Cold Storage Problem

Dosubi Joshua Padjo

04/11/2020

Table of Contents

1. Project Objective

The objective of this report is to use R to solve practical statistical problems on the cold storage problem scenerio as described in the problem statements below. Analysis would be performed on the supplied dataset to provide answers to questions including visualizations. Finally, insights, conclusions and recommendations would be drawn from the results. Below are the questions:

Problem 1: Cold Storage started its operations in Jan 2016. They are in the business of storing Pasteurized Fresh Whole or Skimmed Milk, Sweet Cream, Flavoured Milk Drinks. To ensure that there is no change of texture, body appearance, separation of fats the optimal temperature to be maintained is between 2 - 4 C.

In the first year of business, they outsourced the plant maintenance work to a professional company with stiff penalty clauses. It was agreed that if it was statistically proven that the probability of temperature going outside the 2 - 4 C during the one-year contract 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. The average temperature data at date level is given in the file "Cold_Storage_Temp_Data.csv"

1. Find mean cold storage temperature for Summer, Winter and Rainy Season (3 marks)
2. Find overall mean for the full year (3 marks)
3. Find Standard Deviation for the full year (3 marks)
4. Assume Normal distribution, what is the probability of temperature having fallen below 2 C? (6 marks)
5. Assume Normal distribution, what is the probability of temperature having gone above 4 C? (6 marks)
6. What will be the penalty for the AMC Company? (7 marks)
7. Perform a one-way ANOVA test to determine if there is a significant difference in Cold Storage temperature between rainy, summer and winter seasons and comment on the findings. (9 marks)

Problem 2: 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 the last 35 days' temperatures. As a safety measure, the Supervisor decides to be vigilant to maintain the temperature at 3.9 C or below.

Assume 3.9 C as the upper acceptable value for mean temperature and at alpha = 0.1. Do you feel that there is a need for some corrective action in the Cold Storage Plant or is it that the problem is from the procurement side from where Cold Storage is getting the Dairy Products? The data of the last 35 days is in "Cold_Storage_Mar2018.csv"

1. Which Hypothesis test shall be performed to check if corrective action is needed at the cold storage plant? Justify your answer. (8 marks)
2. State the Hypothesis, perform hypothesis test and determine p-value (11 marks)
3. Give your inference (4 marks)

2. R-Setup

```
#=======================================================================
#
# Practical Data Analysis - Cold storage Problem
#
#=======================================================================
```

2.1 Environment Set up and Data Import

2.1.1 Install necessary Packages and Invoke Libraries

```
# Environment Set up and Data Import

# Invoking Libraries
library(tidyverse) # contains ggplot2,dplyr,forcats,lubridate etc
library(gridExtra) # Needed for plotting multiple ggplot graphs side-by-side
library(knitr) # Necessary to generate sourcecodes from a .Rmd File
```

2.1.2 Set up working Directory

```
# Setup Working Directory
setwd("C:/Users/USER/Documents/El-PaDJo/R programming language/4-Project 2")
```

2.1.3 Import and Read the Dataset

```
# Import and Read Input File
cold_storage_temp_data = read.csv("Cold_Storage_Temp_Data.csv")
cold_storage_mar2018_data = read.csv("Cold_Storage_Mar2018.csv")
```

2.1.4 Global options settings and Function Definitions

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

# Function to calculate mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

2.2 Preliminary analysis Functions

In order for us to get familiar with the cold storage data, below are the functions we would be using to get overview information:

1. **dim()**: this gives us the dimension of the dataset provided. knowing our dimension gives us an idea of how large the data is helping us discern what analysis methods would suffice.
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.
3. **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

4. **str()**: this shows us the structure of the dataset. This function is essential as it helps us to determine if there are datatype mismatches specifically (in a very brief format) so that we handle these ASAP to avoid wrong results from our analysis.
5. **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.) from which we can make insights before even diving into the data themselves for analysis.
6. **sd()**: this helps us calculate the standard deviation of a numeric column in the supplied dataset.
7. **var()**: this helps us calculate the variance of a numeric column in the supplied dataset.
8. **getmode()**: This ia a custom built function as shown in the functions definitions section above that helps us calculate the mode of a numeric column in the supplied dataset.
9. **ggplot()**: this is used for plotting different graph types to further read meaning into the dataset.

