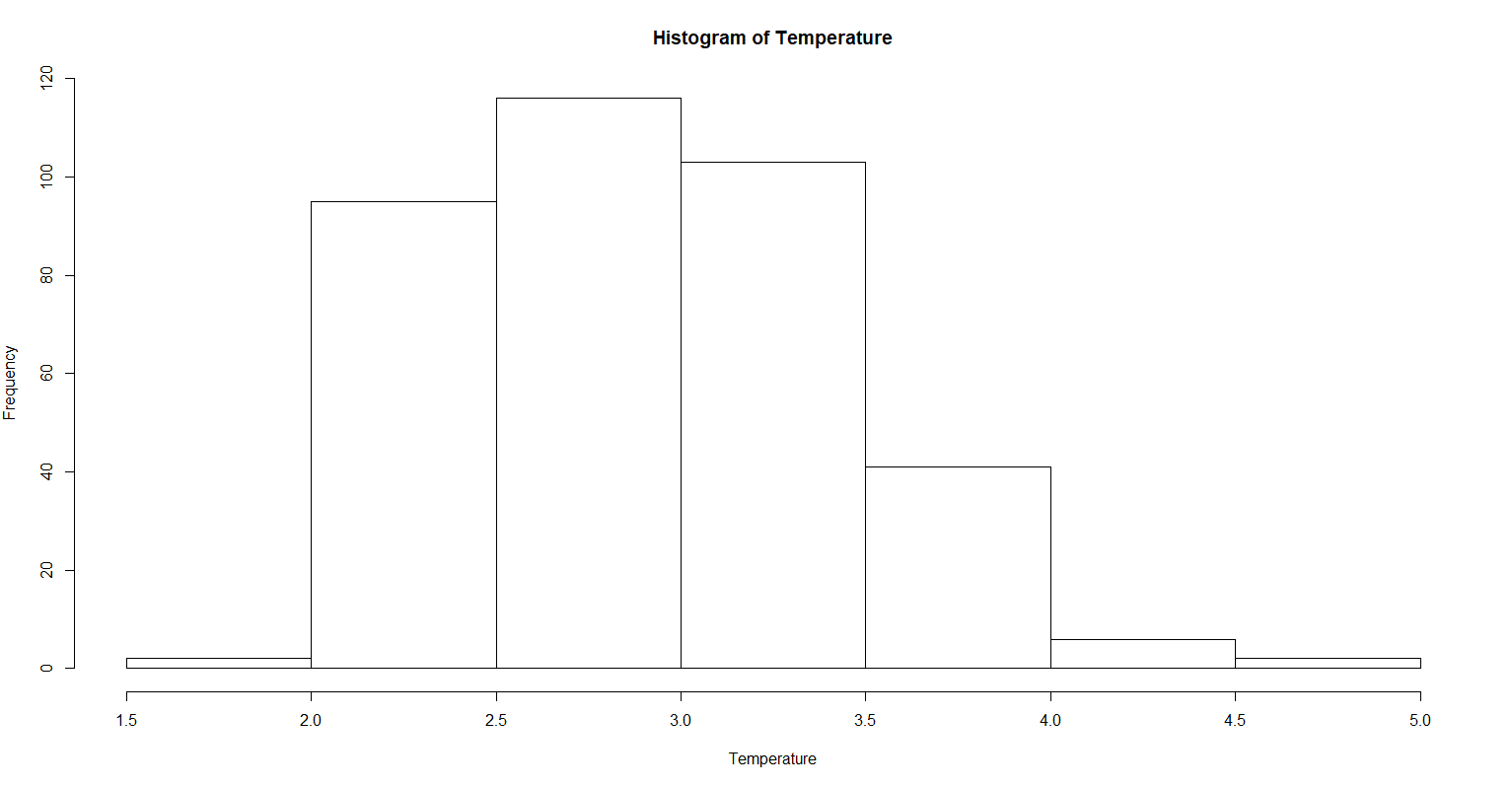
ggcorr(storage\_temp, label= TRUE )

