Cold Storage Project

Benedict Egwuchukwu

7/10/2020

Contents

1. Project Objective

The objective of the project is to analyse the two datasets provided concerning the Cold Storage Company after outsourcing their plant maintenance work to a professional company in the first year of business. Answers are to be provided to the questions asked and this will include the insights unearthed during the course of solving the problem, visualizations to support the results and the recommendations based on the analyses conducted.

2. Assumptions

The datasets provided contain details relating to the opeartions of the Cold Storage Company. There are two datasets, Cold_Storage_Temp_Data.csv and Cold_Storage_Mar2018.csv. I assume Cold_Storage_Temp_Data.csv represents the population while Cold_Storage_Mar2018.csv represents the sample.

The assumptions made in this report are as follows:

- Dataset is clean and has no errors in entries.
- There is no relationship between independent and dependent variables.
- The data has a normal distribution.

3. Environment Set up and Data Import

3.1 Install necessary packages and load libraries

```
# Environment Set up and Data Import

# Invoking Libraries
library(readr) # To import csv files
library(ggplot2) # To create plots
library(gridExtra) # To plot multiple ggplot graphs in a grid
library(car) # for Levenetest
library(dplyr) # To manipulate dataset
library(knitr) # Necessary to generate sourcecodes from a .Rmd File
library(markdown) # To convert to HTML
library(rmarkdown) # To convret analyses into high quality documents
```

3.2 Set up working Directory

```
# Set working directory
setwd("C:/Users/egwuc/Desktop/PGP-DSBA-UT Austin/Fundamental of Business Statistics/Week 4 - Project/")
```

3.3 Import Dataset

```
# Read input file
cold_storage_temp <- read_csv("Cold_Storage_Temp_Data.csv")
```

```
## Parsed with column specification:
## cols(
##   Season = col_character(),
##   Month = col_character(),
##   Date = col_double(),
##   Temperature = col_double()
## )
```

```
# Read input file
cold_storage_mar <- read_csv("Cold_Storage_Mar2018.csv")
```

```
## Parsed with column specification:
## cols(
##   Season = col_character(),
##   Month = col_character(),
##   Date = col_double(),
##   Temperature = col_double()
## )
```

3.4 Global Options Setting

```
# Global options settings
options(scipen = 999) # turn off scientific notation like 1e+06
```

4. Variable Identification

In order to get familiar with the Cold Storage data, I would be using the following functions to get an overview

1. dim(): this gives us the dimension of the dataset provided. Knowing the data dimension gives us an idea of how large the data is. 2. head(): this shows the first 6 rows(observations) of the dataset. It is essential for us to get a glimpse of the dataset in a tabular format without revealing the entire dataset if we are to properly analyse the data.
2. tail(): this shows the last 6 rows(observations) of the dataset. Knowing what the dataset looks like at the end rows also helps us ensure the data is consistent.
3. str(): this shows us the structure of the dataset. It helps us determine the datatypes of the features and identify if there are datatype mismatches, so that we handle these ASAP to avoid inappropriate results from our analysis.
4. summary(): this provides statistical summaries of the dataset. This function is important as we can quickly get statistical summaries (mean,median, quartiles, min, frequencies/counts, max values etc.) which can help us derive insights even before diving deep into the data.
5. View(): helps to look at the entire dataset at a glance.

4.1 Variable Identification - Insights

4.1.1 Cold_Storage_Temp_Data Dataset

Insights from dim():

```
# Variable identification
# check dimension of dataset
dim(cold_storage_temp)
```

```
## [1] 365   4
```

- The dataset has 365 rows and 4 columns.

Insights from head():

```
# check first 6 rows(observations) of dataset
head(cold_storage_temp)
```

```
## # A tibble: 6 x 4
##   Season Month  Date Temperature
##   <chr>  <chr> <dbl>       <dbl>
## 1 Winter Jan       1         2.3
## 2 Winter Jan       2         2.2
## 3 Winter Jan       3         2.4
## 4 Winter Jan       4         2.8
## 5 Winter Jan       5         2.5
## 6 Winter Jan       6         2.4
```

- The season and month variables are characters.
- The date and temperature variables are numeric.

