# USE OF WEATHER DATA TO FORECAST NEXT DAY’S WEATHER

## INTRODUCTION

This project work uses weather data for the Australian Capital Territory, Canberra

to forecast the next day’s weather. R was used for the data analysis.

## # Data input

The data must be read into R first to determine the observations and variables. The dimension of the data has 242 observations with 22 variables. The head displays the first six rows of the data (weather-2022)

```
#Q1
dat <- read.csv("C:/Users/pelza/OneDrive/Desktop/Machine Learning/weather-202
2.csv", encoding= 'UTF-8', check.names=FALSE, header = F)
dim(dat)

## [1] 242  22

head(dat)

##    V1        V2                          V3                          V4
## 1          Date Minimum temperature (\xb0C) Maximum temperature (\xb0C)
## 2 JAN 2022-01-1                        10.6                        30.5
## 3 JAN 2022-01-2                        13.3                        32.6
## 4 JAN 2022-01-3                        13.6                        28.9
## 5 JAN 2022-01-4                        12.6                        27.6
## 6 JAN 2022-01-5                        13.6                        26.1
##              V5               V6               V7
## 1 Rainfall (mm) Evaporation (mm) Sunshine (hours)
## 2             0             <NA>             <NA>
## 3             0             <NA>             <NA>
## 4             5             <NA>             <NA>
## 5           0.2             <NA>             <NA>
## 6             0             <NA>             <NA>
##                                 V8                                V9
## 1 Direction of maximum wind gust  Speed of maximum wind gust (km/h)
## 2                              NNW                                31
## 3                               SE                                59
## 4                                E                                46
## 5                              ESE                                35
## 6                              ESE                                43
##                          V10                     V11
V12
## 1 Time of maximum wind gust 9am Temperature (\xb0C) 9am relative humidity
(%)
```

```
## 2                     10:35                    19.1
68
## 3                     17:39                    22.8
64
## 4                     15:49                    21.9
57
## 5                     13:19                    18.4
88
## 6                     12:53                    20.5
66
##                        V13                V14                   V15
## 1 9am cloud amount (oktas) 9am wind direction 9am wind speed (km/h)
## 2                     <NA>                  N                     4
## 3                     <NA>                  N                     9
## 4                     <NA>                 SE                     7
## 5                        8                 SE                    13
## 6                        8                 SE                    11
##                      V16                     V17                       V18
## 1 9am MSL pressure (hPa) 3pm Temperature (\xb0C) 3pm relative humidity (%)
## 2                 1013.8                    29.8                        35
## 3                 1009.9                    31.4                        25
## 4                 1010.6                    27.4                        48
## 5                 1011.9                    26.9                        48
## 6                 1012.7                    25.1                        61
##                        V19                V20                   V21
## 1 3pm cloud amount (oktas) 3pm wind direction 3pm wind speed (km/h)
## 2                     <NA>                 NW                    13
## 3                        6                  W                    11
## 4                        2                ENE                    24
## 5                     <NA>                  E                    17
## 6                        8                ESE                    26
##                      V22
## 1 3pm MSL pressure (hPa)
## 2                 1008.2
## 3                 1006.3
## 4                 1007.8
## 5                 1009.5
## 6                 1010.5
```

```
str(dat)
```

```
## 'data.frame':    242 obs. of  22 variables:
##  $ V1 : chr  "" "JAN" "JAN" "JAN" ...
##  $ V2 : chr  "Date" "2022-01-1" "2022-01-2" "2022-01-3" ...
##  $ V3 : chr  "Minimum temperature (\xb0C)" "10.6" "13.3" "13.6" ...
##  $ V4 : chr  "Maximum temperature (\xb0C)" "30.5" "32.6" "28.9" ...
##  $ V5 : chr  "Rainfall (mm)" "0" "0" "5" ...
##  $ V6 : chr  "Evaporation (mm)" NA NA NA ...
##  $ V7 : chr  "Sunshine (hours)" NA NA NA ...
##  $ V8 : chr  "Direction of maximum wind gust " "NNW" "SE" "E" ...
##  $ V9 : chr  "Speed of maximum wind gust (km/h)" "31" "59" "46" ...
##  $ V10: chr  "Time of maximum wind gust" "10:35" "17:39" "15:49" ...
##  $ V11: chr  "9am Temperature (\xb0C)" "19.1" "22.8" "21.9" ...
##  $ V12: chr  "9am relative humidity (%)" "68" "64" "57" ...
##  $ V13: chr  "9am cloud amount (oktas)" NA NA NA ...
##  $ V14: chr  "9am wind direction" "N" "N" "SE" ...
##  $ V15: chr  "9am wind speed (km/h)" "4" "9" "7" ...
##  $ V16: chr  "9am MSL pressure (hPa)" "1013.8" "1009.9" "1010.6" ...
##  $ V17: chr  "3pm Temperature (\xb0C)" "29.8" "31.4" "27.4" ...
##  $ V18: chr  "3pm relative humidity (%)" "35" "25" "48" ...
##  $ V19: chr  "3pm cloud amount (oktas)" NA "6" "2" ...
##  $ V20: chr  "3pm wind direction" "NW" "W" "ENE" ...
##  $ V21: chr  "3pm wind speed (km/h)" "13" "11" "24" ...
##  $ V22: chr  "3pm MSL pressure (hPa)" "1008.2" "1006.3" "1007.8" ...
```

# Data Cleaning and Preparation

```
# Q1. Removing Time of Maximum Wind Gust
df<-dat[,-c(10)]
```

#Comment

```
# "df<-dat [,-c(10)]"is used to remove the 10th column Time of Maximum Wind

Gust

#Check the Dimension,and str again
dim(df)

## [1] 242  21

str(df)

## 'data.frame':    242 obs. of  21 variables:
##  $ V1 : chr  "" "JAN" "JAN" "JAN" ...
##  $ V2 : chr  "Date" "2022-01-1" "2022-01-2" "2022-01-3" ...
##  $ V3 : chr  "Minimum temperature (\xb0C)" "10.6" "13.3" "13.6" ...
##  $ V4 : chr  "Maximum temperature (\xb0C)" "30.5" "32.6" "28.9" ...
##  $ V5 : chr  "Rainfall (mm)" "0" "0" "5" ...
##  $ V6 : chr  "Evaporation (mm)" NA NA NA ...
##  $ V7 : chr  "Sunshine (hours)" NA NA NA ...
##  $ V8 : chr  "Direction of maximum wind gust " "NNW" "SE" "E" ...
##  $ V9 : chr  "Speed of maximum wind gust (km/h)" "31" "59" "46" ...
##  $ V11: chr  "9am Temperature (\xb0C)" "19.1" "22.8" "21.9" ...
##  $ V12: chr  "9am relative humidity (%)" "68" "64" "57" ...
##  $ V13: chr  "9am cloud amount (oktas)" NA NA NA ...
##  $ V14: chr  "9am wind direction" "N" "N" "SE" ...
##  $ V15: chr  "9am wind speed (km/h)" "4" "9" "7" ...
##  $ V16: chr  "9am MSL pressure (hPa)" "1013.8" "1009.9" "1010.6" ...
##  $ V17: chr  "3pm Temperature (\xb0C)" "29.8" "31.4" "27.4" ...
##  $ V18: chr  "3pm relative humidity (%)" "35" "25" "48" ...
##  $ V19: chr  "3pm cloud amount (oktas)" NA "6" "2" ...
##  $ V20: chr  "3pm wind direction" "NW" "W" "ENE" ...
##  $ V21: chr  "3pm wind speed (km/h)" "13" "11" "24" ...
##  $ V22: chr  "3pm MSL pressure (hPa)" "1008.2" "1006.3" "1007.8" ...
# Q2. Rename Dataset Column
```

