PROJECT 1 REPORT

Itunu Olufemi Abdullah Akinde Idris Ajibade

September 27, 2021

Contents

1 Problem 1 Data Scraping

For this problem, we used the downloaded data set from D2L as the URL method returned an "HTTP status was 403 Forbidden" error.

```
#load in the data set
data1 <- read.csv("~/APSUGradSchool/_Fall 2021/MATH 5310/week 4/_Project 1/weather-2021.csv",
                  check.names = F)

dim(data1)
```

```
## [1] 255  22
```

The dimension of the data set is 255 rows by 22 columns.

```
head(data1)
```

```
##            Date Minimum temperature (°C) Maximum temperature (°C) Rainfall (mm)
## 1 JAN 2021-01-1                     14.4                     19.7           0.0
## 2 JAN 2021-01-2                     13.0                     22.3           0.0
## 3 JAN 2021-01-3                     15.7                     26.7           5.8
## 4 JAN 2021-01-4                     14.6                     24.8           7.2
## 5 JAN 2021-01-5                     14.1                     28.0           5.4
## 6 JAN 2021-01-6                     13.0                     24.7           0.2
##   Evaporation (mm) Sunshine (hours) Direction of maximum wind gust
## 1               NA               NA                            ESE
## 2               NA               NA                              E
## 3               NA               NA                            NNW
## 4               NA               NA                            WNW
## 5               NA               NA                             NW
## 6               NA               NA                            ESE
##   Speed of maximum wind gust (km/h) Time of maximum wind gust
## 1                                35                     12:39
## 2                                37                     16:02
## 3                                31                     13:28
## 4                                56                     13:28
## 5                                39                     13:16
## 6                                46                     13:23
##   9am Temperature (°C) 9am relative humidity (%) 9am cloud amount (oktas)
## 1                 15.9                         71                       8
## 2                 16.9                         59                       8
## 3                 17.9                         94                       8
## 4                 21.6                         74                       7
## 5                 19.2                         77                       4
## 6                 18.2                         68                       8
##   9am wind direction 9am wind speed (km/h) 9am MSL pressure (hPa)
## 1                 SE                    17                1021.7
## 2                  E                     7                1015.7
## 3              <NA>                  Calm                1008.1
## 4                 W                     6                1006.3
## 5                 W                     9                1008.0
## 6                SE                    11                1014.0
##   3pm Temperature (°C) 3pm relative humidity (%) 3pm cloud amount (oktas)
## 1                 18.5                         59                       8
## 2                 20.6                         60                       8
## 3                 24.1                         60                       8
## 4                 16.3                         97                       8
## 5                 26.6                         41                       2
## 6                 19.1                         76                       7
##   3pm wind direction 3pm wind speed (km/h) 3pm MSL pressure (hPa)
## 1                ESE                    17                1019.1
## 2                  E                    17                1011.7
```

```
## 3                  NW                      17                  1005.8
## 4                 SSW                      13                  1007.6
## 5                  NW                      20                  1006.2
## 6                 ESE                      33                  1013.4
```

The above shows the first 6 rows of the data set.

```
tail(data1)
```

```
##               Date Minimum temperature (°C) Maximum temperature (°C)
## 250 SEP  2021-09-7                     -3.0                      18.8
## 251 SEP  2021-09-8                      0.4                      16.6
## 252 SEP  2021-09-9                     -0.4                      19.9
## 253 SEP 2021-09-10                      9.1                      21.1
## 254 SEP 2021-09-11                      2.6                      22.5
## 255 SEP 2021-09-12                     10.2                        NA
##     Rainfall (mm) Evaporation (mm) Sunshine (hours)
## 250             0               NA               NA
## 251             0               NA               NA
## 252             0               NA               NA
## 253             0               NA               NA
## 254             0               NA               NA
## 255             0               NA               NA
##     Direction of maximum wind gust  Speed of maximum wind gust (km/h)
## 250                            WNW                                48
## 251                            NNW                                37
## 252                            WNW                                61
## 253                           <NA>                                NA
## 254                             NW                                56
## 255                           <NA>                                NA
##     Time of maximum wind gust 9am Temperature (°C) 9am relative humidity (%)
## 250                       14:15                10.0                        63
## 251                       16:35                 4.8                       100
## 252                       17:48                11.1                        76
## 253                        <NA>                15.5                        50
## 254                       10:01                15.5                        48
## 255                        <NA>                15.2                        51
##     9am cloud amount (oktas) 9am wind direction 9am wind speed (km/h)
## 250                        1                  N                    30
## 251                        8               <NA>                  Calm
## 252                       NA                NNW                    22
## 253                       NA                  N                    20
## 254                       NA                 NW                    28
## 255                       NA                 NW                    22
##     9am MSL pressure (hPa) 3pm Temperature (°C) 3pm relative humidity (%)
## 250                 1026.9                17.5                        40
## 251                 1032.8                16.1                        53
```

```
## 252                 1028.5                 18.9                         37
## 253                 1022.5                 20.3                         39
## 254                 1017.6                 21.8                         27
## 255                 1006.1                 13.4                         57
##     3pm cloud amount (oktas) 3pm wind direction 3pm wind speed (km/h)
## 250                       NA                WNW                     30
## 251                       NA                NNW                     20
## 252                       NA                WNW                     28
## 253                        1                 NW                     33
## 254                       NA                 NW                     37
## 255                        8                 NW                     30
##     3pm MSL pressure (hPa)
## 250                 1024.6
## 251                 1028.1
## 252                 1022.9
## 253                 1020.6
## 254                 1011.9
## 255                 1006.4
```

The above represents the last 6 rows of the data set.

2 Problem 2 Data Cleaning and Preparation

2.1 Problem 2A remove the "Time of maximum wind gust" in the 10th column

```
#remove a second columns of all the rows  [r x c]
data1 <- data1[, -c(10)]

dim(data1)
```

```
## [1] 255  21
```

We have successfully removed the 10th column, our new dimension is now 255 by 21.