Insights from tail():

```
# check last 6 rows(observations) of dataset
tail(cold_storage_temp)
```

```
## # A tibble: 6 x 4
##   Season Month  Date Temperature
##   <chr>  <chr> <dbl>       <dbl>
## 1 Winter Dec      26         2.7
## 2 Winter Dec      27         2.7
## 3 Winter Dec      28         2.3
## 4 Winter Dec      29         2.6
## 5 Winter Dec      30         2.3
## 6 Winter Dec      31         2.9
```

- Values in all fields are consistent in each column.

```
# change season and month to factor variable
cold_storage_temp$Season <- as.factor(cold_storage_temp$Season)
cold_storage_temp$Month <- as.factor(cold_storage_temp$Month)
```

- Season and month are factor variables and provide more meaning to the dataset.

Insights from summary():

```
# get summary of dataset
summary(cold_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.:2.700
##  Winter:123   Jan    : 31   Median :16.00   Median :3.000
##               Jul    : 31   Mean   :15.72   Mean   :3.002
##               Mar    : 31   3rd Qu.:23.00   3rd Qu.:3.300
##               May    : 31   Max.   :31.00   Max.   :4.500
##               (Other):179
```

- The season consists of three factors namely rainy, summer and winter.
- There are 122 days in the rainy season, 120 days in summer and 123 days in winter.
- The month goes from January to December.
- The minimum value of date is 1 and the maximum value of date is 31.
- The min temperature is 1.7 C and max temperature is 4.5 C, outside the stipulated range of 2-4 C.
- The mean and median of numeric variables are not too far apart.

Insights from view():

```
# view the dataset
View(cold_storage_temp)
```

- The data shows temperature across different seasons at a particular point in the month.

4.1.2 Cold_Storage_Mar2018 Dataset

Insights from dim():

```
# Variable identification
# check dimension of dataset
dim(cold_storage_mar)
```

```
## [1] 35  4
```

- The dataset has 35 rows and 4 columns.

Insights from head():

```r
# check first 6 rows(observations) of dataset
head(cold_storage_mar)
```

```
## # A tibble: 6 x 4
##   Season Month  Date Temperature
##   <chr>  <chr> <dbl>       <dbl>
## 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
```

- The season and month variables are characters.
- The date and temperature variables are numeric.

Insights from tail():

```r
# check last 6 rows(observations) of dataset
tail(cold_storage_mar)
```

```
## # A tibble: 6 x 4
##   Season Month  Date Temperature
##   <chr>  <chr> <dbl>       <dbl>
## 1 Summer Mar      12         3.8
## 2 Summer Mar      13         4.2
## 3 Summer Mar      14         4.2
## 4 Summer Mar      15         3.8
## 5 Summer Mar      16         3.9
## 6 Summer Mar      17         3.9
```

- Values in all fields are consistent in each column.

```r
# change season and month to factor variable
cold_storage_mar$Season <- as.factor(cold_storage_mar$Season)
cold_storage_mar$Month <- as.factor(cold_storage_mar$Month)
```

- Season and month are factor variables and provide more meaning to the dataset.

Insights from summary():

```r
# get summary of dataset
summary(cold_storage_mar)
```

```
##     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
```

- The season consists of only one factor, summer.
- There are 35 days in the summer season.
- Only February and March are in the month.
- The minimum value of date is 1 and the maximum value of date is 28.
- The min temperature is 3.8 C and max temperature is 4.6 C.
- The mean and median of numeric variables are close.