**#Comment**: The data set has been renamed as shown below because they were too long.

```
names(df) <- c("Month", "Date", "MinTemp", "MaxTemp", "Rainfall",
"Evaporation", "Sunshine", "WindGustDir", "WindGustSpeed",
"Temp9am", "Humidity9am", "Cloud9am", "WindDir9am",
"WindSpeed9am", "Pressure9am", "Temp3pm", "Humidity3pm",
"Cloud3pm", "WindDir3pm", "WindSpeed3pm", "Pressure3pm")

dim(df);

## [1] 242  21

names(df)

##  [1] "Month"         "Date"          "MinTemp"       "MaxTemp"
##  [5] "Rainfall"      "Evaporation"   "Sunshine"      "WindGustDir"
##  [9] "WindGustSpeed" "Temp9am"       "Humidity9am"   "Cloud9am"
## [13] "WindDir9am"    "WindSpeed9am"  "Pressure9am"   "Temp3pm"
## [17] "Humidity3pm"   "Cloud3pm"      "WindDir3pm"    "WindSpeed3pm"
## [21] "Pressure3pm"
```

*#Q3. Printing out Unique values*

**#Comment:** We printed out the unique values using the codes below:

```
vnames <- colnames(df)
n <- nrow(df)
out <- NULL
for (j in 1:ncol(df)){
  vname <- colnames(df)[j]
  x <- as.vector(df[,j])
  n1 <- sum(is.na(x), na.rm=TRUE)  # NA
  n2 <- sum(x=="NA", na.rm=TRUE) # "NA"
  n3 <- sum(x==" ", na.rm=TRUE)  # missing
  nmiss <- n1 + n2 + n3
  nmiss <- sum(is.na(x))
  ncomplete <- n-nmiss
  out <- rbind(out, c(col.num=j, v.name=vname, mode=mode(x), n.level=length(u
nique(x)),
                      ncom=ncomplete, nmiss= nmiss, miss.prop=nmiss/n))
}
out <- as.data.frame(out)
```

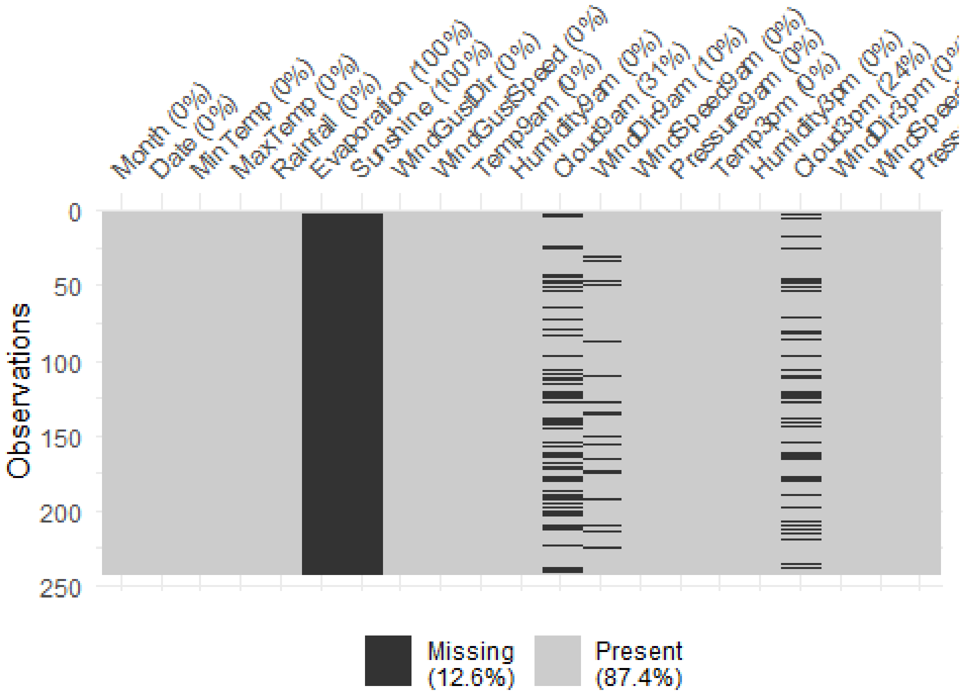
```
row.names(out) <- NULL
out
```

```
##    col.num        v.name      mode n.level ncom nmiss           miss.prop
## 1        1         Month character       9  242     0                   0
## 2        2          Date character     242  242     0                   0
## 3        3       MinTemp character     151  242     0                   0
## 4        4       MaxTemp character     153  241     1 0.00413223140495868
## 5        5      Rainfall character      45  242     0                   0
## 6        6   Evaporation character       2    1   241   0.995867768595041
## 7        7      Sunshine character       2    1   241   0.995867768595041
## 8        8   WindGustDir character      18  241     1 0.00413223140495868
## 9        9 WindGustSpeed character      31  241     1 0.00413223140495868
## 10      10       Temp9am character     146  242     0                   0
## 11      11   Humidity9am character      44  242     0                   0
## 12      12      Cloud9am character      10  168    74   0.305785123966942
## 13      13    WindDir9am character      17  217    25   0.103305785123967
## 14      14  WindSpeed9am character      23  242     0                   0
## 15      15   Pressure9am character     156  242     0                   0
## 16      16       Temp3pm character     146  242     0                   0
## 17      17   Humidity3pm character      67  242     0                   0
## 18      18      Cloud3pm character      10  185    57   0.235537190082645
## 19      19    WindDir3pm character      18  241     1 0.00413223140495868
## 20      20  WindSpeed3pm character      23  242     0                   0
## 21      21   Pressure3pm character     159  242     0                   0
```