3. Solutions to Problems

3.1 Solution to Problem 1

```
# PROBLEM 1:
# Cold Storage started its operations in Jan 2016. They are in the business of
# storing Pasteurized Fresh Whole or Skimmed Milk, Sweet Cream, Flavoured Milk Drinks.
# To ensure that there is no change of texture, body appearance, separation of fats the
# optimal temperature to be maintained is between 2 - 4 C.
# In the first year of business, they outsourced the plant maintenance work to a
# professional company with stiff penalty clauses. It was agreed that if it was
# statistically proven that the probability of temperature going outside the 2 - 4 C
# during the one-year contract 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. The average temperature data at date level is given in
# the file "Cold_Storage_Temp_Data.csv"
```

3.1.1 Preliminary analysis (with Assumptions and Insights)

Insight(s) from dim():

```
# Preliminary analysis for Problem 1
# check dimension of dataset
dim(cold_storage_temp_data)
```

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

- The dataset is not a ‘big’ dataset with 4 columns and 365 rows.
- Sample Size: 365

Insight(s) from head():

```
#see first 6 rows(observations) of dataset
head(cold_storage_temp_data)
```

```
##   Season Month Date Temperature
## 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
```

- "Winter" being present in the dataset suggests that the country of operation isnt Africa.
- No. of Samples: 1 *(its obvious)*

Insight(s) from tail():

```
#see last 6 rows(observations) of dataset
tail(cold_storage_temp_data)
```

```
##     Season Month Date Temperature
## 360 Winter   Dec   26         2.7
## 361 Winter   Dec   27         2.7
## 362 Winter   Dec   28         2.3
## 363 Winter   Dec   29         2.6
## 364 Winter   Dec   30         2.3
## 365 Winter   Dec   31         2.9
```

- This dataset clearly contains data for the complete year with daily temperature averages.

Insight(s) from str():

```
# check structure of dataset
str(cold_storage_temp_data)
```

```
## '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  2.3 2.2 2.4 2.8 2.5 2.4 2.8 3 2.4 2.9 ...
```

- All the columns(variables) have appropriate datatypes.
- Only 3 seasons are featured here; "Rainy", "Summer" and "Winter" (as shown in head()). I guess "Autumn" does not count as a seperate season here…

Insight(s) from summary():

```
# get summary of dataset
summary(cold_storage_temp_data)
```

```
##     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
```

- From the minimum and maximum values of 1.7 and 4.5 respectively, as shown here, it is more than certain that the temperature went outside the 2 - 4 C given as the benchmark. We just need to know to what extent which is why we would find the probability of going below and above the given figures.
- For Temperature, which is our Point of focus here, The mean (3.002) and median (3.000) are not the same but are super close with negligible difference of 0.002 which suggests that the distribution tends to be normal. Also, this is not suprising as we can be sure that this distribution would tend to normal because it is a didtribution of averages as stated in the problem definition(Central Limit Theorem). With further analysis, we would confirm this 100%.

Insight(s) from sd():

```
# get standard deviation of concerned column: Temperature
sd_temp_1yr_data = sd(cold_storage_temp_data$Temperature)
sd_temp_1yr_data
```

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

Insight(s) from var():

```
# get variance of concerned column: Temperature
var_temp_1yr_data = var(cold_storage_temp_data$Temperature)
var_temp_1yr_data
```

```
## [1] 0.2169994
```

Insight(s) from getmode():

```
# get mode of concerned column: Temperature
getmode(cold_storage_temp_data$Temperature)
```

```
## [1] 3.1
```