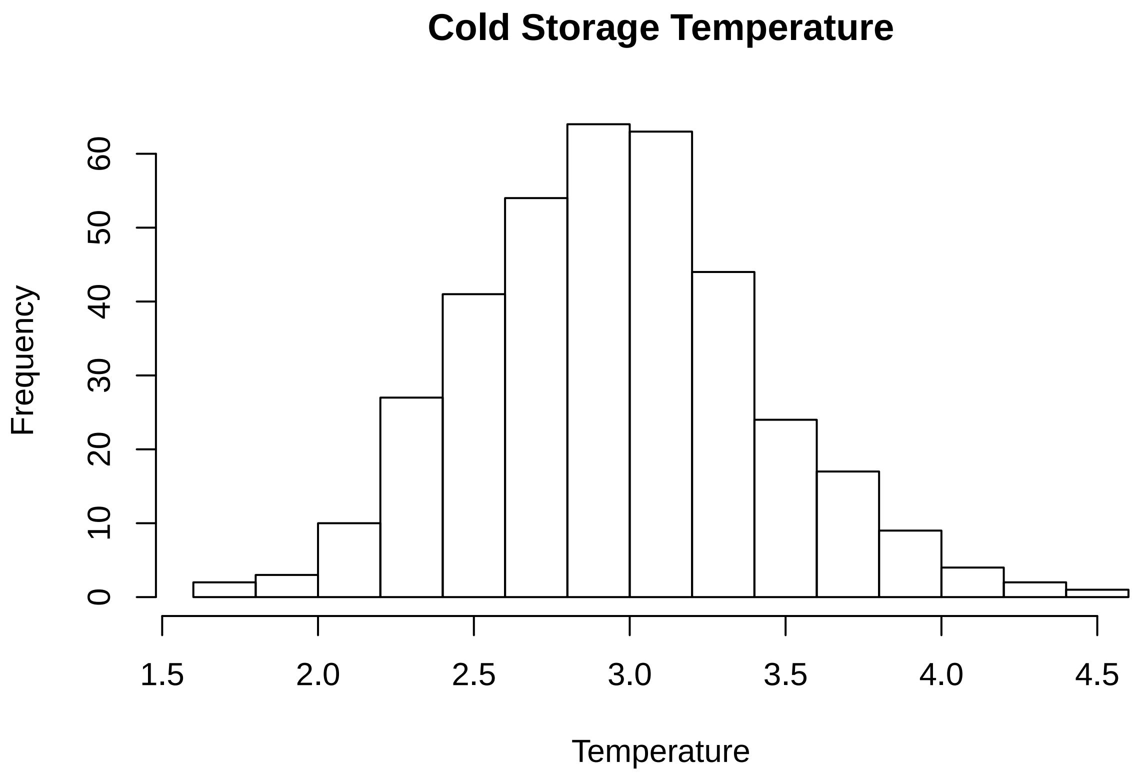
Insights from view():

```
# view the dataset
View(cold_storage_mar)
```

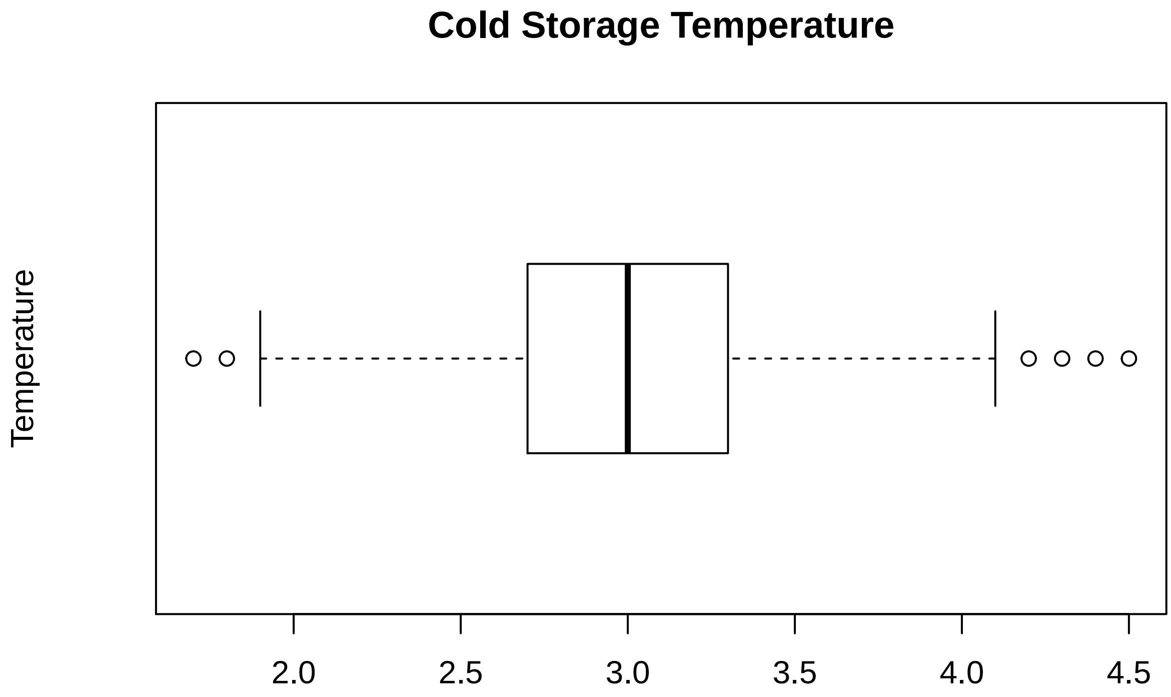
- The data shows temperature in the summer season for the month of February and March.

