Student-Athlete performance analysis

M.G.Manjusha

2024-07-18

Importing libraries

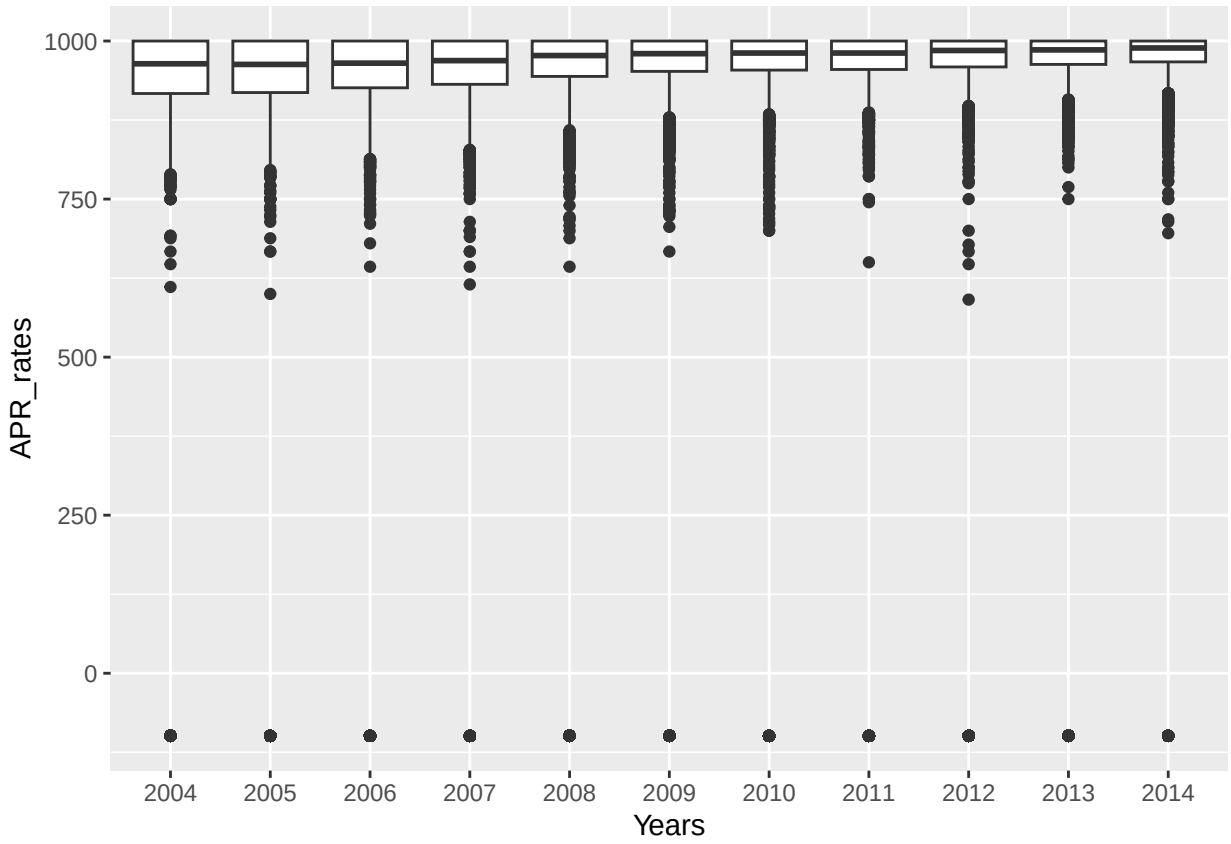
```
#importing libraries
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)
library(reprex)
library(stringr)
```

Loading and tidying data

```
path<-"/Users/mgmanjusha/Documents/NEU/Sem-1/IDMP/Hw-pratice/HW2/NCAA-D1-APR-2003-14/DS0001/26801-0001-
data<-read_tsv(path)
tidy_data<-select(data,c(SCL_UNITID,SCL_NAME,SPORT_CODE,SPORT_NAME,ACADEMIC_YEAR,ends_with("1000")))
#Creating a data frame with Years and Apr_rate values corresponding to them
df <- tidy_data |> select(ends_with("1000")) |>
  pivot_longer(cols = ends_with("1000"),
                   names_to = "Years",
                   values_to = "APR_rates")

df$Years = gsub("APR_RATE_","",df$Years)
df$Years = gsub("_1000","",df$Years)

ggplot(df, aes(x=Years,y=APR_rates))+
  geom_boxplot()
```

