Mini Project #2

Kalide Endale

10/12/2021

Loading packages I will need for this project

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v purrr   0.3.4     v stringr 1.4.0
## v dplyr   1.0.7     v forcats 0.5.1

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

Import dataset into R

```
dir1 <- "/Users/Kalide/Documents/Northeastern/Introduction to Data Management and Processing"
dir2 <- "Homeworks and Exercsies/NCAA-D1-APR-2003-14/DS0001/26801-0001-Data.csv"
path <- file.path(dir1, dir2)
NCAA_D1_APR_2003_14 <- read_csv(path, na = "-99")
```

```
## Warning: One or more parsing issues, see `problems()` for details

## Rows: 6511 Columns: 76

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): SCL_NAME, SPORT_NAME, CONFNAME_14, D1_FB_CONF_14
## dbl (68): SCL_UNITID, SPORT_CODE, ACADEMIC_YEAR, SCL_DIV_14, SCL_SUB_14, SCL...
## lgl  (4): DATA_TAB_GENERALINFO, DATA_TAB_MULTIYRRATE, DATA_TAB_ANNUALRATE, D...

##
## 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.
```

Creating tidied data frame with columns for School name, School ID, Sport code, Sport name, APR Year, APR score

```
TidyNCAADF <- NCAA_D1_APR_2003_14 %>%
  select("SCL_NAME","SCL_UNITID", "SPORT_CODE", "SPORT_NAME",starts_with("APR_RATE"))

colnames(TidyNCAADF)[5:15] <- c("2014", "2013", "2012", "2011", "2010",
                                "2009", "2008", "2007", "2006", "2005", "2004")
TidyNCAADF <- TidyNCAADF[, c(1,2,3,4,15,14,13,12,11,10,9,8,7,6,5)]

tibble(TidyNCAADF)
```

```
## # A tibble: 6,511 x 15
##    SCL_NAME  SCL_UNITID SPORT_CODE SPORT_NAME `2004` `2005` `2006` `2007` `2008`
##    <chr>          <dbl>      <dbl> <chr>       <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Alabama ~     100654         20 Women's B~   1000   1000    875    958   1000
##  2 Alabama ~     100654         14 Men's Tra~    938    926    903     NA     NA
##  3 Alabama ~     100654          4 Football       871    921    941    919    930
##  4 Alabama ~     100654          1 Baseball       975    917    923    953    938
##  5 Alabama ~     100654         19 Women's B~    960   1000   1000   1000    967
##  6 Alabama ~     100654         33 Women's T~   1000    850    938   1000   1000
##  7 Alabama ~     100654          2 Men's Bas~    950    909    923    964    915
##  8 Alabama ~     100654         34 Women's T~    938   1000    982     NA     NA
##  9 Alabama ~     100654         35 Women's T~    938   1000    983    955    898
## 10 Alabama ~     100654         31 Women's S~    960    963   1000    974    984
## # ... with 6,501 more rows, and 6 more variables: 2009 <dbl>, 2010 <dbl>,
## #   2011 <dbl>, 2012 <dbl>, 2013 <dbl>, 2014 <dbl>
```

Creating a pivot_longer version of the tidied data frame to consolidate all APR data into one column

```
NCAA_D1_APR <- pivot_longer(TidyNCAADF, cols = 5:15, names_to = "APR_YEAR", values_to = "APR")

tibble(NCAA_D1_APR)
```

```
## # A tibble: 71,621 x 6
##    SCL_NAME               SCL_UNITID SPORT_CODE SPORT_NAME      APR_YEAR   APR
##    <chr>                       <dbl>      <dbl> <chr>           <chr>    <dbl>
##  1 Alabama A&M University     100654         20 Women's Bowling 2004      1000
##  2 Alabama A&M University     100654         20 Women's Bowling 2005      1000
##  3 Alabama A&M University     100654         20 Women's Bowling 2006       875
##  4 Alabama A&M University     100654         20 Women's Bowling 2007       958
##  5 Alabama A&M University     100654         20 Women's Bowling 2008      1000
##  6 Alabama A&M University     100654         20 Women's Bowling 2009      1000
##  7 Alabama A&M University     100654         20 Women's Bowling 2010       950
##  8 Alabama A&M University     100654         20 Women's Bowling 2011      1000
##  9 Alabama A&M University     100654         20 Women's Bowling 2012      1000
## 10 Alabama A&M University     100654         20 Women's Bowling 2013      1000
## # ... with 71,611 more rows
```