4.2 Preliminary Analysis

```
# Population graph
histogram_coldstoragetemp <- hist(cold_storage_temp$Temperature, xlab = "Temperature", main = "Cold Sto
```

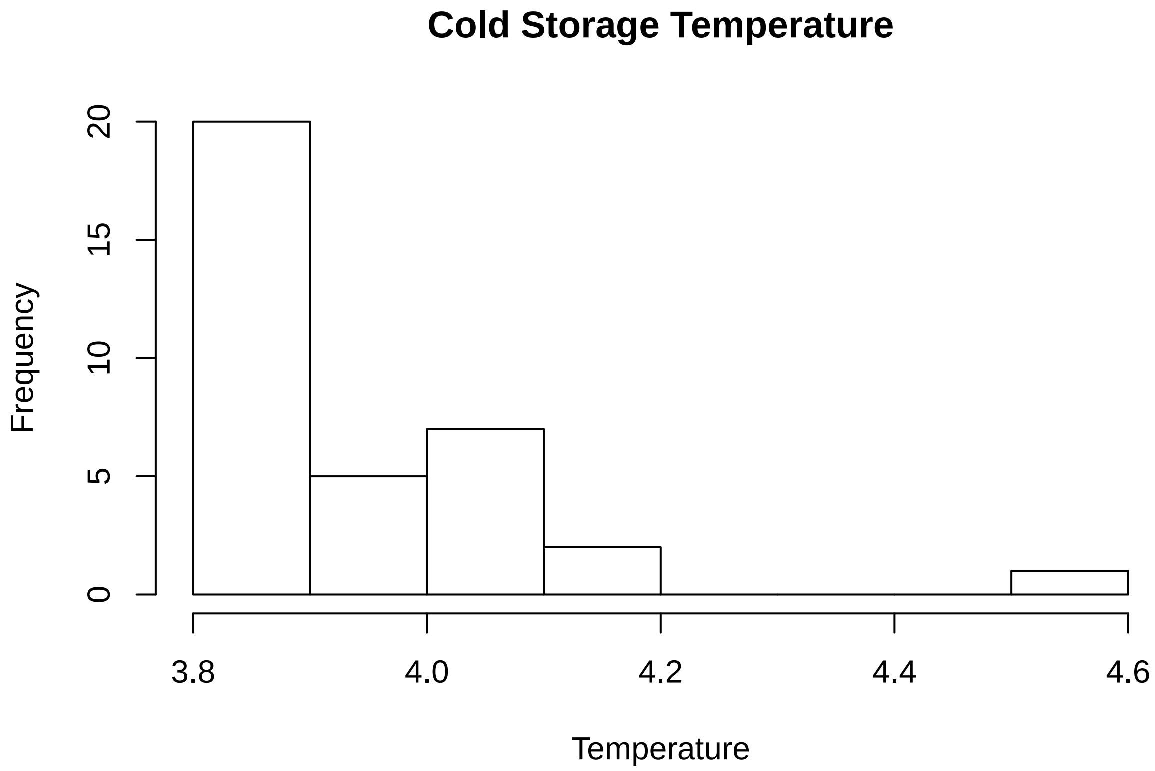


```
boxplot_coldstoragetemp <- boxplot(cold_storage_temp$Temperature, ylab = "Temperature", main = "Cold St
```

