Practice Set 2 (PART - A)

Problems 1–2 correspond to the “Miniposter” assignment on Canvas, and ask you to provide the code that

you used to import, tidy, and visualize the dataset that you used for your “Miniposter”

Problem 1

Find a dataset that is personally interesting to you. It may be a publicly-available dataset, or dataset for which you have permission to use and share results. There are many places online to find publicly-available dataset, and simply searching Google for your preferred topic plus “public dataset” may provide many hits.

This should be the same dataset that you use for the “Miniposter” assignment. It does not have to be the same dataset you will use for your team project later in the semester.Import the dataset into R, tidy the dataset (if necessary), and print the first several lines of the dataset. Describe the dataset and its variables. Comment on whether you had to tidy the dataset, and how you tidied the data (if you did).

Values in the dataset and the explaining what was involved for this visualization

This is a dataset of over 16000 entries.

Name- Name of the game Released

Year of Release- Year on which the respective game was released

Global_Sales- Global Sales of the specific game

NA_Sales, JP_Sales, EU_Sales, Other_Sales are the sales from various regions of the globe.

Platform- The platform in which the game was released.

Publisher/ Developer- The company which developed the game.

```
library("readr")
library("tidyr")
library("dplyr")
```

imputation can be done for the years as it was noticed

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

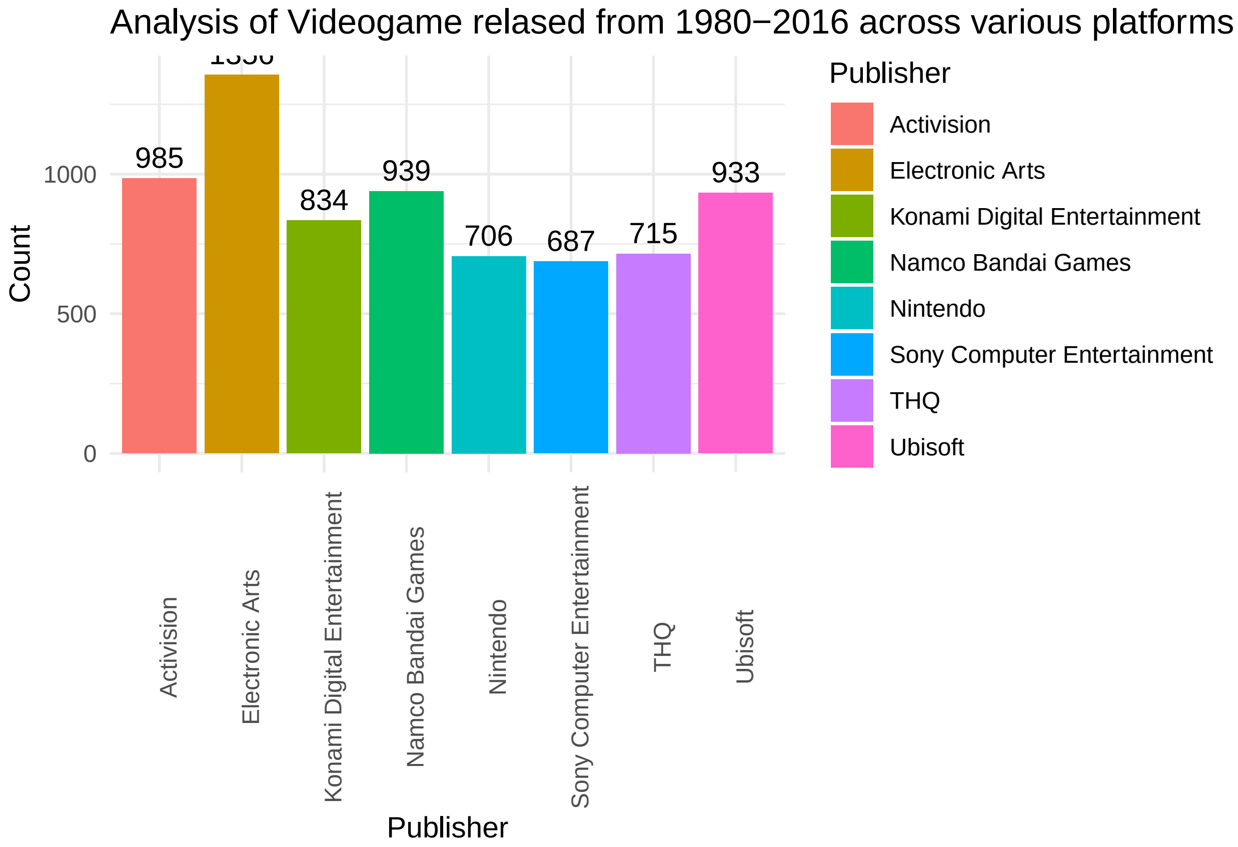
```r
library("ggplot2")


#Source of file:-
#kaggle.com/sidtwr/videogames-sales-dataset?select=Video_Games_Sales_as_at_22_Dec_2016.csv


#Variable Definition:

#video_games_sales_data stores the dataset as it is.
video_games_sales_data<-read_csv("D:/Documents/Sales_2016.csv")
```

```
## Rows: 16719 Columns: 16
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): Name, Platform, Year_of_Release, Genre, Publisher, Developer, Rating
## dbl (9): NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Critic_Sco...
```

```
##
## 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.
```

```r
vs<-video_games_sales_data %>%
  group_by(Publisher) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))

#For this visualization we are considering only publishers who released over 650 games
vs<-vs %>%
  filter(vs$Count>650)

#Plot the graph x =
ggplot(vs, aes(x=Publisher,y=Count, fill=Publisher)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = Count), vjust = -0.5) +
  labs(title="Analysis of Videogame relased from 1980-2016 across various platforms") +
  theme_minimal() +theme(axis.text.x = element_text(angle = 90))
```

