Capstone Data Wrangling

Will Pauls

February 18, 2019

## Data Wrangling

For my Capstone project I will be pulling from the Teams and Salaries datasets from the Lahman package. Performing data wrangling tasks and merging them into a new dataset called baseball

## Steps

1. Load in Teams dataset with necessary variables

# make team data frame with necessary variables  
teams = Teams  
teams = subset(teams, select = c(yearID, lgID, teamID, Rank, G, W, L, DivWin, WCWin))

1. Since Salary info dates back only to 1985 I filtered the Teams dataset from 1985 and on to match

# grab teams from 1985 and on to match with years of payroll information  
teams = filter(teams, yearID >=1985)

1. I added a column in teams for winning %

# add winning % column  
teams = mutate(teams, winpercent = round((W/(W+L)), digits = 3))

1. I added a column with a boolean data type (Y/N) of whether that team made the playoffs

# add a playoffs column  
teams$DivWin[is.na(teams$DivWin)] <- "N"  
teams$WCWin[is.na(teams$WCWin)] <- "N"  
teams <- mutate(teams, playoffs = if\_else(DivWin == 'Y' | WCWin == 'Y', 'Y', 'N'))  
teams = subset(teams, select = -c(DivWin, WCWin))

1. I loaded in the Salaries dataset

# add salaries to a data frame  
sals = read\_csv("Salaries.csv")

1. I summed up team payroll by grouping by year and team

# Get team payroll for each year  
sals <- sals %>%   
 group\_by(yearID, teamID) %>%   
 summarise(payroll = sum(salary))

1. I standardized team names in each data set to prepare for merge

#Standardize team names   
sals$teamID[sals$teamID=="CHN"] <- "CHC"  
sals$teamID[sals$teamID=="CHA"] <- "CHW"  
sals$teamID[sals$teamID=="KCA"] <- "KCR"  
sals$teamID[sals$teamID=="LAN"] <- "LAD"  
sals$teamID[sals$teamID=="NYN"] <- "NYM"  
sals$teamID[sals$teamID=="NYA"] <- "NYY"  
sals$teamID[sals$teamID=="SDN"] <- "SDP"  
sals$teamID[sals$teamID=="SFN"] <- "SFG"  
sals$teamID[sals$teamID=="SLN"] <- "STL"  
sals$teamID[sals$teamID=="TBA"] <- "TBR"  
sals$teamID[sals$teamID=="WSN"] <- "WAS"  
  
teams$teamID <- as.character(teams$teamID)  
teams$teamID[teams$teamID=="CHN"] <- "CHC"  
teams$teamID[teams$teamID=="CHN"] <- "CHC"  
teams$teamID[teams$teamID=="CHA"] <- "CHW"  
teams$teamID[teams$teamID=="KCA"] <- "KCR"  
teams$teamID[teams$teamID=="LAN"] <- "LAD"  
teams$teamID[teams$teamID=="NYN"] <- "NYM"  
teams$teamID[teams$teamID=="NYA"] <- "NYY"  
teams$teamID[teams$teamID=="SDN"] <- "SDP"  
teams$teamID[teams$teamID=="SFN"] <- "SFG"  
teams$teamID[teams$teamID=="SLN"] <- "STL"  
teams$teamID[teams$teamID=="TBA"] <- "TBR"  
teams$teamID[teams$teamID=="WSN"] <- "WAS"  
teams$teamID <- as.factor(teams$teamID)

1. I merged the teams and salaries datasets into a new one I named baseball

# merge data into baseball  
baseball <- merge(teams, sals, by=c("yearID", "teamID"))  
  
baseball$teamID[baseball$teamID=="ML4"] <- "MIL"  
baseball$yearID <- as.factor(baseball$yearID)  
baseball$playoffs <- as.factor(baseball$playoffs)

