Midterm II

Fall 2018

Due: Wednesday, November 14, 2018

##### 1. First write code that loads the workspace “Midterm II.RData”. This should read five different tibbles into your workspace.

load("C:/Users/rdn3.BYU/Documents/GitHub/r4ds/Midterm II/Midterm II.RData")

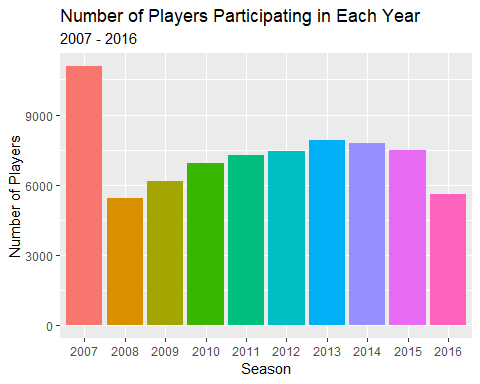
##### 2. Write code that recreates the tibbles “player” and “player\_attributes” by importing the files “player.csv” and “player\_attributes.csv”. Compare the tibbles that you create with those in the Midterm II workspace to make sure that they are the same.

##### 3. Write code that recreates the season factor variable in the player\_attributes tibble. Season is defined as the year from the match\_date variable. Make sure your code converts your variable into a factor.

seasons <- 2007:2016  
player\_attributes <- player\_attributes %>%   
 mutate(season = year(match\_date) %>% factor(levels = seasons))

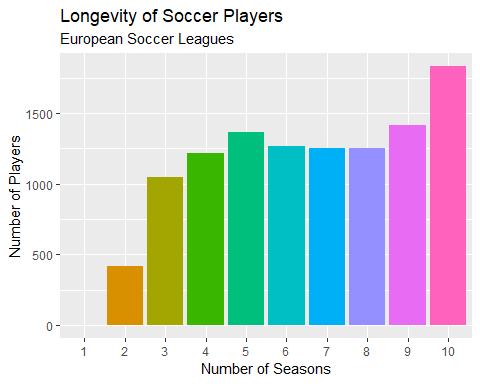
##### 4. We want to determine the total number of players that participate in any given season by using group\_by, summarise, and geom\_bar. Write code that creates the following bar chart:

player\_attributes %>%   
 group\_by(player\_id, season) %>%   
 summarise(count = n()) %>%   
 ggplot(aes(x = season, fill = season)) +  
 geom\_bar(show.legend = FALSE) +  
 labs(  
 title = "Number of Players Participating in Each Year",  
 subtitle = "2007 - 2016",  
 x = "Season",  
 y = "Number of Players"  
 )



##### 5. Explain the code in the script file that creates the following graph:

# Number of seasons played  
number\_of\_seasons <- player\_attributes %>%   
 group\_by(player\_id, season) %>%   
 count() %>%   
 group\_by(player\_id) %>%   
 count()  
colnames(number\_of\_seasons) <- c("player\_id", "number\_of\_seasons")  
number\_of\_seasons <- number\_of\_seasons %>%   
 mutate(number\_of\_seasons = factor(number\_of\_seasons, levels = 1:10))  
  
number\_of\_seasons %>%   
 ggplot(aes(x = number\_of\_seasons, fill = number\_of\_seasons)) +  
 geom\_bar(show.legend = FALSE) +  
 labs(  
 title = "Longevity of Soccer Players",  
 subtitle = "European Soccer Leagues",  
 x = "Number of Seasons",  
 y = "Number of Players"  
 )



##### 6. Write a function named age that will create a variable age for each player based on the birthday of the player and the day of the match.

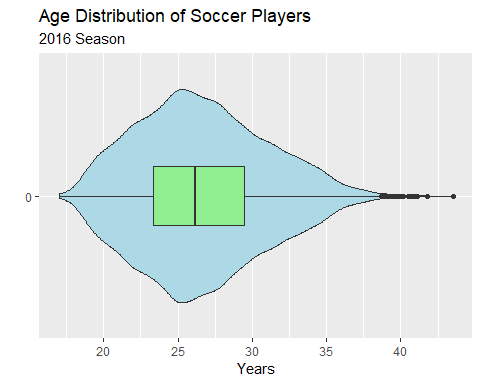
### Function to calculate the age as of the player for each match  
age <- function(birthday, match\_date){  
 (birthday %--% match\_date) / years(1)  
}

##### 7. Interpret the code in the script file that creates the tibble player\_attributes\_complete.

player\_attributes\_complete <- player\_attributes %>%   
 left\_join(player, key = "player\_id") %>%   
 mutate(age = age(birthday, match\_date))

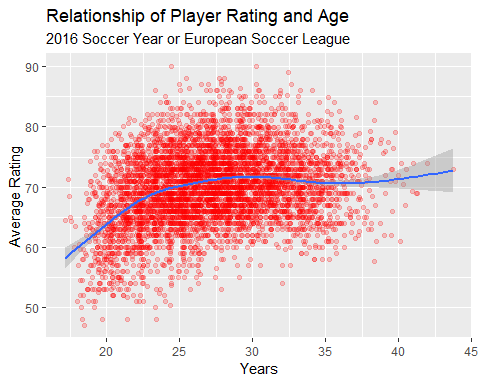
##### 8. We are interested in the distribution of ages of the players that participated in the 2016 season. Recreate the player\_attributes\_2016 tibble and then write code that creates a violin plot for the ages of the players who participated in the 2016 season.

# Violin plot of the ages of those that participated in 2016  
player\_attributes\_2016 <- player\_attributes\_complete %>%   
 filter(year(match\_date) == 2016)  
  
player\_attributes\_2016 %>%   
 ggplot(aes(x = factor(0), y = age)) +  
 geom\_violin(fill = "lightblue") +  
 geom\_boxplot(width = 0.25, fill = "lightgreen") +  
 coord\_flip() +  
 labs(  
 title = "Age Distribution of Soccer Players",  
 subtitle = "2016 Season",  
 x = "",  
 y = "Years"  
 )



##### 9. Explain the code in the script file that creates the following graph:

# Relationship of Age and Average Rating for 2016 players  
player\_attributes\_2016 %>%   
 group\_by(player\_id) %>%   
 summarize(  
 average\_rating = mean(rating)  
 ) %>%   
 left\_join(player, key = "player\_id") %>%   
 mutate(age = age(birthday, ymd(20160701))) %>%   
 ggplot(aes(x = age, y = average\_rating)) +  
 geom\_point(position = "jitter", alpha = 1/5, shape = 19, color = "red") +  
 geom\_smooth() +  
 labs(  
 title = "Relationship of Player Rating and Age",  
 subtitle = "2016 Soccer Year or European Soccer League",  
 x = "Years",  
 y = "Average Rating"  
 )



##### 10. Interpret the following code in the script file that compares the time series for average age and average rating.

# Time series of average age and average rating  
player\_attributes\_complete %>%   
 group\_by(season) %>%   
 summarize(  
 Age = mean(age),  
 Rating = mean(rating, na.rm = TRUE)  
 ) %>%   
 gather(Age, Rating, key = "measure", value = "amount") %>%   
 ggplot(aes(y = amount, color = measure)) +  
 geom\_point(aes(x = season), show.legend = FALSE) +  
 geom\_path(aes(x = as.numeric(season)), show.legend = FALSE) +  
 facet\_grid(rows = vars(measure), scales = "free\_y") +  
 labs(  
 title = "Trends in Average Age and Average Rating",  
 subtitle = "European Soccer League (Kaggle Dataset",  
 x = "",  
 y = ""  
 )

