homework

Sean Li and Ben Thorpe

Due before

The main goals of this homework is to let you practice data loading, data cleaning, and finding useful information. It will mimic the process you will go through in a real sports analysis endeavor.

Remember, this is important because often times in the project, half the work is getting the right data!

Homework Instructions

Complete the exercises below through creating a new RMD file in the same repository you had your first homework done in. You are also welcome to create a new repository to do this homework too.

Data for this homework can be found through the class repository. You can obtain it through downloading the file directly from the repository on Github or downloading the data from https://www.basketball-reference.com/leagues/NBA_2022_per_game.html

Once you are done, commit and push one last time to make sure all your changes are tracked. Then fill out the google form below.

TURN IT IN HERE: https://forms.gle/PVqBb7JhaaKnTvAXA

Exercises

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

```
library(openxlsx)
```

Loading in the Data

We will be working with 2021-2022 NBA Player Stats Data.

1. Load in the data (make sure datafile is in repository folder!!)

```
data <- read.xlsx(xlsxFile = "data/nbadata.xlsx")
```

2. Take a look at the data using glimpse()... what seems to be a bit off about the player names?
3. String Parsing

```
names <- strsplit(data$Player, "\\\\")
data <- data %>%
  separate(Player, "Player",  "\\\\") %>%
  group_by(Player) %>%
  summarise(threes = mean(`3PA`))
```

```
## Warning: Expected 1 pieces. Additional pieces discarded in 609 rows [1, 2, 3, 4,
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

Lets try to fix our issue with the names of each player.

Training Wheels

Here a few exercises that are more guided and easier to complete.

1. How many players are on the Charlotte Hornets?
2. How many shooting guards do the Milwaukee Bucks have?
3. Who is leading the league in turnovers per game?
4. How many 3 point attempts does the average NBA player attempt per game? How many more does Steph Curry attempt?

grepl('Stephen Curry', text)

Rumble.

These are questions made to have you think how to problem solve. I would recommend breaking down some of these problems into several steps. For example, to find the top 5 22-year old players who have high assist to turnover ratios this would be my thinking: "first i need to get all players who are 22 years old who have played at least 10 games. Then I need to make a new variable to track assist to turnover ratio. Then I need to sort the names by descending according to that ratio. Then I need to display the first 5 unique names."

If you are stuck, first google, then reach out for help.

5. How many players are in this dataset? (HINT: its not just the number of rows)
6. Find Atlantic Division (Nets, 76ers, Raptors, Celtics, Knicks) teams' top 3 point shooter based on 3PT percentage that has attempted at least 3 a game.

7. What is the average age on the Minnesota Timberwolves?

8. Which Power Forward (PF) has the largest differential between their offensive and defensive rebounding stats per game?

CHALLENGE: Create a report of the 2021-2022 Phoenix Suns describing the following: leaders in all 5 box score categories (points, rebounds, assists, steals, blocks), most frequent starting lineup, and most underrated player. For the last part to determine the underrated player, I want you to pick your statistical criterion to determine "underratedness" and give an analysis on why you choose that player.

install.packages("hoopR")