2.2 Problem 2B rename the dataset

```
names(data1) <- c("Month", "Date", "MinTemp", "MaxTemp", "Rainfall",
                  "Evaporation", "Sunshine", "WindGustDir", "WindGustSpeed",
                  "Temp9am", "Humidity9am", "Cloud9am", "WindDir9am",
                  "WindSpeed9am", "Pressure9am", "Temp3pm", "Humidity3pm",
                  "Cloud3pm", "WindDir3pm", "WindSpeed3pm", "Pressure3pm")
```

2.3 Problem 2C print out unique values

```
data <- data1
vnames <- colnames(data)
n <- nrow(data)
out <- NULL
for (j in 1:ncol(data)){
  vname <- colnames(data)[j]
  x <- as.vector(data[,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(unique(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  255     0                   0
## 2        2          Date character     255  255     0                   0
## 3        3       MinTemp   numeric     152  255     0                   0
## 4        4       MaxTemp   numeric     156  254     1 0.00392156862745098
## 5        5      Rainfall   numeric      46  255     0                   0
## 6        6   Evaporation   logical       1    0   255                   1
## 7        7      Sunshine   logical       1    0   255                   1
## 8        8   WindGustDir character      17  253     2 0.00784313725490196
## 9        9 WindGustSpeed   numeric      33  253     2 0.00784313725490196
## 10      10       Temp9am   numeric     146  255     0                   0
## 11      11   Humidity9am   numeric      56  255     0                   0
## 12      12      Cloud9am   numeric       9  147   108   0.423529411764706
## 13      13    WindDir9am character      17  218    37   0.145098039215686
## 14      14  WindSpeed9am character      23  255     0                   0
## 15      15   Pressure9am   numeric     173  255     0                   0
## 16      16       Temp3pm   numeric     152  255     0                   0
## 17      17   Humidity3pm   numeric      74  255     0                   0
## 18      18      Cloud3pm   numeric       9  156    99   0.388235294117647
## 19      19    WindDir3pm character      17  252     3  0.0117647058823529
## 20      20  WindSpeed3pm character      19  207    48   0.188235294117647
## 21      21   Pressure3pm   numeric     174  255     0                   0
```

’ There are 21 variables with their respective datatypes for 255 values. This includes:

1. Month : which consists of character datatype and have 9 unique values.
2. Date : which consists of character datatype and have 255 unique values. This is because each date is unique
3. MinTemp : which consists of numeric datatype and have 152 unique values.
4. MaxTemp : which consists of numeric datatype and have 156 unique values.
5. Rainfall : which consists of numeric datatype and have 46 unique values.
6. Evaporation : which consists of logical datatype and have 1 unique values.
7. Sunshine : which consists of logical datatype and have 1 unique values.
8. WindGustDir : which consists of character datatype and have 17 unique values.
9. WindGustSpeed: which consists of integer datatype and have 33 unique values
10. Temp9am : which consists of numeric datatype and have 146 unique values
11. Humidity9am : which consists of integer datatype and have 56 unique values
12. Cloud9am : which consists of integer datatype and have 9 unique values
13. WindDir9am : which consists of character datatype and have 17 unique values.
14. WindSpeed9am : which consists of character datatype and have 23 unique values.This should be integer
15. Pressure9am : which consists of numeric datatype and have 173 unique values
16. Temp3pm : which consists of numeric datatype and have 152 unique values
17. Humidity3pm : which consists of integer datatype and have 74 unique values
18. Cloud3pm : which consists of integer datatype and have 9 unique values
19. WindDir3pm : which consists of character datatype and have 17 unique values
20. WindSpeed3pm : which consists of character datatype and have 19 unique values. This should be integer
21. Pressure3pm : which consists of numeric datatype and have 174 unique values