INTRODUCTION: In 2004, the NCAA developed a metric that measures a team's academic success in order to hold institutions accountable for the academic progress of their student athletes. This metric was called Academic Progress Rate(APR). APR measures student athletes that receive athletically related financial aid and assigns points for grades and retention. A perfect team score is 1000 and a score below 930 (equivalent to 50% graduation rate) means teams could face severe penalties by the NCAA–if score doesn't improve. The cumulative team score is what will be used to measure the academic performance of a sports team. This scoring index helps the NCAA reward institutions for high academic performances and penalizes institutions that don't prepare their student athletes for life post college. In my analysis below, i will try and look into at how APR scoring has improved over time and if some sport teams are generally higher academic performers than others.

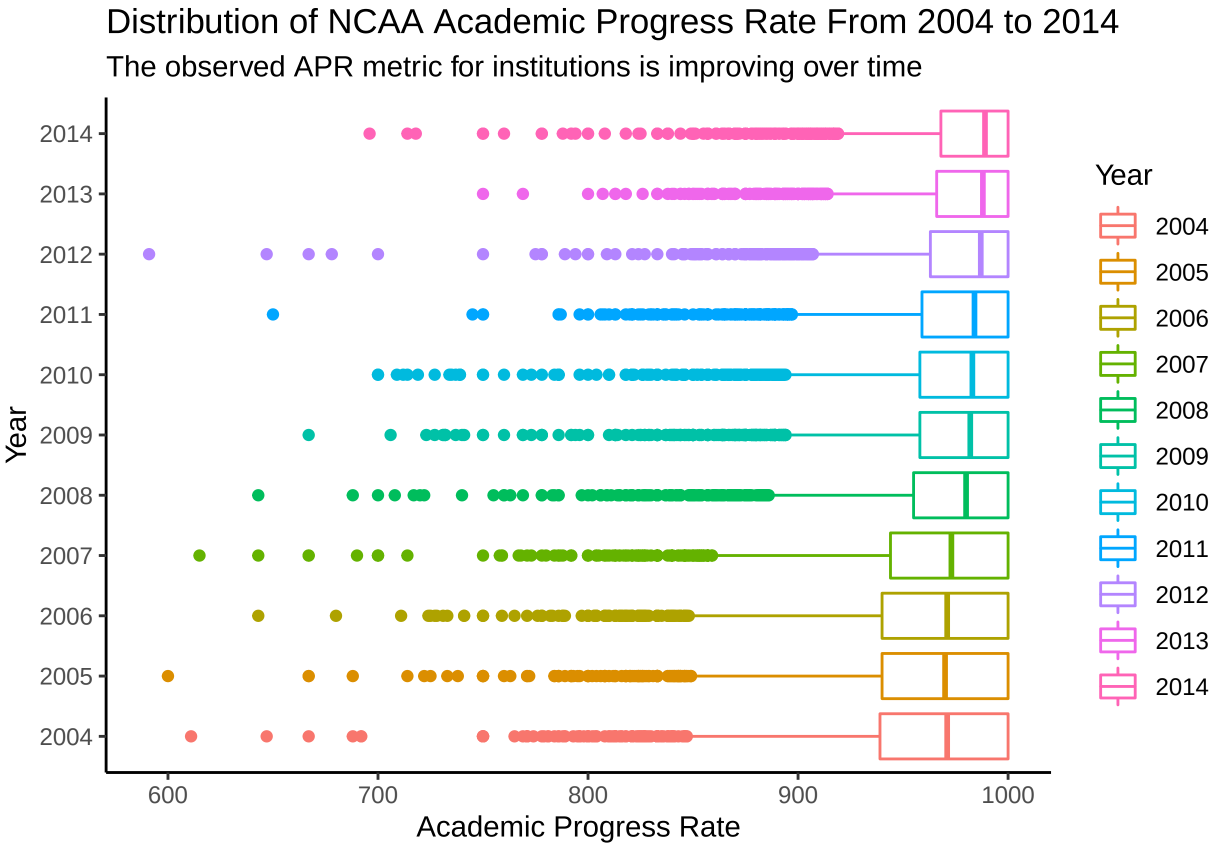
In the box and whisker plots below, I attempt to visualize the distribution of APR scores over time.

