Table of Contents

1. Project Objective
2. Assumptions
3. Exploratory Data Analysis – Step by step approach.
4. Invoke Libraries
5. Setup working directory
6. Import and read the Dataset
7. Variable Identification
8. Univariate Analysis
9. Bi-Variate Analysis
10. Missing Value Identification
11. Outlier Identification
12. Variable Transformation/Feature Creation
13. Conclusion
14. Appendix – Source Code (\*Please refer to the source code for derived answers)

Project Objective

The objective of cold storage case study is twofold. In the first dataset Cold\_Storage\_Temp\_Data.csv” set we will be doing exploratory analysis along with an attempt to efficiently determine the penalty for AMC company given that in a one year contract the probability of temperature of the cold storage does not exceed and fall short of the permissible limit of 2 - 4 Deg C.

In the second dataset “Cold\_Storage\_Mar2018.csv”, there has been a feedback from the end consumer that the dairy products are going sour and often smelling, we will implement z-test and t-test to test our hypothesis and effectively determine if the problem is from the procurement side or if some corrective measure has to be implemented at the cold storage plant.

Assumptions

* Temperature in the first data set is normally distributed.
* The derived sample data set is considered to be normally distributed.
* In the second data set, the upper acceptable temperature range is 3.9 deg C. (Mu:3.9)
* Alpha or level of significance is 0.1.
* Confidence level is 0.9.
* Standard deviation for z-test will be the calculated value from the first dataset.

Exploratory Data Analysis

## Invoking Libraries

library(ggplot2)  
library(lattice)  
library(rmarkdown)

## Setting up the working directory

setwd("C:/Users/Hp/Desktop/R Programming")  
getwd()

## [1] "C:/Users/Hp/Desktop/R Programming"

## Importing data

storage <- read.csv("Cold\_Storage\_Temp\_Data.csv")

## 

## Structure

str(storage)

## '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 ...

This dataset is a dataframe with 2 factor variables, 1 int variable and 1 numeric variable. There are three 3 levels to the season variable (Rainy, Summer and Winter). Month variable has 12 levels which are the months of a year. There are a total of 365 rows and columns.

## Summary

summary(storage)

## 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

This summary is useful to pinpoint the mean temperature of the entire year. Number of days in every season is also verified.

## Summary of dataset grouped by indices/factor “season”

by(storage, INDICES = Season, FUN = summary)

## Season: Rainy  
## Season Month Date Temperature   
## Rainy :122 Aug :31 Min. : 1.00 Min. :1.700   
## Summer: 0 Jul :31 1st Qu.: 8.00 1st Qu.:2.500   
## Winter: 0 Jun :30 Median :16.00 Median :2.900   
## Sep :30 Mean :15.75 Mean :3.039   
## Apr : 0 3rd Qu.:23.00 3rd Qu.:3.300   
## Dec : 0 Max. :31.00 Max. :5.000   
## (Other): 0   
## --------------------------------------------------------   
## Season: Summer  
## Season Month Date Temperature   
## Rainy : 0 Mar :31 Min. : 1.00 Min. :2.500   
## Summer:120 May :31 1st Qu.: 8.00 1st Qu.:2.800   
## Winter: 0 Apr :30 Median :15.50 Median :3.200   
## Feb :28 Mean :15.53 Mean :3.153   
## Aug : 0 3rd Qu.:23.00 3rd Qu.:3.400   
## Dec : 0 Max. :31.00 Max. :4.100   
## (Other): 0   
## --------------------------------------------------------   
## Season: Winter  
## Season Month Date Temperature   
## Rainy : 0 Dec :31 Min. : 1.00 Min. :1.900   
## Summer: 0 Jan :31 1st Qu.: 8.00 1st Qu.:2.400   
## Winter:123 Oct :31 Median :16.00 Median :2.600   
## Nov :30 Mean :15.88 Mean :2.701   
## Apr : 0 3rd Qu.:23.50 3rd Qu.:2.900   
## Aug : 0 Max. :31.00 Max. :3.900   
## (Other): 0

## Mean temperature in rainy season

RainySeason <- storage[Season == "Rainy", ]  
mean(RainySeason$Temperature)

## [1] 3.039344

## Mean temperature in summer season

Summerseason <- storage[Season == "Summer", ]  
mean(Summerseason$Temperature)