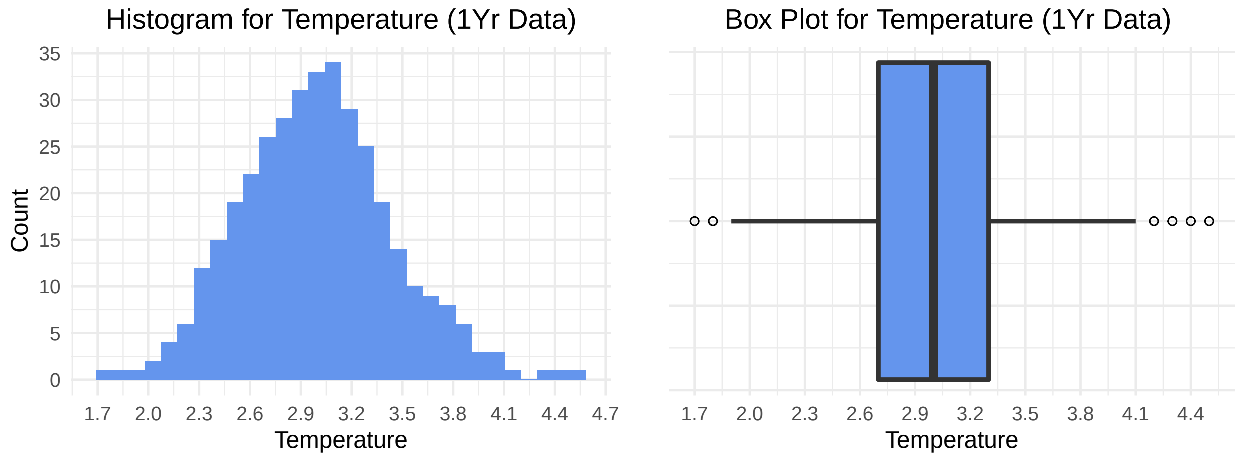
- the modal value(3.1) gotten here is not too far off from both the mean(3.002) and the median(3.000) which implies that it still tends towards a Normal distribution. we would need to see how it looks in a curve for final verification.

Insight(s) from ggplot():

```r
# Histogram for Temperature Variable (1Yr Data)
temperature_histogram =
  ggplot(cold_storage_temp_data, aes(x = cold_storage_temp_data$Temperature)) +
  geom_histogram(fill = "cornflowerblue", bins = 30) +
  labs(title="Histogram for Temperature (1Yr Data)", x="Temperature", y="Count") +
  scale_x_continuous(breaks=seq(1.4,max(cold_storage_temp_data$Temperature)+0.3,by=0.3))+
  scale_y_continuous(breaks = seq(0, 40, by = 5)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box Plot for temperature Variable (1Yr Data)
temperature_boxplot =
  ggplot(cold_storage_temp_data, aes(x = 0, y = cold_storage_temp_data$Temperature)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color="black", fill="cornflowerblue")+
  labs(title = "Box Plot for Temperature (1Yr Data) ", x = "", y="Temperature") +
  scale_y_continuous(breaks=seq(1.4,max(cold_storage_temp_data$Temperature)+0.3,by=0.3))+
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(temperature_histogram,temperature_boxplot,nrow=1,ncol= 2)
```



* From the Histogram, looking at the super pronounced Bell curve, we can see this is a Normal Distribution, however, its a tiny little bit left skewed as the density seems to be greater to the left nd to the right. *

Assumptions for Problem 1:

1. The dataset is a normal distribution

3.1.2 Question 1: Find mean cold storage temperature for Summer, Winter and Rainy Season

```
# Q1: calculate Temperature mean value for Summer, Winter and Rainy Season
by( cold_storage_temp_data$Temperature,INDICES = cold_storage_temp_data$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
```

- Temperature Mean Values-> Rainy Season: 3.087705, Summer Season: 3.1475, Winter Season: 2.776423

3.1.3 Question 2: Find overall mean for the full year

```
# Q2: calculate Temperature mean value for the year
mean_temp_1yr_data = mean(cold_storage_temp_data$Temperature)
mean_temp_1yr_data
```

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

- Temperature Mean Value-> Full Year: 3.002466

3.1.4 Question 3: Find Standard Deviation for the full year

```
# Q3: calculate Temperature Standard deviation for the year
sd_temp_1yr_data = sd(cold_storage_temp_data$Temperature)
sd_temp_1yr_data
```