DOTPLOT OF SEASON AND TEMPERATURE



BARGRAPH OF MONTH AND SEASON



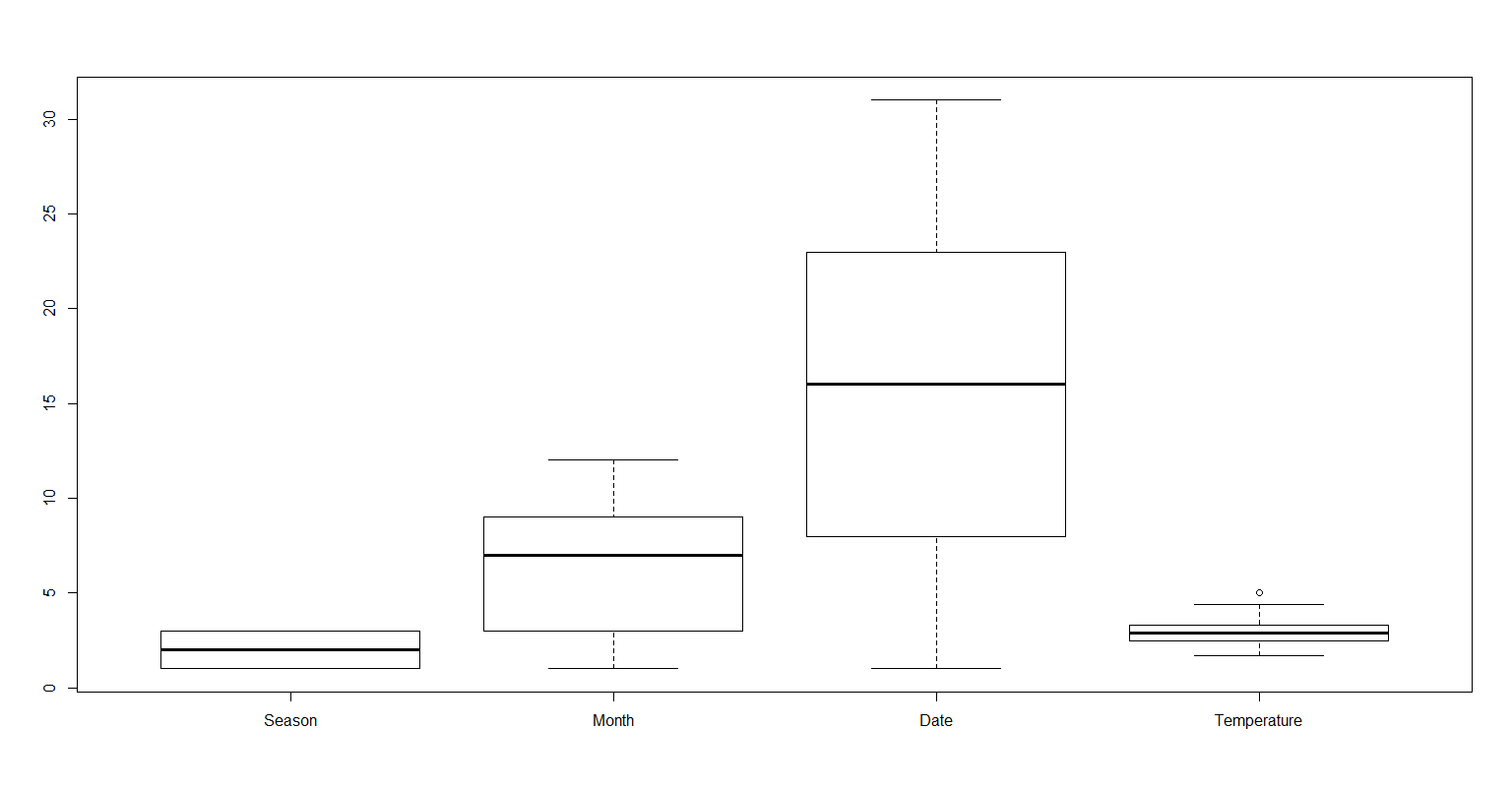


**OUTLIERS DETECTION**

#To check the outliers in the data set

RCODE:

boxplot(storage\_temp)



**CONCLUSIONS/SOLUTONS**

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

**ANS**. The subsets of each season are created and the respective temperatures are assigned to different seasons and the mean is calculated using the mean function.

