Project I: Data Preparation1

Chitra karki

9/7/2021

Instructions: While discussions with classmates are allowed and encouraged, please try to work on the project independently and direct your questions to the TA/instructor. Please interpret your analysis results using concise and clear language and focusing on interesting findings. You are requested to use R Markdown to solve this project. Submissions are to be uploaded as a single (knitted) PDF file that contains both the code, your comments/conclusions and relevant output/figures, etc.

In this project, we work with t he 2021 Australian daily weather database available online at http://www.bom.gov.au/climate/dwo/. We will first read the data into R and then prepare data by going through a number of issues so that a clean data set can be used for forecasting next day's weather. Finally, some simple exploratory data analysis (EDA) will be conducted.

1.**(Data Input)** For simplicity, we shall focus on the weather data only at Canberra, which is the Australian Capital Territory. Data for the latest month can be found at http://www.bom.gov.au/climate/dwo/IDCJDW2801.latest.shtml However, we will use the CSV text file. See the link by scrolling down the page. There are many files like this at the same website for different months over the years. With the following R code, we read in the 2021 daily weather data (available on Blackboard) from January 1st till August 31st into one dataframe via a simple loop.

```
# READ DAILY WEATHER DATA IN 2021
setwd("C:/Users/chitr/OneDrive - University of Texas at El Paso/data_science/sem1-fall-2021/stat5474_datamining-d
r_poko/hw/project-1")
dat <- NULL
i0 = NULL
current.month <- 9
for (i in 1:(current.month - 1)){
i0 <- ifelse(i<10, paste("0", i, sep=""), i)
mth <- paste("2021", i0, sep="")
bom <- paste("IDCJDW2801.", mth, ".csv", sep="")
dat.i <- read.csv(bom, skip=6, check.names=FALSE,
na.strings = c("NA", "", " "), stringsAsFactors = FALSE)
dat.i[, 1] <-toupper(month.abb[i])
dat <- rbind(dat,dat.i)
}
```

Along the way, we have created one month variable with values {JAN, FEB, . .. AUG }. Report the dimension of the data you obtain and include the first few data lines in your report.

```
dim(dat)
```

```
## [1] 243  22
```

The data we created has 243 rows and 22 colums.

few data lines

```
head(dat)
```

	Date	Minimum temperature (°C)	Maximum temperature (°C)	Rainfall (mm)	Evaporation (mm)
	<chr><chr>	<dbl>	<dbl>	<dbl>	<lgl>
1	JAN 2021-01-1	14.4	19.7	0.0	NA
2	JAN 2021-01-2	13.0	22.3	0.0	NA
3	JAN 2021-01-3	15.7	26.7	5.8	NA
4	JAN 2021-01-4	14.6	24.8	7.2	NA
5	JAN 2021-01-5	14.1	28.0	5.4	NA
6	JAN 2021-01-6	13.0	24.7	0.2	NA

6 rows | 1-7 of 23 columns

2. **(Data Cleaning and Preparation)** We next clean the data and make it ready for the subsequent analysis. Specifically, we shall address the following issues:

a. First print out the (sorted) unique values or levels of every variable in the data set. You may obtain the frequency table by treating all variables as if categorical. To do so, you may consider using table(x, useNA="ifany"), where the useNA="ifany" option allows us to see missing values.

```
#colnames(dat)[1] = "Month" # naming the first column
#var = colnames(dat)
#for (i in 1:length(var)) {table(dat[,i],useNA = "ifany")
lapply(dat,table)
```

```
## [[1]]
##
## APR AUG FEB JAN JUL JUN MAR MAY
##  30  31  28  31  31  30  31  31
##
## $Date
##
##  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
##          1          1          1          1          1
##
## $`Minimum temperature (°C)`
##
## -6.3   -6 -5.4 -5.2 -5.1   -5 -4.9 -4.8 -4.7 -4.1 -3.8 -3.7 -3.6 -3.5 -3.4 -3.3
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    3    1
## -3.1   -3 -2.9 -2.8 -2.7 -2.6 -2.5 -2.2 -2.1   -2 -1.9 -1.8 -1.7 -1.6 -1.2 -1.1
##    1    1    1    1    1    1    1    1    1    1    2    1    4    2    3    1
##   -1 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.1    0  0.2  0.3  0.5  0.6  0.7  0.9    1
##    2    1    1    1    1    1    3    1    2    1    1    3    3    1    2    1
##  1.1  1.4  1.8  1.9  2.1  2.3  2.4  2.5  2.6  2.8  3.2  3.3  3.4  3.5  3.6  3.7
##    1    2    2    4    2    2    3    2    1    3    3    1    4    3    3    1
##  3.8  3.9    4  4.1  4.2  4.3  4.4  4.7  4.8  4.9    5  5.4  5.5  5.6  5.7  5.8
##    3    2    2    1    1    1    3    1    3    2    2    4    1    1    2    5
##  5.9    6  6.1  6.2  6.4  6.6  6.7    7  7.1  7.2  7.3  7.4  7.6    8  8.1  8.2
##    2    1    3    1    1    3    2    1    1    2    1    2    1    2    1    2
##  8.3  8.4  8.7    9  9.2  9.7  9.8  9.9 10.1 10.2 10.3 10.4 10.5 10.7 10.9 11.2
##    2    2    1    1    1    1    1    3    2    1    1    1    3    2    2    1
## 11.4 11.5 11.6   12 12.1 12.6 12.8 12.9   13 13.2 13.3 13.4 13.6 13.7 13.9   14
##    1    2    1    2    3    1    2    1    3    2    1    1    2    2    1    1
## 14.1 14.2 14.3 14.4 14.6 14.8 14.9   15 15.5 15.6 15.7 15.9 16.3 16.4 16.7 17.1
##    4    1    1    2    2    1    2    2    1    2    2    1    1    1    1    1
## 17.5 18.1 18.5 18.8
##    1    1    1    1
##
## $`Maximum temperature (°C)`
##
##  6.3  8.2  8.3  8.8  8.9  9.3  9.4  9.5  9.6  9.7   10 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
## 11.3 11.4 11.5 11.7 11.8 11.9   12 12.1 12.3 12.4 12.5 12.6 12.7 12.8 12.9   13
##    3    1    1    1    1    1    2    2    3    3    4    3    1    2    2    3
## 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 14.9
##    3    3    1    3    1    2    1    1    3    2    1    7    3    1    1    3
##   15 15.2 15.3 15.4 15.7 15.8 15.9 16.1 16.3 16.7 16.9   17 17.1 17.4 17.5 17.6
##    1    1    1    1    2    5    3    1    2    1    2    1    1    3    1    2
## 17.9   18 18.2 18.3 18.5 18.6 18.7 18.8 18.9 19.1 19.3 19.4 19.7 19.8 19.9 20.1
##    2    1    1    2    2    3    1    2    1    1    1    3    2    4    1    1
## 20.4 20.5 20.6 20.7 20.8   21 21.3 21.4 21.5 21.6 21.7 21.9 22.1 22.2 22.3 22.4
##    1    1    2    4    1    1    1    2    1    2    1    1    1    1    2    2
## 22.6 23.3 23.4 23.5 23.6 24.1 24.2 24.3 24.4 24.5 24.7 24.8 24.9   25 25.1 25.2
##    2    1    1    1    2    1    2    2    2    1    2    2    2    2    2    1
## 25.3 25.6 25.7   26 26.2 26.3 26.5 26.7 26.9   27 27.1 27.2 27.5 27.7 27.9   28
##    1    1    1    2    1    1    1    1    1    3    1    1    2    3    1    2
## 28.1 28.2 28.3 29.2 29.3 29.4 29.5 30.1 30.3 30.5 30.8 30.9 31.8 32.5 32.7 32.9
##    2    1    2    1    1    1    1    3    1    1    1    1    1    1    1    1
## 33.7 34.1 34.5 35.9 37.5   38
##    1    1    1    1    1    1
##
## $`Rainfall (mm)`
##
##    0  0.2  0.4  0.6  0.8    1  1.2  1.4  1.8    2  2.2  2.4  2.6  2.8    3  3.2
##  153   28    2    3    2    2    1    1    2    1    1    2    1    3    1    1
##  3.6    4  4.2  4.8    5  5.4  5.8  7.2  7.6  7.8    8  8.2    9  9.4  9.8 10.2
##    1    1    1    2    3    3    2    1    1    1    1    1    1    1    1    1
## 10.4 10.8   11 11.2 12.4 19.2 22.2 22.8 23.8 25.4 28.4 29.4 30.6
##    1    2    1    3    1    1    1    1    1    1    1    1    2
##
## $`Evaporation (mm)`
## < table of extent 0 >
##
## $`Sunshine (hours)`
## < table of extent 0 >
##
## $`Direction of maximum wind gust `
##
##   E ENE ESE   N  NE NNE NNW  NW   S  SE SSE SSW  SW   W WNW WSW
##  26  13  11  18   5   3  35  61   8  12   4   7   2   4  32   2
##
## $`Speed of maximum wind gust (km/h)`
##
## 13 15 17 19 20 22 24 26 28 30 31 33 35 37 39 41 43 44 46 48 50 52 54 56 57 59
##  2  3  6  9  3  2  7  7 11 12 14 11 14 27 17  9 11 14  9 10  7  7  8  5  2  4
## 61 63 67 69 70 72
##  2  1  1  1  4  3
##
## $`Time of maximum wind gust`
##
## 00:00 00:10 00:17 00:29 00:40 00:58 01:03 01:08 01:19 01:38 02:52 03:19 04:06
##     1     1     1     1     1     1     1     1     1     1     1     1     1
## 04:22 04:28 04:41 04:56 05:01 05:06 05:07 05:16 05:27 05:31 06:09 06:15 07:10
##     1     1     1     1     1     1     1     1     1     1     1     1     1
## 07:25 07:44 07:58 08:54 08:57 09:04 09:05 09:10 09:13 09:15 09:26 09:45 09:46
##     1     1     1     1     2     1     1     1     1     1     1     1     1
## 09:49 09:51 09:54 09:55 10:02 10:28 10:59 11:03 11:04 11:06 11:10 11:11 11:20
##     1     1     1     1     1     1     1     1     1     1     1     1     1
## 11:26 11:30 11:33 11:37 11:45 11:48 11:49 11:59 12:00 12:02 12:04 12:07 12:08
##     1     1     1     1     1     2     