2.4 Problem 2D Show the frequency table of each value, as if categorical

```
#use apply column by column
data1 <- data
apply(data1, 2, FUN = function(X){table(X, useNA="ifany")})
```

```
## $Month
## X
## APR AUG FEB JAN JUL JUN MAR MAY SEP
##  30  31  28  31  31  30  31  31  12
##
## $Date
## X
##  2021-01-1 2021-01-10 2021-01-11 2021-01-12 2021-01-13 2021-01-14 2021-01-15
##          1          1          1          1          1          1          1
## 2021-01-16 2021-01-17 2021-01-18 2021-01-19  2021-01-2 2021-01-20 2021-01-21
##          1          1          1          1          1          1          1
## 2021-01-22 2021-01-23 2021-01-24 2021-01-25 2021-01-26 2021-01-27 2021-01-28
##          1          1          1          1          1          1          1
## 2021-01-29  2021-01-3 2021-01-30 2021-01-31  2021-01-4  2021-01-5  2021-01-6
##          1          1          1          1          1          1          1
```

```
##  2021-01-7  2021-01-8  2021-01-9  2021-02-1 2021-02-10 2021-02-11 2021-02-12
##          1          1          1          1          1          1          1
## 2021-02-13 2021-02-14 2021-02-15 2021-02-16 2021-02-17 2021-02-18 2021-02-19
##          1          1          1          1          1          1          1
##  2021-02-2 2021-02-20 2021-02-21 2021-02-22 2021-02-23 2021-02-24 2021-02-25
##          1          1          1          1          1          1          1
## 2021-02-26 2021-02-27 2021-02-28  2021-02-3  2021-02-4  2021-02-5  2021-02-6
##          1          1          1          1          1          1          1
##  2021-02-7  2021-02-8  2021-02-9  2021-03-1 2021-03-10 2021-03-11 2021-03-12
##          1          1          1          1          1          1          1
## 2021-03-13 2021-03-14 2021-03-15 2021-03-16 2021-03-17 2021-03-18 2021-03-19
##          1          1          1          1          1          1          1
##  2021-03-2 2021-03-20 2021-03-21 2021-03-22 2021-03-23 2021-03-24 2021-03-25
##          1          1          1          1          1          1          1
## 2021-03-26 2021-03-27 2021-03-28 2021-03-29  2021-03-3 2021-03-30 2021-03-31
##          1          1          1          1          1          1          1
##  2021-03-4  2021-03-5  2021-03-6  2021-03-7  2021-03-8  2021-03-9  2021-04-1
##          1          1          1          1          1          1          1
## 2021-04-10 2021-04-11 2021-04-12 2021-04-13 2021-04-14 2021-04-15 2021-04-16
##          1          1          1          1          1          1          1
## 2021-04-17 2021-04-18 2021-04-19  2021-04-2 2021-04-20 2021-04-21 2021-04-22
##          1          1          1          1          1          1          1
## 2021-04-23 2021-04-24 2021-04-25 2021-04-26 2021-04-27 2021-04-28 2021-04-29
##          1          1          1          1          1          1          1
##  2021-04-3 2021-04-30  2021-04-4  2021-04-5  2021-04-6  2021-04-7  2021-04-8
##          1          1          1          1          1          1          1
##  2021-04-9  2021-05-1 2021-05-10 2021-05-11 2021-05-12 2021-05-13 2021-05-14
##          1          1          1          1          1          1          1
## 2021-05-15 2021-05-16 2021-05-17 2021-05-18 2021-05-19  2021-05-2 2021-05-20
##          1          1          1          1          1          1          1
## 2021-05-21 2021-05-22 2021-05-23 2021-05-24 2021-05-25 2021-05-26 2021-05-27
##          1          1          1          1          1          1          1
## 2021-05-28 2021-05-29  2021-05-3 2021-05-30 2021-05-31  2021-05-4  2021-05-5
##          1          1          1          1          1          1          1
##  2021-05-6  2021-05-7  2021-05-8  2021-05-9  2021-06-1 2021-06-10 2021-06-11
##          1          1          1          1          1          1          1
## 2021-06-12 2021-06-13 2021-06-14 2021-06-15 2021-06-16 2021-06-17 2021-06-18
##          1          1          1          1          1          1          1
## 2021-06-19  2021-06-2 2021-06-20 2021-06-21 2021-06-22 2021-06-23 2021-06-24
##          1          1          1          1          1          1          1
## 2021-06-25 2021-06-26 2021-06-27 2021-06-28 2021-06-29  2021-06-3 2021-06-30
##          1          1          1          1          1          1          1
##  2021-06-4  2021-06-5  2021-06-6  2021-06-7  2021-06-8  2021-06-9  2021-07-1
##          1          1          1          1          1          1          1
## 2021-07-10 2021-07-11 2021-07-12 2021-07-13 2021-07-14 2021-07-15 2021-07-16
##          1          1          1          1          1          1          1
## 2021-07-17 2021-07-18 2021-07-19  2021-07-2 2021-07-20 2021-07-21 2021-07-22
##          1          1          1          1          1          1          1
```

```
## 2021-07-23 2021-07-24 2021-07-25 2021-07-26 2021-07-27 2021-07-28 2021-07-29
##          1          1          1          1          1          1          1
##  2021-07-3 2021-07-30 2021-07-31  2021-07-4  2021-07-5  2021-07-6  2021-07-7
##          1          1          1          1          1          1          1
##  2021-07-8  2021-07-9  2021-08-1 2021-08-10 2021-08-11 2021-08-12 2021-08-13
##          1          1          1          1          1          1          1
## 2021-08-14 2021-08-15 2021-08-16 2021-08-17 2021-08-18 2021-08-19  2021-08-2
##          1          1          1          1          1          1          1
## 2021-08-20 2021-08-21 2021-08-22 2021-08-23 2021-08-24 2021-08-25 2021-08-26
##          1          1          1          1          1          1          1
## 2021-08-27 2021-08-28 2021-08-29  2021-08-3 2021-08-30 2021-08-31  2021-08-4
##          1          1          1          1          1          1          1
##  2021-08-5  2021-08-6  2021-08-7  2021-08-8  2021-08-9  2021-09-1 2021-09-10
##          1          1          1          1          1          1          1
## 2021-09-11 2021-09-12  2021-09-2  2021-09-3  2021-09-4  2021-09-5  2021-09-6
##          1          1          1          1          1          1          1
##  2021-09-7  2021-09-8  2021-09-9
##          1          1          1
##
## $MinTemp
## X
## -0.1 -0.3 -0.4 -0.5 -0.6 -0.7 -0.8 -1.0 -1.1 -1.2 -1.6 -1.7 -1.8 -1.9 -2.0 -2.1
##    1    3    2    1    1    1    1    2    1    3    2    4    1    2    1    1
## -2.2 -2.5 -2.6 -2.7 -2.8 -2.9 -3.0 -3.1 -3.3 -3.4 -3.5 -3.6 -3.7 -3.8 -4.1 -4.7
##    1    1    1    1    1    1    2    1    1    3    1    1    1    1    1    1
## -4.8 -4.9 -5.0 -5.1 -5.2 -5.4 -6.0 -6.3  0.0  0.2  0.3  0.4  0.5  0.6  0.7  0.9
##    1    1    1    1    1    1    1    1    3    1    1    1    3    3    1    2
##  1.0  1.1  1.4  1.8  1.9  2.1  2.3  2.4  2.5  2.6  2.8  3.0  3.2  3.3  3.4  3.5
##    1    1    2    2    4    2    2    4    2    2    3    1    3    1    4    3
##  3.6  3.7  3.8  3.9  4.0  4.1  4.2  4.3  4.4  4.7  4.8  4.9  5.0  5.4  5.5  5.6
##    3    1    3    2    2    1    1    1    3    1    4    2    2    4    1    