library(dplyr)

library(ggplot2)

arrange(), filter(),select(), slice(),

summarise(), rename()

Center: mean(), median()

Spread: sd(), IQR(), mad()

Range: min(), max(), quantile()

Position: first(), last(), nth(),

Count: n(), n_distinct()

Logical: any(), all()

Other single table verbs: arrange, filter, mutate, select, slice

"playerID: Player ID code. Factor yearID: Year. Factor teamID: Team. factor lgID: League. Factor: AA AL FL NL PL UA AB: At bats. Numeric G: Games: number of games by a player. Numeric R: Runs. Numeric HR: Homeruns. Numeric SH: Sacrifice hits. Numeric "

rename(data, g=no.of games)

Step 1

data <- read.csv("Dataset/lahman-batting.csv") data

Step 2

data%>%select(c(playerID, yearID, AB, teamID, lgID, G, R, HR, SH))%>% arrange(yearID)

Step 2

data%>%select(c(playerID, yearID, AB, teamID, lgID, G, R, HR, SH))%>% arrange(desc(yearID))

data1 <- data %>% filter(teamID == "ATL" | teamID == "BOS" | teamID == "CHA") head(data1)

" data: Dataset used to construct the summary statistics group_by(lgID): Compute the summary by grouping the variable `lgID summarise(mean_run = mean(HR)): Compute the average homerun" data %>% group_by(lgID) %>% summarise(mean_run = mean(HR))

Mean

ex1 <- data %>% group_by(yearID) %>% summarise(mean_game_year = mean(G)) head(ex1)

Plot the graph

ggplot(ex1, aes(x = yearID, y = mean_game_year)) + geom_line() + theme_classic() + labs(x = "Year", y = "Average games played", title = paste("Average games played from 1871 to 2016"))

Sum

data %>% group_by(lgID) %>% summarise(sum_homerun_league = sum(HR))

Min and max

data %>% group_by(playerID) %>% summarise(min_G = min(G), max_G = max(G))

count observations--The number of observations in the current group

data %>% group_by(playerID) %>% summarise(number_year = n()) %>% arrange(desc(number_year))

first and last

data %>% group_by(playerID) %>% summarise(first_appearance = first(yearID), last_appearance = last(yearID))

nth

data %>% group_by(teamID) %>% summarise(second_game = nth(yearID, 3)) %>% arrange(second_game)

distinct values

data %>% group_by(teamID) %>% summarise(number_player = n_distinct(playerID)) %>% arrange(desc(number_player))

Multiple groups

data %>% group_by(yearID, teamID) %>% summarise(mean_games = mean(G)) %>% arrange(desc(teamID, yearID))

Filter

data %>% filter(yearID > 2002) %>% group_by(yearID) %>% summarise(mean_game_year = mean(G))

Ungroup the data

data %>% filter(HR > 0) %>% group_by(playerID) %>% summarise(average_HR_game = sum(HR) / sum(G)) %>% ungroup() %>% summarise(total_average_homerun = mean(average_HR_game))