**RCODE:**

summer\_temp = storage\_temp[which(Season== "Summer"),"Temperature"]

Winter\_temp = storage\_temp[which(Season== "Winter"),"Temperature"]

Rainy\_temp = storage\_temp[which(Season== "Rainy"),"Temperature"]

mean\_summer\_temp= mean(summer\_temp)

mean\_Winter\_temp= mean(Winter\_temp)

mean\_Rainy\_temp= mean(Rainy\_temp)

**OUTPUT:**

> mean\_summer\_temp

[1] 3.153333

> mean\_Winter\_temp

[1] 2.700813

> mean\_Rainy\_temp

[1] 3.039344

Mean of the temperatures of the three seasons , calculated individually.

* 1. **Find overall mean for the full year.**

**ANS.** A subset is created which contains all the temperatures of the dataset given. Then we use the in-built mean function to calculate the mean of the entire year.

**RCODE:**

all\_temp= storage\_temp[,"Temperature"]

mean\_of\_entire=mean(all\_temp)

print(paste0("The mean of the entire year is= ",mean\_of\_entire ))

**OUTPUT:**

[1] "The mean of the entire year is= 2.9627397260274"

The mean of the entire year is calculated by using the temperatures of the 365 days.

* 1. **Find Standard Deviation for the full year.**

**ANS.** We have calculated the STANDARD DEVIATION of the entire year by using the in-built function of R namely, sd().

**RCODE:**

sd\_of\_entire= sd(all\_temp)

print(paste0("The standard deviation of all the temperatures is ", sd\_of\_entire))

**OUTPUT:**

[1] "The standard deviation of all the temperatures is 0.508589031488563"

We have thus, calculated the standard deviation of the entire year.

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

**ANS.** To calculate the probability of the given normal distribution, we have used the function named, pnorm(). We wish to calculate the probability of falling below 2 Deg Celsius, there =fore, we set the attribute of lower.tail =TRUE.

RCODE:

np1= pnorm(2,mean = 2.96274,sd= .508589, lower.tail = TRUE )

print(paste0("The probability of the temperature having fallen below 2 deg C is =", np1))

**OUTPUT:**

[1] "The probability of the temperature having fallen below 2 deg C is =0.0291814158657375"

As a result, we get the probability of falling below 2 DEG.

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

**ANS.** To calculate the probability of the given normal distribution, we have used the function named, pnorm(). We wish to calculate the probability of temperature having gone above 4 deg C, therefore, we set the attribute of lower.tail =FALSE.

**RCODE:**

np2= pnorm(4,mean = 2.96274,sd= .508589, lower.tail = FALSE )

print(paste0("The probability of temperature having gone above 4 deg C is =", np2))

**OUTPUT:**

[1] "The probability of temperature having gone above 4 deg C is =0.0207007877727783"

As a result, we get the probability of temperature lying about 4 deg C.

* 1. **What will be the penalty for the AMC Company?**

**ANS.** We use the pnorm() function to find the probability of temperatures falling below 2 Deg and the those lying above 4 Deg. We’ll add the probabilities of both the cases and that would give us the probability of the temperature lying outside the range of 2-4 Degrees.

**RCODE:**

NP= np1+np2

print(paste0("the probability that the temperature will go out of the 2-4 Degrees range is = ",NP))

print(paste0("The percentage will be = ", NP\*100))

#this is greater than 2.5 and less than 5

#The penalty to the AMC will be= 10%

**OUTPUT:**

[1] "the probability that the temperature will go out of the 2-4 Degrees range is = 0.0498822036385158"

[1] "The percentage will be = 4.98822036385158"

> #this is greater than 2.5 and less than 5

> #The penalty to the AMC will be= 10%

As the combined probability of the temperatures lying outside the range of 2-4 is equal to 4.988220, the penalty = 10%.

**2.1. State the Hypothesis, do the calculation using z test.**

**ANS.** We’ll consider:

NULL HYPTHESIS, H0 = The temperature of the cold storage= 3.9 Deg C

ALTERNATIVE HYPOTHESIS, H1 = The temperature of the cold storage >3.9 Deg C

THIS IS A RIGHT-TAILED TEST.

* We will calculate Z-value from the formula = **z = (x – μ) / (σ / √n)**
* We will take the critical value of Z and compare the calculated and the critical values of Z to accept or reject the Null Hypothesis.
* We will calculate the P value for the right tail and compare it with alpha.