What we can see is that since the observation of this metric[2004] to 2014, the median APR score has gradually improved over time. Additionally, it seems like the entire box and whisker plot has consolidated over the years(both outliers and interquartile range).

Extra: The data points outside of the interquartile range were approaching the interquartile range over the years this metric was observed. Furthermore, the interquartile range itself was also wider in the earlier years of this metric and began to consolidate to become smaller over time. Simply stated, this tells us that the APR scores are getting better over time for schools. This bodes well for the NCAA since institutions are making more and more effort to make sure their athletes are excelling in the classroom.

```
ggplot(NCAA_D1_APR, aes(x = APR, y = APR_YEAR, color = APR_YEAR)) +
  geom_boxplot() +
  labs(x = "Academic Progress Rate",
       y = "Year",
       title = "Distribution of NCAA Academic Progress Rate From 2004 to 2014",
       subtitle = "The observed APR metric for institutions is improving over time",
       color = "Year") +
  theme_classic()
```

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



Transform the tidied dataset to remove mixed sports, and create a column indicating the gender division of each sport. Sport codes 1-18 are men's, and 19-37 are women's

```
length(unique(NCAA_D1_APR$SPORT_CODE))
```

```
## [1] 38
```

```
NCAA_D1_APR2 <- NCAA_D1_APR[!(NCAA_D1_APR$SPORT_CODE=="38"), ]
NCAA_D1_APR2 <- NCAA_D1_APR2 %>%
  mutate(SPORT_TYPE = ifelse(SPORT_CODE %in% c(1:18),"Male", "Female"))

tibble(NCAA_D1_APR2)
```

```
## # A tibble: 71,379 x 7
##    SCL_NAME         SCL_UNITID SPORT_CODE SPORT_NAME   APR_YEAR    APR SPORT_TYPE
##    <chr>                 <dbl>      <dbl> <chr>        <chr>     <dbl> <chr>
##  1 Alabama A&M Uni~     100654         20 Women's Bow~ 2004       1000 Female
##  2 Alabama A&M Uni~     100654         20 Women's Bow~ 2005       1000 Female
##  3 Alabama A&M Uni~     100654          20 Women's Bow~ 2006       875 Female
##  4 Alabama A&M Uni~     100654          20 Women's Bow~ 2007       958 Female
##  5 Alabama A&M Uni~     100654          20 Women's Bow~ 2008      1000 Female
##  6 Alabama A&M Uni~     100654          20 Women's Bow~ 2009      1000 Female
##  7 Alabama A&M Uni~     100654          20 Women's Bow~ 2010       950 Female
##  8 Alabama A&M Uni~     100654          20 Women's Bow~ 2011      1000 Female
##  9 Alabama A&M Uni~     100654          20 Women's Bow~ 2012      1000 Female
## 10 Alabama A&M Uni~     100654          20 Women's Bow~ 2013      1000 Female
## # ... with 71,369 more rows
```