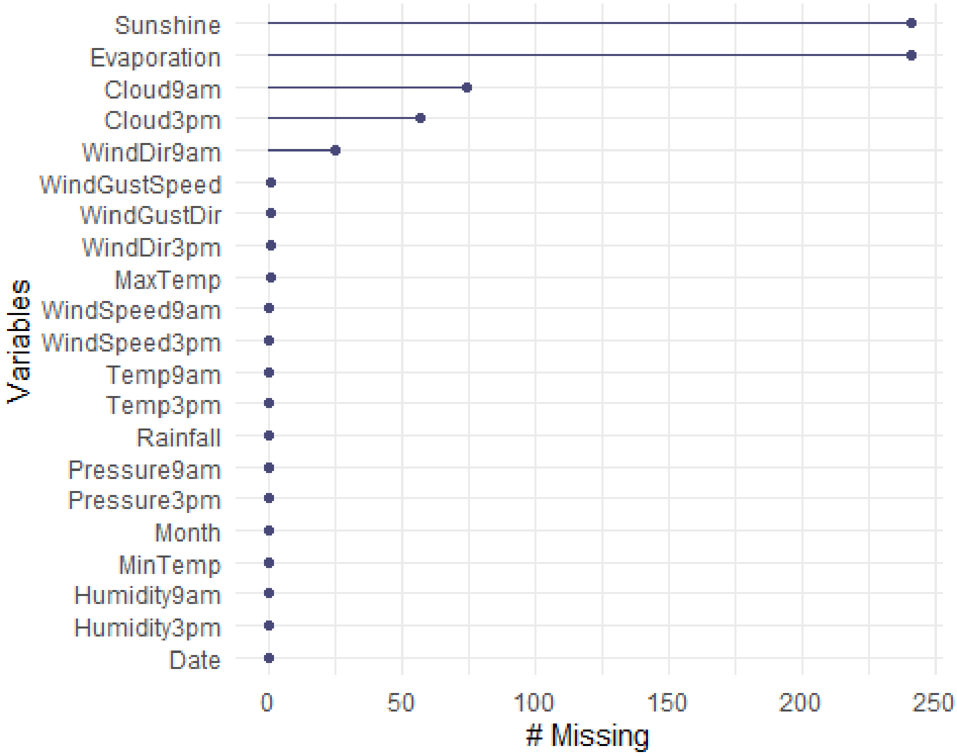
```
colMeans(is.na(df))
```

```
##         Month          Date       MinTemp       MaxTemp      Rainfall 
##   0.000000000   0.000000000   0.000000000   0.004132231   0.000000000 
##   Evaporation      Sunshine   WindGustDir WindGustSpeed       Temp9am 
##   0.995867769   0.995867769   0.004132231   0.004132231   0.000000000 
##   Humidity9am      Cloud9am    WindDir9am  WindSpeed9am   Pressure9am 
##   0.000000000   0.305785124   0.103305785   0.000000000   0.000000000 
##       Temp3pm   Humidity3pm      Cloud3pm    WindDir3pm  WindSpeed3pm 
##   0.000000000   0.000000000   0.235537190   0.004132231   0.000000000 
##   Pressure3pm 
##   0.000000000
```

```
######## Missing values by visualization ###############
library(naniar)
vis_miss(df)
```



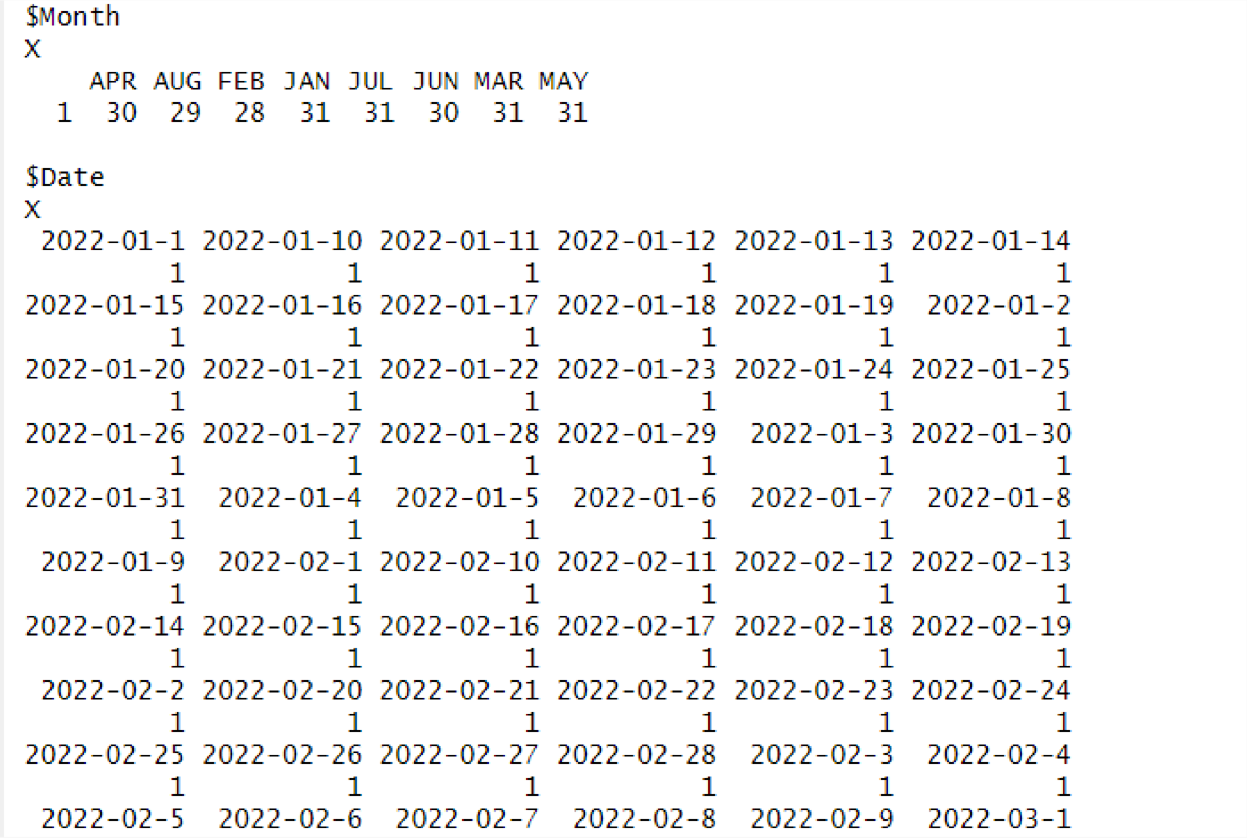
```r
gg_miss_var(df)
```



```
################### Imputing the data ###################
#install.packages("mice")
#library(mice, quietly=TRUE)
#fit.mice <- mice(df, m=1, maxit=50, method='pmm', seed=5474, printFlag=FALSE
)
#df <- complete(fit.mice, 1)

#df <- na.omit(df)
#dim(df)
```

```
#Q4. Showing Frequency Table of Each Value
#apply(df, 2, FUN = function(X){table(X, useNA="ifany")})
```

