S5110 Homework 2 - Solutions

Kylie riel Bemis

22 January 2024

Instructions

Your solutions should include all of the code necessary to answer the problems. ll of your code should run (assuming the data is available). ll plots should be generated using `ggplot2`. *Make sure that you answer all parts of the problem.*

Submit your solutions on Canvas by the deadline displayed online. For full credit, your submission must include exactly two files:

- R Markdown (.Rmd)
- Knitted PDF report (.pdf)

Problems must appear in order, and problem numbers must be clearly marked. ny written responses should appear outside of code blocks and use Markdown for text formatting. Code comments are encouraged, but will be ignored for grading purposes. Solutions that are especially difficult to grade due to poor formatting will not receive full credit.

 ll solutions to the given problems must be your own work. If you use third-party code for ancillary tasks, you **must** cite them.

```
# lo d p ck ges
library(readr)
library(tidyr)
library(dplyr)
library(ggplot2)
```

Part

Problems 1–2 ask you to find a dataset that is personally interesting to you. It may be a publicly-available dataset, or a 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. Here some additional resources to get you started:

- US Government datasets (https://catalog.data.gov/dataset)
- Center for Disease Control (CDC) data (https://data.cdc.gov)
- Bureau of Labor Statistics (https://www.bls.gov/data/)
- N S datasets (https://nssdc.gsfc.nasa.gov)
- World Bank Open Data (https://data.worldbank.org)
- Kaggle Datasets (https://www.kaggle.com/datasets)

Pro lem 1

Import the dataset into R. Perform any preprocessing (tidying and cleaning) necessary for visualizing the data.

Cite the source for the dataset. Describe the data, the variables, and any preprocessing steps you performed.

nswers will vary

Pro lem 2

Visualize something interesting to you from the dataset using `ggplot2`. Comment on what the visualization shows and any key conclusions.

nswers will vary

Part B

Problems 3–5 use data on NC student-athlete academic performance. Download the data files from "NC -D1- PR-2003-14.zip" on Piazza. The files include the codebook and tab-delimited data for team-level cademic Progress Rates (PRs) of Division I student-athletes from 2003-2014.

team's PR 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 N s.

Pro lem 3

Create a tidy data frame that includes columns for:

- School ID
- School name
- Sport code
- Sport name
- Year
- PR

ll other columns can be discarded.

Use your tidied dataset to visualize the distributions of PRs over time. How does the distribution of PRs change year-to-year from 2004 to 2014?

Hint: The `tidyr::st rts_with()` and `stringr::str_sub()` functions may be useful.

Solution

First we import the dataset using `re d_tsv()`, reading "-99" as missing values.

```
dir <- "~/Documents/Northeastern/Courses/DS5110/Content/Data"
path <- file.path(dir, "NC  -D1- PR-2003-14/DS0001/26801-0001-Data.tsv")
d1_raw <- read_tsv(path, na=c("", "-99"))
```