2     1     1     1     1     1     1
## 12:14 12:16 12:21 12:22 12:24 12:25 12:29 12:34 12:35 12:36 12:39 12:40 12:41
##     1     1     1     1     2     1     1     1     2     1     1     1     1
## 12:42 12:43 12:46 12:58 12:59 13:00 13:05 13:06 13:08 13:09 13:15 13:16 13:19
##     1     1     1     2     1     1     1     1     3     1     1     3     1
## 13:21 13:22 13:23 13:25 13:26 13:28 13:29 13:33 13:37 13:41 13:42 13:43 13:45
##     1     2     1     1     1     2     2     4     1     2     1     1     1
## 13:46 13:47 13:50 13:51 13:52 13:57 13:59 14:01 14:03 14:04 14:08 14:09 14:14
##     1     1     1     1     2     2     2     2     1     1     1     1     2
## 14:17 14:18 14:19 14:23 14:24 14:29 14:33 14:35 14:36 14:40 14:42 14:44 14:45
##     1     1     1     2     2     1     1     1     1     1     1     2     1
## 14:50 14:53 14:54 14:55 14:56 14:58 15:00 15:01 15:02 15:05 15:10 15:16 15:17
##     1     2     1     2     1     1     1     1     3     1     1     1     1
## 15:20 15:21 15:26 15:27 15:31 15:33 15:37 15:38 15:41 15:42 15:43 15:44 15:45
##     1     1     1     1     1     2     1     1     2     1     1     1     1
## 15:46 15:47 15:51 15:54 15:55 15:59 16:02 16:06 16:12 16:15 16:16 16:24 16:30
##     1     1     1     1     1     1     1     1     1     1     1     1     1
## 16:46 16:47 16:49 16:55 17:09 17:21 17:22 17:27 17:43 17:57 17:58 18:05 18:06
##     1     1     1     1     1     1     1     1     1     1     1     1     1
## 18:07 18:15 18:23 18:26 18:27 18:34 18:35 18:46 18:51 18:55 18:56 19:05 19:24
##     1     1     2     1     1     1     1     1     1     1     1     1     1
## 19:27 19:52 20:00 20:04 20:06 20:10 20:25 20:26 20:59 21:45 21:51 22:05 22:28
##     1     1     1     1     2     1     1     1     1     1     1     1     1
## 22:30 23:00
##     1     1
##
## $`9am Temperature (°C)`
##
## -0.8    0  0.3  0.4  0.6  0.8  0.9    1  1.2  1.3  1.4  1.8    2  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  3.1  3.4  3.7  3.8    4  4.1  4.3  4.4  4.5  4.6  5.2  5.4  5.5  5.6
##    1    2    1    2    1    1    5    1    1    1    2    1    2    4    2    2
##  5.7  5.8  5.9    6  6.2  6.3  6.5  6.6  6.7  6.8  6.9    7  7.2  7.3  7.6  7.8
##    1    2    1    2    3    2    1    2    3    2    2    2    1    2    1    3
##    8  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  9.6
##    2    2    3    2    2    3    2    6    3    1    1    5    1    1    1    2
##  9.8  9.9 10.2 10.3 10.4 10.5 10.6 10.7 10.8   11 11.1 11.2 11.3 11.4 11.5 11.8
##    3    3    3    2    1    2    2    1    1    1    1    1    1    2    1    1
##   12 12.2 12.3 12.4 12.7 12.9   13 13.2 13.3 13.4 13.8 14.2 14.4 14.5 14.7   15
##    1    1    4    2    1    1    2    1    1    2    3    2    2    1    1    1
## 15.1 15.2 15.3 15.4 15.5 15.6 15.8 15.9   16 16.3 16.4 16.5 16.6 16.7 16.8 16.9
##    3    1    2    4    1    2    1    2    1    1    3    1    4    4    2    3
##   17 17.1 17.2 17.3 17.5 17.8 17.9   18 18.1 18.2 18.3 18.4 18.8   19 19.1 19.2
##    1    1    1    1    2    2    1    1    1    3    2    2    1    1    3    3
## 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 25.4 26.3
##    1    1    3    1    1    1    1    1    1    1    1    1    1    1    1    1
##
## $`9am relative humidity (%)`
##
##  34  37  42  45  49  50  52  53  55  56  57  58  59  60  61  62  63  64  65  66
##   1   2   1   1   2   4   1   1   1   3   3   1   4   3   7   2   8   3   6   3
##  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86
##   2   3   4   5   1   1   6   6   5   4   7   7   6   2   4   8   3   8   4   7
##  87  88  89  90  91  92  93  94  95  96  97  98  99 100
##   6   7   5   3   5   4   1   7   6   2   2   1  25  19
##
## $`9am cloud amount (oktas)`
##
##  1  2  3  4  5  6  7  8
##  8  9  4  4  6  5 18 88
##
## $`9am wind direction`
##
##   E ENE ESE   N  NE NNE NNW  NW   S  SE SSE SSW  SW   W WNW WSW
##  13   8  13  26   2   1  25  20  19  21  25  13   7   5   8   1
##
## $`9am wind speed (km/h)`
##
##   11   13   15   17   19    2   20   22   24   26   28   30   31   33   35   37
##   13   16   12   12    7   16    8    6    4    3    3    7    2    1    1    1
##   39    4   43    6    7    9 Calm
##    2   17    2   19   33   22   36
##
## $`9am MSL pressure (hPa)`
##
##  996.2  996.9  998.7 1001.8 1002.3 1003.2 1003.4   1005 1005.5 1005.7 1006.1
##      1      1      1      1      1      1      1      1      1      1      1
## 1006.3 1006.5 1006.6 1006.8   1007 1007.2   1008 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 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 1013.1 1013.2 1013.3 1013.5 1013.6 1013.8 1013.9   1014
##      2      1      3      1      1      1      1      3      4      1      1
## 1014.1 1014.4 1014.5 1014.9 1015.2 1015.4 1015.5 1015.6 1015.7 1015.9   1016
##      1      1      3      4      1      2      2      2      2      1      3
## 1016.1 1016.3 1016.4 1016.6 1016.8   1017 1017.1 1017.2 1017.3 1017.5 1017.6
##      3      1      2      1      3      1      1      1      1      3      1
## 1017.9 1018.1 1018.2 1018.3 1018.6 1018.7 1018.8 1018.9   1019 1019.1 1019.2
##      1      1      1      2      3      1      1      1      1      1      1
## 1019.3 1019.4 1019.5 1019.6 1019.7 1019.8   1020 1020.1 1020.2 1020.3 1020.6
##      1      3      1      3      1      1      2      2      1      1      1
## 1020.8 1020.9 1021.3 1021.5 1021.6 1021.7 1021.8   1022 1022.1 1022.2 1022.3
##      1      1      1      2      1      1      1      2      2      2      2
## 1022.4 1022.5 1022.6 1022.7   