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

- Temperature Standard Deviation-> Full Year: 0.4658319

3.1.5 Question 4: Assume Normal distribution, what is the probability of temperature having fallen below 2 C?

```
# Q4: On assumption of Normal Distribution, calculate probability of temperature < 2 C
norm_prob_below_2 = pnorm(2,mean=mean_temp_1yr_data,sd=sd_temp_1yr_data)
norm_prob_below_2
```

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

- On assumption of Normal Distribution, probability of temperature < 2 C = 0.01569906 ~ 1.6%

3.1.6 Question 5: Assume Normal distribution, what is the probability of temperature having gone above 4 C?

```r
# Q5: On assumption of Normal Distribution, calculate probability of temperature > 4 C
norm_prob_above_4 = 1 - pnorm(4,mean=mean_temp_1yr_data,sd=sd_temp_1yr_data)
norm_prob_above_4
```

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

- On assumption of Normal Distribution, probability of temperature > 4 C = 0.01612075 ~ 1.6% *(Since Probability generally totals 1, by subtracting the probability less than 4 from 1, we have the probability greater than 4. Thats how we got our value.)*

3.1.7 Question 6: What will be the penalty for the AMC Company?

```r
# Q6: On assumption of Normal Distribution, calculate probability
# of temperature > 4 C OR temperature < 2 C
norm_prob_outside_2_4 = norm_prob_below_2 + norm_prob_above_4
norm_prob_outside_2_4
```

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

- On assumption of Normal Distribution, probability of 2 C > temperature > 4 C = 0.03181981 ~ 3.2% *(By union of mutually exclusive events, we simply add probabilities of temperature below 2 and above 4 to get the probability of being outside 2 - 4 C.)*
- **AMC Company Penalty**: 10% of AMC (annual maintenance contract) *(2.5% < 3.2% < 5%)*

(It was stated from the problem definition that: "It was agreed that if it was statistically proven that the probability of temperature going outside the 2 - 4 C during the one-year contract was above 2.5% and less than 5% then the penalty would be 10% of AMC (annual maintenance contract).")

3.1.8 Question 7: Perform a one-way ANOVA test to determine if there is a significant difference in Cold Storage temperature between rainy, summer and winter seasons and comment on the findings

```r
# Q7: calculate a one-way ANOVA
aov_temp_1yr_data=aov(cold_storage_temp_data$Temperature~cold_storage_temp_data$Season,
                      data=cold_storage_temp_data)

# test for differences between rainy, summer and winter Seasons
TukeyHSD(aov_temp_1yr_data)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = cold_storage_temp_data$Temperature ~ cold_storage_temp_data$Season, data = cold_s
##
```

```
## $`cold_storage_temp_data$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
```

- **Summer-Rainy**: VERY Significant Diffences between both (pvalue=0.5376924 ~ 53.8%)
- **Winter-Rainy**: No Significant Diffences between both (pvalue=0.0000002 ~ 0.0%)
- **Winter-Summer**: No Significant Diffences between both (pvalue=0.0000000 ~ 0.0%)
- **Comments**: The result of the above comparisms implies that between the rainy and summer seasons the difference in temperature values are very high.

3.1.9 Conclusion

We've been able to use R to statistically prove that the probability of temperature going outside the 2 - 4 C during the one-year contract was above 2.5% and less than 5% thereby determining the penalty of AMC to be 10% of AMC (annual maintenance contract).

3.2 Solution to Problem 2

```
# PROBLEM 2:
# 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 the last 35 days' temperatures. As a safety measure, the Supervisor decides
# to be vigilant to maintain the temperature at 3.9 C or below.
# Assume 3.9 C as the upper acceptable value for mean temperature and at alpha = 0.1.
# Do you feel that there is a need for some corrective action in the Cold Storage
# Plant or is it that the problem is from the procurement side from where Cold
# Storage is getting the Dairy Products? The data of the last 35 days is in
# "Cold_Storage_Mar2018.csv"
```

3.2.1 Preliminary analysis (with Assumptions and Insights)

Insight(s) from dim():

```
# Preliminary analysis for problem 2
# check dimension of dataset
dim(cold_storage_mar2018_data)
```