**#Comment**: A small section of the frequency table is shown below:



```
$Month
X
    APR AUG FEB JAN JUL JUN MAR MAY
  1  30  29  28  31  31  30  31  31

$Date
X
 2022-01-1 2022-01-10 2022-01-11 2022-01-12 2022-01-13 2022-01-14
         1          1          1          1          1          1
2022-01-15 2022-01-16 2022-01-17 2022-01-18 2022-01-19  2022-01-2
         1          1          1          1          1          1
2022-01-20 2022-01-21 2022-01-22 2022-01-23 2022-01-24 2022-01-25
         1          1          1          1          1          1
2022-01-26 2022-01-27 2022-01-28 2022-01-29  2022-01-3 2022-01-30
         1          1          1          1          1          1
2022-01-31  2022-01-4  2022-01-5  2022-01-6  2022-01-7  2022-01-8
         1          1          1          1          1          1
 2022-01-9  2022-02-1 2022-02-10 2022-02-11 2022-02-12 2022-02-13
         1          1          1          1          1          1
2022-02-14 2022-02-15 2022-02-16 2022-02-17 2022-02-18 2022-02-19
         1          1          1          1          1          1
 2022-02-2 2022-02-20 2022-02-21 2022-02-22 2022-02-23 2022-02-24
         1          1          1          1          1          1
2022-02-25 2022-02-26 2022-02-27 2022-02-28  2022-02-3  2022-02-4
         1          1          1          1          1          1
 2022-02-5  2022-02-6  2022-02-7  2022-02-8  2022-02-9  2022-03-1
```

```
df <- df[-1, ]


#changing the value Calm to 0
```

#Comment: The character “Calm” was changed to integer “0” for data analysis purposes for the variable “dat$WindSpeed9am”, which had that issue or problem.

```
df[df$WindSpeed9am == "Calm", ]$WindSpeed9am <- 0
df$WindSpeed9am
```

```
##   [1] "4"  "9"  "7"  "13" "11" "17" "9"  "9"  "13" "11" "13" "20" "15" "9" 
"7" 
##  [16] "9"  "7"  "9"  "15" "20" "11" "7"  "11" "11" "6"  "2"  "13" "2"  "0" 
"11"
##  [31] "4"  "0"  "13" "20" "20" "31" "20" "13" "17" "6"  "6"  "0"  "13" "9" 
"0" 
##  [46] "6"  "4"  "0"  "7"  "9"  "7"  "9"  "13" "11" "9"  "13" "6"  "6"  "11
" "13"
##  [61] "15" "20" "11" "6"  "17" "28" "19" "44" "17" "13" "7"  "6"  "9"  "9" 
"4" 
##  [76] "7"  "9"  "17" "9"  "11" "7"  "6"  "2"  "11" "9"  "0"  "6"  "13" "11
" "17"
##  [91] "35" "22" "22" "24" "2"  "9"  "9"  "9"  "15" "2"  "7"  "4"  "6"  "7" 
"6" 
## [106] "9"  "2"  "0"  "0"  "17" "7"  "11" "9"  "7"  "7"  "9"  "9"  "2"  "20
" "26"
## [121] "4"  "6"  "2"  "2"  "13" "0"  "26" "4"  "7"  "7"  "9"  "9"  "0"  "0" 
"13"
## [136] "19" "19" "39" "7"  "11" "7"  "11" "9"  "0"  "9"  "7"  "2"  "0"  "0" 
"9" 
## [151] "35" "24" "7"  "0"  "26" "39" "35" "24" "20" "28" "24" "17" "19" "0" 
"2" 
## [166] "20" "17" "7"  "6"  "6"  "0"  "0"  "0"  "7"  "30" "17" "11" "15" "9" 
"7" 
## [181] "2"  "6"  "20" "24" "33" "26" "13" "2"  "20" "0"  "9"  "4"  "4"  "7" 
"24"
## [196] "6"  "2"  "31" "24" "19" "7"  "0"  "13" "4"  "11" "11" "24" "0"  "19
" "0" 
## [211] "6"  "0"  "26" "2"  "28" "28" "15" "20" "9"  "7"  "9"  "6"  "0"  "24
" "13"
## [226] "28" "31" "20" "2"  "20" "26" "20" "6"  "26" "30" "4"  "17" "9"  "4" 
"2" 
## [241] "2"
```

```
#Q.5 converting the column to numeric
#change WindSpeed data type from character to numeric

df$WindSpeed9am <- as.numeric(df$WindSpeed9am)
df$WindSpeed9am
```

```
##   [1]  4  9  7 13 11 17  9  9 13 11 13 20 15  9  7  9  7  9 15 20 11  7 11
11  6
##  [26]  2 13  2  0 11  4  0 13 20 20 31 20 13 17  6  6  0 13  9  0  6  4  0
7  9
##  [51]  7  9 13 11  9 13  6  6 11 13 15 20 11  6 17 28 19 44 17 13  7  6  9
9  4
```

```
##  [76]  7  9 17  9 11  7  6  2 11  9  0  6 13 11 17 35 22 22 24  2  9  9  9
15  2
## [101]  7  4  6  7  6  9  2  0  0 17  7 11  9  7  7  9  9  2 20 26  4  6  2
2 13
## [126]  0 26  4  7  7  9  9  0  0 13 19 19 39  7 11  7 11  9  0  9  7  2  0
0  9
## [151] 35 24  7  0 26 39 35 24 20 28 24 17 19  0  2 20 17  7  6  6  0  0  0
7 30
## [176] 17 11 15  9  7  2  6 20 24 33 26 13  2 20  0  9  4  4  7 24  6  2 31
24 19
## [201]  7  0 13  4 11 11 24  0 19  0  6  0 26  2 28 28 15 20  9  7  9  6  0
24 13
## [226] 28 31 20  2 20 26 20  6 26 30  4 17  9  4  2  2
```

