R Coursework 3

Name: Congye Wang Student ID: 35427962

Oct 31th, 2020

```
# Import Packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

Question 1

```
# Load Dataset
df_1 <- read.csv("Australia_severe_storms_1975-2015.csv", header = T)
# Print the dimensions of data_1
nrow(df_1)
```

```
## [1] 14457
```

```
ncol(df_1)
```

```
## [1] 14
```

```
dim(df_1)
```

```
## [1] 14457    14
```

As a result, the number of the data_1 frame's columns is 14, and that of rows is 14457.

Question 2

```
# Clean dataset
df_2 <- df_1 %>%
  select(-ID) %>%
  filter(Database != "Waterspout")
# Print the dimensions of data_2
nrow(df_2)
```

```
## [1] 14417
```

```
ncol(df_2)
```

```
## [1] 13
```

```
dim(df_2)
```

```
## [1] 14417    13
```

```
# Print 6th rows
df_2 %>%
  select(seq(1, length(names(df_2)) - 6)) %>%
  head()
```

```
##   Event.ID Database        Date.Time   Nearest.town State Latitude Longitude
## 1    20812     Wind 23/11/1975 07:00         SYDNEY   NSW -33.8834  151.2167
## 2    20813  Tornado 02/12/1975 14:00         BARHAM   NSW -35.6333  144.1333
## 3    20814     Wind 09/01/1976 08:50 COFF'S HARBOUR   NSW -30.3167  153.1167
## 4    20815     Hail 16/02/1976 14:00      BANKSTOWN   NSW -33.8834  151.2167
## 5    20816     Rain 25/10/1976 14:00          BOOMI   NSW -28.4333  152.6167
## 6    20817     Hail 08/11/1976 14:00          YOUNG   NSW -34.3167  148.3000
```

As a result, the number of the data_2 frame's columns is 13, and that of rows is 14417.

Question 3

```
map_tz <- function(s, t) {

  if (str_detect(t, "[B,b][R,r][O,o][K,k][E,e][N,n].+[H,h][I,i][L,l]{2}") == TRUE) {

    z <- "Australia/Broken_Hill"

  } else {

    z <- switch (s,
      QLD = "Australia/Queensland",
      NSW = "Australia/NSW",
      VIC = "Australia/Victoria",
      SA = "Australia/South",
      WA = "Australia/West",
      TAS = "Australia/Tasmania",
      NT = "Australia/North",
      ACT = "Australia/ACT"
    )

  }
```

```
}

df_2$tz <- mapply(map_tz, s = df_2$State, t = df_2$Nearest.town)
```

Question 4

```
dt_Australia <- mapply(as_datetime, x = df_2$Date.Time, format = '%d/%m/%Y %H:%M',
                       tz = df_2$tz)
df_2$date.time.utc <- as.POSIXct(dt_Australia, tz = "UTC", origin = "1970-01-01")

df_2 %>%
  select(-c(Comments, X, X.1, X.2, X.3, X.4)) %>%
  head()
```

```
##   Event.ID Database        Date.Time   Nearest.town State Latitude Longitude
## 1    20812     Wind 23/11/1975 07:00         SYDNEY   NSW -33.8834  151.2167
## 2    20813  Tornado 02/12/1975 14:00         BARHAM   NSW -35.6333  144.1333
## 3    20814     Wind 09/01/1976 08:50 COFF'S HARBOUR   NSW -30.3167  153.1167
## 4    20815     Hail 16/02/1976 14:00      BANKSTOWN   NSW -33.8834  151.2167
## 5    20816     Rain 25/10/1976 14:00          BOOMI   NSW -28.4333  152.6167
## 6    20817     Hail 08/11/1976 14:00          YOUNG   NSW -34.3167  148.3000
##              tz       date.time.utc
## 1 Australia/NSW 1975-11-22 20:00:00
## 2 Australia/NSW 1975-12-02 03:00:00
## 3 Australia/NSW 1976-01-08 21:50:00
## 4 Australia/NSW 1976-02-16 03:00:00
## 5 Australia/NSW 1976-10-25 04:00:00
## 6 Australia/NSW 1976-11-08 03:00:00
```

Question 5

```
df_5 <- df_2
dt_my <- str_extract(df_5$Date.Time, '\\d{2}/\\d{4}')
df_5$year <- str_extract(dt_my, '\\d{4}')
df_5$month <- str_extract(dt_my, '^\\d{2}')
df_5 %>%
  select(-c(Comments, X, X.1, X.2, X.3, X.4)) %>%
  head()
```