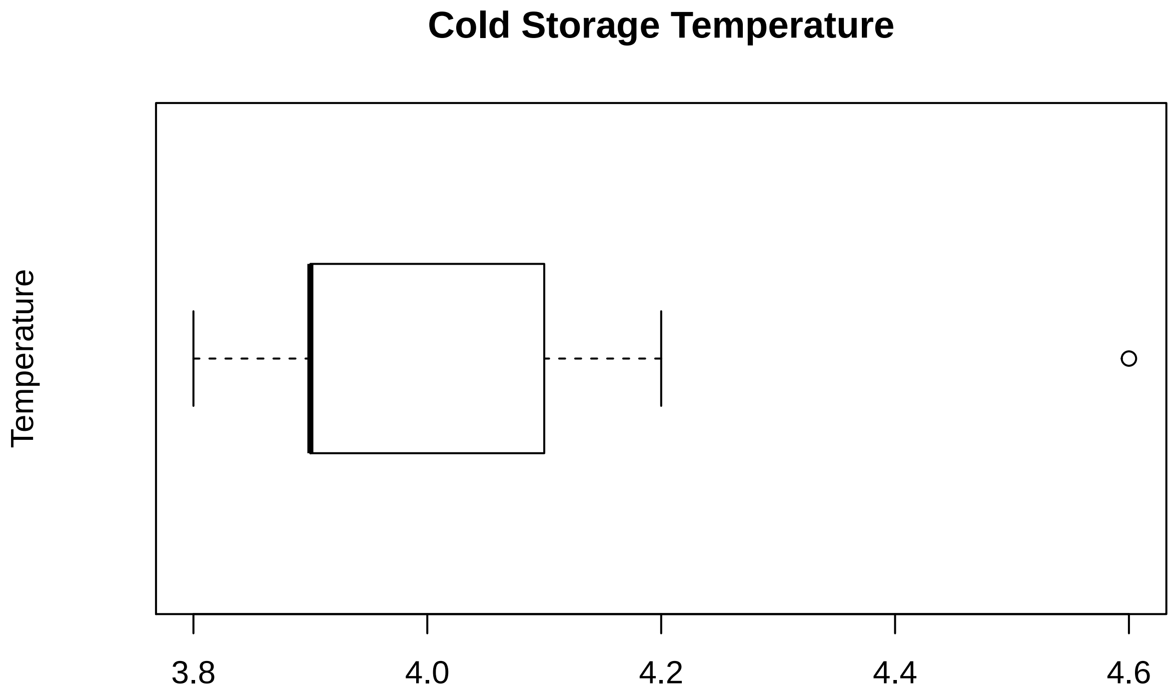


- The population seems to be normally distributed.
- Mean and Median Values are equal.

```
# Sample graph
histogram_coldstoragemar <- hist(cold_storage_mar$Temperature, xlab = "Temperature", main = "Cold Stora
```



```
boxplot_coldstoragemar <- boxplot(cold_storage_mar$Temperature, ylab = "Temperature", main = "Cold Stora
```



- The sample seems is to skewed to the right.
- The mean and median are not much different.

5. Solutions to Problem 1

5.1 Mean of Cold Storage Temperature for Summer, Winter and Rainy Season

```
mean_of_season <- cold_storage_temp %>% group_by(Season) %>% summarise(Mean.Temperature = mean(Temperatu
print(mean_of_season)
```

```
## # A tibble: 3 x 2
##   Season Mean.Temperature
##   <fct>             <dbl>
## 1 Rainy              3.09
## 2 Summer             3.15
## 3 Winter             2.78
```

- The mean of the rainy season is 3.087705 degree celsius.
- The mean of the summer season is 3.147500 degree celsius.
- The mean of the winter season is 2.776423 degree celsius.

5.2 Overall Mean for the Full Year

```
mean_of_fullyear <- mean(cold_storage_temp$Temperature)
print(mean_of_fullyear)
```

```
## [1] 3.002466
```

- The mean temperature for the full year is 3.002466 degree celsius.

5.3 Standard Deviation for the Full Year

```
sd_of_fullyear <- sd(cold_storage_temp$Temperature)
print(sd_of_fullyear)
```

```
## [1] 0.4658319
```

- The standaed deviation for the full year is 0.4658319 degree celsius.

5.4 Probability of Temperature Below 2 Degree Celsius

```
prob_below_2C <- pnorm(2, mean = mean_of_fullyear, sd = sd_of_fullyear)
print(prob_below_2C)
```

```
## [1] 0.01569906
```

- The probability of temperature going below 2 degree celsius is 0.01569906.

5.5 Probability of Temperature Above 4 Degree Celsius

```
prob_above_4C <- 1 - pnorm(4, mean = mean_of_fullyear, sd = sd_of_fullyear)
print(prob_above_4C)
```

```
## [1] 0.01612075
```

- The probability of temperature going above 4 degree celsius is 0.01612075.

5.6 Penalty for the AMC Company

```
penalty <- prob_below_2C + prob_above_4C
print(penalty)
```

```
## [1] 0.03181981
```

```
if(penalty > 0.025 && penalty < 0.05) {
  print("Penalty is 10%")
} else if(Penalty > 0.05) {
  print("Penalty is 25%")
} else {
  print("No need to pay penalty")
}
```

```
## [1] "Penalty is 10%"
```

The penalty for the AMC company will be 10%.