1023 1023.2 1023.3 1023.4 1023.5 1023.7 1023.8
##      1      2      5      1      2      1      1      1      1      4      1
## 1024.1 1024.3 1024.5 1024.6 1024.9   1025 1025.1 1025.3 1025.4 1025.5 1025.6
##      2      2      2      1      1      1      1      2      3      1      2
## 1025.8 1025.9   1026 1026.1 1026.2 1026.3 1026.4 1026.5 1026.7 1026.8   1027
##      1      1      1      1      1      1      2      1      2      1      2
## 1027.1 1027.2 1027.3 1027.6 1028.1 1028.4 1029.3 1029.4 1029.5 1029.6 1029.7
##      1      3      1      1      2      1      1      1      3      1      1
##   1030 1030.1 1030.2 1030.7 1030.8 1030.9   1031 1031.5 1032.2 1032.3 1032.6
##      1      1      1      1      1      1      1      1      1      1      1
## 1033.4 1033.8 1034.3 1036.6
##      1      1      1      1
##
## $`3pm Temperature (°C)`
##
##  5.8    6  6.9    7  7.4  7.5  7.7  7.9    8  8.5  8.7    9  9.2  9.3  9.5  9.6
##    2    1    1    1    1    1    1    1    2    1    2    1    1    3    1    2
##  9.9 10.1 10.3 10.4 10.5 10.7 10.8 10.9   11 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 12.1 12.2 12.3 12.4 12.5 12.6 12.7 12.8   13 13.1 13.2 13.3 13.4 13.6
##    2    6    1    1    3    1    3    2    2    2    3    2    1    1    5    3
## 13.8   14 14.2 14.3 14.4 14.5 14.6 14.7 14.8 14.9   15 15.1 15.3 15.4 15.7   16
##    3    3    1    2    1    3    2    1    2    1    2    1    2    1    1    2
## 16.1 16.3 16.6 16.8   17 17.1 17.2 17.3 17.6 17.7 17.8 17.9   18 18.1 18.2 18.3
##    3    1    1    1    2    1    2    1    1    1    1    1    5    2    1    1
## 18.4 18.5 18.6 18.7 18.9   19 19.1 19.2 19.4 19.5 19.8   20 20.1 20.3 20.6 20.7
##    2    3    1    1    1    2    3    1    1    2    2    1    1    2    1    3
## 20.8   21 21.1 21.4 21.5 21.6 21.9 22.3 22.5 22.6 22.7 22.8 22.9 23.1 23.2 23.4
##    2    2    1    1    3    2    2    1    1    1    4    2    2    2    1    2
##   24 24.1 24.4 24.6 24.7   25 25.1 25.2 25.4 25.6 25.9   26 26.1 26.3 26.4 26.5
##    1    2    2    1    1    1    2    1    1    1    1    1    2    1    1    1
## 26.6 26.7 26.8   27 27.2 27.4 27.6 27.8 28.1 28.3 28.4   29 29.4 29.7 31.3 31.5
##    4    1    1    1    1    1    1    1    2    2    2    2    1    1    2    1
##   32 32.8 34.6 36.1 36.2
##    1    1    1    1    1
##
## $`3pm relative humidity (%)`
##
## 12 16 18 20 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
##  1  2  2  1  2  2  3  1  2  1  2  1  1  2  3  7  1  8  1  2  2 10  5 11  8  7
## 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 66 67 68 69 70
##  3  8  1  5  7  6  4  8  6  6  3  5  6  3  4  3 10  2  3  3  7  7  3  3  2  1
## 71 72 73 74 75 76 77 81 82 83 84 86 90 91 92 93 94 96 97 99
##  3  2  1  2  1  1  1  1  3  2  1  1  2  1  2  3  2  1  2  4
##
## $`3pm cloud amount (oktas)`
##
##  1  2  3  4  5  6  7  8
## 19  7  7 11  9  5 21 71
##
## $`3pm wind direction`
##
##   E ENE ESE   N  NE NNE NNW  NW   S  SE SSE SSW  SW   W WNW WSW
##  13  11  10  14   5   6  47  45   7   5   6   6   4   8  47   6
##
## $`3pm wind speed (km/h)`
##
##   11   13   15   17   19   20   22   24   26   28   30   31   33   35   37   39
##   15   20   20   29   13   18   10   14   17   12   17    9    5    3    1    2
##    4   41   43    6    7    9 Calm
##    1    1    1    2    4   26    3
##
## $`3pm MSL pressure (hPa)`
##
##    995  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 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 1006.8   1007 1007.1 1007.2 1007.4 1007.5 1007.6 1007.8 1007.9   1008
##      1      1      1      1      1      1      1      2      1      1      1
## 1008.3 1008.5 1008.7 1008.8 1008.9   1009 1009.1 1009.2 1009.3 1009.4 1009.6
##      1      1      1      1      1      1      1      2      1      3      2
## 1009.7 1009.8 1009.9   1010 1010.1 1010.4 1010.5 1010.7 1010.8 1010.9 1011.1
##      4      2      1      1      1      1      3      1      2      1      1
## 1011.5 1011.7 1011.8 1011.9   1012 1012.1 1012.2 1012.3 1012.4 1012.5 1012.7
##      1      3      1      1      1      2      1      1      1      2      1
##   1013 1013.1 1013.2 1013.3 1013.4 1013.5 1013.7 1014.1 1014.2 1014.3 1014.4
##      1      3      4      1      1      1      1      1      4      2      2
## 1014.6 1014.7 1014.8   1015 1015.1 1015.2 1015.6 1015.9   1016 1016.1 1016.3
##      1      2      1      1      1      1      2      1      3      3      2
## 1016.4 1016.5 1016.6 1016.8 1016.9   1017 1017.1 1017.3 1017.4 1017.5 1017.6
##      1      1      2      2      1      1      2      2      1      3      1
## 1017.7 1017.8   1018 1018.1 1018.6 1018.7 1018.8 1018.9   1019 1019.1 1019.2
##      1      3      2      3      1      1      3      1      2      1      4
## 1019.4 1019.5 1019.6 1019.7 1019.8   1020 1020.3 1020.4 1020.6 1020.8   1021
##      1      1      1      2      2      4      1      2      1      1      1
## 1021.1 1021.3 1021.4 1021.6 1021.8 1021.9   1022 1022.1 1022.3 1022.4 1022.5
##      1      1      2      1      2      1      1      1      1      1      2
## 1022.6 1022.7 1022.8   1023 1023.1 1023.2 1023.3 1023.5 1023.7 1023.9   1024
##      2      1      1      1      4      1      3      1      1      1      1
## 1024.2 1024.3 1024.5 1024.8 1025.2 1025.3 1025.5 1025.7 1026.1 1026.5 1026.6
##      1      2      1      1      2      1      1      2      2      2      1
## 1026.7 1026.9   1027 1027.7 1027.8 1028.1 1028.3 1028.4 1028.6 1029.7 1031.7
##      2      1      4      1      1      1      1      1      1      1      1
##   1032 1033.2
##      1      1
```

```
#}
```