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

- The dataset is not a 'big' dataset with 4 columns and 35 rows.
- Sample Size: 35

Insight(s) from head():

```r
#see first 6 rows(observations) of dataset
head(cold_storage_mar2018_data)
```

```
##   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
```

- No. of Samples: 1 *(its obvious)*

Insight(s) from tail():

```r
#see last 6 rows(observations) of dataset
tail(cold_storage_mar2018_data)
```

```
##    Season Month Date Temperature
## 30 Summer   Mar   12         3.8
## 31 Summer   Mar   13         4.2
## 32 Summer   Mar   14         4.2
## 33 Summer   Mar   15         3.8
## 34 Summer   Mar   16         3.9
## 35 Summer   Mar   17         3.9
```

- This dataset contains data for temperature averages in some days in february (from head()) and some days in march.

Insight(s) from str():

```r
# check structure of dataset
str(cold_storage_mar2018_data)
```

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

- All the columns(variables) have appropriate datatypes.

Insight(s) from summary():

```
# get summary of dataset
summary(cold_storage_mar2018_data)
```

```
##     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
```

- From the minimum and maximum values of 3.8 and 4.6 respectively, as shown here, it is more than certain that the temperature went outside the 2 - 4 C given as the benchmark. We just need to know to what extent which is why we would find the probability of going below and above the given figures.
- For Temperature, which is our Point of focus here, The mean (3.974) and median (3.900) are not the same but are super close which suggests that the distribution tends to be normal. Also, this is not suprising as we can be sure that this distribution would tend to normal because it is a didtribution of averages as stated in the problem definition(Central Limit Theorem). With further analysis, we would confirm this 100%.

Insight(s) from sd():

```
# get standard deviation of concerned column: Temperature
sd_temp_mar_data = sd(cold_storage_mar2018_data$Temperature)
sd_temp_mar_data
```

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

Insight(s) from var():

```
# get variance of concerned column: Temperature
var_temp_mar_data = var(cold_storage_mar2018_data$Temperature)
var_temp_mar_data
```

```
## [1] 0.0254958
```

Insight(s) from getmode():

```
# get mode of concerned column: Temperature
getmode(cold_storage_mar2018_data$Temperature)
```

```
## [1] 3.9
```

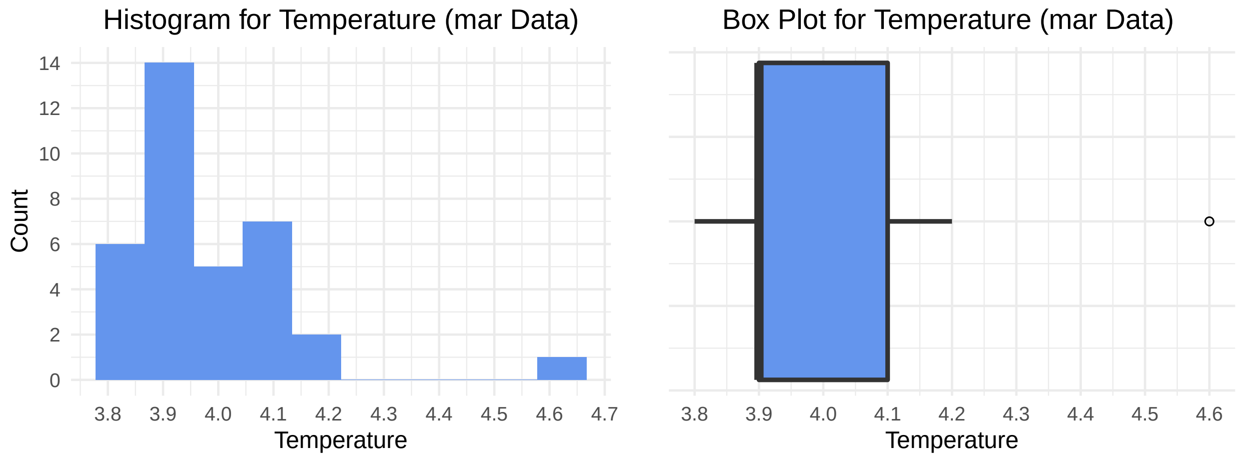
- the modal value(3.9) gotten here is not too far off from both the mean and the median which implies that it still tends towards a Normal distribution. we would need to see how it looks in a curve for final verification.

