DATA 612 HOMEWORK 5

Johnson ODEJIDE

2022-10-15

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --  
## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1  
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
library(ggplot2)  
# install.packages("Lahman")  
library(Lahman)  
  
# =======================================================================  
# Answer 1  
# =======================================================================  
is\_tibble(Batting)

## [1] FALSE

# =======================================================================  
# Answer 2  
#   
# Batting has 110495 rows and 22 Columns  
# =======================================================================  
dim(Batting)

## [1] 110495 22

# =======================================================================  
# Answer 3  
# =======================================================================  
as\_tibble(Batting)

## # A tibble: 110,495 x 22  
## playerID yearID stint teamID lgID G AB R H X2B X3B HR  
## <chr> <int> <int> <fct> <fct> <int> <int> <int> <int> <int> <int> <int>  
## 1 abercda01 1871 1 TRO NA 1 4 0 0 0 0 0  
## 2 addybo01 1871 1 RC1 NA 25 118 30 32 6 0 0  
## 3 allisar01 1871 1 CL1 NA 29 137 28 40 4 5 0  
## 4 allisdo01 1871 1 WS3 NA 27 133 28 44 10 2 2  
## 5 ansonca01 1871 1 RC1 NA 25 120 29 39 11 3 0  
## 6 armstbo01 1871 1 FW1 NA 12 49 9 11 2 1 0  
## 7 barkeal01 1871 1 RC1 NA 1 4 0 1 0 0 0  
## 8 barnero01 1871 1 BS1 NA 31 157 66 63 10 9 0  
## 9 barrebi01 1871 1 FW1 NA 1 5 1 1 1 0 0  
## 10 barrofr01 1871 1 BS1 NA 18 86 13 13 2 1 0  
## # ... with 110,485 more rows, and 10 more variables: RBI <int>, SB <int>,  
## # CS <int>, BB <int>, SO <int>, IBB <int>, HBP <int>, SH <int>, SF <int>,  
## # GIDP <int>

# =======================================================================  
# Answer 4  
#   
# 12 Players hit more than 30 home runs in 1991  
# =======================================================================  
Batting %>%  
 select(playerID, teamID, yearID, HR) %>%  
 filter(HR > 30, yearID == 1991) -> HR1991  
HR1991

## playerID teamID yearID HR  
## 1 cansejo01 OAK 1991 44  
## 2 cartejo01 TOR 1991 33  
## 3 dawsoan01 CHN 1991 31  
## 4 fieldce01 DET 1991 44  
## 5 gantro01 ATL 1991 32  
## 6 johnsho01 NYN 1991 38  
## 7 mcgrifr01 SDN 1991 31  
## 8 ripkeca01 BAL 1991 34  
## 9 tartada01 KCA 1991 31  
## 10 tettlmi01 DET 1991 31  
## 11 thomafr04 CHA 1991 32  
## 12 willima04 SFN 1991 34

# Number of players that hit more than 30 home runs in 1991  
# nrow(HR1991)  
count(HR1991)

## n  
## 1 12

# =======================================================================  
# Answer 5  
# =======================================================================  
Batting %>%  
 filter(yearID == 1991) %>%  
 group\_by(yearID) %>%  
 summarise(MeanHR1991 = mean(HR))

## # A tibble: 1 x 2  
## yearID MeanHR1991  
## <int> <dbl>  
## 1 1991 3.12

# =======================================================================  
# Answer 6  
#  
# 12 Players hit more than 30 home runs in 1990  
# =======================================================================  
Batting %>%  
 select(playerID, teamID, yearID, HR) %>%  
 filter(HR > 30, yearID == 1990) -> HR1990  
HR1990

## playerID teamID yearID HR  
## 1 bondsba01 PIT 1990 33  
## 2 bonilbo01 PIT 1990 32  
## 3 cansejo01 OAK 1990 37  
## 4 fieldce01 DET 1990 51  
## 5 gantro01 ATL 1990 32  
## 6 grubeke01 TOR 1990 31  
## 7 mcgrifr01 TOR 1990 35  
## 8 mcgwima01 OAK 1990 39  
## 9 mitchke01 SFN 1990 35  
## 10 sandbry01 CHN 1990 40  
## 11 strawda01 NYN 1990 37  
## 12 willima04 SFN 1990 33