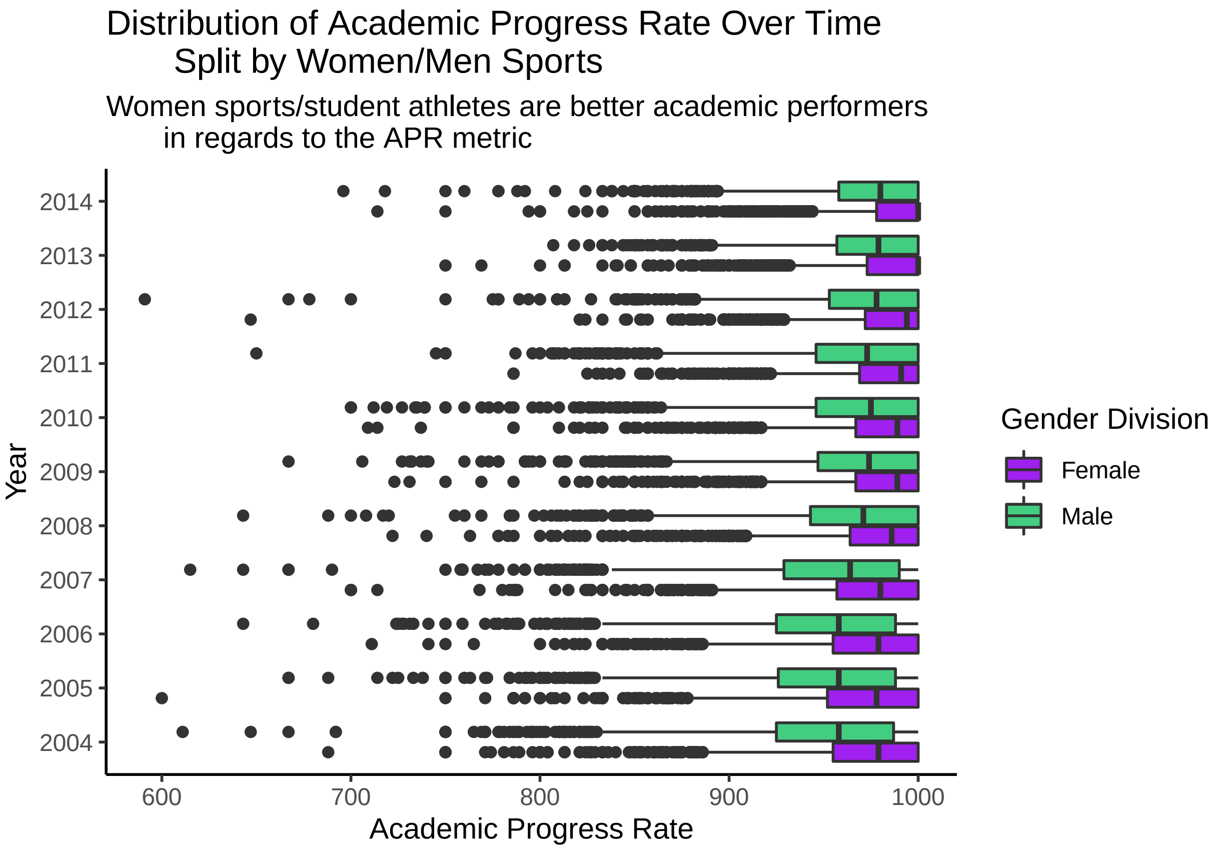
When the dataset is broken down by gender division, another theme begins to surface. The box and whisker plot shows that Women sports teams are generally better academic performers than their male counter parts(for each year: 2004 - 2014). Additionally, the distribution(described by the length of the box and whisker plot) for males is wider while the distribution for females is tighter. However, we can also say that both Males and Females have both improved their median APR score over time.

```
c <- ggplot(NCAA_D1_APR2, aes(x= APR, y=APR_YEAR, fill = SPORT_TYPE)) +
  geom_boxplot() +
  labs(x = "Academic Progress Rate",
       y = "Year",
       title = "Distribution of Academic Progress Rate Over Time
       Split by Women/Men Sports",
       subtitle = "Women sports/student athletes are better academic performers
       in regards to the APR metric") +
  theme_classic()

plot2 <-c + scale_fill_manual(name = "Gender Division", values = c("purple", "seagreen3"))

plot2
```