Insight(s) from ggplot():

```
# Histogram for Temperature Variable (mar Data)
temperature_histogram2 =
  ggplot(cold_storage_mar2018_data, aes(x = cold_storage_mar2018_data$Temperature)) +
  geom_histogram(fill = "cornflowerblue", bins = 10) +
  labs(title="Histogram for Temperature (mar Data)", x="Temperature", y="Count") +
  scale_x_continuous(breaks=seq(3.7,max(cold_storage_mar2018_data$Temperature)+0.1,by=0.1))+
  scale_y_continuous(breaks = seq(0, 20, by = 2)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box Plot for temperature Variable (mar Data)
temperature_boxplot2 =
  ggplot(cold_storage_mar2018_data, aes(x = 0, y = cold_storage_mar2018_data$Temperature)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color="black", fill="cornflowerblue")+
  labs(title = "Box Plot for Temperature (mar Data) ", x = "", y="Temperature") +
  scale_y_continuous(breaks=seq(3.7,max(cold_storage_mar2018_data$Temperature)+0.1,by=0.1))+
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(temperature_histogram2,temperature_boxplot2,nrow=1,ncol= 2)
```



* From the Histogram, The data looks right skewed..

Assumptions for Problem 2:

1. The dataset constitutes a sample from a normal population.
2. The observations are independent of each other.
3. The temperature variable is numeric and continuous.

3.2.2 Question 1: Which Hypothesis test shall be performed to check if corrective action is needed at the cold storage plant? Justify your answer.

- We are going to use **a one Sample upper tailed T-test** as our Hypothesis test.
- We use a One sample T-Test because the
    1. The dataset is a small sample dataset,
    2. The dataset itself is provided,
    3. The true mean was provided as 3.9 C,
    4. We can safely assume that this dataset is a sample from a normal population,
- It is one tailed (upper-tailed) because we only want to confirm that the average means does not exceed 3.9 C *(To the right of the 3.9 point of the histogram shown above)*.

```
# Q1: Which Hypothesis test shall be performed to check if corrective action is
# needed at the cold storage plant?
# ANS: A one Sample upper tailed T-test
```

3.2.3 Question 2: State the Hypothesis, perform hypothesis test and determine p-value

1. The hypothesis is stated as follows:

- H0: E[Temperature] = 3.9 C
- HA: E[Temperature] > 3.9 C

```
# Q2: State the Hypothesis, perform hypothesis test and determine p-value
#   H0: E[Temperature] = 3.9 C
#   HA: E[Temperature] > 3.9 C
```

2. To perform the Hypothesis test, we use R’s t.test function with the dataset, true mean, and alternative hypothesis value as inputs:

```
# perform upper tailed t.test on mar2018 data with true mean provided
t.test(cold_storage_mar2018_data$Temperature,mu=3.9,alternative='greater')
```

```
##
##  One Sample t-test
##
## data:  cold_storage_mar2018_data$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
```

3. **P-value:** 0.004711 ~ 0.5%

3.2.4 Question 3: Give your inference

- From the p-value of the t.test, we can see that its value (0.5%) is less than the level of significance of (0.1 ~ 10%) provided in the problem definition.
- Therefore, we **reject the hypothesis** which means that there is a need for some corrective action in the Cold Storage Plant. The problem is NOT from the procurement side from where Cold Storage is getting the Dairy Products.

```
# Since pvalue is less than level of significance, we reject the hypothesis.
# There is need for some corrective action in the Cold Storage Plant.
# The problem is NOT from the procurement side from where Cold Storage is
# getting the Dairy Products.
```

3.2.5 Conclusion

We've been able to use R to statistically determine that there is a need for the company to put in place some corrective action in the cold storage plant in an attempt to fix the issue of "dairy products going sour and often smelling".

4. Conclusion

Based on datasets provided alongside the cold storage problem definition, we have been able to successfully use R to carry out a successful statistical analysis and visualizations to generate answers to provided questions and provide useful insights.