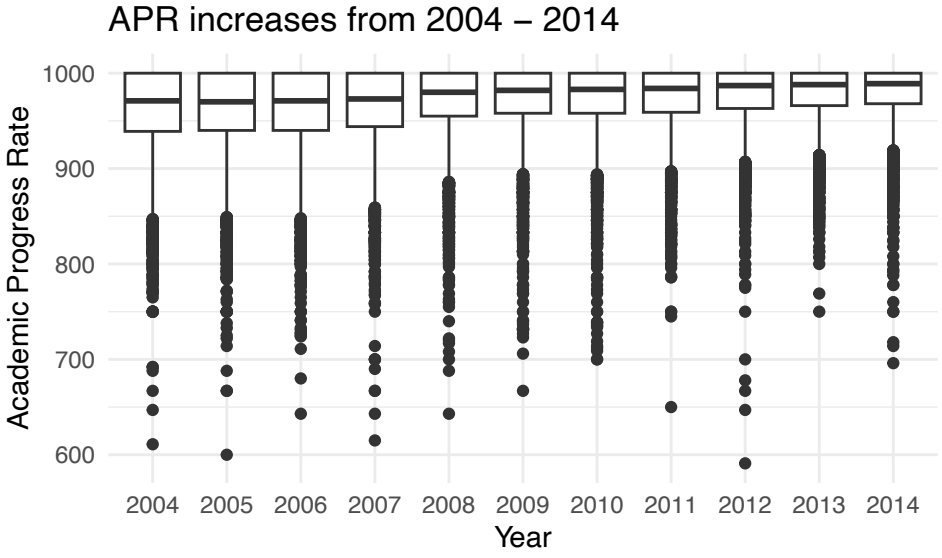
The PR variable is spread across multiple columns with column names encoding the year. We need to use `pivot_longer()` to transform these columns into a tidy representation, and then extract the year from the column names.

```
d1 <- d1_raw %>%
  select(-D T _T B_GENER LINFO) %>%
  pivot_longer(cols=starts_with(" PR_R TE"),
               names_to="YE R", values_to=" PR") %>%
  select(SCL_UNITID, SCL_N ME, SPORT_CODE, SPORT_N ME, YE R,  PR) %>%
  mutate(YE R=as.numeric(stringr::str_sub(YE R, start=10, 13)))
d1
```

```
## #   tibble: 71,621 x 6
##    SCL_UNITID SCL_N ME                SPORT_CODE SPORT_N ME       YE R    PR
##         <dbl> <chr>                        <dbl> <chr>           <dbl> <dbl>
##  1     100654  labama  &M University         20 Women's Bowling  2014  1000
##  2     100654  labama  &M University         20 Women's Bowling  2013  1000
##  3     100654  labama  &M University         20 Women's Bowling  2012  1000
##  4     100654  labama  &M University         20 Women's Bowling  2011  1000
##  5     100654  labama  &M University         20 Women's Bowling  2010   950
##  6     100654  labama  &M University         20 Women's Bowling  2009  1000
##  7     100654  labama  &M University         20 Women's Bowling  2008  1000
##  8     100654  labama  &M University         20 Women's Bowling  2007   958
##  9     100654  labama  &M University         20 Women's Bowling  2006   875
## 10     100654  labama  &M University         20 Women's Bowling  2005  1000
## # i 71,611 more rows
```

Finally, we use boxplots to visualize the distribution of PRs over time.

```
ggplot(d1) +
  geom_boxplot(aes(x=as.factor(YE R), y= PR)) +
  labs(x="Year", y=" cademic Progress Rate",
       title=" PR increases from 2004 - 2014") +
  theme_minimal()
```



In general, average PRs are increasing over time.

Pro lem 4

We would like to compare PRs 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.)

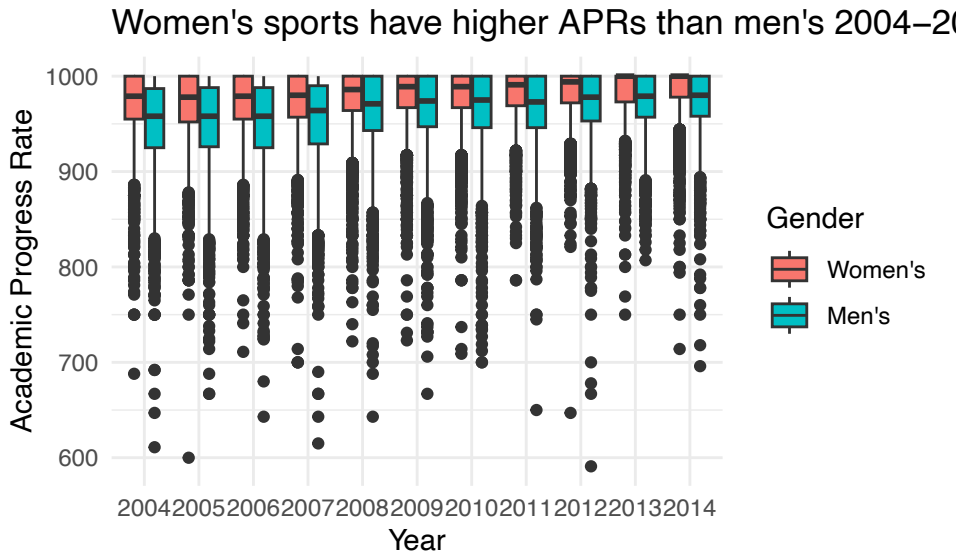
Visualize the distributions of PRs over time again, but broken down by gender division. How do the median PRs compare between men's and women's sports? Does this relationship hold true across each year from 2004 to 2014?