1
##  5.7  5.8  5.9  6.0  6.1  6.2  6.4  6.6  6.7  7.0  7.1  7.2  7.3  7.4  7.5  7.6
##    2    5    2    1    4    1    1    3    2    1    1    2    1    2    1    1
##  8.0  8.1  8.2  8.3  8.4  8.7  9.0  9.1  9.2  9.7  9.8  9.9 10.1 10.2 10.3 10.4
##    2    1    2    2    2    1    1    1    1    1    1    3    2    2    1    1
## 10.5 10.7 10.9 11.2 11.4 11.5 11.6 12.0 12.1 12.6 12.8 12.9 13.0 13.2 13.3 13.4
##    3    2    2    1    1    2    1    2    3    1    2    1    3    2    1    1
## 13.6 13.7 13.9 14.0 14.1 14.2 14.3 14.4 14.6 14.8 14.9 15.0 15.5 15.6 15.7 15.9
##    2    2    1    1    4    1    1    2    2    1    2    2    1    2    2    1
## 16.3 16.4 16.7 17.1 17.5 18.1 18.5 18.8
##    1    1    1    1    1    1    1    1
##
## $MaxTemp
## X
##  6.3  8.2  8.3  8.8  8.9  9.3  9.4  9.5  9.6  9.7 10.0 10.3 10.4 10.5 10.7 10.8
##    1    1    1    1    1    2    1    1    1    1    1    2    2    1    1    1
## 11.2 11.3 11.4 11.5 11.7 11.8 11.9 12.0 12.1 12.3 12.4 12.5 12.6 12.7 12.8 12.9
##    1    3    1    1    1    1    1    2    2    3    3    4    3    1    3    2
```

```
## 13.0 13.2 13.3 13.5 13.6 13.7 13.8 13.9 14.1 14.2 14.3 14.4 14.5 14.6 14.7 14.8
##    3    3    3    1    3    1    2    1    1    3    2    1    7    3    1    1
## 14.9 15.0 15.2 15.3 15.4 15.7 15.8 15.9 16.1 16.3 16.6 16.7 16.9 17.0 17.1 17.4
##    3    1    1    1    1    2    6    3    1    2    1    1    2    1    1    3
## 17.5 17.6 17.9 18.0 18.2 18.3 18.5 18.6 18.7 18.8 18.9 19.1 19.3 19.4 19.7 19.8
##    1    2    2    1    1    2    2    3    1    3    1    1    1    3    2    4
## 19.9 20.1 20.4 20.5 20.6 20.7 20.8 21.0 21.1 21.3 21.4 21.5 21.6 21.7 21.9 22.1
##    2    1    1    1    2    4    1    1    1    1    2    2    2    1    1    1
## 22.2 22.3 22.4 22.5 22.6 22.7 23.3 23.4 23.5 23.6 24.1 24.2 24.3 24.4 24.5 24.7
##    1    2    2    1    2    2    1    1    1    2    1    2    2    2    1    2
## 24.8 24.9 25.0 25.1 25.2 25.3 25.6 25.7 26.0 26.2 26.3 26.5 26.7 26.9 27.0 27.1
##    2    2    2    2    1    1    1    1    2    1    1    1    1    1    3    1
## 27.2 27.5 27.7 27.9 28.0 28.1 28.2 28.3 29.2 29.3 29.4 29.5 30.1 30.3 30.5 30.8
##    1    2    3    1    2    2    1    2    1    1    1    1    3    1    1    1
## 30.9 31.8 32.5 32.7 32.9 33.7 34.1 34.5 35.9 37.5 38.0 <NA>
##    1    1    1    1    1    1    1    1    1    1    1    1
##
## $Rainfall
## X
##  0.0  0.2  0.4  0.6  0.8  1.0  1.2  1.4  1.8  2.0  2.2  2.4  2.6  2.8  3.0  3.2
##  162   28    2    3    2    2    1    1    2    1    1    2    1    3    1    1
##  3.6  4.0  4.2  4.8  5.0  5.4  5.8  7.2  7.6  7.8  8.0  8.2  9.0  9.4  9.8 10.2
##    1    1    1    2    5    3    2    1    1    1    1    1    1    1    1    1
## 10.4 10.8 11.0 11.2 12.4 19.2 22.2 22.8 23.8 25.4 28.4 29.4 30.6 39.4
##    1    2    1    3    1    1    1    1    1    1    1    1    2    1
##
## $Evaporation
## X
## <NA>
##  255
##
## $Sunshine
## X
## <NA>
##  255
##
## $WindGustDir
## X
##    E  ENE  ESE    N   NE  NNE  NNW   NW    S   SE  SSE  SSW   SW    W  WNW  WSW
##   26   13   11   21    5    3   37   63    9   12    4    7    2    4   34    2
## <NA>
##    2
##
## $WindGustSpeed
## X
##   13   15   17   19   20   22   24   26   28   30   31   33   35   37   39   41
##    2    3    6    9    3    3    7    7   11   12   14   12   14   28   17   11
##   43   44   46   48   50   52   54   56   57   59   61   63   67   69   70   72
```

```
##   12   14   10   11    7    7    8    6    2    4    3    1    1    1    4    3
## <NA>
##    2
##
## $Temp9am
## X
## -0.8  0.0  0.3  0.4  0.6  0.8  0.9  1.0  1.2  1.3  1.4  1.8  2.0  2.2  2.3  2.7
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##  2.8  3.0  3.1  3.4  3.7  3.8  4.0  4.1  4.3  4.4  4.5  4.6  4.8  5.2  5.4  5.5
##    1    2    1    2    1    1    5    1    1    1    2    1    1    2    4    2
##  5.6  5.7  5.8  5.9  6.0  6.2  6.3  6.5  6.6  6.7  6.8  6.9  7.0  7.2  7.3  7.6
##    2    1    2    1    2    3    2    1    2    3    2    2    2    1    2    1
##  7.8  8.0  8.1  8.2  8.3  8.4  8.5  8.6  8.7  8.8  8.9  9.1  9.2  9.3  9.4  9.5
##    3    2    2    3    3    2    3    2    7    3    1    1    5    1    1    1
##  9.6  9.8  9.9 10.0 10.2 10.3 10.4 10.5 10.6 10.7 10.8 11.0 11.1 11.2 11.3 11.4
##    2    3    3    2    3    2    1    2    2    1    1    1    2    1    1    2
## 11.5 11.8 12.0 12.2 12.3 12.4 12.7 12.9 13.0 13.2 13.3 13.4 13.8 14.2 14.4 14.5
##    1    1    2    2    4    2    1    1    2    1    1    2    3    2    2    1
## 14.7 15.0 15.1 15.2 15.3 15.4 15.5 15.6 15.8 15.9 16.0 16.3 16.4 16.5 16.6 16.7
##    2    1    3    2    2    4    3    2    1    2    1    1    3    1    4    4
## 16.8 16.9 17.0 17.1 17.2 17.3 17.5 17.8 17.9 18.0 18.1 18.2 18.3 18.4 18.8 19.0
##    2    3    1    1    1    1    2    2    1    1    1    3    2    2    1    1
## 19.1 19.2 19.5 19.7 19.9 20.3 20.5 20.8 21.1 21.2 21.5 21.6 21.7 22.1 22.4 25.0
##    3    3    1    1    3    1    1    1    1    1    1    1    1    1    1    1
## 25.4 26.3
##    1    1
##
## $Humidity9am
## X
##  34  37  42  45  48  49  50  51  52  53  55  56  57  58  59  60  61  62  63  64
##   1   2   1   1   1   2   5   1   1   1   1   3   3   1   4   3   7   2   9   5
##  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84
##   6   3   2   3   4   5   1   1   7   6   5   5   7   8   6   2   4   8   3   8
##  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
##   4   8   6   7   5   3   5   4   1   7   6   2   2   1  26  20
##
## $Cloud9am
## X
##   1   2   3   4   5   6   7   8 <NA>
##   9   9   4   4   6   7  18  90  108
##
## $WindDir9am
## X
##    E  ENE  ESE    N   NE  NNE  NNW   NW    S   SE  SSE  SSW   SW    W  WNW  WSW
##   13    8   14   28    2    1   29   22   19   21   25   15    7    5    8    1
## <NA>
##   37
##
```

```
## $WindSpeed9am
## X
##   11   13   15   17   19    2   20   22   24   26   28   30   31   33   35   37
##   13   16   12   12    7   16    9    8    5    3    5    8    2    1    1    1
##   39    4   43    6    7    9 Calm
##    2   19    2   21   33   22   37
##
## $Pressure9am
## X
##  996.2  996.9  998.7 1001.8 1002.3 1003.2 1003.4 1005.0 1005.5 1005.7 1006.1
##      1      1      1      1      1      1      1      1      1      1      2
## 1006.3 1006.5 1006.6 1006.8 1007.0 1007.2 1008.0 1008.1 1008.2 1008.6 1009.1
##      2      1      1      2      1      2      1      1      1      1      1
## 1009.3 1009.4 1009.7 1009.8 1009.9 1010.1 1010.2 1010.3 1010.4 1010.8 1010.9
##      1      1      1      1      2      2      1      2      1      2      2
## 1011.1 1011.3 1011.5 1011.6 1011.7 1012.0 1012.2 1012.3 1012.4 1012.6 1012.7
##      1      1      1      1      1      2      1      1      2      1      1
## 1012.8 1012.9 1013.0 1013.1 1013.2 1013.3 1013.5 1013.6 1013.8 1013.9 1014.0
##      2      1      3      1      1      1      1      3      4      1      1
## 1014.1 1014.4 1014.5 1014.7 1014.9 1015.2 1015.4 1015.5 