```
##   Event.ID Database        Date.Time   Nearest.town State Latitude Longitude
## 1    20812     Wind 23/11/1975 07:00         SYDNEY   NSW -33.8834  151.2167
## 2    20813  Tornado 02/12/1975 14:00         BARHAM   NSW -35.6333  144.1333
## 3    20814     Wind 09/01/1976 08:50 COFF'S HARBOUR   NSW -30.3167  153.1167
## 4    20815     Hail 16/02/1976 14:00      BANKSTOWN   NSW -33.8834  151.2167
## 5    20816     Rain 25/10/1976 14:00          BOOMI   NSW -28.4333  152.6167
## 6    20817     Hail 08/11/1976 14:00          YOUNG   NSW -34.3167  148.3000
##              tz       date.time.utc year month
## 1 Australia/NSW 1975-11-22 20:00:00 1975    11
## 2 Australia/NSW 1975-12-02 03:00:00 1975    12
## 3 Australia/NSW 1976-01-08 21:50:00 1976    01
## 4 Australia/NSW 1976-02-16 03:00:00 1976    02
```

```
## 5 Australia/NSW 1976-10-25 04:00:00 1976    10
## 6 Australia/NSW 1976-11-08 03:00:00 1976    11
```

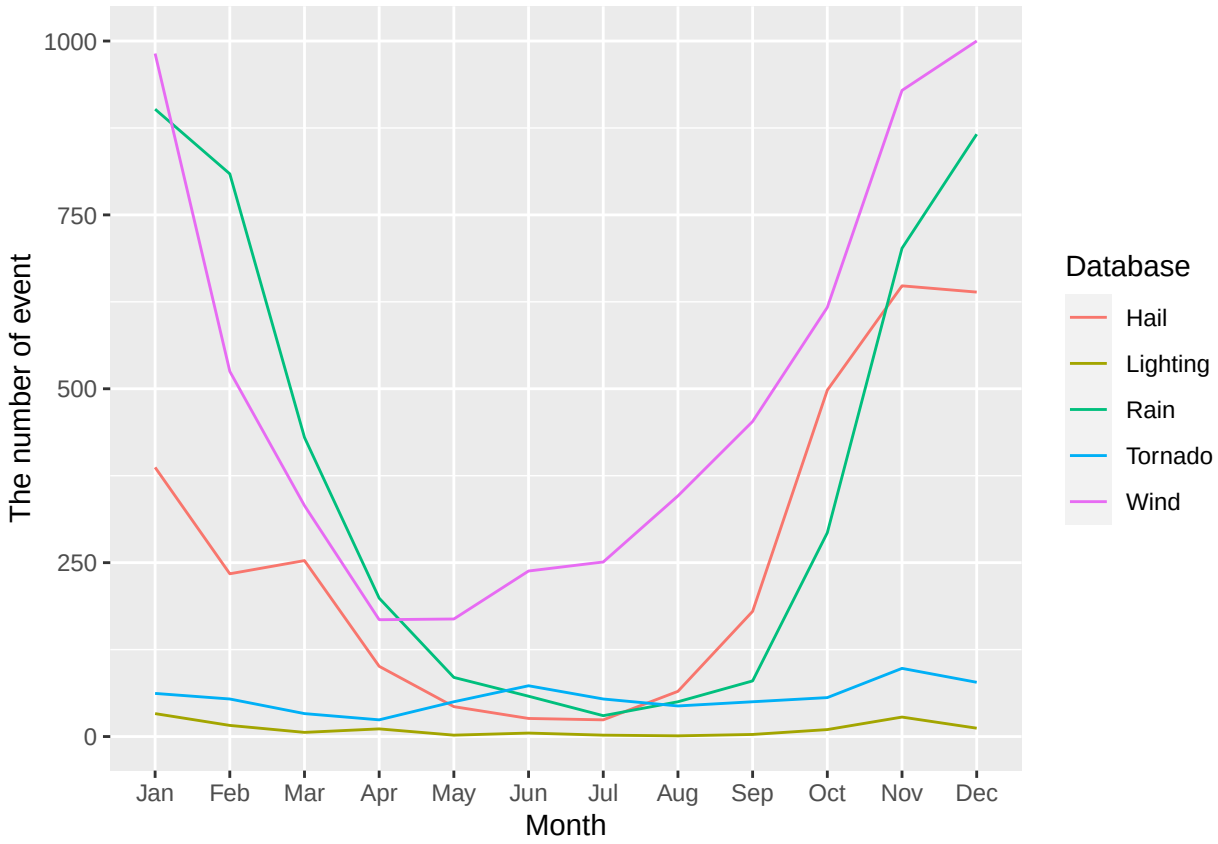
Question 6

i