## [1] 3.153333

## Mean temperature in winter season

Winterseason <- storage[Season == "Winter", ]  
mean(Winterseason$Temperature)

## [1] 2.700813

## Overall mean temperature for the entire year

Mean <- mean(Temperature)  
Mean

## [1] 2.96274

## Standard Deviation of temperature for the entire year

StandardDeviation <- sd(Temperature)  
StandardDeviation

## [1] 0.508589

## Probability of temperature falling below 2 deg C

pnorm(2,Mean,StandardDeviation)

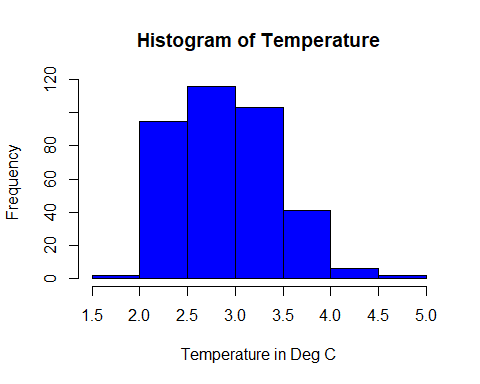
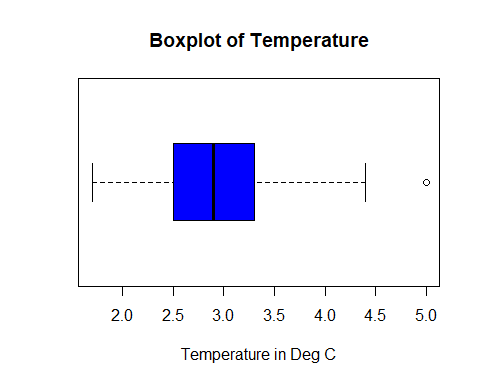
## [1] 0.02918146

## Probability of temperature going above 4 deg C

1 - pnorm(4,Mean,StandardDeviation)

## [1] 0.02070077

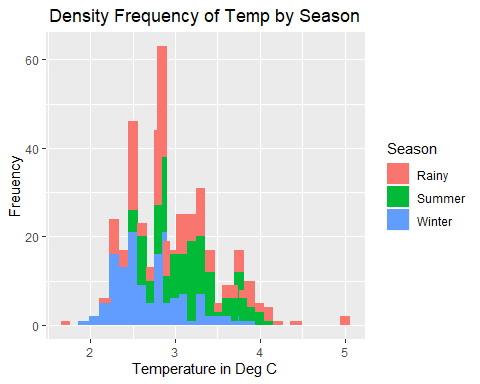
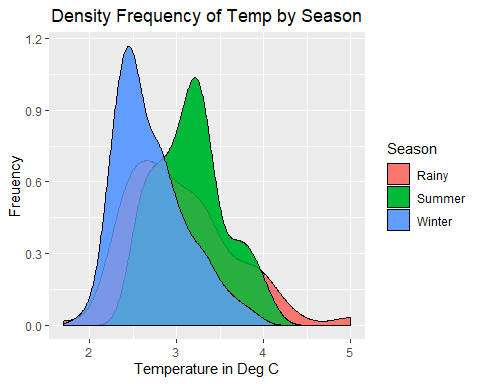
## Univariate Analysis



**Fig 1.1 Fig 1.2**

Fig 1.1 represents a boxplot of temperature for the entire year. The mean is understood to be around 2.9 along with some outliers.

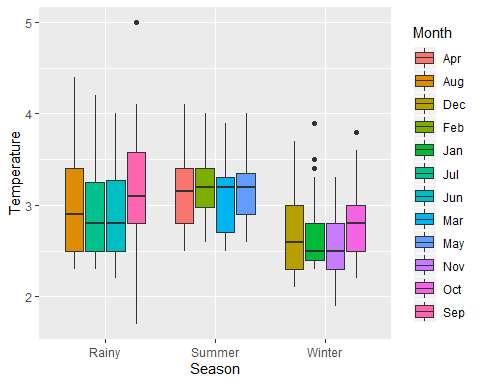
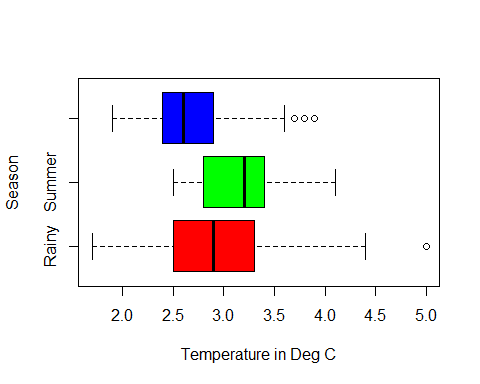
Fig 1.2 is a histogram of temperature for the entire year. We can interpret that 2.5 - 3.0 deg C is usually the most frequent occurring temperature range on an average. In other words, chance of temperature being in 2.5-3.0 deg C range is more likely than the other distributed ranges.