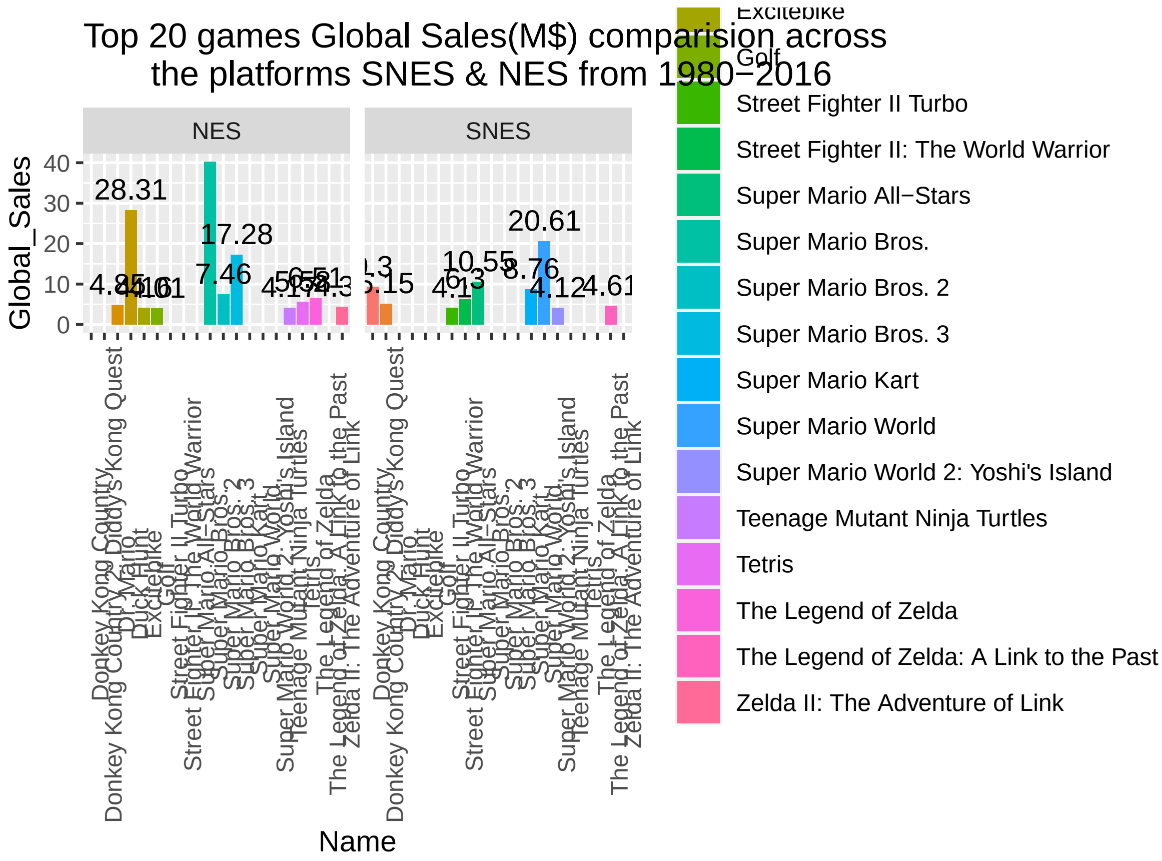


Among the top 8 Publishers from the list of game developers, Electronic Arts have developed the most number of games while Sony Computer Entertainment developed the least.

