Assignment-2

Aditya Singh

2022-10-03

PART A

Question 1

Below is the Pokemon Dataset I used for miniposter. Lets import the dataset I made use of tiny-url to shorten the url.

```
#importing some useful libraries
library(ggplot2)
library(reshape2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fmsb)
df <- read.csv("https://tinyurl.com/2p9da7z3") #shorted url
head(df)
```

```
##   X.                  Name Type.1 Type.2 Total HP Attack Defense Sp..Atk
## 1  1             Bulbasaur  Grass Poison   318 45     49      49      65
## 2  2               Ivysaur  Grass Poison   405 60     62      63      80
## 3  3              Venusaur  Grass Poison   525 80     82      83     100
## 4  3 VenusaurMega Venusaur  Grass Poison   625 80    100     123     122
## 5  4            Charmander   Fire          309 39     52      43      60
## 6  5            Charmeleon   Fire          405 58     64      58      80
##   Sp..Def Speed Generation Legendary
## 1      65    45          1     False
## 2      80    60          1     False
## 3     100    80          1     False
## 4     120    80          1     False
## 5      50    65          1     False
## 6      65    80          1     False
```

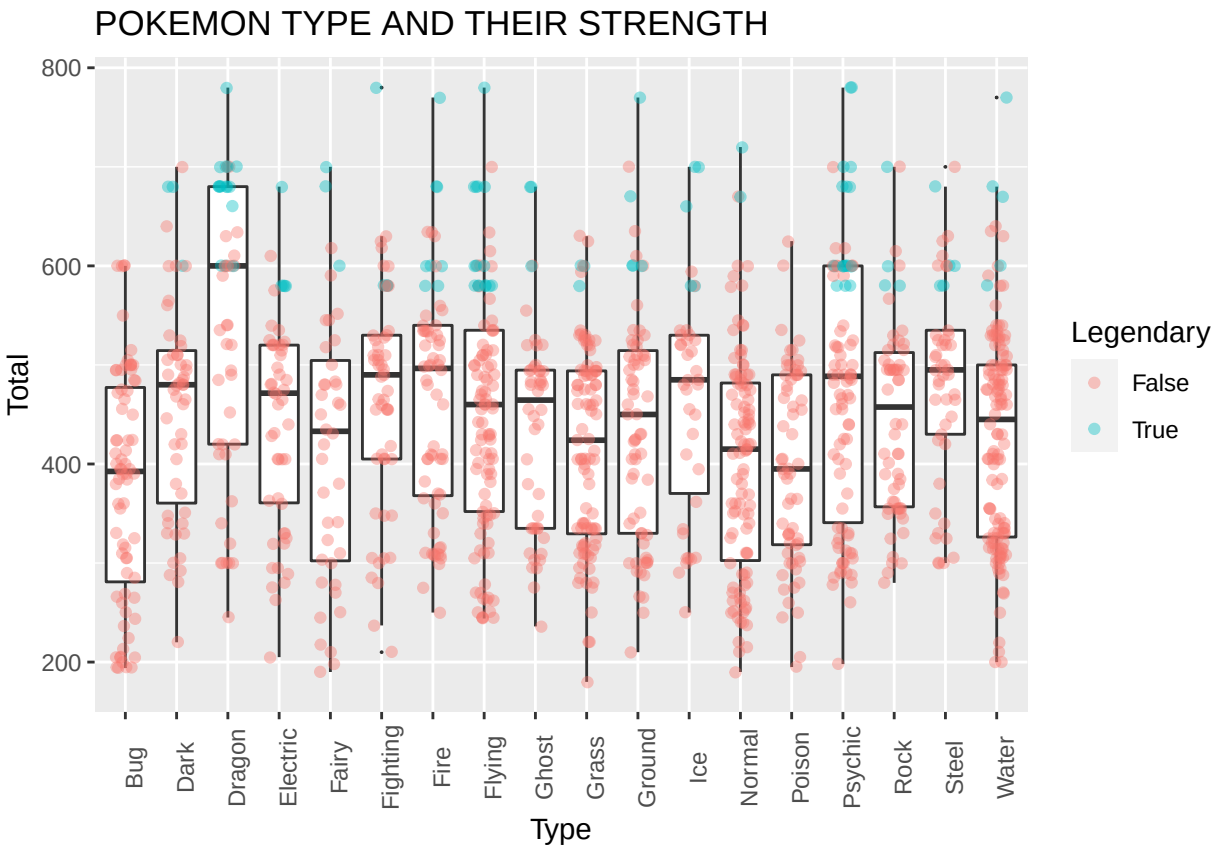
For a Dataset to be tidy in general following conditions should be satisfied.