**Fig 1.3 Fig 1.4**

Fig 1.3 and Fig 1.4 is a density distribution and histogram of temperature for all three seasons. Winter season temperature is likely to be in the lower range whereas summer temperature is likely to be in the mid-upper range. Rainy seasons are more unpredictable with extreme temperatures are likely on some days

## Bivariate Analysis



**Fig 2.1 Fig 2.2**

Fig 2.1 is a boxplot representation of temperature distribution for all three seasons. The mean temperature for summer, winter and rainy season is 3.15, 2,70 and 3.04 Deg C respectively. Winter season has the most outliers followed by rainy season. In other words, it is more likely that there would be an extreme temperature from what is considered to be average temperature range for winter and rainy seasons.

Fig 2.2 and 2.3 shows us that entire year is divided into 3 seasons with 4 months each. We can more accurately find out which month the extremities occurred(outliers) along with the season. Also temperature mean for every month can be compared. For example, we can conclude that lowest mean temperatures observed were in the months of Jan and Nov.

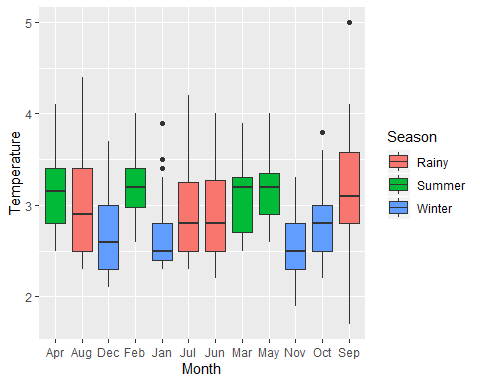


Fig 2.3

**Problem 2:**

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

## Hypothesis

## H0:Mu => 3.9 , Ha:Mu < 3.9

## z-test

Xbar1 <- 2.96  
sigm <- 0.50  
Mu <- 3.9  
n <- 35  
zstat1 <- (Xbar1 - Mu)/(sigm/(n^0.5))  
zstat1

## [1] -11.12223

zcriticlvalue <- qnorm(0.1)  
zcriticlvalue

## [1] -1.281552

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

## Hypothesis

# H0:Mu => 3.9 , Ha:Mu < 3.9

# t-test

Xbar <- mean(ColdStorageMarch2018$Temperature)  
sigma <- sd(ColdStorageMarch2018$Temperature)  
Mu <- 3.9  
n <- 35  
tstat <- (Xbar - Mu)/(sigma/(n^0.5))  
tstat

## [1] 2.752359

Pvalue <- pt(tstat,34)  
Pvalue

## [1] 0.9952888

## Alternative way to calculate pvalue using t-test

t.test(ColdStorageMarch2018$Temperature, mu = 3.9 , alternative = "less" , conf.level = 0.9)

##   
## One Sample t-test  
##   
## data: ColdStorageMarch2018$Temperature  
## t = 2.7524, df = 34, p-value = 0.9953  
## alternative hypothesis: true mean is less than 3.9  
## 90 percent confidence interval:  
## -Inf 4.00956  
## sample estimates:  
## mean of x   
## 3.974286

1. Give your inference after doing both the tests.

The conclusion one may derive from running a one sample **z-test** is that since the z-computed is -11.12 which falls under the z-critical range of -1.28 we **reject the null hypothesis which is H0=>3.9**. This is an indication that the temperature was maintained below 3.9 deg C and did not exceed the upper acceptable limit. We can safely conclude that the **problem is from the procurement side were the dairy product are being obtained from.**

**T-test** indicates that the p-value is 0.9953 which is again more than our significance level (alpha) of 0.1. **We can effectively say that we fail to reject the null hypothesis which is H0=>3.9** and the temperature within the cold storage exceeded the upper permissible limit resulting in smelling and sour dairy products. **Corrective measures and actions are required at the cold storage.**