Problem 2

Use ggplot2 to create visualizations to identify interesting or unexpected relationships in the dataset. After performing your analysis, present your results by creating an attractive "Miniposter" slide using PowerPoint, Keynote, or similar program. Submit your slide to the "Miniposter" assignment on Canvas. In your homework solutions, reproduce the plots from your "Miniposter" figures, and provide your interpretations of them.

```
#vs_filtered has the entries of games from SNES and NES Platforms and are among top 20 most
#sold games
vs_filtered<-video_games_sales_data %>%
filter(Platform=="NES" | Platform=="SNES") %>%
arrange(desc(Global_Sales)) %>%
head(n=20)


#Graph plotted to showcase the difference in the total number of sales and comparing them
ggplot(vs_filtered, aes(x=Name,y=Global_Sales, fill=Name)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = Global_Sales), vjust = -0.5) +
  facet_wrap(~Platform)+
  labs(title="Top 20 games Global Sales(M$) comparision across
       the platforms SNES & NES from 1980-2016") +
  theme(axis.text.x = element_text(angle = 90))
```



On comparing the global sales(top 20) across the platforms of NES and SNES, Super Mario Bros is the highest sold game from the lot.

Part - B

Problems 3–5 use data on NCAA student-athlete academic performance. Download the data files from "NCAA-D1-APR-2003-14.zip" on Piazza. The files include the codebook and tab-delimited data for team-level

Academic Progress Rates (APRs) of Division I student-athletes from 2003-2014. A team's APR is calculated out of a maximum score of 1000 points, and takes into account a team's academic eligibility and retention, to derive an overall cohort rate of academic progress. Import the dataset into R using the readr package, making sure that any missing data codes are imported as NAs.

```
#Importing required packages
library("tidyr")
library("dplyr")
library("stringr")
library("ggplot2")
```

Problem 3

Create a tidy data frame that includes columns for: School ID, School name, Sport code, Sport name Year, APR. All other columns can be discarded. Use your tidied dataset to visualize the distributions of APRs over time. How does the distribution of APRs change year-to-year from 2004 to 2014?

```
ncaa_original<-read_tsv("D:/Documents/NCAA_Student_Dataset.tsv")
```