**#Q6**

```
#Q6.Define a variable called “RainToday” and create an additional
variable called “RainTomorrow”
```

**#Comment:** Here we created variable called RainToday with the ifelse statement condition that assigns 1 when Rainfall is >= 1mm and 0 if Rainfall is < 1. The adddditional variable termed “RainTomorrow” was created by shifting RainToday one day forward or upward.

```
RainToday: 1 if Rainfall > 1 mm, otherwise 0
df$RainToday <- ifelse(df$Rainfall > 1, 1, 0)
#RainTomorrow by shifting RainToday one day forward
df$RainTomorrow <- c(df$RainToday[2:nrow(df)], NA)
#Deleting NA data columns
df <- df[, !(names(df) %in% c("Evaporation", "Sunshine", "WindDir9am"))]



numeric_columns <- c(3, 4,5,7,8,9,10,11,12,13,14,19,18,16,15)
df[, numeric_columns] <- lapply(df[, numeric_columns], function(x) as.numeric
(as.character(x)))
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
str(df)
```

```
## 'data.frame':    241 obs. of  20 variables:
##  $ Month        : chr  "JAN" "JAN" "JAN" "JAN" ...
##  $ Date         : chr  "2022-01-1" "2022-01-2" "2022-01-3" "2022-01-4" ...
##  $ MinTemp      : num  10.6 13.3 13.6 12.6 13.6 17.9 16.8 14.2 17.2 16.6 .
..
##  $ MaxTemp      : num  30.5 32.6 28.9 27.6 26.1 27.9 23.5 28.1 27 31.8 ...
##  $ Rainfall     : num  0 0 5 0.2 0 1.6 35.4 13.2 0 0 ...
##  $ WindGustDir  : chr  "NNW" "SE" "E" "ESE" ...
##  $ WindGustSpeed: num  31 59 46 35 43 50 35 43 37 28 ...
##  $ Temp9am      : num  19.1 22.8 21.9 18.4 20.5 21.4 19.7 17.7 19.8 22.8 .
```

```
..
##  $ Humidity9am  : num  68 64 57 88 66 78 95 99 74 76 ...
##  $ Cloud9am     : num  NA NA NA 8 8 7 4 8 8 7 ...
##  $ WindSpeed9am : num  4 9 7 13 11 17 9 9 13 11 ...
##  $ Pressure9am  : num  1014 1010 1011 1012 1013 ...
##  $ Temp3pm      : num  29.8 31.4 27.4 26.9 25.1 24.9 21.4 27.6 25.5 30.4 .
..
##  $ Humidity3pm  : num  35 25 48 48 61 65 76 45 60 33 ...
##  $ Cloud3pm     : num  NA 6 2 NA 8 5 8 1 2 7 ...
##  $ WindDir3pm   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WindSpeed3pm : chr  "13" "11" "24" "17" ...
##  $ Pressure3pm  : num  1008 1006 1008 1010 1010 ...
##  $ RainToday    : num  0 0 1 0 0 1 1 1 0 0 ...
##  $ RainTomorrow : num  0 1 0 0 1 1 1 0 0 0 ...
```

```r
dim(df)
```

```
## [1] 241  20
```

```r
#Our target variable is categorical, hence, we converted it to a factor.
df$RainToday <- as.factor(df$RainToday)
df$RainTomorrow <- as.factor(df$RainTomorrow)
```

## #Q7. Save cleaned data set

```r
write.csv(df, file="Cleaned_Weather-2022.csv", row.names =FALSE)
#Weather_csv<- read.csv("Weather.csv", header = TRUE, sep = ",")
```

# #Exploratory Data Analysis

```r
set.seed(1000)
tab  <- table(df$Month, df$Cloud9am, useNA="no");
tab1 <- table(df$Month, df$Cloud3pm, useNA = "no")
tab2 <- table(df$Month, df$WindGustSpeed, useNA = "no")
tab3 <- table(df$Month, df$Humidity3pm, useNA = "no")
tab4 <- table(df$Month, df$Humidity9am, useNA = "no")
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  tab
## p-value = 0.7546
## alternative hypothesis: two.sided
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  tab1
## p-value = 0.92
## alternative hypothesis: two.sided

## 
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
## 
## data:  tab2
## p-value = 0.003998
## alternative hypothesis: two.sided

## 
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
## 
## data:  tab3
## p-value = 0.2824
## alternative hypothesis: two.sided

## 
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
## 
## data:  tab4
## p-value = 0.06797
## alternative hypothesis: two.sided

## Warning in chisq.test(tab): Chi-squared approximation may be incorrect

## 
##  Pearson's Chi-squared test
## 
## data:  tab
## X-squared = 40.247, df = 49, p-value = 0.8091

## Warning in chisq.test(tab1): Chi-squared approximation may be incorrect

## 
##  Pearson's Chi-squared test
## 
## data:  tab1
## X-squared = 36.02, df = 49, p-value = 0.9162

## Warning in chisq.test(tab2): Chi-squared approximation may be incorrect

## 
##  Pearson's Chi-squared test
## 
## data:  tab2
## X-squared = 245.06, df = 196, p-value = 0.009902

## Warning in chisq.test(tab3): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab3
## X-squared = 473.63, df = 455, p-value = 0.2639

## Warning in chisq.test(tab4): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  tab4
## X-squared = 300.94, df = 294, p-value = 0.3778
```

### #Comments on the explanatory Data Analysis

```
#We see from the frequency tables above (tab ~ tab4), that there were some re
lationships between the selected variables and months of the year. To further
justify the extent of relationship, we test for the hypothesis using an alpha
level of 0.1 for fisher's test   and chi-square test to determine the associa
tion between categorical variable and binary outcome.

#For the Fisher’s test, we see that;
#tab, tab1 and tab3 all have p-values greater than the alpha level of 0.1. He
nce, we accept the null hypothesis and reject the alternate hypothesis. We co
nclude that:

#For tab: there is no association between month and cloud
#For tab 1: there is no association between month and cloud
#For tab 3: there is no association between month and humidity3pm

#We also observed for tab 2 and tab 4 that their p- values were less than the
alpha level 0f 0.1. Hence, we reject the null hypothesis and accept the alter
nate hypothesis. We can conclude for tab 2 and tab 4 that:

#For tab 2: there is an association between month and windgustspeed
#For tab 4: there is an association between month and humidity9am.

#Chi square test
#When it came to the chi square test, tab, tab1 tab3 and tab4 all have p-valu
es greater than the alpha level of 0.1. Hence, we accept the null hypothesis
and reject the alternate hypothesis. We conclude that: tab, tab1, tab3 and ta
b4 all do not have an association between month and the variables (cloud,clou
d, humidity3pm and humidity9am respectively).

#For tab2: We see that the p- values was less than the alpha level 0f 0.1. He
nce, we reject the null hypothesis and accept the alternate hypothesis. We ca
n conclude tab2 has an association between month and windgustspeed.

cor(tab)        #correlation coefficient gives us  matrix,
##            1          2           3           4          5           6
## 1  1.0000000  0.0000000  0.25819889  0.25819889 -0.2236068 -0.48038446
## 2  0.0000000  1.0000000 -0.18257419 -0.18257419  0.7905694  0.56613852
## 3  0.2581989 -0.1825742  1.00000000  0.06666667 -0.4618802 -0.53748385
## 4  0.2581989 -0.1825742  0.06666667  1.00000000  0.0000000 -0.12403473
## 5 -0.2236068  0.7905694 -0.46188022  0.00000000  1.0000000  0.42966892
## 6 -0.4803845  0.5661385 -0.53748385 -0.12403473  0.4296689  1.00000000
## 7 -0.1581139 -0.3354102 -0.57154761  0.24494897  0.0000000  0.05063697
## 8 -0.7337994  0.0000000  0.18946619  0.24359938  0.2344036  0.18464591
##             7           8
## 1 -0.15811388 -0.73379939
## 2 -0.33541020  0.00000000
## 3 -0.57154761  0.18946619
## 4  0.24494897  0.24359938
## 5  0.00000000  0.23440362
## 6  0.05063697  0.18464591
## 7  1.00000000  0.09944903
## 8  0.09944903  1.00000000
```