**Appendix – Source Code**

ColdStorage

knitr::opts\_knit$set(root.dir = 'C:/Users/Hp/Desktop/R Programming')

## Invoking Libraries

library(ggplot2)  
library(lattice)  
library(rmarkdown)  
library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.1

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Seting up the working directory

setwd("C:/Users/Hp/Desktop/R Programming")  
getwd()

## [1] "C:/Users/Hp/Desktop/R Programming"

## Importing data

storage <- read.csv("Cold\_Storage\_Temp\_Data.csv")

## Top 15 entries on the dataset

head(storage , 15)

## 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  
## 11 Winter Jan 11 2.4  
## 12 Winter Jan 12 2.5  
## 13 Winter Jan 13 2.6  
## 14 Winter Jan 14 2.8  
## 15 Winter Jan 15 3.4

## Bottom 15 entries of the dataset

tail(storage , 15)

## Season Month Date Temperature  
## 351 Winter Dec 17 2.5  
## 352 Winter Dec 18 2.5  
## 353 Winter Dec 19 2.5  
## 354 Winter Dec 20 2.5  
## 355 Winter Dec 21 2.6  
## 356 Winter Dec 22 3.3  
## 357 Winter Dec 23 3.0  
## 358 Winter Dec 24 3.7  
## 359 Winter Dec 25 3.2  
## 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

## Dimensions of dataset

dim(storage)

## [1] 365 4

## Column names of the dataset

colnames(storage)

## [1] "Season" "Month" "Date" "Temperature"

## Structure of the dataset

str(storage)

## '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(storage)

## 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

## Referencing column names without $

attach(storage)

## Summary of dataset grouped by indices/factor “season”

by(storage, INDICES = Season, FUN = summary)

## Season: Rainy  
## Season Month Date Temperature   
## Rainy :122 Aug :31 Min. : 1.00 Min. :1.700   
## Summer: 0 Jul :31 1st Qu.: 8.00 1st Qu.:2.500   
## Winter: 0 Jun :30 Median :16.00 Median :2.900   
## Sep :30 Mean :15.75 Mean :3.039   
## Apr : 0 3rd Qu.:23.00 3rd Qu.:3.300   
## Dec : 0 Max. :31.00 Max. :5.000   
## (Other): 0   
## --------------------------------------------------------   
## Season: Summer  
## Season Month Date Temperature   
## Rainy : 0 Mar :31 Min. : 1.00 Min. :2.500   
## Summer:120 May :31 1st Qu.: 8.00 1st Qu.:2.800   
## Winter: 0 Apr :30 Median :15.50 Median :3.200   
## Feb :28 Mean :15.53 Mean :3.153   
## Aug : 0 3rd Qu.:23.00 3rd Qu.:3.400   
## Dec : 0 Max. :31.00 Max. :4.100   
## (Other): 0   
## --------------------------------------------------------   
## Season: Winter  
## Season Month Date Temperature   
## Rainy : 0 Dec :31 Min. : 1.00 Min. :1.900   
## Summer: 0 Jan :31 1st Qu.: 8.00 1st Qu.:2.400   
## Winter:123 Oct :31 Median :16.00 Median :2.600   
## Nov :30 Mean :15.88 Mean :2.701   
## Apr : 0 3rd Qu.:23.50 3rd Qu.:2.900   
## Aug : 0 Max. :31.00 Max. :3.900   
## (Other): 0

## Mean temperature in rainy season

RainySeason <- storage[Season == "Rainy", ]  
mean(RainySeason$Temperature)

## [1] 3.039344

## Mean temperature in summer season

Summerseason <- storage[Season == "Summer", ]  
mean(Summerseason$Temperature)

## [1] 3.153333

## Mean temperature in winter season

Winterseason <- storage[Season == "Winter", ]  
mean(Winterseason$Temperature)

## [1] 2.700813

## Overall mean temperature for the entire year

Mean <- mean(Temperature)  
Mean

## [1] 2.96274

## Standard Deviation of temperature for the entire year

StandardDeviation <- sd(Temperature)  
StandardDeviation

## [1] 0.508589

## Maximum & Minimum temperature in a year

max(Temperature)