**RCODE:**

read.csv("Cold\_Storage\_Mar2018.csv")

data= read.csv("Cold\_Storage\_Mar2018.csv")

str(data)

summary(data)

mean\_of\_sample= mean(data[,"Temperature"])

z\_value = ((mean\_of\_sample- 3.9)/sd\_of\_entire)\*35^0.5

print(paste0("The COMPUTED Z-Value is= " , z\_value))

Z\_critical= qnorm(1-0.1)

print(paste0("The CRITICAL Z-Value is= " , Z\_critical))

?pnorm

# At ALPHA=1%, the critical Z-value is= 1.2815

p\_value1=1-pnorm(z\_value)

print(paste0("The P-Value is= ", p\_value1))

print("The value of Alpha is= 0.01")

print("The Critical value of Z > the Computed Z,the P-Value > alpha.

So, we FAIL TO REJECT THE NULL HYPOTHESIS")

# we can also compare by the critical value of the mean, Xc= 4.0100 .This value is derived from the critical value of Zc

#Since our Xc > X.

#The NULL HYPOTHESIS cannot be rejected.

#The temperature < 3.9 at the cold storage and the PROBLEM IS AT THE RETAILER/OUTLET SITE.

**OUTPUT:**

[1] "The COMPUTED Z-Value is= 0.864116575956302"

[1] "The CRITICAL Z-Value is= 1.2815515655446"

[1] "The P-Value is= 0.193761927023461"

[1] "The value of Alpha is= 0.01"

[1] "The Critical value of Z > the Computed Z, the P-Value > alpha. So, we FAIL TO REJECT THE NULL HYPOTHESIS"

**CRITICAL Z-VALUE > COMPUTED Z-VALUE,**

**AND P-VALUE > ALPHA**.

Therefore, we have FAILED TO REJECT THE NULL HYPOTHESIS.

Since, we have failed to reject the Null hypothesis, we have to **agree to the statement that the Cold storage maintained the temperature equal to or below 3.9 Deg C.**

The Alternative Hypothesis is rejected signifying that **the fault was at the PROCUREMENT SITE** form where the cold storage was getting the products.

**2.2. State the Hypothesis, do the calculation using t-test.**

**ANS.** We’ll consider:

NULL HYPTHESIS, H0 = The temperature of the cold storage= 3.9 Deg C

ALTERNATIVE HYPOTHESIS, H1 = The temperature of the cold storage >3.9 Deg C

THIS IS A RIGHT-TAILED TEST.

* We will calculate T-STAT from the formula  **= (x – μ) / (σ / √n), where σ is the standard deviation of the sample.**
* We will take the critical value of T-stat and compare the calculated and the critical values of T-Stat to accept or reject the Null Hypothesis.
* We will calculate the P value for the right tail and compare it with alpha.

**RCODE:**

sd\_of\_sample= sd(data[,"Temperature"])

print(paste0("The estimated standard deviation of the population is = " , sd\_of\_sample))

mean\_of\_sample

mean\_of\_entire

?t.test

print("We are using a direct function to calculate the answers, that is t.test")

t.test(data$Temperature, alternative = "greater", mu= 3.9, conf.level = 0.9)

print(paste0("P-value through t.test()= ",.0047))

T\_critical= qt(1-.1/2,34)

print(paste0("The value of the CRITICAL T-Stat is= ", T\_critical))

print("Now, we are calculating the values of T-stat manually")

tstat=(mean\_of\_sample- 3.9)/(sd\_of\_sample/(35^0.5))

print(paste0("The value of the COMPUTED T-Stat is= ", tstat))

p\_value2=1-pt(tstat,34)

print(paste0("The P-Value is = ", p\_value2))

print(paste0("The value of Alpha = ", 0.1))

print("The Critical value of T-Stat < the Computed T-Stat,the P-Value < alpha. So, we REJECT THE NULL HYPOTHESIS")

# The NULL HYPOTHESIS IS REJECTED and the temperature at the cold storage is at fault and its temperature is MORE THAN 3.9

**OUTPUT:**

[1] "The estimated standard deviation of the population is = 0.159674037712233"

[1] "We are using a direct function to calculate the answers, that is t.test"

[1] "P-value through t.test()= 0.0047"

[1] "The value of the CRITICAL T-Stat is= 1.69092425518685"

[1] "Now, we are calculating the values of T-stat manually"

[1] "The value of the COMPUTED T-Stat is= 2.75235860980021"

[1] "The P-Value is = 0.00471119770213257"

[1] "The value of Alpha = 0.1"

[1] "The Critical value of T-Stat < the Computed T-Stat,the P-Value < alpha. So, we REJECT

THE NULL HYPOTHESIS"

**CRITICAL T-VALUE > COMPUTED Z-VALUE,**

**AND P-VALUE > ALPHA**.

Therefore, we have to REJECT THE NULL HYPOTHESIS.