**#Comment**

For a continuous predictor variable and Binary outcome: #Here, we used the Wilcoxon rank sum test (two-sample t test) to check the association between continuous variable and binary outcome

```
#df$RainToday <- as.numeric(df$RainToday)
#df$MaxTemp <- as.numeric(df$MaxTemp)
# Convert RainTomorrow to numeric if it's not already
#df$RainTomorrow <- as.numeric(df$RainTomorrow)

# Remove rows with missing values in MaxTemp or RainTomorrow
#df <- df[!is.na(df$MaxTemp) & !is.na(df$RainTomorrow), ]

str(df)
```

```
## 'data.frame':    241 obs. of  20 variables:
##  $ Month        : chr  "JAN" "JAN" "JAN" "JAN" ...
##  $ Date         : chr  "2022-01-1" "2022-01-2" "2022-01-3" "2022-01-4" ...
##  $ MinTemp      : num  10.6 13.3 13.6 12.6 13.6 17.9 16.8 14.2 17.2 16.6 .
..
##  $ MaxTemp      : num  30.5 32.6 28.9 27.6 26.1 27.9 23.5 28.1 27 31.8 ...
##  $ Rainfall     : num  0 0 5 0.2 0 1.6 35.4 13.2 0 0 ...
##  $ WindGustDir  : chr  "NNW" "SE" "E" "ESE" ...
##  $ WindGustSpeed: num  31 59 46 35 43 50 35 43 37 28 ...
##  $ Temp9am      : num  19.1 22.8 21.9 18.4 20.5 21.4 19.7 17.7 19.8 22.8 .
..
##  $ Humidity9am  : num  68 64 57 88 66 78 95 99 74 76 ...
##  $ Cloud9am     : num  NA NA NA 8 8 7 4 8 8 7 ...
##  $ WindSpeed9am : num  4 9 7 13 11 17 9 9 13 11 ...
##  $ Pressure9am  : num  1014 1010 1011 1012 1013 ...
##  $ Temp3pm      : num  29.8 31.4 27.4 26.9 25.1 24.9 21.4 27.6 25.5 30.4 .
```

```
..
##  $ Humidity3pm  : num  35 25 48 48 61 65 76 45 60 33 ...
##  $ Cloud3pm     : num  NA 6 2 NA 8 5 8 1 2 7 ...
##  $ WindDir3pm   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WindSpeed3pm : chr  "13" "11" "24" "17" ...
##  $ Pressure3pm  : num  1008 1006 1008 1010 1010 ...
##  $ RainToday    : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 2 2 1 1 ...
##  $ RainTomorrow : Factor w/ 2 levels "0","1": 1 2 1 1 2 2 2 1 1 1 ...
```

```r
str(df)
```

```
## 'data.frame':    241 obs. of  20 variables:
##  $ Month        : chr  "JAN" "JAN" "JAN" "JAN" ...
##  $ Date         : chr  "2022-01-1" "2022-01-2" "2022-01-3" "2022-01-4" ...
##  $ MinTemp      : num  10.6 13.3 13.6 12.6 13.6 17.9 16.8 14.2 17.2 16.6 .
..
##  $ MaxTemp      : num  30.5 32.6 28.9 27.6 26.1 27.9 23.5 28.1 27 31.8 ...
##  $ Rainfall     : num  0 0 5 0.2 0 1.6 35.4 13.2 0 0 ...
##  $ WindGustDir  : chr  "NNW" "SE" "E" "ESE" ...
##  $ WindGustSpeed: num  31 59 46 35 43 50 35 43 37 28 ...
##  $ Temp9am      : num  19.1 22.8 21.9 18.4 20.5 21.4 19.7 17.7 19.8 22.8 .
..
##  $ Humidity9am  : num  68 64 57 88 66 78 95 99 74 76 ...
##  $ Cloud9am     : num  NA NA NA 8 8 7 4 8 8 7 ...
##  $ WindSpeed9am : num  4 9 7 13 11 17 9 9 13 11 ...
##  $ Pressure9am  : num  1014 1010 1011 1012 1013 ...
##  $ Temp3pm      : num  29.8 31.4 27.4 26.9 25.1 24.9 21.4 27.6 25.5 30.4 .
..
##  $ Humidity3pm  : num  35 25 48 48 61 65 76 45 60 33 ...
##  $ Cloud3pm     : num  NA 6 2 NA 8 5 8 1 2 7 ...
##  $ WindDir3pm   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WindSpeed3pm : chr  "13" "11" "24" "17" ...
##  $ Pressure3pm  : num  1008 1006 1008 1010 1010 ...
##  $ RainToday    : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 2 2 1 1 ...
##  $ RainTomorrow : Factor w/ 2 levels "0","1": 1 2 1 1 2 2 2 1 1 1 ...
```

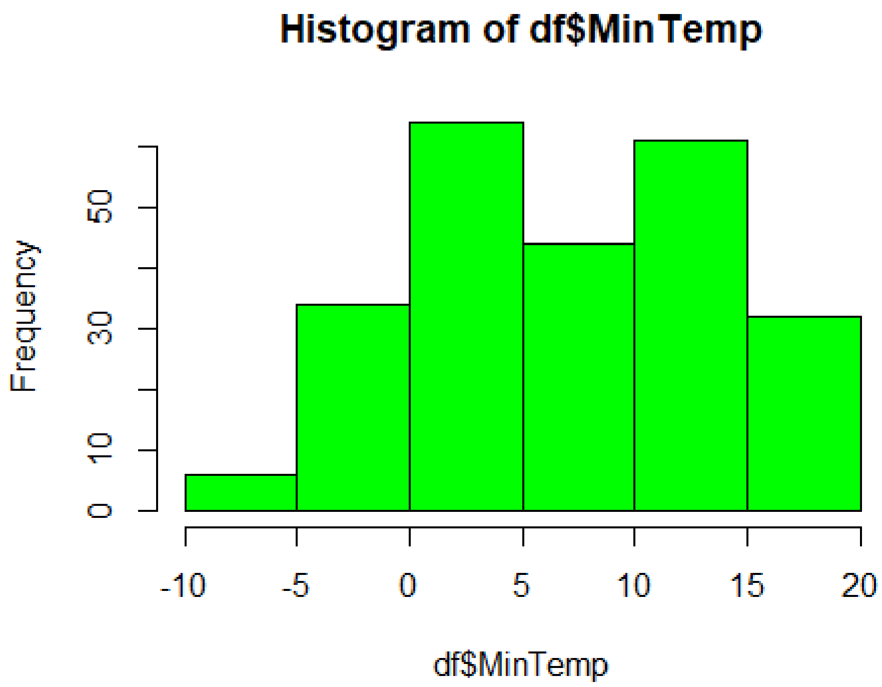
```r
wilcox.test(df$MaxTemp ~ df$RainTomorrow, alternative = "two.sided")
```