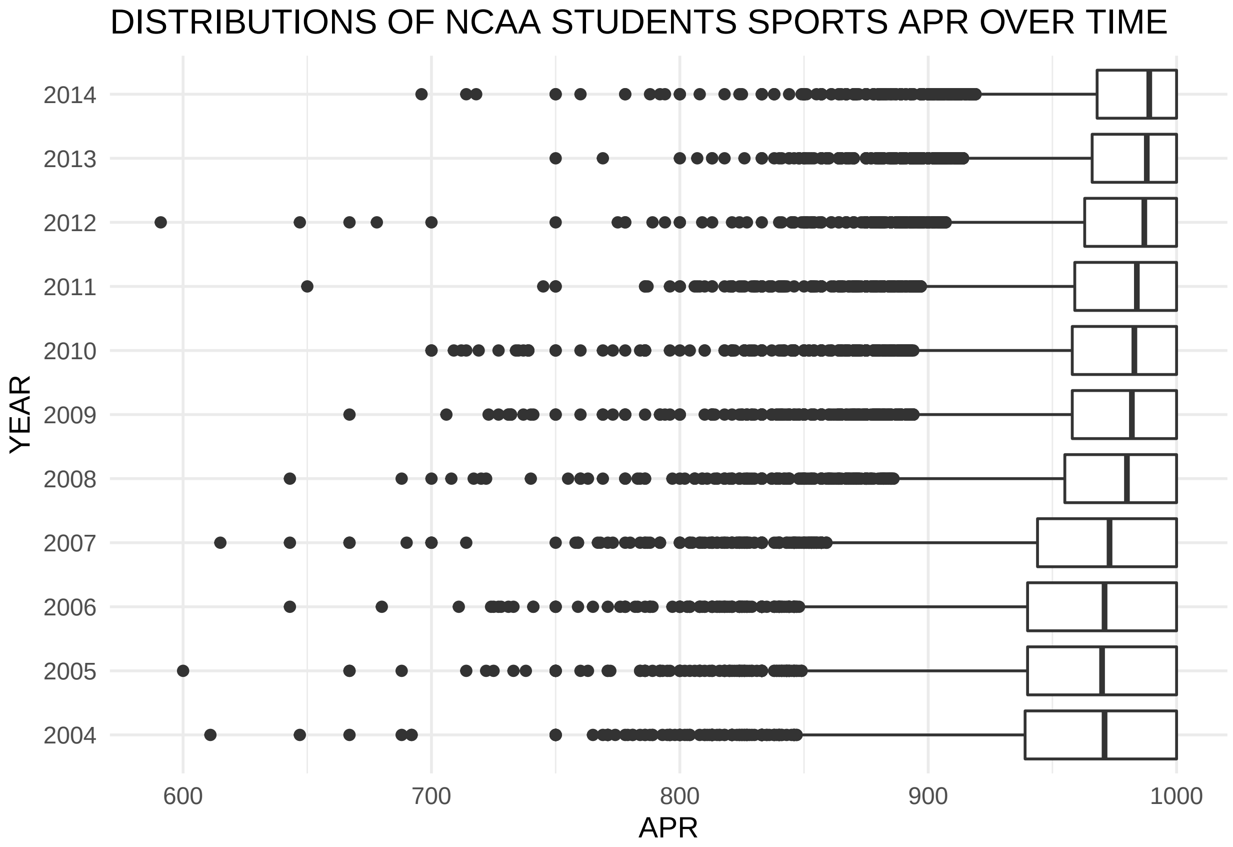
```
## Rows: 6511 Columns: 76

## -- 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.
```

```
#summary(ncaa_original)
ncaa_apr_cols<-names( select(ncaa_original,starts_with("APR")))
#Transposing Column Name and Values (APR and Year)
ncaa_tidy<-pivot_longer(ncaa_original,ncaa_apr_cols,names_to ="YEAR", values_to = "APR")
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(ncaa_apr_cols)` instead of `ncaa_apr_cols` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
#Discarding unnecessary fields
ncaa_tidy<-select(ncaa_tidy,SCL_UNITID,SCL_NAME,SPORT_CODE,SPORT_NAME,YEAR,APR)
#Subsetting the year from column name
ncaa_tidy$YEAR<-str_sub(ncaa_tidy$YEAR,10,13)
#Data tidying by removing -99's and replacing with NA
ncaa_tidy$APR[ncaa_tidy$APR==-99] <- NA
#Plot the graph x-axis= APR value, y-axis= Year
ggplot(ncaa_tidy, aes(x=APR, y=YEAR )) +
  geom_boxplot() +
  labs(title="DISTRIBUTIONS OF NCAA STUDENTS SPORTS APR OVER TIME",
       x="APR", y="YEAR") +theme_minimal()
```

```
## Warning: Removed 4732 rows containing non-finite values (stat_boxplot).
```



The interquartile range of the APR values reduced and the APR values have increased over the decade and the number of outliers seems to have reduced too.