Hint: The `ifelse()` function may be useful.

Solution

```
d1_gender <- d1 %>%
  filter(SPORT_CODE != 38) %>%
  mutate(GENDER = ifelse(SPORT_CODE >= 19, "Women's", "Men's"),
         GENDER = factor(GENDER, c("Women's", "Men's")))

ggplot(d1_gender) +
  geom_boxplot(aes(x=as.factor(YE R), y= PR, fill=GENDER)) +
  labs(x="Year", y=" cademic Progress Rate", fill="Gender",
       title="Women's sports have higher PRs than men's 2004-2014") +
  theme_minimal()
```



On average, women's teams have a higher PR than men's teams. This holds true every year from 2004 to 2014.

Pro lem 5

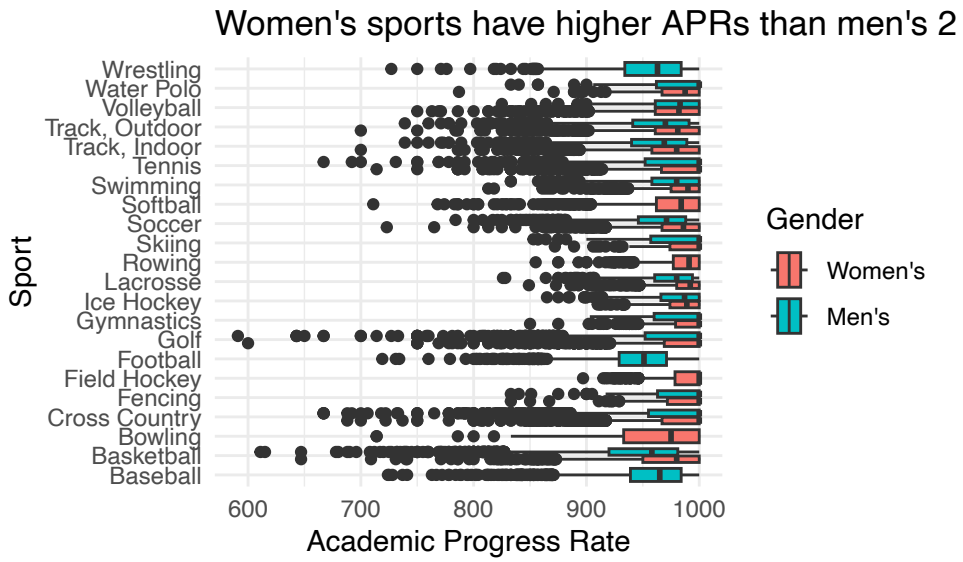
We would like to further visualize PR by both gender and specific sports. Process the the sport names to remove the "Men's" and "Women's" prefixes so that we can compare men's and women's teams within each sport. Then visualize the distribution of PR for both men's and women's teams for each sport. re there sports where men's and women's teams have similar PRs?

Hint: The `stringr::str_remove()` function may be useful.

Solution

```
d1_gender %>%
  mutate(SPORT_N ME = stringr::str_remove(SPORT_N ME, "Women's "),
         SPORT_N ME = stringr::str_remove(SPORT_N ME, "Men's ")) %>%
  ggplot() +
  geom_boxplot(aes(x= PR, y=SPORT_N ME, fill=GENDER)) +
  labs(x=" cademic Progress Rate", y="Sport", fill="Gender",
```

```
      title="Women's sports have higher  PRs than men's 2004-2014") +
theme_minimal()
```



The men's and women's teams have similar PR distributions for water polo, volleyball, tennis, and ice hockey.