{r} #a = list() #for (i in 1:length(names(dat))) { # a[i]= table(dat[,i],useNA = "ifany") } #table(dat$`Maximum temperature (°C)`,useNA = "ifany")

Then inspect for suspicious or problematic records. Comment on issues that you find out. For example, you may find that variable called "9am wind speed (km/h)" has records with value Calm. This value should be changed to 0 for the analysis purpose. What other variables have similar problems? (b) The variable called "Time of maximum wind gust" in the 10-th column may not be useful for weather forecast purpose and its character-valued time info is difficult to enter any model. So let's remove it from the data set.

```
dat <- dat[, -c(10)]
```

c. The variable names are too long. Rename the data set as follows.

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

```
## [1] 243  21
```

```
names(dat)
```

```
##  [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"
```

d. Variables that have Calm as a value would be automatically treated by R as 'character' or categorical, which is not the right way. For these variables, first change their Calm values as 0 and then change their types into 'numerical' using the function as.numeric().

```
dat[dat == "Calm"]= as.numeric(0)
```

e. Define a variable called RainToday based on Rainfall so that RainToday is 1 if Rainfall is greater than 1mm and 0 otherwise. Namely, if it rains less than 1 mm in a day, then we report that as no rain.

```
dat$RainToday = ifelse(dat$Rainfall> 1,1,0)
```

Next, define a variable called RainTomorrow by shifting RainToday one day forward:

```
dat$RainTomorrow <- c(dat$RainToday[2:nrow(dat)], NA)
```

Now the data is pretty much ready for some EDA and weather forest usage, where the 0-1 binary variable RainTomorrow is the target or response variable.

f. Save a Rdata copy of the data set you have prepared using the save() function that you obtain but do not submit it with your project report.

```
save(dat,file = "dat.Rdata")
```

3. **(Exploratory Data Analysis)** Perform some EDA of your choice on the data set that you obtained. As a general principal, numerically you may consider two-way contingency table plus χ2 test (or Fisher exact test) for independence to assess the association between a categorical variable with the binary outcome and consider two-sample t test (or the nonparametric Wilcoxon rank sum test) to assess the association between a continuous variable with the binary outcome. Graphical tools can be used as well, e.g., histogram and/or boxplot for continuous 2 variables and bar plot and/or mosaic plot for categorical variables (possibly grouped by the binary outcome) among many other choices.

```
# chi-square test for selected catoregorical variables with binary output
chisq.test(dat$WindDir3pm,dat$RainTomorrow)
```

```
## Warning in chisq.test(dat$WindDir3pm, dat$RainTomorrow): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  dat$WindDir3pm and dat$RainTomorrow
## X-squared = 21.577, df = 15, p-value = 0.1194
```

```
chisq.test(dat$WindSpeed9am,dat$RainTomorrow) # not significant
```

```
## Warning in chisq.test(dat$WindSpeed9am, dat$RainTomorrow): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  dat$WindSpeed9am and dat$RainTomorrow
## X-squared = 25.781, df = 22, p-value = 0.2612
```

```
chisq.test(dat$Cloud9am,dat$RainTomorrow) # not significant
```