Problem 4

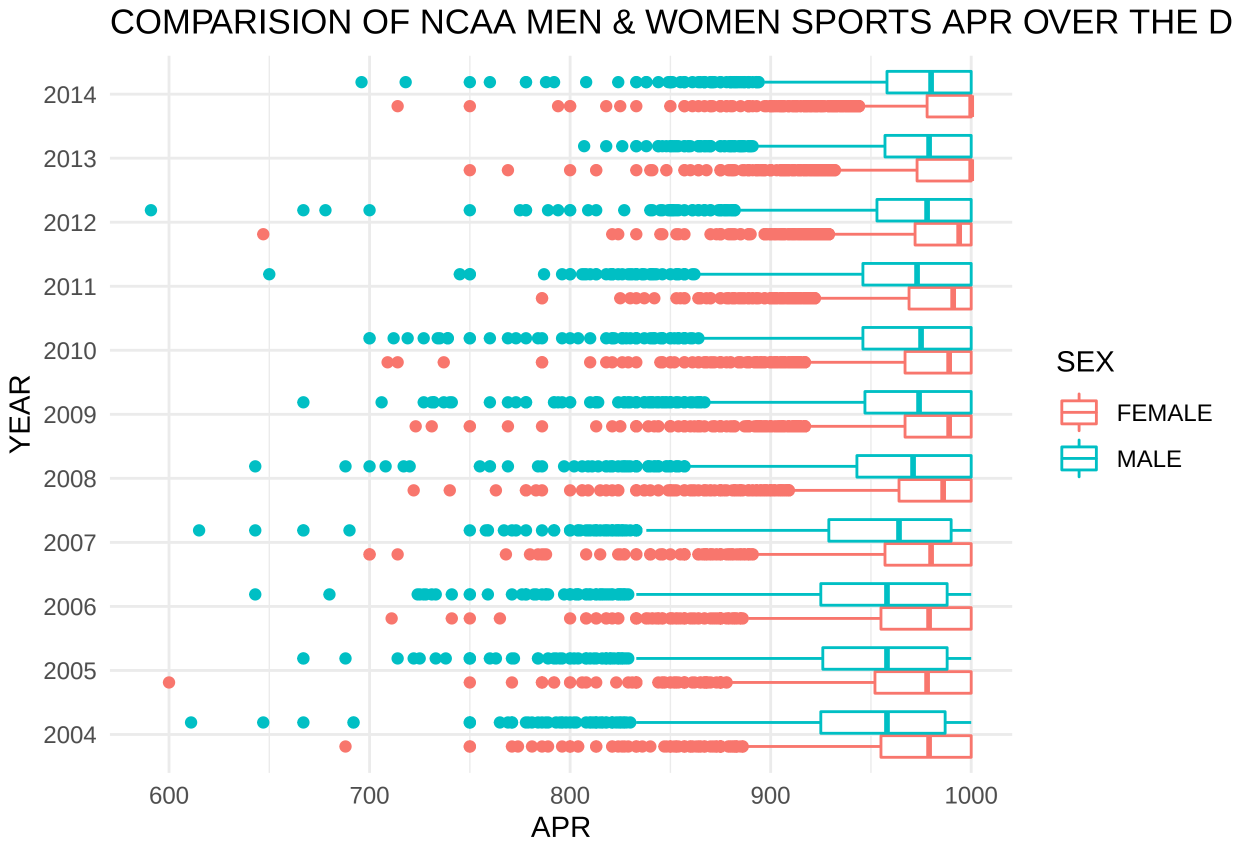
We would like to compare APRs between men's and women's sports. Transform your tidied dataset to remove mixed sports, and create a column indicating the gender division of each sport. (You may assume sport codes 1-18 are men's, and 19-37 are women's.) Visualize the distributions of APRs over time again, but broken down by gender division. How do the average APRs compare between men's and women's sports? Does this relationship hold true across each year from 2004 to 2014?

```
unique(ncaa_tidy$SPORT_CODE)
```

```
##  [1] 20 14  4  1 19 33  2 34 35 31  6 13 21 29 36 15 11 25  8  3 12 26 32 22 38
## [26] 18 37 16 28 17 24  7 30 10  9  5 23 27
```

```
#Assuming that the Sports code with 38 is a mixed gender game, it is not considered for
#analysis Mens Sports codes are from 1-18 while the 19-37 are for women
ncaa_tidy<-ncaa_tidy[ncaa_tidy$SPORT_CODE !=38,]

#Adding a new column Sex for simplicity in analysis
ncaa_tidy_modified <- ncaa_tidy %>%
  mutate(SEX = if_else(SPORT_CODE >= 1 & SPORT_CODE <=18, "MALE", "FEMALE"))

#Plot the graph x-axis= APR Values, y-axis= Year, color is changed as per the gender
ggplot(ncaa_tidy_modified, aes(x=APR, y=YEAR,color=SEX )) +
  geom_boxplot() +
  labs(title="COMPARISION OF NCAA MEN & WOMEN SPORTS APR OVER THE DECADE",
       x="APR", y="YEAR") +  theme_minimal()
```

```
## Warning: Removed 4696 rows containing non-finite values (stat_boxplot).
```