1015.6 1015.7 1015.9
##      1      1      3      1      4      1      2      2      2      2      1
## 1016.0 1016.1 1016.3 1016.4 1016.6 1016.8 1017.0 1017.1 1017.2 1017.3 1017.5
##      3      3      1      2      1      3      1      1      1      1      3
## 1017.6 1017.9 1018.1 1018.2 1018.3 1018.6 1018.7 1018.8 1018.9 1019.0 1019.1
##      2      1      1      1      2      3      1      1      1      1      1
## 1019.2 1019.3 1019.4 1019.5 1019.6 1019.7 1019.8 1020.0 1020.1 1020.2 1020.3
##      1      1      3      1      4      1      1      2      2      1      1
## 1020.6 1020.8 1020.9 1021.3 1021.5 1021.6 1021.7 1021.8 1022.0 1022.1 1022.2
##      1      1      1      1      2      1      1      1      2      2      2
## 1022.3 1022.4 1022.5 1022.6 1022.7 1023.0 1023.2 1023.3 1023.4 1023.5 1023.7
##      2      1      3      5      1      2      1      1      1      1      4
## 1023.8 1024.1 1024.3 1024.5 1024.6 1024.9 1025.0 1025.1 1025.3 1025.4 1025.5
##      1      2      2      2      1      1      1      1      2      3      1
## 1025.6 1025.8 1025.9 1026.0 1026.1 1026.2 1026.3 1026.4 1026.5 1026.7 1026.8
##      2      2      2      1      1      1      1      2      1      2      1
## 1026.9 1027.0 1027.1 1027.2 1027.3 1027.6 1028.1 1028.4 1028.5 1029.3 1029.4
##      1      2      1      3      1      1      2      1      1      1      1
## 1029.5 1029.6 1029.7 1030.0 1030.1 1030.2 1030.7 1030.8 1030.9 1031.0 1031.5
##      3      1      1      1      1      3      1      1      1      1      1
## 1032.2 1032.3 1032.6 1032.8 1033.4 1033.8 1034.3 1036.6
##      1      1      1      1      1      1      1      1
##
## $Temp3pm
## X
##  5.8  6.0  6.9  7.0  7.4  7.5  7.7  7.9  8.0  8.5  8.7  9.0  9.2  9.3  9.5  9.6
##    2    1    1    1    1    1    1    1    3    1    2    1    1    3    1    2
##  9.9 10.1 10.3 10.4 10.5 10.7 10.8 10.9 11.0 11.1 11.2 11.3 11.4 11.5 11.6 11.7
##    1    1    1    1    1    2    1    2    1    2    1    3    2    2    1    3
```

```
## 11.9 12.0 12.1 12.2 12.3 12.4 12.5 12.6 12.7 12.8 13.0 13.1 13.2 13.3 13.4 13.6
##    2    6    1    2    3    1    3    2    2    2    3    2    1    1    6    3
## 13.8 14.0 14.2 14.3 14.4 14.5 14.6 14.7 14.8 14.9 15.0 15.1 15.3 15.4 15.7 16.0
##    3    3    1    2    1    3    2    1    2    1    2    1    2    2    1    2
## 16.1 16.3 16.6 16.8 17.0 17.1 17.2 17.3 17.5 17.6 17.7 17.8 17.9 18.0 18.1 18.2
##    4    1    1    1    2    1    2    1    1    1    1    1    1    5    2    1
## 18.3 18.4 18.5 18.6 18.7 18.9 19.0 19.1 19.2 19.4 19.5 19.8 20.0 20.1 20.3 20.6
##    1    2    3    1    1    2    2    3    1    1    2    2    1    1    3    1
## 20.7 20.8 21.0 21.1 21.3 21.4 21.5 21.6 21.8 21.9 22.3 22.5 22.6 22.7 22.8 22.9
##    4    2    2    1    1    1    3    2    1    2    2    1    1    4    2    2
## 23.1 23.2 23.4 24.0 24.1 24.4 24.6 24.7 25.0 25.1 25.2 25.4 25.6 25.9 26.0 26.1
##    2    1    2    1    2    2    1    1    1    2    1    1    1    1    1    2
## 26.3 26.4 26.5 26.6 26.7 26.8 27.0 27.2 27.4 27.6 27.8 28.1 28.3 28.4 29.0 29.4
##    1    1    1    4    1    1    1    1    1    1    1    2    2    2    2    1
## 29.7 31.3 31.5 32.0 32.8 34.6 36.1 36.2
##    1    2    1    1    1    1    1    1
##
## $Humidity3pm
## X
##  12  16  18  20  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37
##   1   2   2   1   2   2   3   1   3   2   2   2   1   2   3   7   1   8   1   3
##  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57
##   2  11   6  11   8   7   4   8   1   5   7   6   4   8   6   7   3   6   6   4
##  58  59  60  61  62  63  64  66  67  68  69  70  71  72  73  74  75  76  77  81
##   4   3  10   2   3   3   7   7   3   3   2   1   3   2   1   2   1   1   1   1
##  82  83  84  86  88  90  91  92  93  94  96  97  99 100
##   3   2   1   1   1   2   1   2   3   2   1   2   4   1
##
## $Cloud3pm
## X
##    1    2    3    4    5    6    7    8 <NA>
##   21    7    7   11    9    5   21   75   99
##
## $WindDir3pm
## X
##    E  ENE  ESE    N   NE  NNE  NNW   NW    S   SE  SSE  SSW   SW    W  WNW  WSW
##   13   11   10   17    5    6   49   49    8    5    6    6    4    8   49    6
## <NA>
##    3
##
## $WindSpeed3pm
## X
##   11   13   15   17   19   20   22   24   26   28   30   31   33   37   39    7
##   18   20    8   29   13   20   10   15   18   13   19    3    6    2    2    1
##    9 Calm <NA>
##    7    3   48
##
## $Pressure3pm
```

```
## X
##  995.0  998.7  999.1  999.6 1000.6 1001.4 1001.7 1001.9 1002.8 1003.1 1003.9
##      1      1      1      1      1      1      1      1      1      1      1
## 1004.1 1004.5 1005.0 1005.2 1005.5 1005.6 1005.7 1005.8 1005.9 1006.1 1006.2
##      2      1      1      1      1      1      1      1      1      1      2
## 1006.3 1006.4 1006.8 1007.0 1007.1 1007.2 1007.4 1007.5 1007.6 1007.8 1007.9
##      1      1      1      1      1      1      1      1      2      1      1
## 1008.0 1008.3 1008.5 1008.7 1008.8 1008.9 1009.0 1009.1 1009.2 1009.3 1009.4
##      1      1      1      1      1      1      1      1      2      1      3
## 1009.6 1009.7 1009.8 1009.9 1010.0 1010.1 1010.4 1010.5 1010.7 1010.8 1010.9
##      2      4      2      1      1      1      1      3      1      2      1
## 1011.1 1011.5 1011.7 1011.8 1011.9 1012.0 1012.1 1012.2 1012.3 1012.4 1012.5
##      1      1      3      1      2      1      2      1      1      1      2
## 1012.7 1013.0 1013.1 1013.2 1013.3 1013.4 1013.5 1013.7 1014.1 1014.2 1014.3
##      1      1      3      4      1      1      1      1      1      4      2
## 1014.4 1014.6 1014.7 1014.8 1015.0 1015.1 1015.2 1015.3 1015.6 1015.9 1016.0
##      2      1      2      1      1      1      1      1      2      2      3
## 1016.1 1016.3 1016.4 1016.5 1016.6 1016.8 1016.9 1017.0 1017.1 1017.3 1017.4
##      3      2      1      1      2      2      1      1      2      2      1
## 1017.5 1017.6 1017.7 1017.8 1018.0 1018.1 1018.6 1018.7 1018.8 1018.9 1019.0
##      3      1      1      3      2      3      1      1      3      1      2
## 1019.1 1019.2 1019.4 1019.5 1019.6 1019.7 1019.8 1020.0 1020.1 1020.3 1020.4
##      1      4      1      1      1      2      2      4      1      1      2
## 1020.6 1020.8 1021.0 1021.1 1021.3 1021.4 1021.6 1021.8 1021.9 1022.0 1022.1
##      2      1      1      1      1      2      1      2      1      1      1
## 1022.3 1022.4 1022.5 1022.6 1022.7 1022.8 1022.9 1023.0 1023.1 1023.2 1023.3
##      1      1      2      2      1      1      1      1      4      1      3
## 1023.5 1023.7 1023.9 1024.0 1024.2 1024.3 1024.5 1024.6 1024.8 1025.2 1025.3
##      1      1      1      1      2      2      1      1      1      2      1
## 1025.5 1025.7 1026.0 1026.1 1026.5 1026.6 1026.7 1026.9 1027.0 1027.3 1027.7
##      1      2      1      2      2      1      2      1      4      1      1
## 1027.8 1028.1 1028.3 1028.4 1028.6 1029.7 1031.7 1032.0 1033.2
##      1      2      1      1      1      1      1      1      1
```