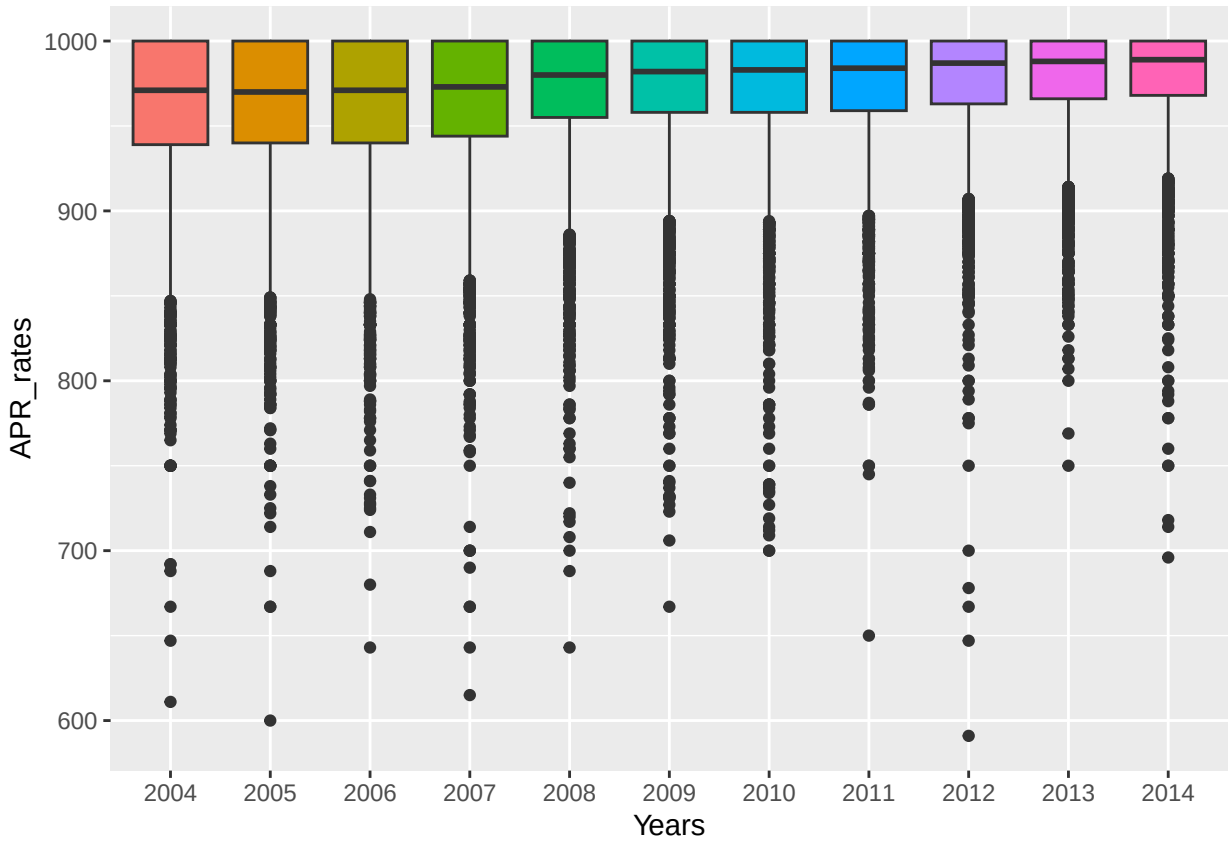


There are some negative values as outliers in the APR rates which need to be removed.

Cleaning the data

```
clean_df <- df |>
       filter(!(APR_rates<0))

#running boxplot again
ggplot(clean_df, aes(x=Years,y=APR_rates))+
  geom_boxplot(aes(fill=Years))+
  theme(legend.position = "none")
```



Observations from the data:

- As we can see from the plot, the APR(Academic Progress Rates) is gradually increasing throughout the years.
- There is no drastic change from 2004-2014. We can observe that from the medians plotted.
- There is a slight increase of median every year but, we see a noticeable change in every 4 years.
- We notice this change in the years of 2008 and 2012.
- The median APRs from 2004 to 2007 are similar and it has a slight increase in 2008 and remains similar till 2012 and increases in 2013 again. From this we can say that there is a noticeable change in APRs every 4 years.