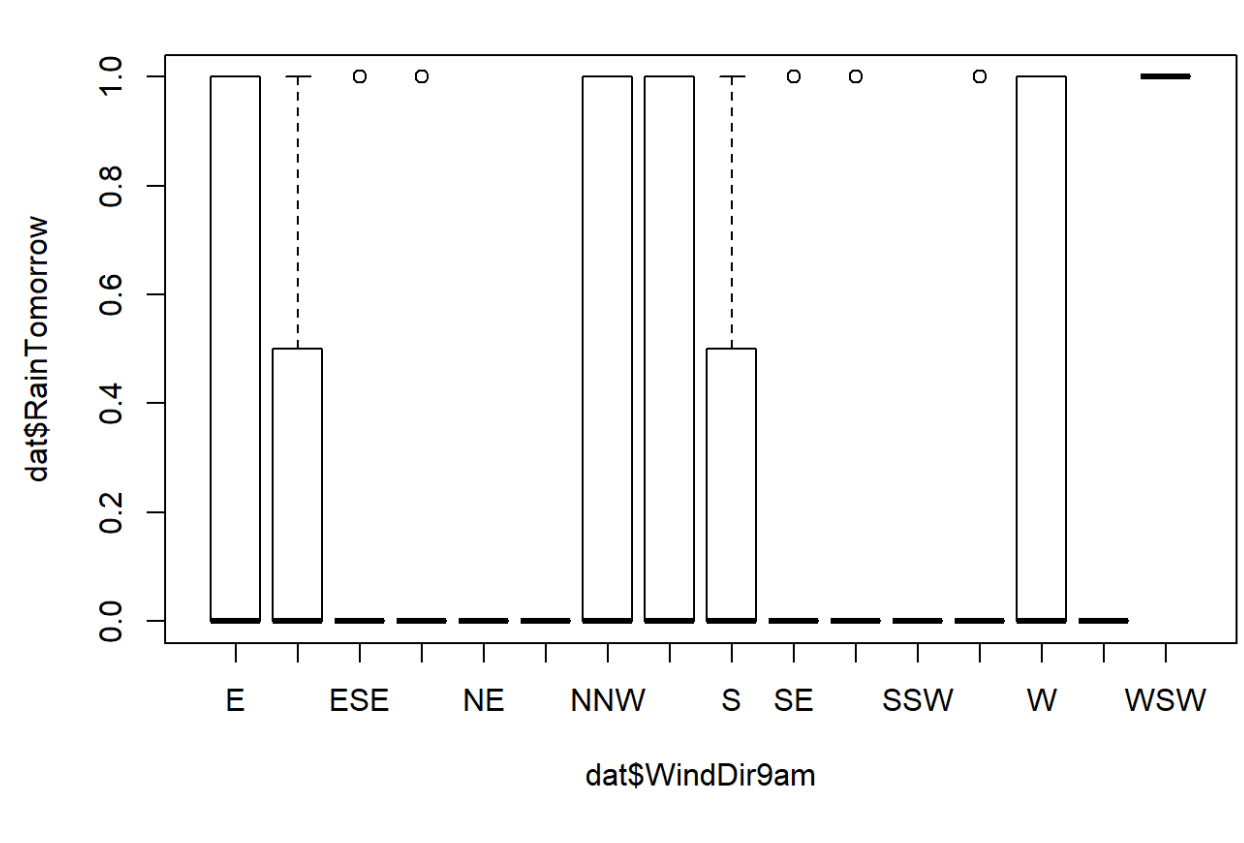
```
## Warning in chisq.test(dat$Cloud9am, dat$RainTomorrow): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  dat$Cloud9am and dat$RainTomorrow
## X-squared = 7.921, df = 7, p-value = 0.3396
```

