Name: **Krishna Chaitanya**

Introduction to R data.table

# Introduction

For this assignment our aim is to describe the data set using R commands.

Below are the different types of functions that are applied on the baseball dataset, click here for each function.

[Download the baseball data set from website.](#_Downloading_Data_set)

[Loading data from CSV and converting to data.table](#_Loading_Data_into)

[Running Summary Functions on data set](#_Summary_Functions)

[Running Chaining Functions on Data set](#_Chaining_Functions)

[Running subset functions on Dataset](#_Subset_Functions)

# Downloading Data set

Downloading comma delimited version data set from below URL:

http://www.seanlahman.com/baseball-archive/statistics/

Once file is downloaded, unzip it to a directory and copy the salaries.csv file to the present working directory.

# Loading Data into R

Load the dataset into R & verify data

Command to Execute:

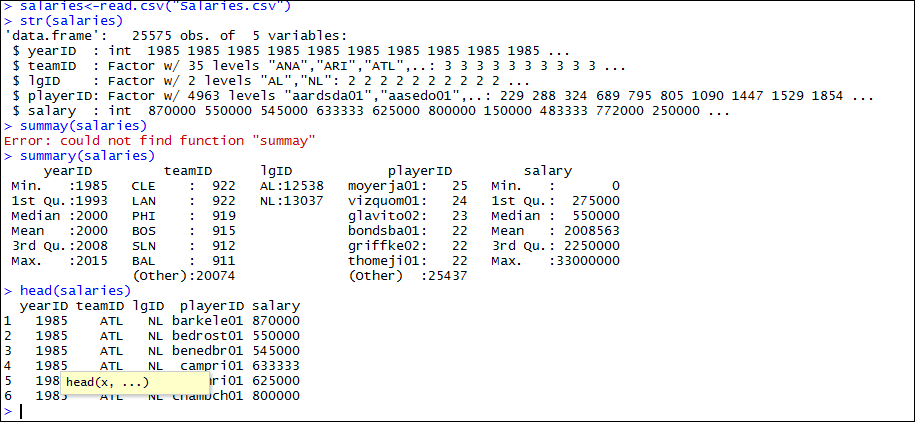
salaries<-read.csv("Salaries.csv") #Read the data as csv and store into salaries dataframe.

str(salaries) #This command will compactly display the internal structure of the data set

summary(salaries) #This command with output the summary of the data frame.

head(salaries) #This command will give you first 5 lines of the dataframe.

Below are the screen shots for the same.



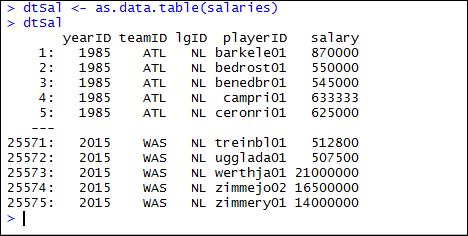
Converting data.frame to data.table.

install.packages('data.table') #Installing data.table package

library('data.table') #using data.table library

dtSal <- as.data.table(salaries) #converts dataframe to data.table

print(dtSal) # allows you to view the first & last few lines

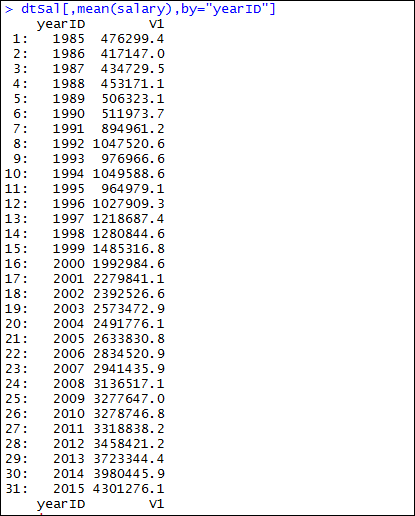


[BACK TO MENU](#_Introduction)

# Summary Functions

Finding mean of all the salaries of all the player’s in each year

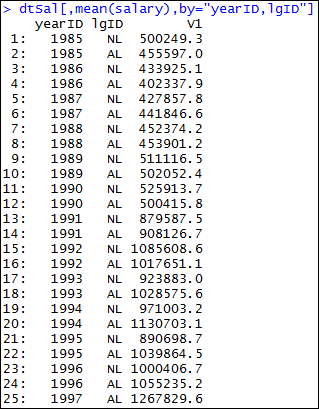
dtSal[,mean(salary),by="yearID"] #finding mean of salaries in each year



Group by yearID and lgID and compute means on salary

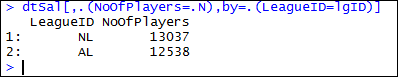
dtSal[,mean(salary),by="yearID,lgID"] #compute mean w.r.t yearID and lgID

Below is the screen shot



Compute No Of Players in each league ID.