We have also provided tailored conclusions to each problem in each problems respective section.

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

5. Appendix A – Source Code

```
#Generate the .R file to hold the source code
purl("", documentation = 0)
```

```
## [1] ".R"
```

```
#=======================================================================
#
# Practical Data Analysis - Cold storage Problem
#
#=======================================================================

# Environment Set up and Data Import

# Invoking Libraries
library(tidyverse) # contains ggplot2,dplyr,forcats,lubridate etc
library(gridExtra) # Needed for plotting multiple ggplot graphs side-by-side
library(knitr) # Necessary to generate sourcecodes from a .Rmd File

# Setup Working Directory
setwd("C:/Users/USER/Documents/El-PaDJo/R programming language/4-Project 2")

# Import and Read Input File
cold_storage_temp_data = read.csv("Cold_Storage_Temp_Data.csv")
cold_storage_mar2018_data = read.csv("Cold_Storage_Mar2018.csv")


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

# Function to calculate mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# PROBLEM 1:
# Cold Storage started its operations in Jan 2016. They are in the business of
# storing Pasteurized Fresh Whole or Skimmed Milk, Sweet Cream, Flavoured Milk Drinks.
# To ensure that there is no change of texture, body appearance, separation of fats the
# optimal temperature to be maintained is between 2 - 4 C.
# In the first year of business, they outsourced the plant maintenance work to a
# professional company with stiff penalty clauses. It was agreed that if it was
# statistically proven that the probability of temperature going outside the 2 - 4 C
# during the one-year contract 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. The average temperature data at date level is given in
# the file "Cold_Storage_Temp_Data.csv"


# Preliminary analysis for Problem 1
```

```r
# check dimension of dataset
dim(cold_storage_temp_data)

#see first 6 rows(observations) of dataset
head(cold_storage_temp_data)

#see last 6 rows(observations) of dataset
tail(cold_storage_temp_data)

# check structure of dataset
str(cold_storage_temp_data)

# get summary of dataset
summary(cold_storage_temp_data)

# get standard deviation of concerned column: Temperature
sd_temp_1yr_data = sd(cold_storage_temp_data$Temperature)
sd_temp_1yr_data

# get variance of concerned column: Temperature
var_temp_1yr_data = var(cold_storage_temp_data$Temperature)
var_temp_1yr_data

# get mode of concerned column: Temperature
getmode(cold_storage_temp_data$Temperature)

# Histogram for Temperature Variable (1Yr Data)
temperature_histogram =
  ggplot(cold_storage_temp_data, aes(x = cold_storage_temp_data$Temperature)) +
  geom_histogram(fill = "cornflowerblue", bins = 30) +
  labs(title="Histogram for Temperature (1Yr Data)", x="Temperature", y="Count") +
  scale_x_continuous(breaks=seq(1.4,max(cold_storage_temp_data$Temperature)+0.3,by=0.3))+
  scale_y_continuous(breaks = seq(0, 40, by = 5)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box Plot for temperature Variable (1Yr Data)
temperature_boxplot =
  ggplot(cold_storage_temp_data, aes(x = 0, y = cold_storage_temp_data$Temperature)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color="black", fill="cornflowerblue")+
  labs(title = "Box Plot for Temperature (1Yr Data) ", x = "", y="Temperature") +
  scale_y_continuous(breaks=seq(1.4,max(cold_storage_temp_data$Temperature)+0.3,by=0.3))+
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(temperature_histogram,temperature_boxplot,nrow=1,ncol= 2)

# Q1: calculate Temperature mean value for Summer, Winter and Rainy Season
by( cold_storage_temp_data$Temperature,INDICES = cold_storage_temp_data$Season,FUN=mean)
```

```
# Q2: calculate Temperature mean value for the year
mean_temp_1yr_data = mean(cold_storage_temp_data$Temperature)
mean_temp_1yr_data