1. Every column is a variable.
2. Every row is an observation.
3. Every cell is a single value.

The dataset was already tidy since it satisfies all three condition. I combined the Type 1 and Type 2 characteristic of Pokemons to plot them by their strength and see which pokemon type is stronger in general regardless of it being dual type pokemon. I used `melt` functionality from `dplyr` to achieve the desired result and plotted a `barplot` for visualization

Question 2

```
df2 <- melt(df,id.vars = setdiff(names(df),c('Type.1','Type.2')),value.name  = 'Type')
df2 <- df2[df2$Type != '',]
p <- ggplot(df2, aes(x=Type, y = Total)) + geom_boxplot(outlier.size=0) + geom_jitter(width=0.2,alpha=0
p <- p + theme(axis.text.x = element_text(angle = 90))
p <- p + labs(title = 'POKEMON TYPE AND THEIR STRENGTH')
p
```



Below is code I used to generate list of top 5 strongest Pokemons and plot their attributes in a spider chart.

```r
top_5 <- sort(df$Total,decreasing = TRUE)[1:5] #sorting pokemons as per their strength
strongest_pokemons = filter(df,df$Total %in% top_5) #selecting top 5 pokemons
strongest_pokemons
```

```
##    X.                  Name  Type.1   Type.2 Total  HP Attack Defense Sp..Atk
## 1 150   MewtwoMega Mewtwo X Psychic Fighting   780 106    190     100     154
## 2 150   MewtwoMega Mewtwo Y Psychic            780 106    150      70     194
## 3 382   KyogrePrimal Kyogre   Water            770 100    150      90     180
## 4 383 GroudonPrimal Groudon  Ground     Fire   770 100    180     160     150
## 5 384 RayquazaMega Rayquaza  Dragon   Flying   780 105    180     100     180
##   Sp..Def Speed Generation Legendary
## 1     100   130          1      True
## 2     120   140          1      True
## 3     160    90          3      True
## 4      90    90          3      True
## 5     100   115          3      True
```