We identified 1 suspicious variables apart from "WindSpeed9am". This was "WindSpeed3pm" which also had 3 entries with name "Calm"

2.5 Problem 2E change "Calm" to 0, then format to numeric

```
#fixing Windspeed9am
data1[data1$WindSpeed9am == "Calm", ]$WindSpeed9am <- 0
data1$WindSpeed9am <- as.numeric(data1$WindSpeed9am)
```

```
data1$WindSpeed9am
```

```
##   [1] 17  7  0  6  9 11 19 17  6  7  2  7  2 24 13 11  7  7 15 19  7  7  0  9  2
```

```
##  [26] 35  9 17  9 11 13  6 11 15  4  0 30  2  6 11 13  4 15  7 11 26 19 15 15 11
##  [51]  7  6  6 20 11  2 11 11  6  7 15 13  2  0 17  4  7 13  7  0  6  0 19  9  9
##  [76] 15 11 20 20 22 13 13 30 19  0  0  0  0 13  9  0  7  7  0 11  7  2  7 17 28
## [101] 28  9  7 33 17  4  6  4  2 13 17 13  4  4  9  7  9  7  2  9  2  0  4  6 17
## [126] 20 11  0  7  7 13  6  4 20 43  0 13  7  6  7  9  9  0  7  0 13  0  9 30  6
## [151]  6  0  0  0  2  7  4  4  4  4 17 15 30 17  7  4  7 30  9 39 24  6  0  0 43
## [176] 19 20 13  4  9  7  0 15 17 22  0  9  6  7  7  4  9  0  6 15 17 26 26 20  9
## [201] 22 20  0  9 24 30 19 22 39 22  4  9  9  0 31 30 22  6  0  0  0  0 31 13  7
## [226]  6  2 28 13  2  0  0 15  0 37 15 24  7  7  2  2  2  0  6  4  4 24 28  6 30
## [251]  0 22 20 28 22
```

```
#fixing Windspeed3pm

data1[!is.na(data1$WindSpeed3pm == "Calm"), ]$WindSpeed3pm <- 0
data1$WindSpeed3pm <- as.numeric(data1$WindSpeed3pm)
```

```
data1$WindSpeed3pm
```

```
##   [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##  [26]  0  0  0  0  0  0  0  0  0 NA  0 NA NA NA  0 NA  0 NA NA NA  0  0  0  0 NA
##  [51]  0  0  0  0  0  0  0  0  0 NA  0 NA  0  0 NA  0  0 NA  0  0  0  0  0 NA NA
##  [76]  0  0  0  0  0 NA  0  0 NA  0 NA  0 NA NA NA NA  0  0 NA  0 NA NA NA  0  0
## [101]  0 NA  0 NA  0 NA NA NA  0  0 NA  0  0  0  0  0 NA NA NA  0 NA NA  0 NA  0
## [126]  0  0  0  0  0  0  0  0  0  0 NA  0  0  0  0 NA NA  0 NA  0  0  0 NA NA NA
## [151] NA  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [176]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [201]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [226]  0  0 NA  0  0  0  0  0  0 NA  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [251]  0  0  0  0  0
```

```
data1$RainToday <- ifelse(data1$Rainfall > 1, 1, 0)
data1$RainTomorrow <- c(data1$RainToday[2:nrow(data1)], NA)
```