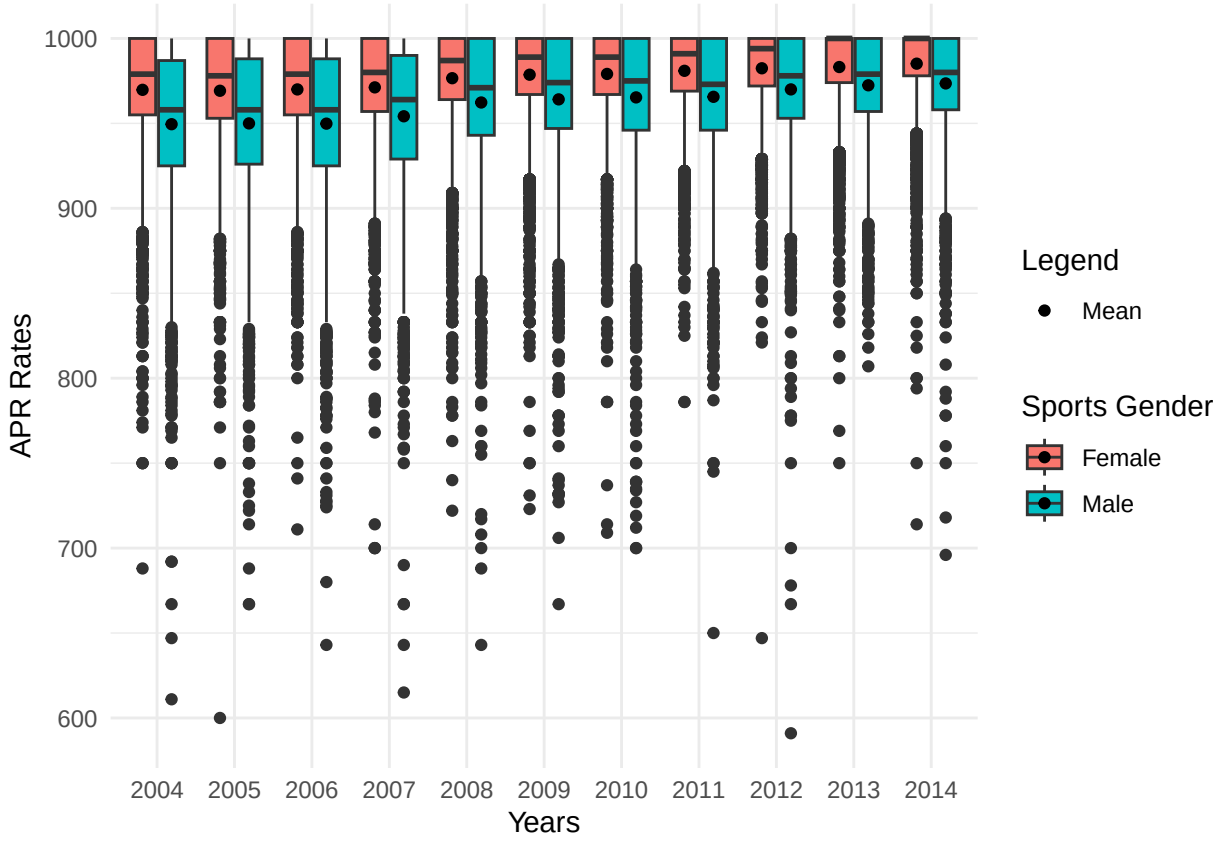
Analyzing data by comparing average APR broken down by Gender

```
tidy_data_gender <- mutate(tidy_data,
                           Sports_Gender = ifelse(SPORT_CODE <= 18, "Male", "Female"))

df <- tidy_data_gender|>
  select(Sports_Gender,ends_with("1000")) |>
  pivot_longer(cols = ends_with("1000"),
               names_to = "Years",
               values_to = "APR_rates")
clean_df <- df |>
  filter(!(APR_rates<0))

clean_df$Years = gsub("APR_RATE_","",clean_df$Years)
clean_df$Years = gsub("_1000","",clean_df$Years)
```