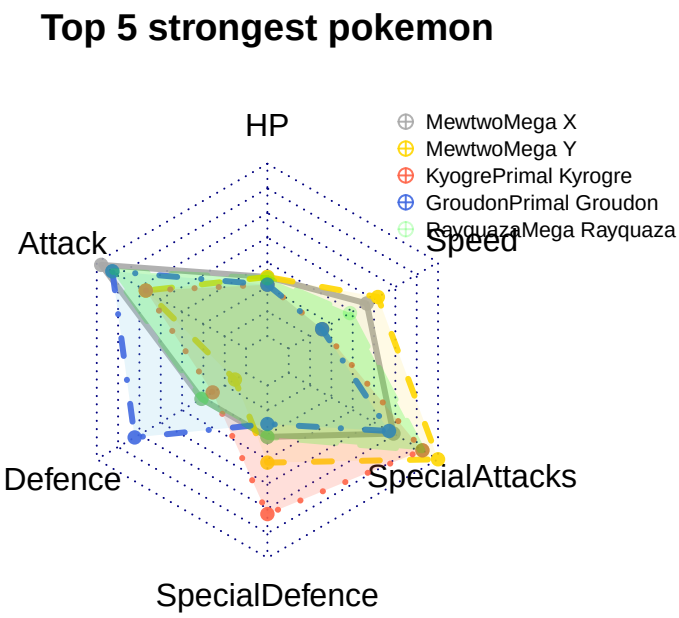
```r
#creating data frame for spider chart
spider <- data.frame(HP = c(19.4, 6.0, 10.6, 10.6, 10.0, 10.0,10.5),
                     Attack = c(19.4, 6.0, 19.0, 15.0,15.0,18.0,18.0),
                     Defence = c(19.4, 6.0, 10.0, 7.0,9.0,16.0,10.0),
                     SpecialDefence = c(19.4, 6.0, 10.0, 12.0,16.0,9.0,10.0),
                     SpecialAttacks = c(19.4, 6.0, 15.4, 19.4,18.0,15.0,18.0),
                     Speed = c(19.4, 6.0, 13.0, 14.0,9.0,9.0,11.5),
                     row.names = c("max", "min", "MewtwoMega X", "MewtwoMega Y",
                                   "KyogrePrimal Kyrogre", "GroudonPrimal Groudon","RayquazaMega Rayquaza

colors_fill <- c(scales::alpha("gray", 0.1),
                 scales::alpha("gold", 0.1),
                 scales::alpha("tomato", 0.2),
                 scales::alpha("skyblue", 0.2),
                 scales::alpha("green", 0.2))

# Define line colors
colors_line <- c(scales::alpha("darkgray", 0.9),
                 scales::alpha("gold", 0.9),
                 scales::alpha("tomato", 0.9),
                 scales::alpha("royalblue", 0.9),
                 scales::alpha("green", 0.2))

# Create plot
radarchart(spider,
           seg = 7,  # Number of axis segments
           title = "Top 5 strongest pokemon",
           pcol = colors_line,
           pfcol = colors_fill,
           plwd = 3)

legend(x=0.6,
       y=1.35,
       legend = rownames(spider[-c(1,2),]),
       bty = "n", col = colors_line, pch=10 , cex = 0.7, pt.cex = 1)
```