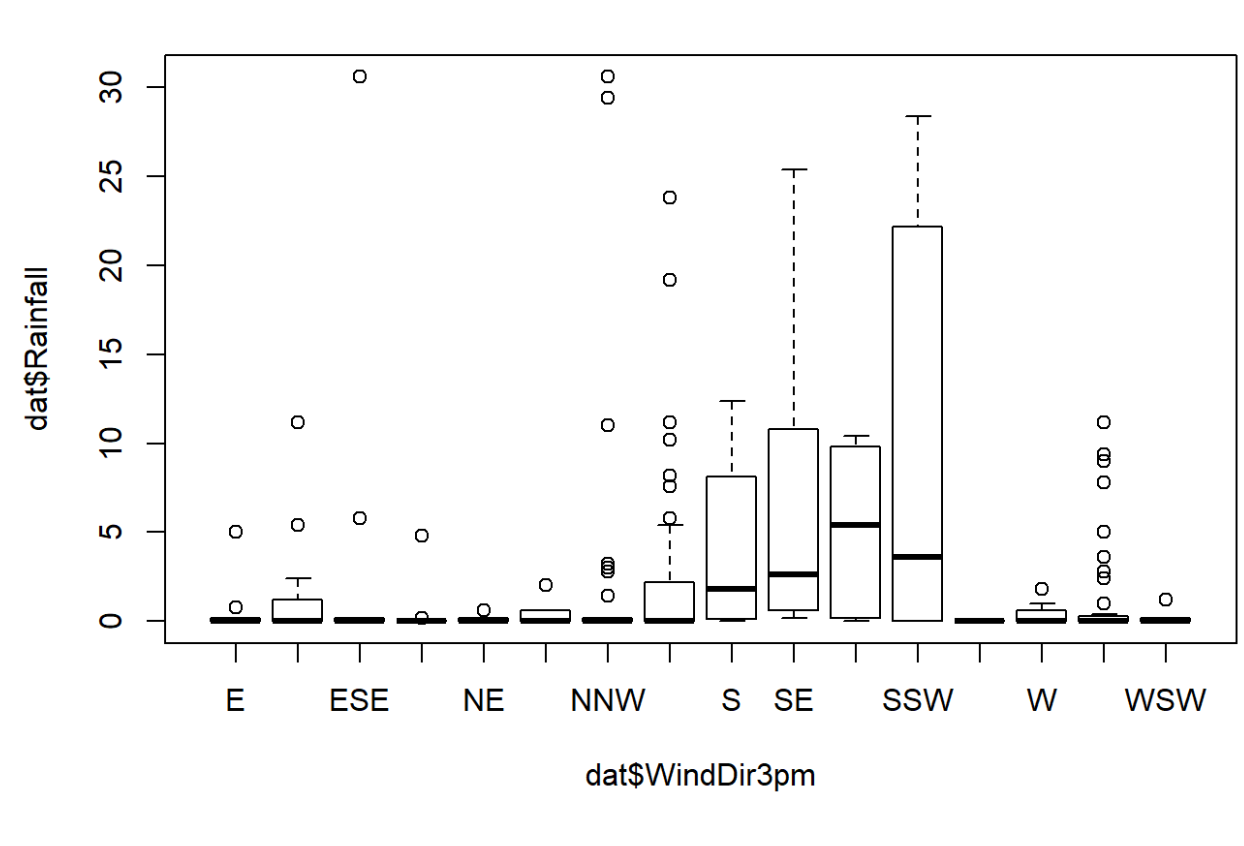
#``` {r} # Fisherman test #fisher.test(dat$WindDir3pm, dat$RainTomorrow) #fisher.test(dat$WindSpeed9am, dat$RainTomorrow) #fisher.test(dat$Cloud9am, dat$RainTomorrow)

#```

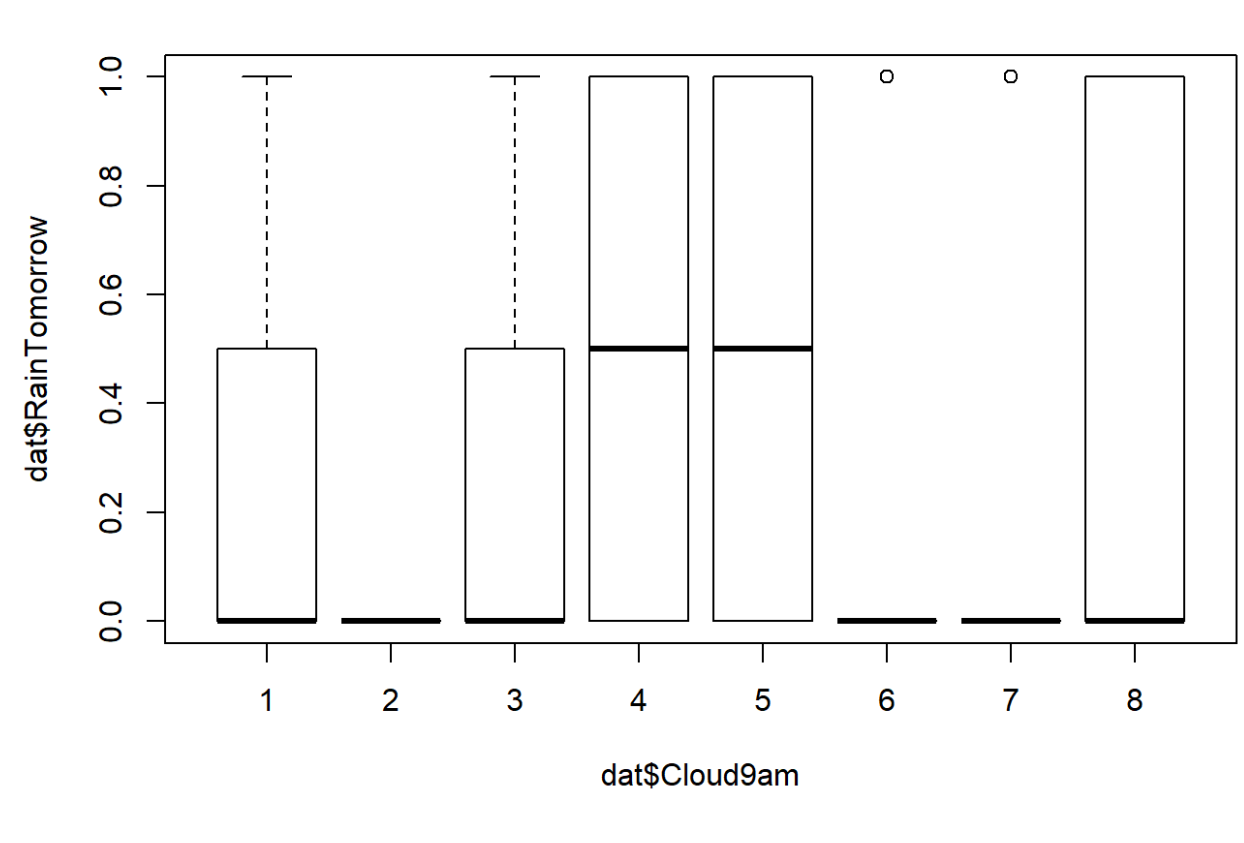
```
# boxplot
boxplot( dat$RainTomorrow~dat$WindDir9am)
```