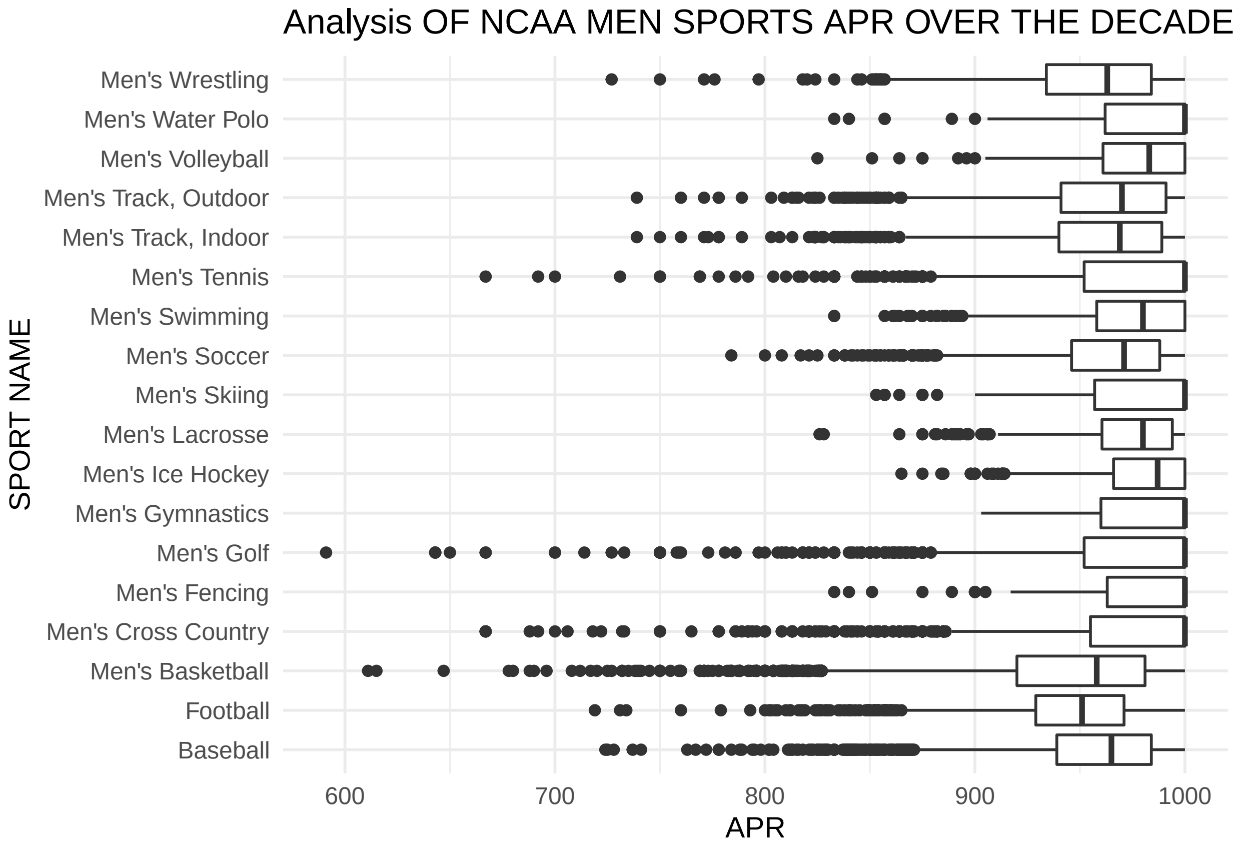
The average of the female APR is higher than that of the males when compared to the men's APR but the overall number of participants in each sport is more in men than women. This relationship holds throughout the decade. However, the outliers might impact the comparision as well.

Problem 5

We would like to further investigate the APRs of different men's teams. Filter your tidied dataset to include only men's sports. Then visualize the distribution of APRs for each sport. Which sports tend to have higher and lower APRs on average?

```
ncaa_tidy_male<-ncaa_tidy[ncaa_tidy$SPORT_CODE <=18 & ncaa_tidy$SPORT_CODE>=1,]
ggplot(ncaa_tidy_male, aes(x=APR, y=SPORT_NAME )) +geom_boxplot() +
  labs(title="Analysis OF NCAA MEN SPORTS APR OVER THE DECADE",
     x="APR", y="SPORT NAME")  + theme_minimal()
```

```
## Warning: Removed 2199 rows containing non-finite values (stat_boxplot).
```



Upon a closer analysis into the mens sports, we can observe that the sports involved in Olympics tend to have higher APR than the other sports in the list and it seems true for the most part.