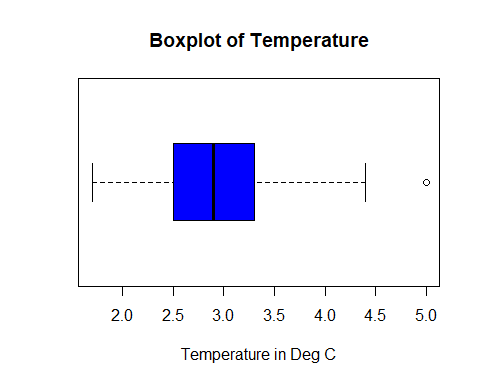
## [1] 5

min(Temperature)

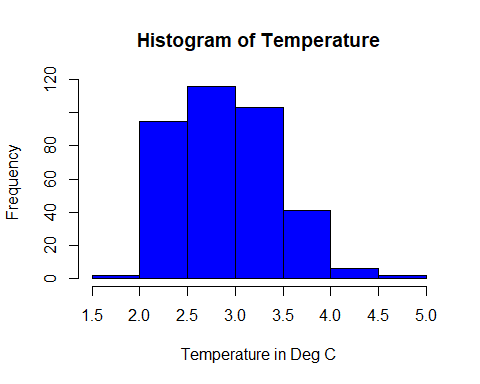
## [1] 1.7

## Univariate Analysis

boxplot(Temperature , xlab = "Temperature in Deg C" , main = "Boxplot of Temperature" , col = "Blue",  
 horizontal = TRUE)

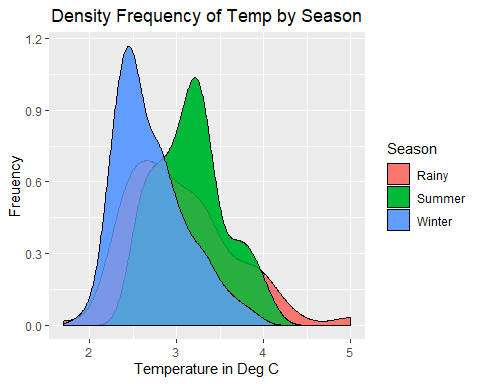


hist(Temperature, xlab = "Temperature in Deg C", col = "Blue")

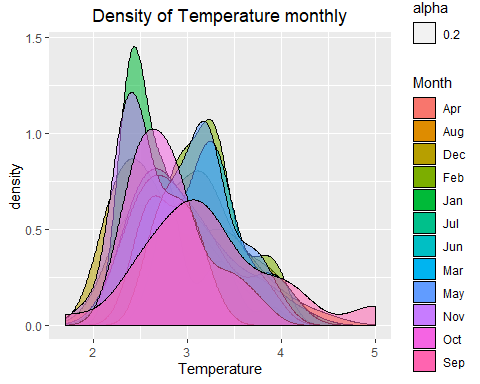


## Univariate Analysis using ggplot2

qplot(Temperature, data = storage , geom = "density" , fill = Season, xlab = "Temperature in Deg C" , ylab = "Freuency") + ggtitle("Density Frequency of Temp by Season") + theme(plot.title = element\_text(hjust = 0.5)) + geom\_density(alpha =0.8)

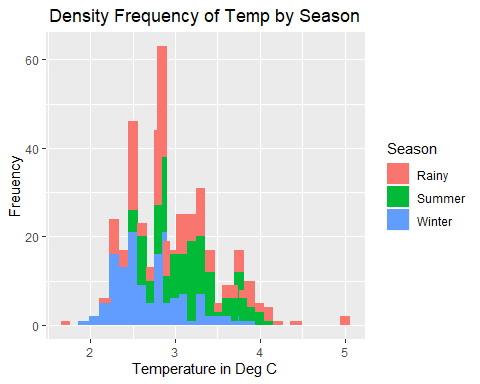


ggplot(storage, aes(x = Temperature, fill = Month, alpha = 0.2)) + geom\_density() + ggtitle("Density of Temperature monthly") + theme(plot.title = element\_text(hjust = 0.5))



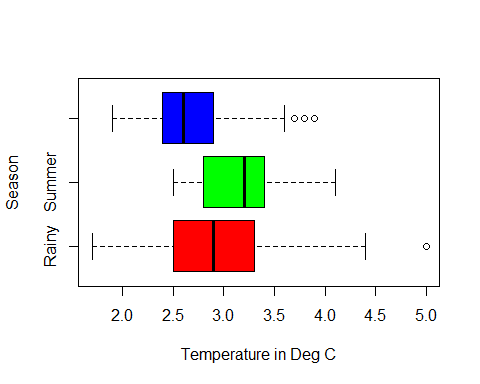
qplot(Temperature, data = storage , fill = Season,   
 xlab = "Temperature in Deg C" , ylab = "Freuency") + ggtitle("Density Frequency of Temp by Season") + theme(plot.title = element\_text(hjust = 0.5)) + geom\_bar()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