1. I created a column in baseball calculating the amount of payroll per team win (Payroll/W)

# calculate Amount of payroll per team win  
baseball <- baseball %>% mutate(dolperwin = payroll/W)

1. I calculated the average payroll in the MLB per year by summing each teams payroll by year and dividing by the number of teams for that year and added it to a new column in baseball

#calculate total payroll per year for entire MLB  
mlb <- baseball %>%   
 group\_by(yearID) %>%   
 summarise(mlbpayroll = sum(as.numeric(payroll)))  
  
# get number of teams for each year  
 numteams <- baseball %>%   
 group\_by(yearID) %>%  
 summarise(numteams =length(yearID))  
  
mlb$numteams <- numteams$numteams  
  
# calculate average payroll per year  
mlb <- mlb %>%   
 mutate(avgpayroll = mlbpayroll/numteams)  
mlb <- subset(mlb, select = -c(numteams))  
  
# merge into baseball data set  
baseball <- merge(baseball, mlb, by="yearID")

1. I divided each teams payroll by the the total MLB payroll to get a % of payroll each team has compared to entire league and added it into a new column

# calculate percentange of total payroll each team has  
baseball <- baseball %>%   
 mutate(percentofmlb = round((payroll/mlbpayroll), digits = 4))

1. I added a column that calculated the average amount of MLB payroll per each win

#calculate average dollars of mlb payroll per win  
wins <- baseball %>%   
 group\_by(yearID) %>%   
 summarise(mlbwin= (sum(W)))  
  
baseball <- merge(baseball, wins, by="yearID")  
baseball <- baseball %>%   
 mutate(mlbdolperwin = mlbpayroll/mlbwin)  
  
baseball <- baseball %>% group\_by(yearID) %>% mutate(payrank = dense\_rank((desc(payroll))))

1. I added a column with the difference in payroll from the previous year for each team

# calculate teams difference in payroll from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(paydiff = payroll - lag(payroll))

1. I created a % change of that difference in a new column

# +/- % change of payroll from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(percentdiff = paydiff/lag(payroll))   
  
baseball$percentdiff <- round(baseball$percentdiff, 4)

1. I created a column with a difference in winning % from the previous year for each team

## calculate change in winning percentage from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(winpercentdiff = winpercent - lag(winpercent))

1. I created a column with a difference in a teams % of MLB payroll from previous year

## calculate change in percent of mlb from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(percentofmlbdiff = percentofmlb - lag(percentofmlb))

1. I created a column with the differencr in Payroll Rank from the previous year

## calculate change in payroll rank from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(payrankdiff = payrank - lag(payrank))

1. I created a column to test if payroll increased for each team that year

# create column to test if payroll increased for team  
baseball$payincreased[baseball$paydiff<=0]<- "No"  
baseball$payincreased[baseball$paydiff>0]<- "Yes"  
  
baseball$payincreased <- as.factor(baseball$payincreased)

1. I created a column to test if winning % increased for each team that year

# create a column to test if winning % increased for team  
baseball$winincreased[baseball$winpercentdiff<=0]<- "No"  
baseball$winincreased[baseball$winpercentdiff>0]<- "Yes"  
  
baseball$winincreased <- as.factor(baseball$winincreased)

1. I created a column to test if teams % of MLB payroll increased from previous year

# create a column to test if percent of mlb increased for team  
baseball$percentofmlbincreased[baseball$percentofmlbdiff<=0]<- "No"  
baseball$percentofmlbincreased[baseball$percentofmlbdiff>0]<- "Yes"  
  
baseball$percentofmlbincreased <- as.factor(baseball$percentofmlbincreased)

1. I created a column to test if a Teams Payroll rank increased from previous year

# create column to test if payroll increased for team  
baseball$payrankincreased[baseball$payrankdiff<=0]<- "No"  
baseball$payrankincreased[baseball$payrankdiff>0]<- "Yes"  
  
baseball$payrankincreased <- as.factor(baseball$payrankincreased)