5.7 Significant Difference in Cold Storage Temperature between Rainy, Summer and Winter Seasons

```
# Apply one-way ANOVA
aovOutput <- aov(Temperature ~ Season, data = cold_storage_temp)
summary(aovOutput)
```

5.7.1 One Way ANOVA

```
##              Df Sum Sq Mean Sq F value          Pr(>F)    
## Season        2   9.70   4.848   25.32 0.0000000000508 ***
## Residuals   362  69.29   0.191                            
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 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). Given the factor has 3 levels, DF corresponding to season is 3 - 1 = 2.
- The significance level is 0.05.
- Total DF is 365 - 3 = 362.
- Mean sum of squares is obtained by dividing the sums of squares by corresponding DF.
- The value of the F-statistic is 25.32 and the p-value is significant.
- Based on the ANOVA test, I therefore reject the null hypothesis that the three seasons mean are identical. At least for one season, mean temperature is different from the rest.

5.7.2 Post-hoc Test A significant one-way ANOVA is generally followed up by Tukey post-hoc tests to perform multiple pairwise comparisons between groups.

```
# Pairwise comparisons
TukeyHSD(x = aovOutput, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = Temperature ~ Season, data = cold_storage_temp)
## 
## $Season
##                      diff         lwr        upr     p adj
## Summer-Rainy   0.05979508 -0.07258434  0.1921745 0.5376924
## Winter-Rainy  -0.31128215 -0.44284519 -0.1797191 0.0000002
## Winter-Summer -0.37107724 -0.50318954 -0.2389649 0.0000000
```

- It can be seen from the output, that only the difference between winter and rainy (adjusted p-value = 0.0000002) & winter and summer (adjusted p-value = 0.0000000) are significant.

6. Solutions to Problem 2

6.1 Hypothesis Test to be Performed to Check if Corrective Action is Needed at the Cold Storage Plant

- For the above question, the Z-Test and T-Test hypothesis test can be used to check if corrective action is needed at the cold storage plant.
- One reason both test can be used is because the sample size is large enough (when n > 30). In this case, n = 35.
- Second, a population and sample dataset was provided. This means the mean, variance and standard deviation for the population and sample can be derived.
- When the population variance is known, the Z-Test can be used. On the other hand, in the absence of a population variance, the T-Test is used.
- The level of significance (alpha) is 0.10.

6.2 State the Hypothesis, perform hypothesis test and determine p-value

```

mu = 3.9
a = 0.90
```

```
# Null Hypothesis H0: E[Temperature] <= 3.9
# Alternative Hypothesis H1: E[Temperature] > 3.9
```

```

mean_of_sample <- mean(cold_storage_mar$Temperature)
print(mean_of_sample)
```

6.2.1 Z-Test Hypothesis Test

```
## [1] 3.974286
```

```
sd_of_sample <- sd(cold_storage_mar$Temperature)
print(sd_of_sample)
```

```
## [1] 0.159674
```

```
std_error <- sd_of_sample / (sqrt(35))
print(std_error)
```

```
## [1] 0.02698984
```

```
z_test <- (mean_of_sample - mu) / std_error
print(z_test)
```

```
## [1] 2.752359
```

```
# Finding the critical value
z_critical <- qnorm(0.90)
print(z_critical)
```

```
## [1] 1.281552
```

```
# P-value of z
pnorm(z_test, lower.tail = FALSE)
```

```
## [1] 0.002958384
```

```

t_test <- t.test(cold_storage_mar$Temperature, mu = 3.9, conf.level = 0.90, alternative = "greater")
print(t_test)
```

6.2.2 T-Test Hypothesis Test

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

6.3 Inferences

- In the two bypothesis tests carried out, given that the p-value was less than the significance level of 0.1%, the null hypothesis was rejected while the alternative hypothesis was accepted. This means that the supervisor did not maintain the temperature in the cold storage equal to or less than 3.9 degrees celsius.

7. Appendix A – Source Code

```
# Environment Set up and Data Import

# Invoking Libraries
library(readr) # To import csv files
library(ggplot2) # To create plots
library(gridExtra) # To plot multiple ggplot graphs in a grid
library(car) # for Levenetest
library(dplyr) # To manipulate dataset
library(knitr) # Necessary to generate sourcecodes from a .Rmd File
library(markdown) # To convert to HTML
library(rmarkdown) # To convret analyses into high quality documents