PART B

Question 3

```
library(tidyverse)
library(readr)
library(dplyr)
library(stringr)
data = read_tsv("26801-0001-Data.tsv") #downloaded data in my local working directory
```

```
## Rows: 6511 Columns: 76
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (4): SCL_NAME, SPORT_NAME, CONFNAME_14, D1_FB_CONF_14
## dbl (69): SCL_UNITID, SPORT_CODE, ACADEMIC_YEAR, SCL_DIV_14, SCL_SUB_14, SCL...
## lgl  (3): DATA_TAB_GENERALINFO, DATA_TAB_MULTIYRRATE, DATA_TAB_ANNUALRATE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

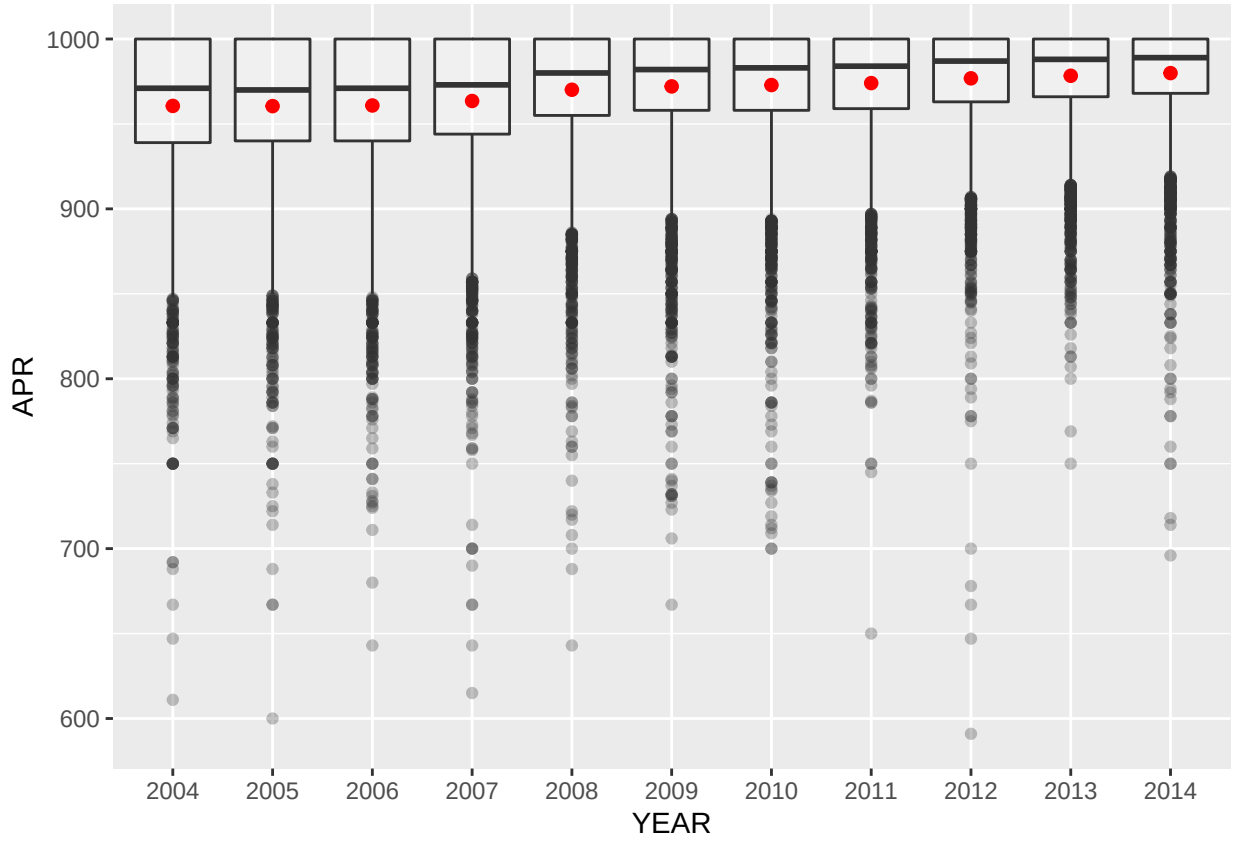
```
#tidying the data
data <- pivot_longer(data,cols = starts_with('APR'), names_to = 'YEAR',values_to = 'APR')
```

```
data$YEAR = str_sub(data$YEAR,start = 10 ,end =  13)

tidy_data <- subset(data,select = c('SCL_UNITID','SCL_NAME','SPORT_CODE','SPORT_NAME','YEAR','APR'))
tidy_data <- filter(tidy_data, tidy_data$APR != -99)