```
df_6 <- count(df_5, month, Database)
```

ii

```
cmd <- c("month.abb[1]")
for (i in 2:12) {
    cmd <- paste(cmd, paste("month.abb[", i, "]", sep = ""), sep = ", ")
}

map_xlab <- function(i) {

  switch_cmd <- paste("z <- switch (i,", cmd, ")", sep = "")
  eval(parse(text = switch_cmd))

  return(z)

}

df_6$x_lab <- sapply(as.integer(df_6$month), map_xlab)

p <- ggplot(df_6, aes(x = reorder(x_lab,
                                  as.integer(month)
                                  ),
                      y = n,
                      group = Database,
                      color = Database)) +
  geom_line() +
  xlab("Month") +
  ylab("The number of event")
p
```



Question 7

i

```
df_7_1 <- df_5 %>%
  mutate(All.comments = paste(Comments, X, X.1, X.2, X.3, X.4))
```

ii

```
df_7_2 <- df_7_1 %>%
  select(Event.ID, Database, State, All.comments, year)
```

iii

```
print(sapply(df_7_2, class))
```

```
##     Event.ID     Database        State All.comments         year
##    "integer"  "character"  "character"  "character"  "character"
```

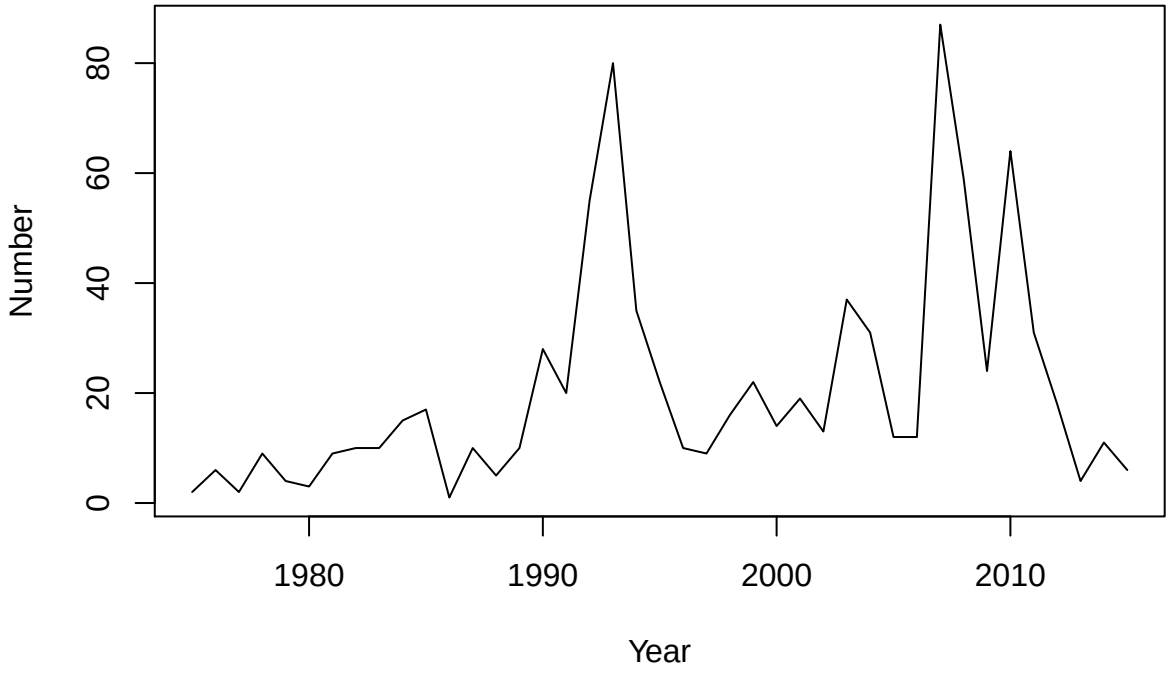
Question 8

i

```
df_8_1 <- df_7_2
df_8_1$flash_flood_indicator <- str_detect(df_7_2$All.comments,
'([F,f][L,l][A,a][S,s][H,h].+[F,f][L,l][O.o]{2}[D,d])|
([F,f][L,l][O.o]{2}[D,d].+[F,f][L,l][A,a][S,s][H,h])')
```

ii