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  df$MaxTemp by df$RainTomorrow
## W = 4901, p-value = 0.6102
```

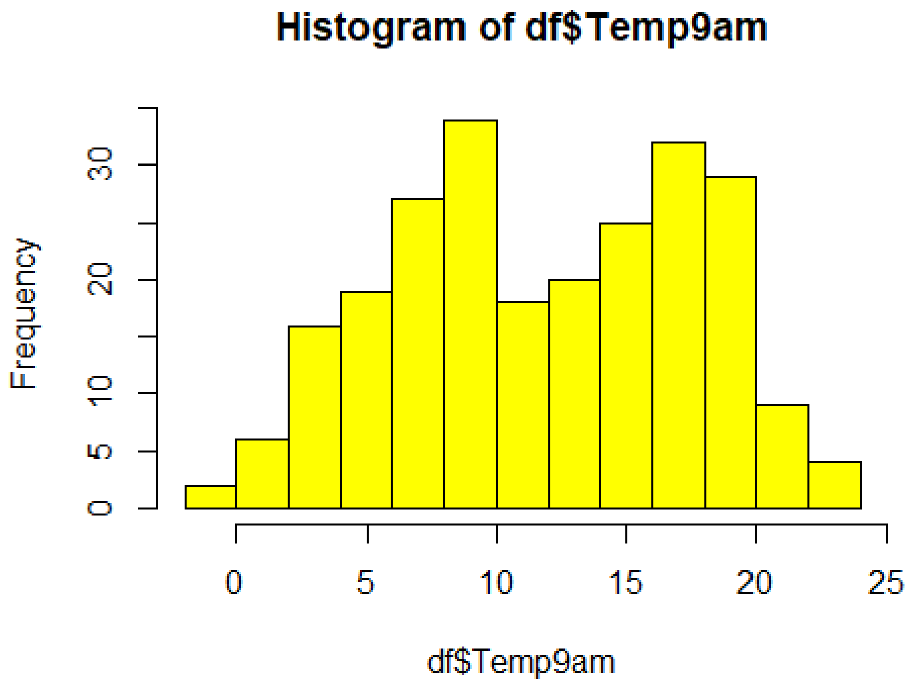
**#Comment:** *The p -value above greater than the alpha level of 0.1, hence accept the null hypothesis and reject the alternate hypothesis. There is no association between MinTemp and RainTommrow*

#Distribution Graphs Of Any Three Continuous Variables

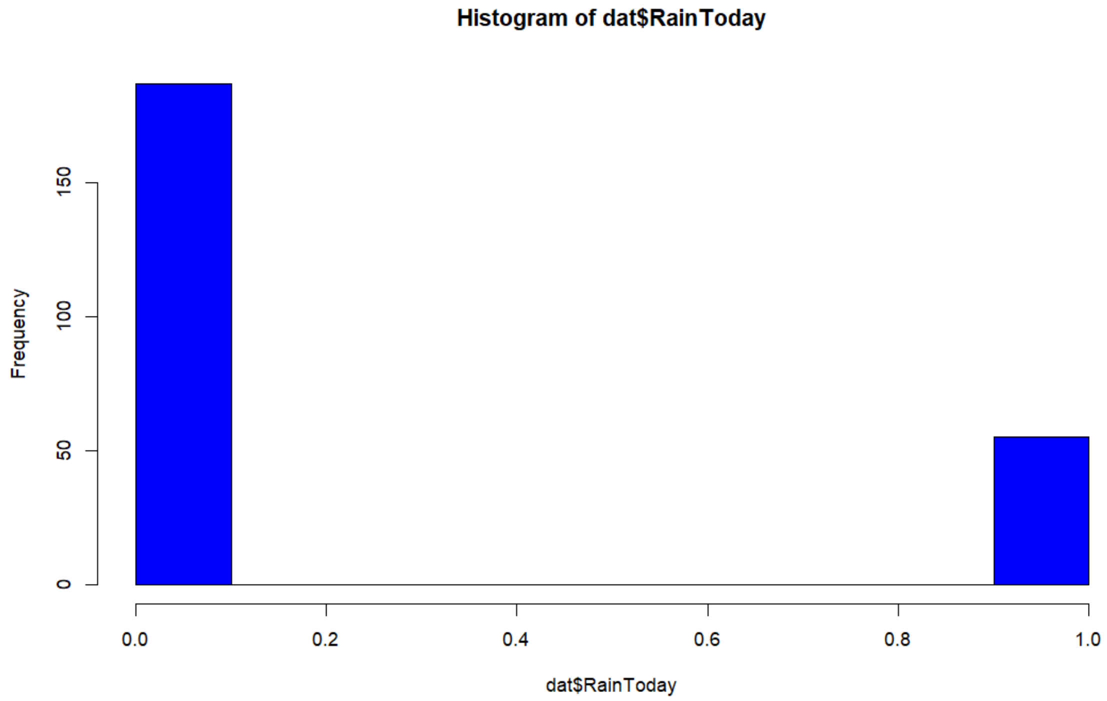
```
#Below are the graphs for the histograms of the MinTemp, Tem9am and Raintoday.