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



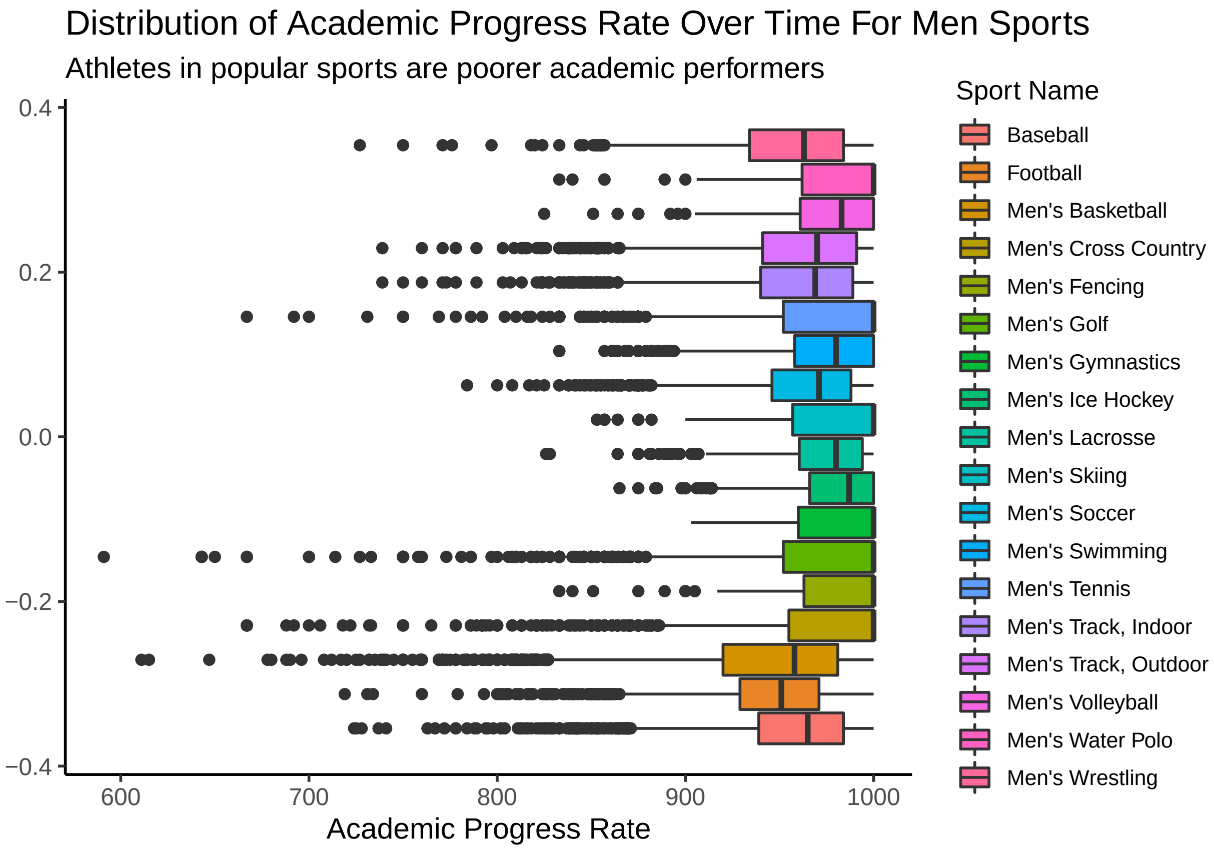
As we investigate the APR scores of different men sport teams, we can see that football, basketball, baseball, indoor track, and outdoor track teams have a lower median APR score while men's cross country, fencing, golf, gymnastics, tennis, and water polo are on the higher-end of the APR median.

source for theme: *https://www.statology.org/ggplot2-legend-size/*

```
NCAA_D1_APR3 <- filter(NCAA_D1_APR2, SPORT_TYPE == "Male")

ggplot(NCAA_D1_APR3, aes(x=APR, fill = SPORT_NAME)) +
  geom_boxplot() +
  labs(x = "Academic Progress Rate",
       title = "Distribution of Academic Progress Rate Over Time For Men Sports",
       subtitle = "Athletes in popular sports are poorer academic performers",
       fill = "Sport Name") +
  theme_classic() +
  theme(legend.key.size = unit(.5, 'cm'), #change legend key size
        legend.key.height = unit(.5, 'cm'), #change legend key height
        legend.key.width = unit(.5, 'cm'), #change legend key width
        legend.title = element_text(size=10), #change legend title font size
        legend.text = element_text(size=8)) #change legend text font size
```

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


Distribution of Academic Progress Rate Over Time For Men Sports
Athletes in popular sports are poorer academic performers
0.4
0.2
0.0
−0.2
−0.4
600
700
800
900
1000
Academic Progress Rate
Sport Name
Baseball
Football
Men's Basketball
Men's Cross Country
Men's Fencing
Men's Golf
Men's Gymnastics
Men's Ice Hockey
Men's Lacrosse
Men's Skiing
Men's Soccer
Men's Swimming
Men's Tennis
Men's Track, Indoor
Men's Track, Outdoor
Men's Volleyball
Men's Water Polo
Men's Wrestling