# Q3: calculate Temperature Standard deviation for the year
sd_temp_1yr_data = sd(cold_storage_temp_data$Temperature)
sd_temp_1yr_data

# Q4: On assumption of Normal Distribution, calculate probability of temperature < 2 C
norm_prob_below_2 = pnorm(2,mean=mean_temp_1yr_data,sd=sd_temp_1yr_data)
norm_prob_below_2

# Q5: On assumption of Normal Distribution, calculate probability of temperature > 4 C
norm_prob_above_4 = 1 - pnorm(4,mean=mean_temp_1yr_data,sd=sd_temp_1yr_data)
norm_prob_above_4

# Q6: On assumption of Normal Distribution, calculate probability
# of temperature > 4 C OR temperature < 2 C
norm_prob_outside_2_4 = norm_prob_below_2 + norm_prob_above_4
norm_prob_outside_2_4

# Q7: calculate a one-way ANOVA
aov_temp_1yr_data=aov(cold_storage_temp_data$Temperature~cold_storage_temp_data$Season,
                      data=cold_storage_temp_data)

# test for differences between rainy, summer and winter Seasons
TukeyHSD(aov_temp_1yr_data)

# PROBLEM 2:
# 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 the last 35 days' temperatures. As a safety measure, the Supervisor decides
# to be vigilant to maintain the temperature at 3.9 C or below.
# Assume 3.9 C as the upper acceptable value for mean temperature and at alpha = 0.1.
# Do you feel that there is a need for some corrective action in the Cold Storage
# Plant or is it that the problem is from the procurement side from where Cold
# Storage is getting the Dairy Products? The data of the last 35 days is in
# "Cold_Storage_Mar2018.csv"


# Preliminary analysis for problem 2
# check dimension of dataset
dim(cold_storage_mar2018_data)

#see first 6 rows(observations) of dataset
head(cold_storage_mar2018_data)

#see last 6 rows(observations) of dataset
tail(cold_storage_mar2018_data)

# check structure of dataset
str(cold_storage_mar2018_data)
```

```r
# get summary of dataset
summary(cold_storage_mar2018_data)

# get standard deviation of concerned column: Temperature
sd_temp_mar_data = sd(cold_storage_mar2018_data$Temperature)
sd_temp_mar_data

# get variance of concerned column: Temperature
var_temp_mar_data = var(cold_storage_mar2018_data$Temperature)
var_temp_mar_data

# get mode of concerned column: Temperature
getmode(cold_storage_mar2018_data$Temperature)

# Histogram for Temperature Variable (mar Data)
temperature_histogram2 =
  ggplot(cold_storage_mar2018_data, aes(x = cold_storage_mar2018_data$Temperature)) +
  geom_histogram(fill = "cornflowerblue", bins = 10) +
  labs(title="Histogram for Temperature (mar Data)", x="Temperature", y="Count") +
  scale_x_continuous(breaks=seq(3.7,max(cold_storage_mar2018_data$Temperature)+0.1,by=0.1))+
  scale_y_continuous(breaks = seq(0, 20, by = 2)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box Plot for temperature Variable (mar Data)
temperature_boxplot2 =
  ggplot(cold_storage_mar2018_data, aes(x = 0, y = cold_storage_mar2018_data$Temperature)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color="black", fill="cornflowerblue")+
  labs(title = "Box Plot for Temperature (mar Data) ", x = "", y="Temperature") +
  scale_y_continuous(breaks=seq(3.7,max(cold_storage_mar2018_data$Temperature)+0.1,by=0.1))+
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(temperature_histogram2,temperature_boxplot2,nrow=1,ncol= 2)

# Q1: Which Hypothesis test shall be performed to check if corrective action is
# needed at the cold storage plant?
# ANS: A one Sample upper tailed T-test


# Q2: State the Hypothesis, perform hypothesis test and determine p-value
#   H0: E[Temperature] = 3.9 C
#   HA: E[Temperature] > 3.9 C


# perform upper tailed t.test on mar2018 data with true mean provided
t.test(cold_storage_mar2018_data$Temperature,mu=3.9,alternative='greater')


# Since pvalue is less than level of significance, we reject the hypothesis.
# There is need for some corrective action in the Cold Storage Plant.
# The problem is NOT from the procurement side from where Cold Storage is
# getting the Dairy Products.


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