# Set working directory
setwd("C:/Users/egwuc/Desktop/PGP-DSBA-UT Austin/Fundamental of Business Statistics/Week 4 - Project/")

# Read input file
cold_storage_temp <- read_csv("Cold_Storage_Temp_Data.csv")

# Read input file
cold_storage_mar <- read_csv("Cold_Storage_Mar2018.csv")

# Global options settings
options(scipen = 999) # turn off scientific notation like 1e+06

# Variable identification
# check dimension of dataset
dim(cold_storage_temp)

# check first 6 rows(observations) of dataset
head(cold_storage_temp)

# check last 6 rows(observations) of dataset
tail(cold_storage_temp)

# change season and month to factor variable
cold_storage_temp$Season <- as.factor(cold_storage_temp$Season)
cold_storage_temp$Month <- as.factor(cold_storage_temp$Month)

# get summary of dataset
summary(cold_storage_temp)

# view the dataset
View(cold_storage_temp)

# Variable identification
# check dimension of dataset
dim(cold_storage_mar)

# check first 6 rows(observations) of dataset
```

```
head(cold_storage_mar)

# check last 6 rows(observations) of dataset
tail(cold_storage_mar)

# change season and month to factor variable
cold_storage_mar$Season <- as.factor(cold_storage_mar$Season)
cold_storage_mar$Month <- as.factor(cold_storage_mar$Month)

# get summary of dataset
summary(cold_storage_mar)

# view the dataset
View(cold_storage_mar)

# Population graph
histogram_coldstoragetemp <- hist(cold_storage_temp$Temperature, xlab = "Temperature", main = "Cold Sto
boxplot_coldstoragetemp <- boxplot(cold_storage_temp$Temperature, ylab = "Temperature", main = "Cold St

# Sample graph
histogram_coldstoragemar <- hist(cold_storage_mar$Temperature, xlab = "Temperature", main = "Cold Storag
boxplot_coldstoragemar <- boxplot(cold_storage_mar$Temperature, ylab = "Temperature", main = "Cold Stora

mean_of_season <- cold_storage_temp %>% group_by(Season) %>% summarise(Mean.Temperature = mean(Temperatu
print(mean_of_season)

mean_of_fullyear <- mean(cold_storage_temp$Temperature)
print(mean_of_fullyear)

sd_of_fullyear <- sd(cold_storage_temp$Temperature)
print(sd_of_fullyear)

prob_below_2C <- pnorm(2, mean = mean_of_fullyear, sd = sd_of_fullyear)
print(prob_below_2C)

prob_above_4C <- 1 - pnorm(4, mean = mean_of_fullyear, sd = sd_of_fullyear)
print(prob_above_4C)

penalty <- prob_below_2C + prob_above_4C
print(penalty)
if(penalty > 0.025 && penalty < 0.05) {
  print("Penalty is 10%")
} else if(Penalty > 0.05) {
  print("Penalty is 25%")
} else {
  print("No need to pay penalty")
}

# Apply one-way ANOVA
aovOutput <- aov(Temperature ~ Season, data = cold_storage_temp)
summary(aovOutput)

# Pairwise comparisons
TukeyHSD(x = aovOutput, conf.level = 0.95)
```

```
mu = 3.9
a = 0.90
# Null Hypothesis H0: E[Temperature] <= 3.9
# Alternative Hypothesis H1: E[Temperature] > 3.9

mean_of_sample <- mean(cold_storage_mar$Temperature)
print(mean_of_sample)

sd_of_sample <- sd(cold_storage_mar$Temperature)
print(sd_of_sample)

std_error <- sd_of_sample / (sqrt(35))
print(std_error)

z_test <- (mean_of_sample - mu) / std_error
print(z_test)

# Finding the critical value
z_critical <- qnorm(0.90)
print(z_critical)

# P-value of z
pnorm(z_test, lower.tail = FALSE)

t_test <- t.test(cold_storage_mar$Temperature, mu = 3.9, conf.level = 0.90, alternative = "greater")
print(t_test)

#=======================================================================
#
# T H E - E N D
#
#=======================================================================
```

Generate .R file from this Rmd. The .R will contain only the R source code.

```r
# Generate the .R file from this .Rmd to hold the source code

purl("Cold-Storage-Project.Rmd", documentation = 0)
```

To create word or pdf report -> click on Knit in the toolbar above, select knit to pdf.