ggplot(data= tidy_data, mapping = aes(x=YEAR, APR),fill= class)+
geom_boxplot(alpha = 0.3)+
labs(y="APR","YEAR")+
stat_summary(fun=mean, geom="point", shape=20, size=3, color="red", fill="red")+
theme(legend.position="none") +
scale_fill_brewer(palette="BuPu")
```



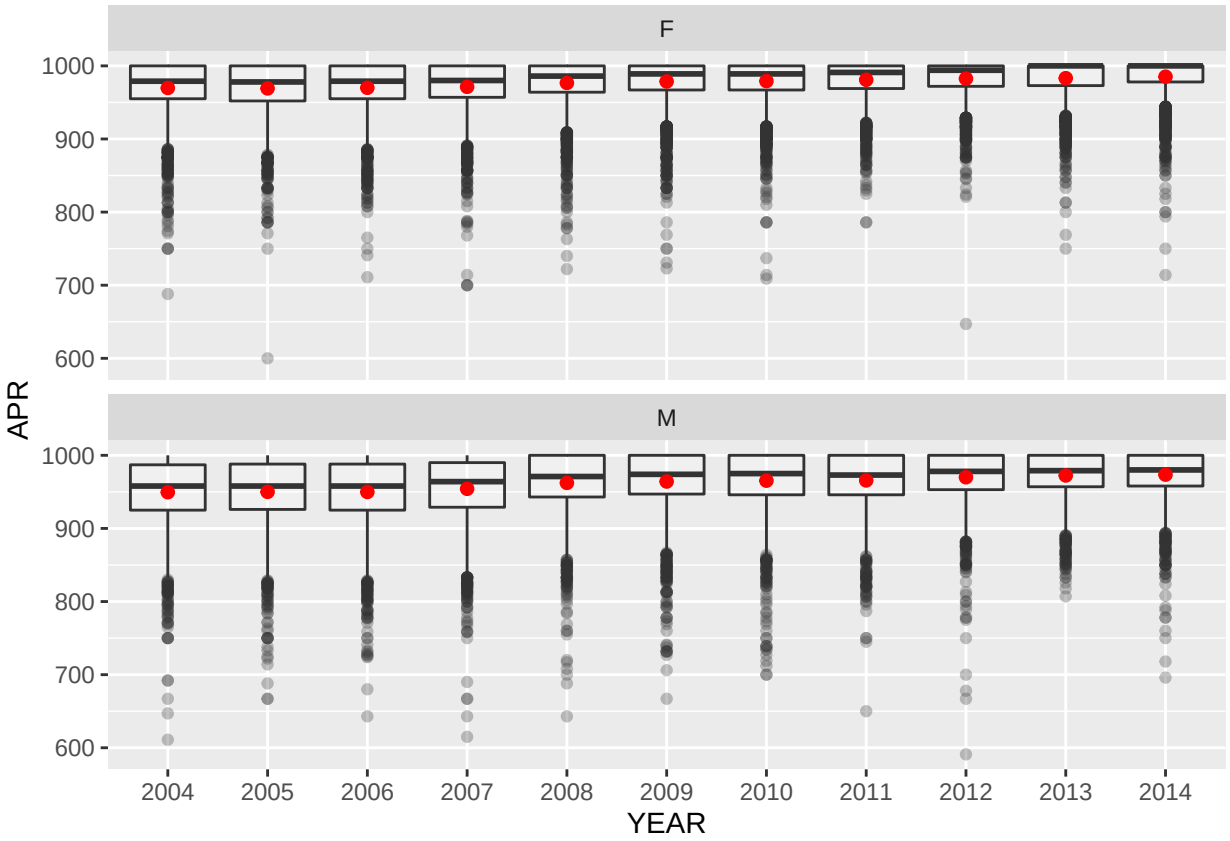
From the Barplot it is pretty evident that the APRs have increased over the years

Question 4

```
tidy_data <- filter(tidy_data, tidy_data$SPORT_CODE != 38)
tidy_data$GENDER = 0
tidy_data$GENDER <- ifelse(tidy_data$SPORT_CODE<=18, 'M','F')