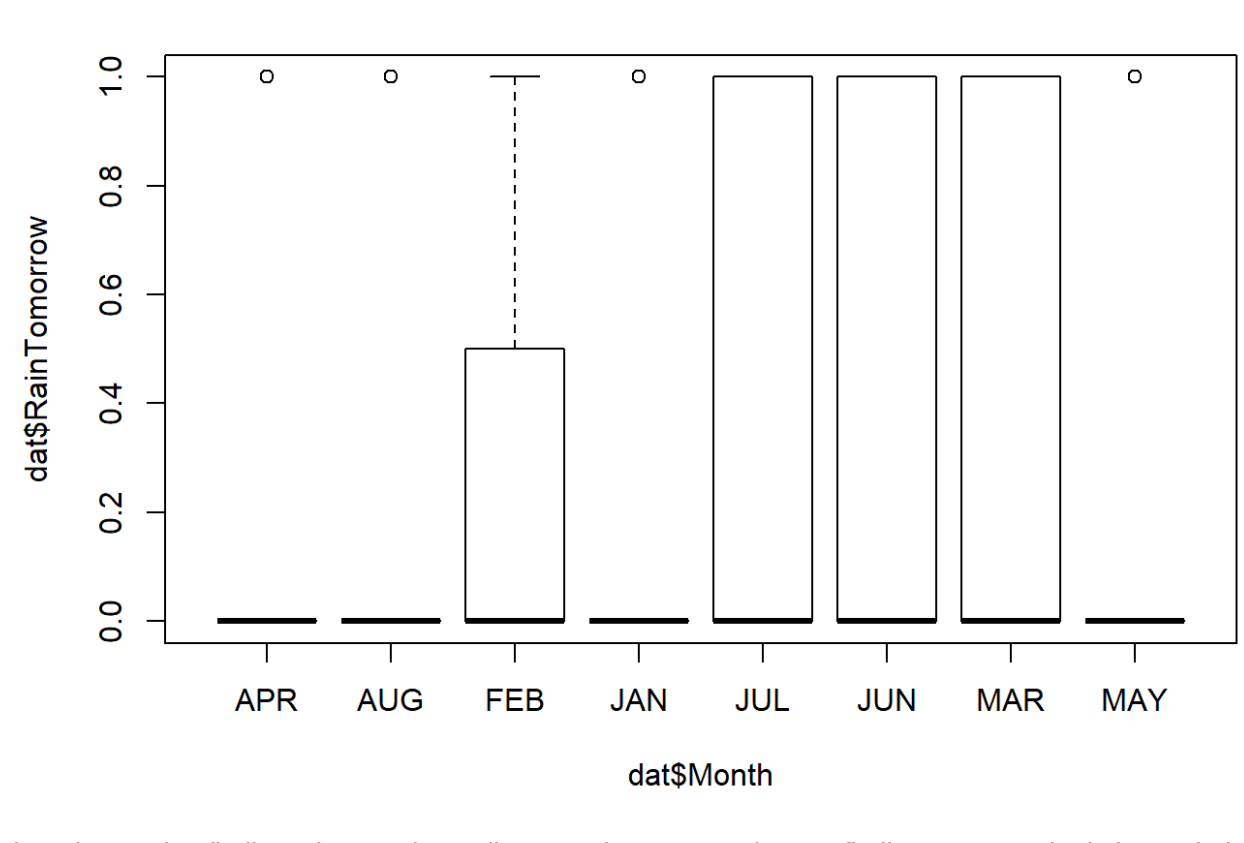

```
boxplot(dat$Rainfall~dat$WindDir3pm)
```




```
boxplot(dat$RainTomorrow~dat$Cloud9am)
```




```
boxplot(dat$RainTomorrow~dat$Month)
```



The deliverable consists of a l ist of t hree interesting findings that you have discovered. For example, one finding one can obtain by exploring t he association between month and RainTomorrow is that the raining likelihood varies with months significantly (p-value = 0.05997 based on Fisher's exact test and test size of 0.1) and April has the least rainy days (1 days with RainTomorrow = 1).

```
tab <- table(dat$Month, dat$RainTomorrow, useNA="no"); tab
```

```
##
##        0  1
##   APR 29  1
##   AUG 24  6
##   FEB 21  7
##   JAN 24  7
##   JUL 22  9
##   JUN 20 10
##   MAR 22  9
##   MAY 27  4
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 11.879, df = 7, p-value = 0.1046
```

```
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 = 0.06197
## alternative hypothesis: two.sided
```

Finally, remember to upload a single PDF file produced with R Markdown containing all code, (relevant) output and plots, accompanying comments as well as summary of your findings and conclusions.