2.6 Problem 2F saving the cleaned dataset

```
#save to csv
#?write.csv()
write.csv(data1, file="C:\\Users\\john\\Documents\\APSUGradSchool\\_Fall 2021\\MATH 5310\\_proj
```

3 Problem 3 Exploratory Data Analysis

```
set.seed(500) #reproducibility
tab <- table(data1$Month, data1$WindGustDir, useNA="no");
tab
```

```
##
##        E ENE ESE  N NE NNE NNW NW  S SE SSE SSW SW  W WNW WSW
##   APR  5   1   0  4  0   0   2  6  0  0   0   2  0  4   6   0
##   AUG  0   0   1  2  1   0   8 11  1  0   0   1  1  0   5   0
##   FEB  6   5   2  2  2   1   0  5  0  1   1   0  0  0   2   1
##   JAN  3   3   4  3  1   1   3 10  0  2   0   0  0  0   1   0
##   JUL  2   0   0  1  0   1   6 11  1  3   1   0  0  0   5   0
##   JUN  1   0   0  4  1   0  11  5  2  2   0   0  0  0   4   0
##   MAR  7   3   2  0  0   0   0  8  2  2   2   0  0  0   5   0
##   MAY  2   1   2  2  0   0   5  5  2  2   0   4  1  0   4   1
##   SEP  0   0   0  3  0   0   2  2  1  0   0   0  0  0   2   0
```

We noticed some strong NNW and NW Winds for June, July and Aug. This means some form of association.

Null Hypothesis H0: is that there is no association between Month and WindGustDir. Alt Hypothesis H1: is that there is a association between Month and WindGustDir.

Assume our confidence level, alpha to be 0.01

```
fisher.test(tab, simulate.p.value =TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  tab
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

Therefore since p-value is less than 0.1, we reject the Null Hypothesis HO and conclude that there is some form of association between Months and the WindGustDir.

```
set.seed(500) #reproducibility
tab <- table(data1$Month, data1$WindGustSpeed, useNA="no");
tab
```

```
##
##       13 15 17 19 20 22 24 26 28 30 31 33 35 37 39 41 43 44 46 48 50 52 54 56
##   APR  0  0  2  1  2  0  4  2  3  0  1  2  2  3  0  1  2  0  0  0  2  0  1  1
##   AUG  0  0  0  0  0  0  0  3  0  1  3  0  3  4  2  2  0  2  0  3  0  2  1  1
##   FEB  0  0  0  0  0  0  0  0  1  2  0  4  4  4  1  1  3  3  1  1  1  0  1  0
```

```
##   JAN  0  0  0  0  0  0  0  0  0  1  3  1  2  3  5  1  0  1  3  1  1  3  0  3
##   JUL  1  0  1  3  0  2  0  0  0  0  1  3  1  5  1  1  1  1  1  1  1  0  2  0
##   JUN  1  2  1  2  0  0  1  0  4  3  1  0  1  3  1  2  2  2  0  1  1  0  0  0
##   MAR  0  0  0  0  0  0  0  1  1  2  3  1  0  5  2  0  3  3  3  2  1  2  1  0
##   MAY  0  1  2  3  1  0  2  1  2  3  2  0  1  0  5  1  0  2  1  1  0  0  2  0
##   SEP  0  0  0  0  0  1  0  0  0  0  0  1  0  1  0  2  1  0  1  1  0  0  0  1
##
##       57 59 61 63 67 69 70 72
##   APR  1  0  0  0  0  0  0  0
##   AUG  0  0  1  0  0  0  2  1
##   FEB  0  0  0  0  0  0  0  1
##   JAN  1  0  0  0  1  0  1  0
##   JUL  0  3  1  0  0  0  0  1
##   JUN  0  0  0  0  0  1  1  0
##   MAR  0  1  0  0  0  0  0  0
##   MAY  0  0  0  1  0  0  0  0
##   SEP  0  0  1  0  0  0  0  0
```

We noticed no form of association between Month and WindGustSpeed.

Null Hypothesis H0: is that there is no association between Month and WindGustSpeed. Alt Hypothesis H1: is that there is a association between Month and WindGustSpeed.

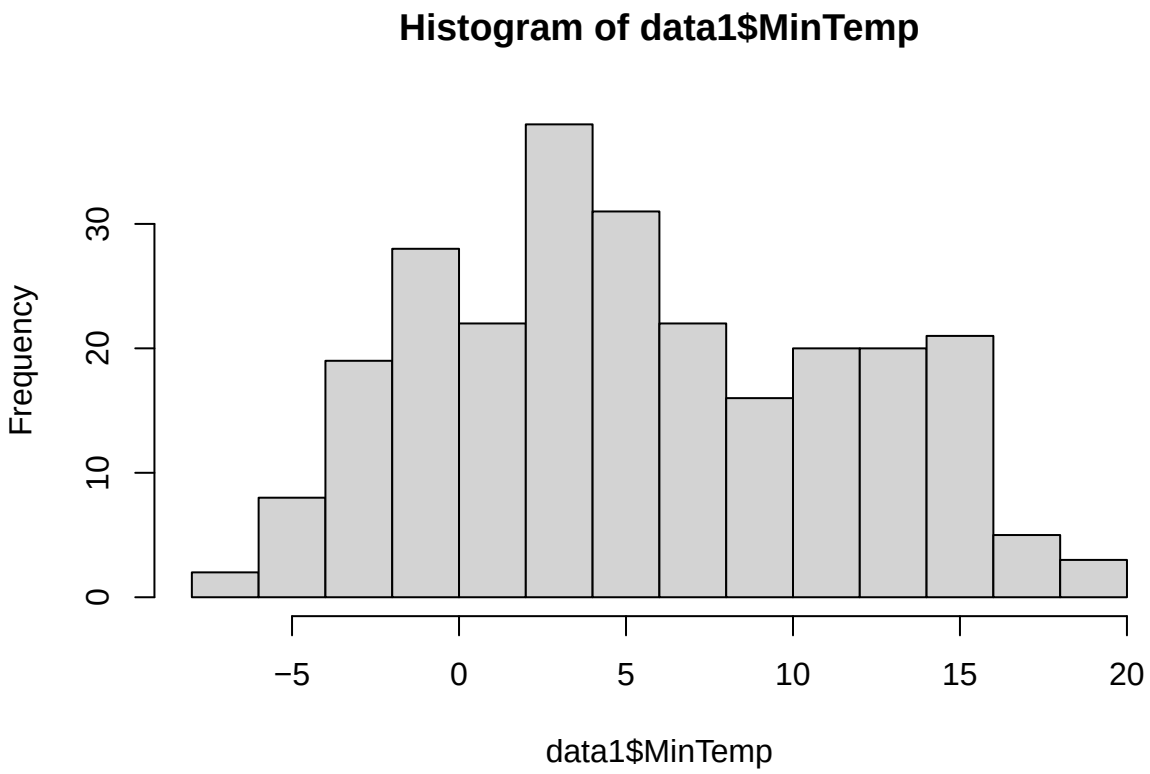
Assume our confidence level, alpha to be 0.01