ggplot(data= tidy_data, mapping = aes(x=YEAR, APR),fill= class)+
geom_boxplot(alpha = 0.3)+
```

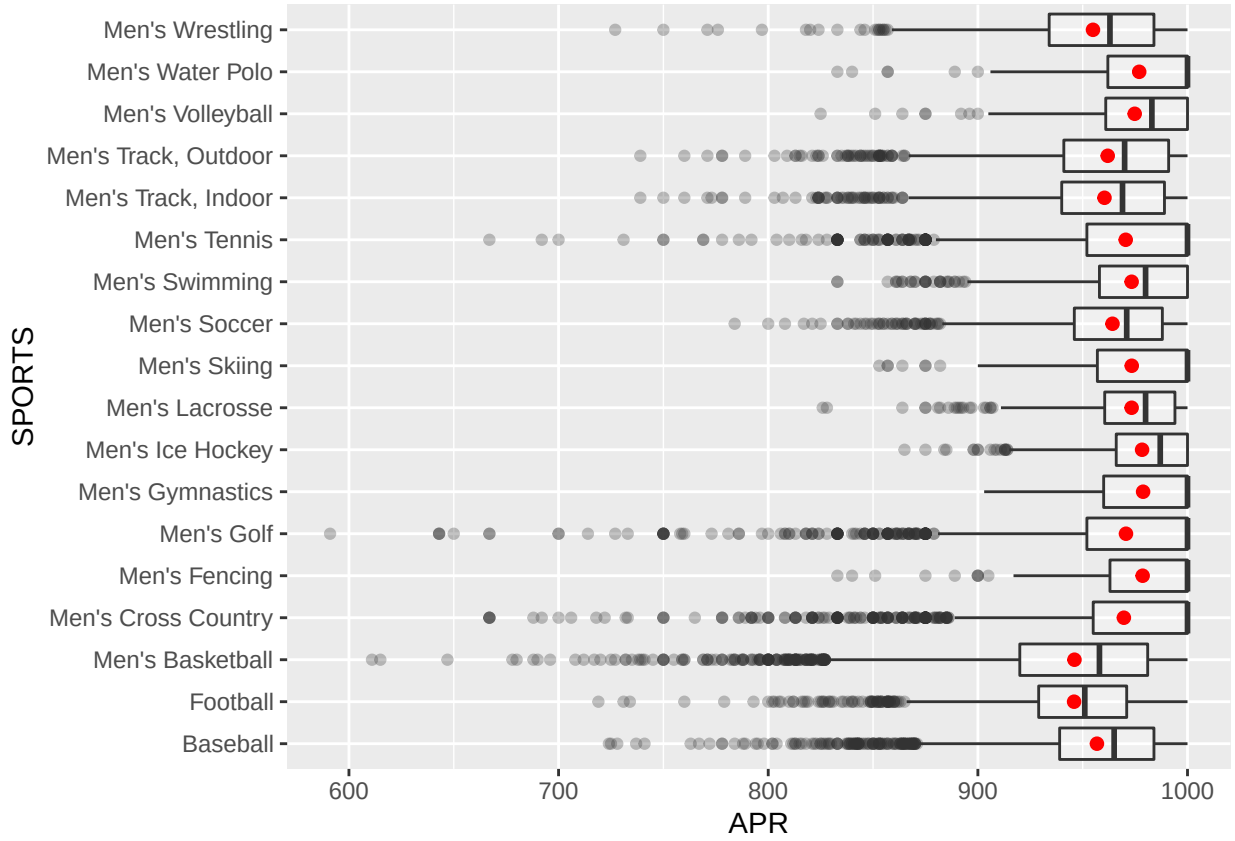
```
labs(y="APR","YEAR")+
stat_summary(fun=mean, geom="point", shape=20, size=3, color="red", fill="red")+
theme(legend.position="none") +
scale_fill_brewer(palette="BuPu")+facet_wrap(~GENDER,nrow=2)
```



Women Athletes in general have performed better then their male counterparts specially in 2012,2013 and 2014.

Question 5

```
tidy_data <- filter(tidy_data,tidy_data$GENDER == 'M')
ggplot(data=tidy_data, mapping = aes(x=SPORT_NAME, y = APR),fill = class)+
geom_boxplot(alpha = 0.3)+
labs(y="APR",x="SPORTS")+
stat_summary(fun=mean, geom="point", shape=20, size=3, color="red", fill="red")+
theme(legend.position="none") +
scale_fill_brewer(palette="BuPu")+coord_flip()
```



The following sports have higher APRs in case of Men, Water Polo, Tennis,Skiing, Gymnastics, Golf, Fencing, Cross Country

While Following Sports have lower APRs on Average, Wrestling, Outdoor and Indoor Track, Basketball, Football.