Since, we have accepted the Null hypothesis, we have to **agree to the statement that the Cold storage maintained the temperature greater than 3.9 Deg C.**

The Alternative Hypothesis is accepted signifying that **the fault was at the COLD STORAGE SITE.**

* 1. **Give your inference after doing both the tests.**

**ANS.** **Z-Test:**

* The Computed Z-Value < Critical Z-Value
* The P-Value > Alpha

This shows that we cannot reject the Null Hypothesis . The Cold Storage had maintained the temperature well below 3.9 Deg C.

And the alternative hypothesis that the temperature was greater than 3.9 Deg C is rejected. Hence inferring that the procurement site is at fault. The Cold Storage shouldn’t be blamed.

**T-Test**

* The Computed T-Stat value > Critical T-Stat value.
* The P-Value < Alpha

This shows that we have to reject the null hypothesis. This means that the temperature at the Cold Storage was more than the upper acceptable temperature of 3.9 Deg C.

And the Alternative Hypothesis that the temperature was greater than 3.9 Deg C. Hence inferring that the Cold Storage was at fault. The procurement site shouldn’t be blamed.

In the Z-Test we use the STANDARD ERROR OF THE POPULATION , whereas

In the T-Test we use the STANDARD ERROR OF THE SAMPLE (Estimated Standard Error).

Therefore Z-Test are better than T-Test.

**Plus, we can also prove it by finding the powers of the test.**

> T\_power=power.t.test(35,delta= 0, sd= 0.15,sig.level = .1, type = "one.sample", alternative = "one.sided")

> T\_power

One-sample t test power calculation

n = 35

delta = 0

sd = 0.15

sig.level = 0.1

power = 0.1

alternative = one.sided

> pwr.norm.test(.0742, 35, sig.level = .1, alternative = "greater")

Mean power calculation for normal distribution with known variance

d = 0.0742

n = 35

sig.level = 0.1

power = 0.1997321

alternative = greater

Power of T-test= 0.1

Power of Z-Test= 0.199

**APPENDICES**

setwd("C:/Users/mitta/Downloads/Cold storage project")

getwd()

install.packages("pwr")

library(readr)

library(Hmisc)

library(stats)

library(pwr)

library(effsize)

storage\_temp= read.csv("Cold\_Storage\_Temp\_Data.csv")

str(storage\_temp)

#DESCRIPTIVE ANALYSIS:

summary(storage\_temp)

attach(storage\_temp)

describe(storage\_temp) # it gives the descriptio of all the variables of our dataset

#To check the outliers in the data set

boxplot(storage\_temp)

#The boxplots of the four variables show that there are no outliers except in the TEMPERATURE variable.

hist(Temperature)

barchart()

?barchart

qplot(Month,fill=Season,data=storage\_temp, geom="bar",

main = "Month VS Season",

xlab = "Month",

ylab = "Season")

qplot(Season, Temperature, data = storage\_temp,

main = "season Vs Temperature",

xlab = "Season",

ylab = "Temperature")

library(corrplot)

library(GGally)

ggcorr(storage\_temp)

ggcorr(storage\_temp, label= TRUE )

#This shows that none of the variables have any correlation with each other.

null=is.na(Temperature)

null

1.1.

?mean

summer\_temp = storage\_temp[which(Season== "Summer"),"Temperature"]

Winter\_temp = storage\_temp[which(Season== "Winter"),"Temperature"]

Rainy\_temp = storage\_temp[which(Season== "Rainy"),"Temperature"]

mean\_summer\_temp= mean(summer\_temp)

mean\_Winter\_temp= mean(Winter\_temp)

mean\_Rainy\_temp= mean(Rainy\_temp)

mean\_summer\_temp

mean\_Winter\_temp

mean\_Rainy\_temp

----------------------------------------------------------------------------------

1.2.

all\_temp= storage\_temp[,"Temperature"]

mean\_of\_entire=mean(all\_temp)

print(paste0("The mean of the entire year is= ",mean\_of\_entire ))