```
chisq.test(tab, simulate.p.value =TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tab
## X-squared = 298, df = NA, p-value = 0.01349
```

Therefore since p-value is greater than 0.1, we accept the Null Hypothesis HO and conclude that there is no form of association between Months and the WindGustSpeed.

```
set.seed(500) #reproducibility
hist(data1$MinTemp, useNA="no")
```



We noticed the histogram of Minimum Temperature does not follow a normal distribution

Null Hypothesis H0: is that there is no association between RainToday and MinTemp. Alt Hypothesis H1: is that there is a association between RainToday and MinTemp.

Assume our confidence level, alpha to be 0.00001

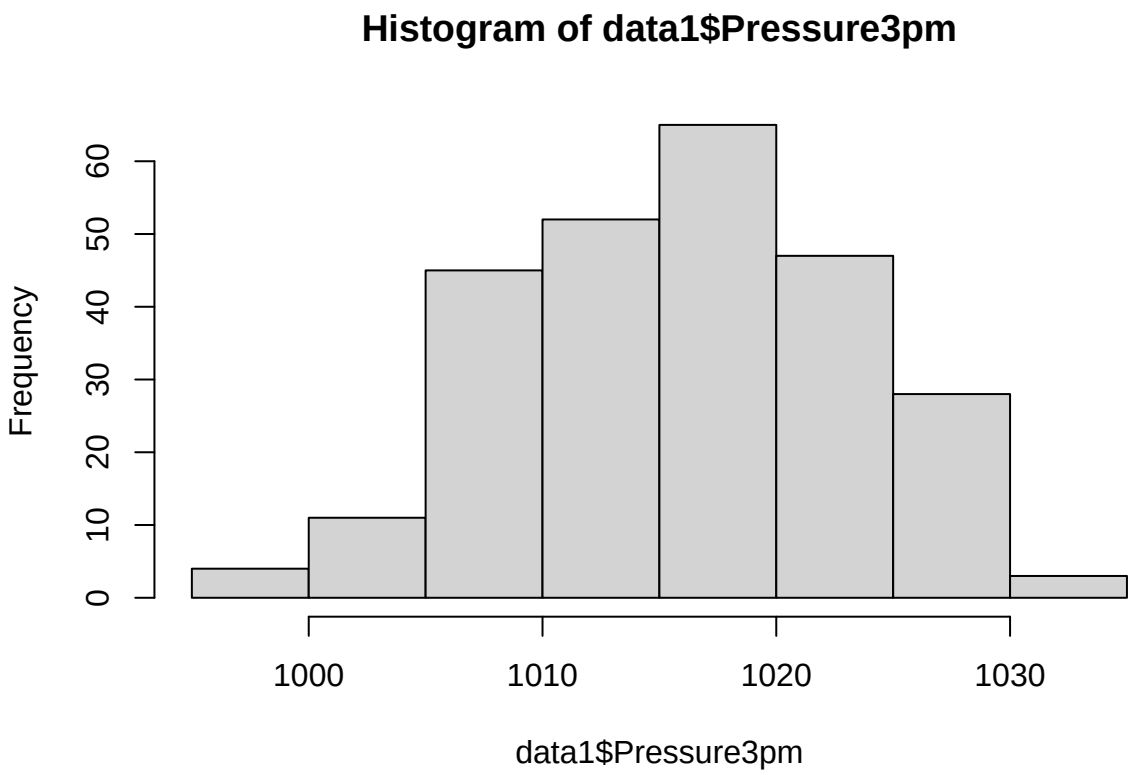
```
wilcox.test(data1$MinTemp~data1$RainToday, data=data1, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data1$MinTemp by data1$RainToday
## W = 3780.5, p-value = 0.0002394
## alternative hypothesis: true location shift is not equal to 0
```

```
#chisq.test(tab, simulate.p.value =TRUE)
```

Therefore since p-value is greater than 0.00001, we accept the Null Hypothesis HO and conclude that there is no association between RainToday and MinTemp.

```
set.seed(500) #reproducibility
hist(data1$Pressure3pm, useNA="no")
```



We noticed the histogram of Minimum Temperature nearly follow a normal distribution

Null Hypothesis H0: is that there is no association between RainToday and Pressure at 3pm. Alt Hypothesis H1: is that there is a association between RainToday and Pressure at 3pm.

Assume our confidence level, alpha to be 0.01

```
t.test(data1$Pressure3pm~data1$RainToday, data=data1, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  data1$Pressure3pm by data1$RainToday
## t = 4.6597, df = 87.621, p-value = 1.123e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
##  2.833121 7.047120
## sample estimates:
## mean in group 0 mean in group 1
##        1017.195        1012.255
```

Therefore since p-value is lesser than 0.00001, we reject the Null Hypothesis HO and conclude that there is some form of association between RainToday and Pressure at 3pm.