```
df_8_2 <- df_8_1 %>%
  filter(flash_flood_indicator == TRUE) %>%
  count(year)
plot(df_8_2$year, df_8_2$n, type = "l", xlab = "Year", ylab = "Number")
```



Question 9

i

```
df_9_1 <- df_8_1

norm_speeds <- str_extract(df_8_1$All.comments,
"(\\d{1,3}\\s?[K,k][N,n][O,o][T,t][S,s])|(\\d{1,3}\\s?[K,k][T,t][S,s])|
(\\d{1,3}\\s?[K,k][T,t])|(\\d{1,3}\\s?[K,k][M,m]/[H,h])")

# The given wind speed is the range value, I would like to average the range.
# E.g. "80-90 km/h" in range_speeds[complete.cases(range_speeds)][1]
range_speeds <- str_extract(df_8_1$All.comments,
"(\\d{1,3}-\\d{1,3}\\s?[K,k][N,n][O,o][T,t][S,s])|(\\d{1,3}-\\d{1,3}\\s?[K,k][T,t][S,s])|
(\\d{1,3}-\\d{1,3}\\s?[K,k][T,t])|(\\d{1,3}-\\d{1,3}\\s?[K,k][M,m]/[H,h])")

lrange_speeds <- str_detect(df_8_1$All.comments,
"(\\d{1,3}-\\d{1,3}\\s?[K,k][N,n][O,o][T,t][S,s])|(\\d{1,3}-\\d{1,3}\\s?[K,k][T,t][S,s])|
(\\d{1,3}-\\d{1,3}\\s?[K,k][T,t])|(\\d{1,3}-\\d{1,3}\\s?[K,k][M,m]/[H,h])")
```

```
lower_range_speeds <- sapply(range_speeds, str_extract, pattern = "^\\d{1,3}")
upper_range_speeds <- str_extract(df_8_1$All.comments,
"(\\d{1,3}exp1(?=-\\d{1,3}\\s?[K,k][N,n][O,o][T,t][S,s]))|
(\\d{1,3}exp1(?=-\\d{1,3}\\s?[K,k][T,t][S,s]))|
(\\d{1,3}(?=-\\d{1,3}\\s?[K,k][T,t]))|
(\\d{1,3}(?=-\\d{1,3}\\s?[K,k][M,m]/[H,h]))")

avg_range_speeds <- rep("", length(range_speeds))

for (i in 1:length(lower_range_speeds)) {

  if (lrange_speeds[i] == TRUE) {

    l <- as.integer(lower_range_speeds[i])
    u <- as.integer(upper_range_speeds[i])

    avg <- as.character(mean(c(l, u)))
    avg_range_speeds[i] <- avg

  } else {

    avg_range_speeds[i] <- NA

  }
}

for (i in which(complete.cases(range_speeds))) {
  norm_speeds[i] <- str_replace(norm_speeds[i], "\\d{1,3}", avg_range_speeds[i])
}

# The given wind speed is in the table string.
# E.g. "049" in table_speeds[complete.cases(table_speeds)][1]
p <- c()
for (i in month.abb) {
  cmd <- paste("((?<=", i, "-", "\\d{4}\\s{0,5}", ")", "\\d{1,3})|", sep = "")
  p <- paste(p, cmd, sep = "")
}
p <- str_sub(p, end = -2)
table_speeds <- str_extract(df_8_1$All.comments, p)
for (i in which(complete.cases(table_speeds))) {
  norm_speeds[i] <- paste(table_speeds[i], "kts")
}

df_9_1$speeds <- norm_speeds
```

ii

```
df_9_2 <- df_9_1[complete.cases(df_9_1[,7]),]
df_9_2$num <- as.integer(str_extract(df_9_2$speeds, '\\d{1,3}'))
df_9_2$unitage <- str_extract(df_9_2$speeds,
'([K,k][N,n][O,o][T,t][S,s])|([K,k][T,t][S,s])|
([K,k][T,t])|([K,k][M,m]/[H,h])')
```

```
knots <- rep(0, length(df_9_2$unitage))
for (i in 1:length(df_9_2$unitage)) {

  if (df_9_2$unitage[i] == "km/h") {

    knots[i] <- round(df_9_2$num[i] / 1.852)

  } else {
    knots[i] <- df_9_2$num[i]
  }

}
df_9_2$knots <- knots
```

iii

```
boxplot(knots ~ State, data = df_9_2, ylab = "Speed")
```