------------------------------------------------------------------------------------

1.3.

sd\_of\_entire= sd(all\_temp)

print(paste0("The standard deviation of all the temperatures is ", sd\_of\_entire))

------------------------------------------------------------------------------------

1.4.

np1= pnorm(2,mean = 2.96274,sd= .508589, lower.tail = TRUE )

print(paste0("The probability of the temperature having fallen below 2 deg C is =", np1))

------------------------------------------------------------------------------------

1.5.

np2= pnorm(4,mean = 2.96274,sd= .508589, lower.tail = FALSE )

print(paste0("The probability of temperature having gone above 4 deg C is =", np2))

-----------------------------------------------------------------------------------

1.6.

np3= pnorm(4,mean = 2.96274,sd= .508589, lower.tail = TRUE )- np1

np3

NP= np1+np2 # Here, NP is the combined probabilities of the temperature lying outside the range of 2-4 Degrees.

print(paste0("the probability that the temperature will go out of the 2-4 Degrees range is = ",NP))

print(paste0("The percentage will be = ", NP\*100))

#this is greater thann 2.5 and less than 5

#The penalty to the AMC will be= 10%

----------------------------------------------------------------------------------------

# 2.1.

# H0= The Temperature was = 3.9 Deg C

# H1= The Temperature was > 3.9 Deg C

read.csv("Cold\_Storage\_Mar2018.csv")

data= read.csv("Cold\_Storage\_Mar2018.csv")

str(data)

summary(data)

mean\_of\_sample= mean(data[,"Temperature"])

z\_value = ((mean\_of\_sample- 3.9)/sd\_of\_entire)\*35^0.5

print(paste0("The COMPUTED Z-Value is= " , z\_value))

Z\_critical= qnorm(1-0.1)

print(paste0("The CRITICAL Z-Value is= " , Z\_critical))

?pnorm

# At ALPHA=1%, the critical Z-value is= 1.2815

...................

p\_value1=1-pnorm(z\_value)

print(paste0("The P-Value is= ", p\_value1))

print("The value of Alpha is= 0.01")

print("The Critical value of Z > the Computed Z,the P-Value > alpha. So, we FAIL TO REJECT THE NULL HYPOTHESIS")

# we can also compare by the critical value of the mean, Xc= 4.0100 .This value is derived from the critical value of Zc

#Since our Xc is greater than X.

#The NULL HYPOTHESIS cannot be rejected.

#The temperature < 3.9 at the cold storage and the PROBLEM IS AT THE RETAILER/OUTLET SITE.

--------------------------------------------------------------------------------------------------------------

2.2.

# H0= The Temperature was = 3.9 Deg C

# H1= The Temperature was > 3.9 Deg C

#

sd\_of\_sample= sd(data[,"Temperature"])

print(paste0("The estimated standard deviation of the population is = " , sd\_of\_sample))

mean\_of\_sample

mean\_of\_entire

?t.test

print("We are using a direct function to calculate the answers, that is t.test")

t.test(data$Temperature, alternative = "greater", mu= 3.9, conf.level = 0.9)

print(paste0("P-value through t.test()= ",.0047))

T\_critical= qt(1-.1/2,34)

print(paste0("The value of the CRITICAL T-Stat is= ", T\_critical))

print("Now, we are calculating the values of T-stat manually")

tstat=(mean\_of\_sample- 3.9)/(sd\_of\_sample/(35^0.5))

print(paste0("The value of the COMPUTED T-Stat is= ", tstat))

p\_value2=1-pt(tstat,34)

print(paste0("The P-Value is = ", p\_value2))

print(paste0("The value of Alpha = ", 0.1))

print("The Critical value of T-Stat < the Computed T-Stat,the P-Value < alpha. So, we REJECT THE NULL HYPOTHESIS")

# The NULL HYPOTHESIS IS REJECTED and the temperature at the cold storage is at fault and its temperature is MORE THAN 3.9

-------------------------------------------------------------------------------------------------------

2.3.

(A z-test uses the population standard error whereas the t-test uses the estimated standard error.

Thus, the z-test is more accurate and more powerful.)

T\_power=power.t.test(35,delta= 0, sd= 0.15,sig.level = .1, type = "one.sample", alternative = "one.sided")

pwr.norm.test(.0742, 35, sig.level = .1, alternative = "greater")