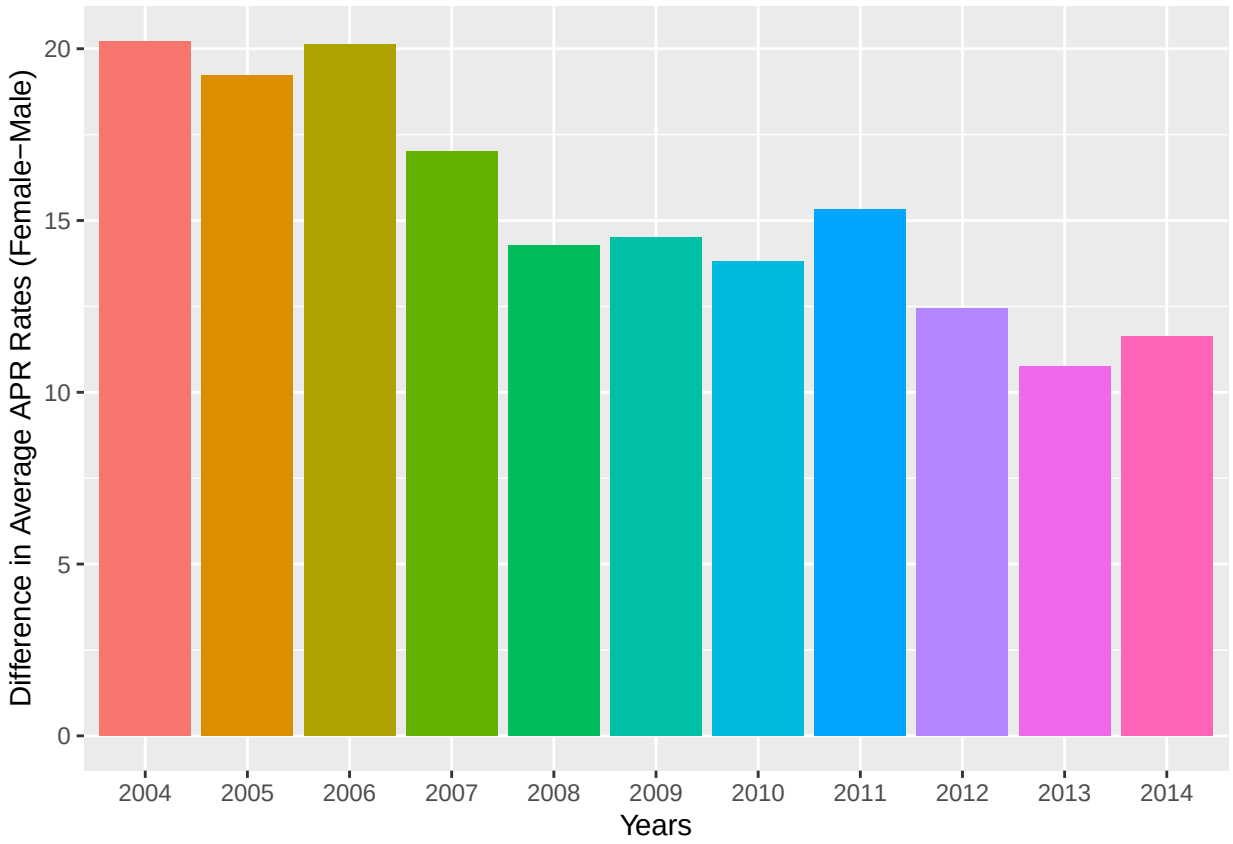
```r
ggplot(clean_df, aes(x=Years,y=APR_rates,fill=Sports_Gender))+
  geom_boxplot()+
  stat_summary(aes(group=Sports_Gender,shape="Mean"),fun=mean,geom="point",size=1.5,
               position=position_dodge(width=0.75))+
  scale_shape_manual(name = "Legend", values = c("Mean" = 19)) +
  labs(fill = "Sports Gender", shape = "Mean") +
  ylab("APR Rates")+
  theme_minimal()
```



The Black circles on the boxplot represent the averages of APR rates. From the graph, we can see that the average APR of Women is more than average APR of Men each year from 2004 to 2014.
We can see the difference between Average Female APR and average male APR from 2004 to 2014 by representing it in a graph as well.

```r
avg_apr <- clean_df |>
  group_by(Years, Sports_Gender) |>
  summarise(avg_apr = mean(APR_rates)) |>
  spread(key = Sports_Gender, value = avg_apr) |>
  mutate(diff = Female-Male)

# Plotting the difference between avg APR rates of male and female each year
ggplot(avg_apr, aes(x = Years, y = diff)) +
  labs(x = "Years", y = "Difference in Average APR Rates (Female-Male)") +
  geom_col(aes(fill=Years))+
  theme(legend.position = "none")
```



A point that can be noted from the graph is that : The difference between the average APRs of Male and Female sports keep decreasing from 2004 to 2014.