hist(df$MinTemp, col="green")
```



```
hist(df$Temp9am, col="yellow")
```

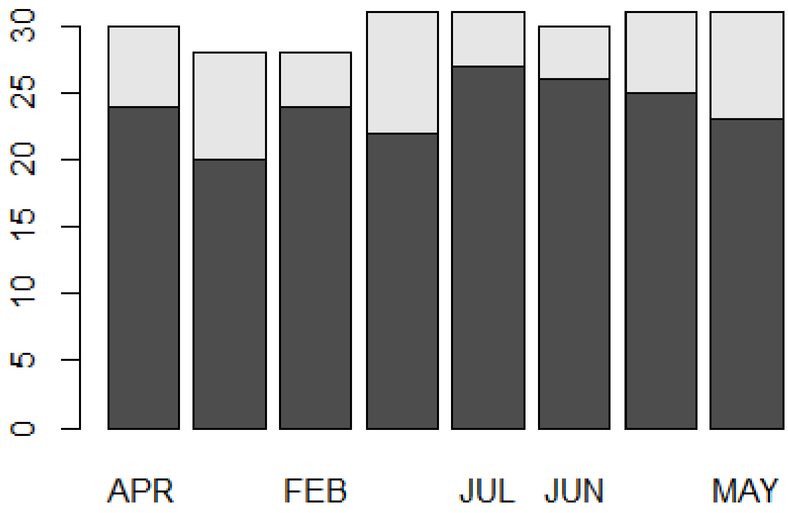


```
hist(df$Raintoday, col="blue")
```

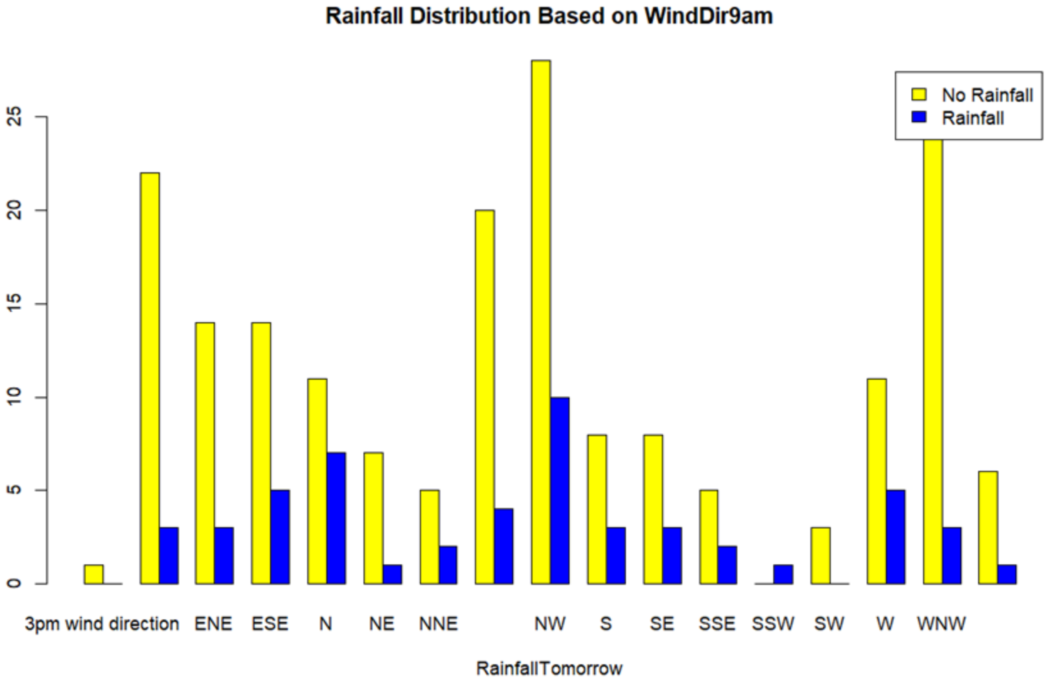


# Distribution Graphs Of Any Three Categorical Variables

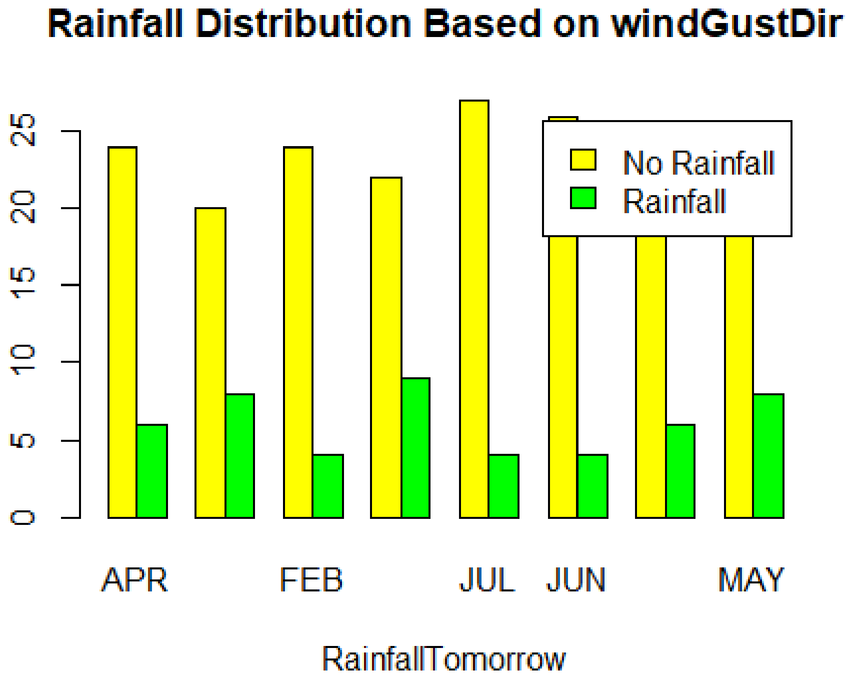
```
#Below are the bar graphs for Rainfall distribution for Windir9am
, WindGustDir, and Winddir3pm.
mytable<-table(df$RainTomorrow, df$Month) # to create table with proportions
mytable

## 
##     APR AUG FEB JAN JUL JUN MAR MAY
##   0  24  20  24  22  27  26  25  23
##   1   6   8   4   9   4   4   6   8

barplot((mytable))
```



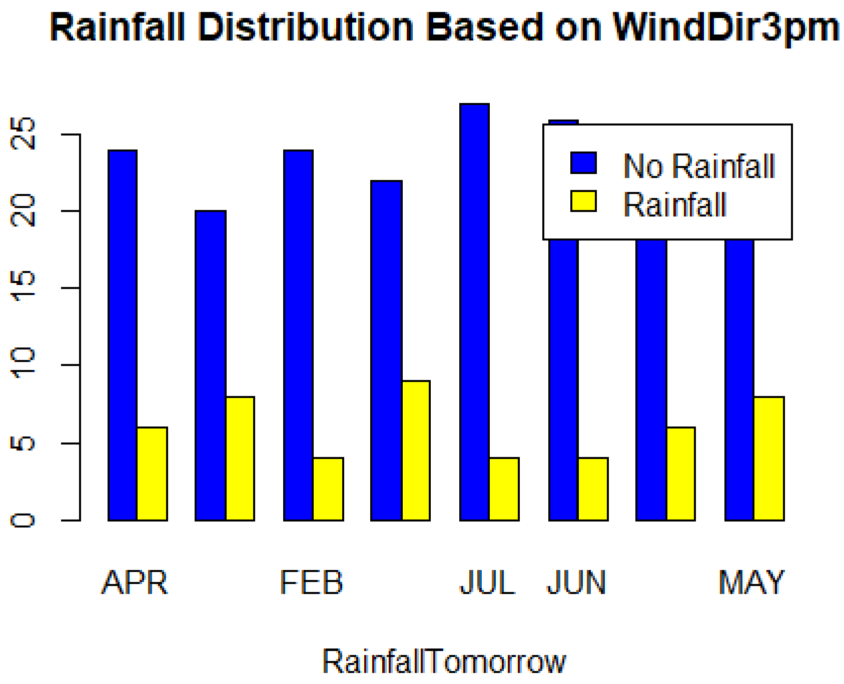
```r
#barplot(mytable*100)
barplot(mytable, main=" Rainfall Distribution Based on windDir9am",
        xlab="RainfallTomorrow", col=c("yellow","blue"),
        legend = rownames(mytable), beside = TRUE)
```



```
#barplot(mytable*100)
barplot(mytable, main="Rainfall Distribution Based on windGustDir",xlab="RainfallTomorrow",col=c("yellow","green"),legend = c('No Rainfall','Rainfall'), beside = TRUE)
```



```
#barplot(mytable*100)
barplot(mytable, main="Rainfall Distribution Based on WindDir3pm",xlab="RainfallTomorrow", col=c("blue","yellow"), legend = c('No Rainfall','Rainfall'), beside = TRUE)
```


Rainfall Distribution Based on WindDir3pm
0
5
10
15
20
25
No Rainfall
Rainfall
APR
FEB
JUL
JUN
MAY
RainfallTomorrow