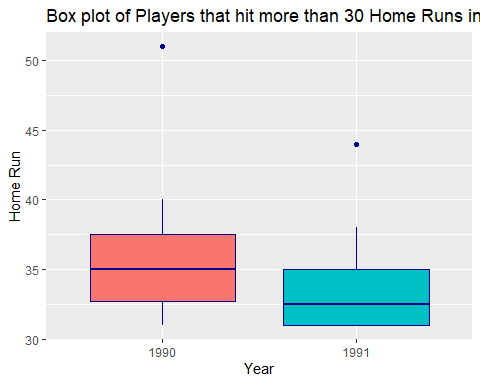
# Number of players that hit more than 30 home runs in 1990  
count(HR1990)

## n  
## 1 12

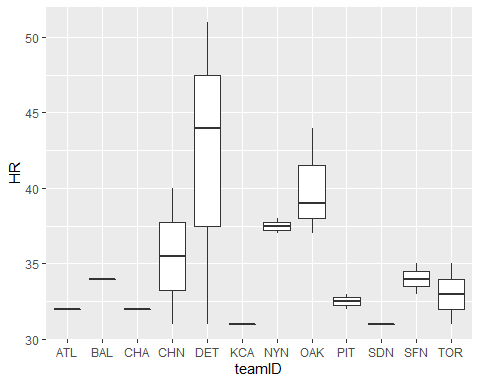
# =======================================================================  
# Answer 7  
# =======================================================================  
Batting %>%  
 filter(yearID == 1990) %>%  
 group\_by(yearID) %>%  
 summarise(MeanHR1990 = mean(HR))

## # A tibble: 1 x 2  
## yearID MeanHR1990  
## <int> <dbl>  
## 1 1990 2.97

# =======================================================================  
# Answer 8  
# =======================================================================  
Batting %>%  
 filter(yearID %in% c(1990, 1991), HR > 30) %>%  
 select(yearID, HR) %>%  
 mutate(yearID = as\_factor(yearID)) %>%  
 ggplot(mapping = aes(x = yearID, y = HR)) +  
 geom\_boxplot(mapping = aes(fill = yearID),   
 color = "darkblue",   
 show.legend = FALSE) +  
 labs(  
 title = "Box plot of Players that hit more than 30 Home Runs in 1990 and 1991",  
 x = "Year",  
 y = "Home Run"  
 )



# =======================================================================  
# Answer 9  
# =======================================================================  
Batting %>%  
 filter(yearID %in% c(1990, 1991), HR > 30) %>%  
 select(teamID, HR) %>%  
 ggplot(mapping = aes(x = teamID, y = HR)) +  
 geom\_boxplot()



# =======================================================================  
# Answer 10  
# =======================================================================  
tribble (~Name, ~Age, ~Department, ~YrsofSrvce, ~EduLevel, ~Salary,  
 "Carlos", 30, "Personnel", 4, "MS", 71500,  
 "Jacob", 26, "Accounting", 6, "BS", 70000,  
 "Elaine", 31, "IT", 4, "BS", 75000,  
 "Alice", 42, "Sales", 5, "BS", 72000,  
 "Juan", 31, "IT", 7, "BS", 68000,  
 "Ray", 28, "Accounting", 5, "MS", 81000,  
 "Kate", 25, "Sales", 4, "BS", 74000,  
 "Leon", 30, "Personnel", 11, "MS", 78000,  
 "Robert", 29, "Accounting", 8, "MS", 77500  
) -> EmployeeData  
EmployeeData

## # A tibble: 9 x 6  
## Name Age Department YrsofSrvce EduLevel Salary  
## <chr> <dbl> <chr> <dbl> <chr> <dbl>  
## 1 Carlos 30 Personnel 4 MS 71500  
## 2 Jacob 26 Accounting 6 BS 70000  
## 3 Elaine 31 IT 4 BS 75000  
## 4 Alice 42 Sales 5 BS 72000  
## 5 Juan 31 IT 7 BS 68000  
## 6 Ray 28 Accounting 5 MS 81000  
## 7 Kate 25 Sales 4 BS 74000  
## 8 Leon 30 Personnel 11 MS 78000  
## 9 Robert 29 Accounting 8 MS 77500

# =======================================================================  
# Answer 11  
# =======================================================================  
ggplot(data = EmployeeData, mapping = aes(x = Name, y = YrsofSrvce)) +  
 geom\_bar(mapping = aes(fill = Name), stat = "identity") +  
 geom\_text(aes(label = YrsofSrvce), vjust = 2) +  
 ggtitle("Years of Service Bar Graph")