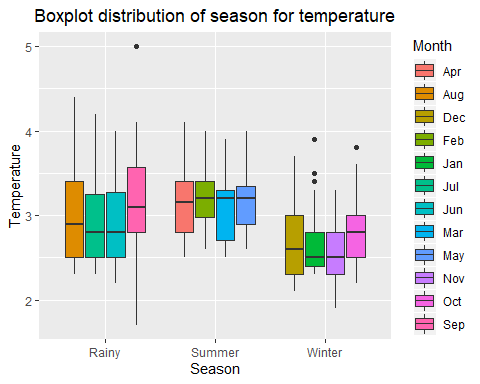


## Bivariate Analysis

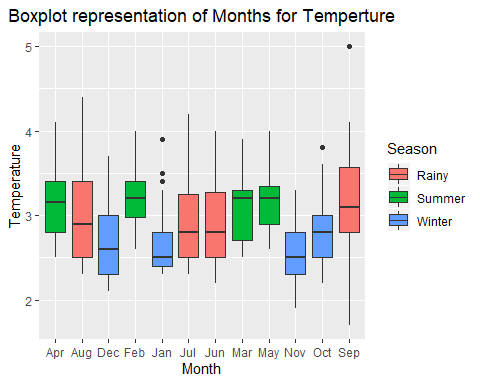
boxplot(Temperature~Season , col = c("red","green","blue"), horizontal = TRUE ,  
 xlab = "Temperature in Deg C", ylab = "Season")



ggplot(storage, aes(x = Season, y = Temperature, fill = Month)) + geom\_boxplot() + ggtitle("Boxplot distribution of season for temperature") + theme(plot.title = element\_text(hjust = 0.5))



ggplot(storage, aes(x = Month, y = Temperature, fill = Season )) + geom\_boxplot() + ggtitle("Boxplot representation of Months for Temperture") + theme(plot.title = element\_text(hjust = 0.5))



## Probability of temperature falling below 2 deg C

pnorm(2,Mean,StandardDeviation)

## [1] 0.02918146

## Probability of temperature going above 4 deg C

1 - pnorm(4,Mean,StandardDeviation)

## [1] 0.02070077

ColdStorageSample

knitr::opts\_knit$set(root.dir = 'C:/Users/Hp/Desktop/R Programming')

## Importing Sample Data

setwd("C:/Users/Hp/Desktop/R Programming")  
getwd()

## [1] "C:/Users/Hp/Desktop/R Programming"

ColdStorageMarch2018 <- read.csv("Cold\_Storage\_Mar2018.csv" , header = TRUE)

## Hypothesis

## H0:Mu => 3.9 , Ha:Mu < 3.9

## z-test

Xbar1 <- 2.96  
sigm <- 0.50  
Mu <- 3.9  
n <- 35  
zstat1 <- (Xbar1 - Mu)/(sigm/(n^0.5))  
zstat1

## [1] -11.12223

zcriticlvalue <- qnorm(0.1)  
zcriticlvalue

## [1] -1.281552

## Hypothesis

# H0:Mu => 3.9 , Ha:Mu < 3.9

# t-test

Xbar <- mean(ColdStorageMarch2018$Temperature)  
sigma <- sd(ColdStorageMarch2018$Temperature)  
Mu <- 3.9  
n <- 35  
tstat <- (Xbar - Mu)/(sigma/(n^0.5))  
tstat

## [1] 2.752359

Pvalue <- pt(tstat,34)  
Pvalue

## [1] 0.9952888

## Alternative way to calculate pvalue using t-test

t.test(ColdStorageMarch2018$Temperature, mu = 3.9 , alternative = "less" , conf.level = 0.9)

##   
## One Sample t-test  
##   
## data: ColdStorageMarch2018$Temperature  
## t = 2.7524, df = 34, p-value = 0.9953  
## alternative hypothesis: true mean is less than 3.9  
## 90 percent confidence interval:  
## -Inf 4.00956  
## sample estimates:  
## mean of x   
## 3.974286