Visualizing the distribution of APR for both men's and women's teams for each sport

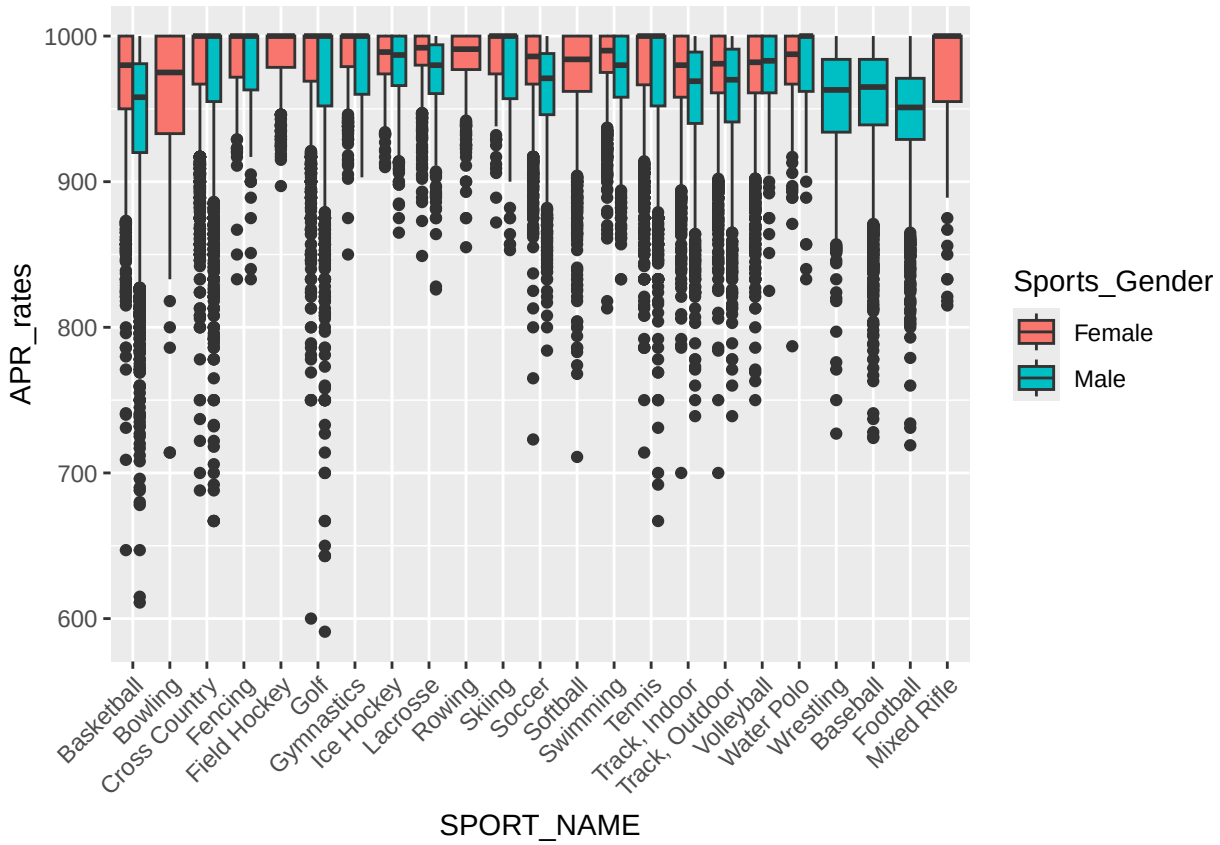
```
df <- tidy_data_gender|>
  select(Sports_Gender,ends_with("1000"),SPORT_NAME) |>
  pivot_longer(cols = ends_with("1000"),
               names_to = "Years",
               values_to = "APR_rates")

clean_df <- df |>
  filter(!(APR_rates<0))

clean_df$Years = gsub("APR_RATE_","",clean_df$Years)
clean_df$Years = gsub("_1000","",clean_df$Years)


clean_df$SPORT_NAME =  str_remove(clean_df$SPORT_NAME,"Women's")
clean_df$SPORT_NAME =  str_remove(clean_df$SPORT_NAME,"Men's")

ggplot(clean_df, aes(x=SPORT_NAME,y=APR_rates))+
  geom_boxplot(aes(fill=Sports_Gender),position = "dodge")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



We can see from the plot that some sports were played only by a single gender. Bowling, field hockey, soft ball, and mixed rifle was played only by women and not men. whereas, Wrestling, baseball, and football were only played by Men and not women. The sports having similar APRs of Men and Women can be noted by observing the median of the APRs:

- Cross country
- Fencing
- Golf
- Gymnastics
- Ice hockey
- Skiing
- Tennis
- Volleyball