dtSal[,(NoOfPlayers=.N),by="lgID"] #.N represents the count in data.table for each lgID

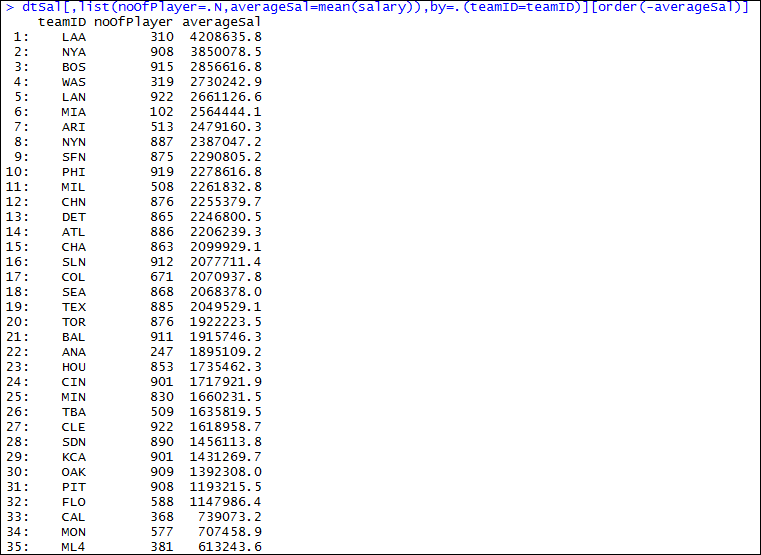


[BACK TO MENU](#_Introduction)

# Chaining Functions

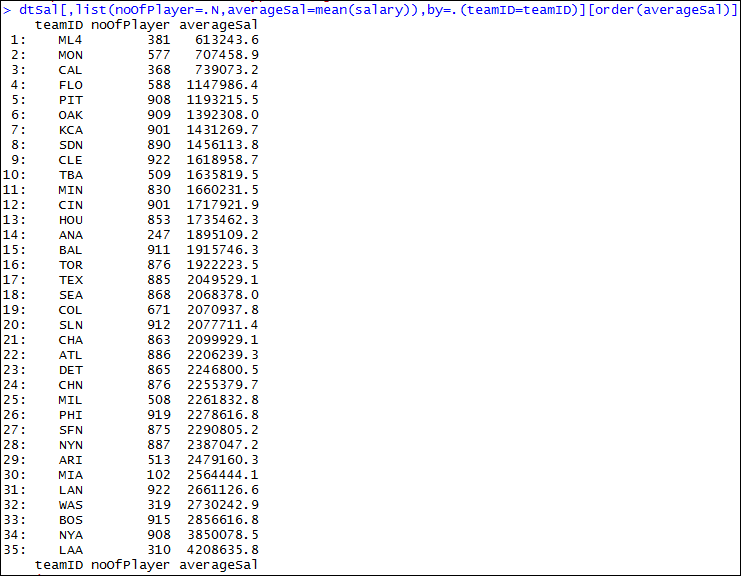
Average Salary of each Team in descending order.

dtSal[,list(noOfPlayer=.N,averageSal=mean(salary)),by=.(teamID=teamID)][order(-averageSal)] #This will output the average salaries for each teamID and order by averageSal in descending order



Average Salary of each Team in ascending order.

dtSal[,list(noOfPlayer=.N,averageSal=mean(salary)),by=.(teamID=teamID)][order(averageSal)] #This will output the average salaries for each teamID and order by averageSal in ascending order

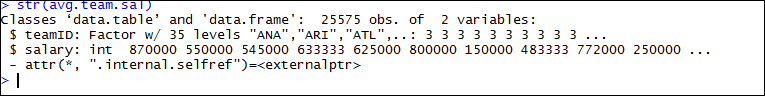


[BACK TO MENU](#_Introduction)

# Subset Functions

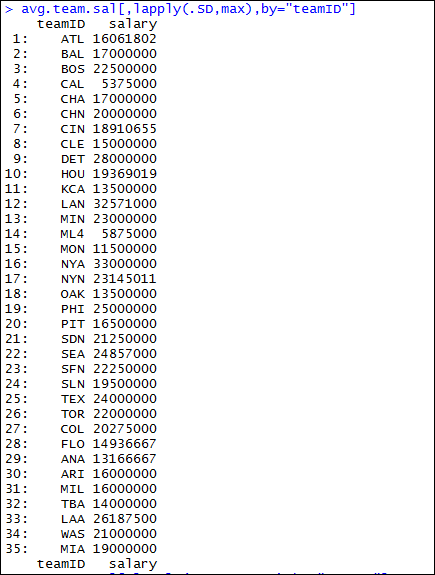
Now let us prepare the dataset salaries for each teamID to compute subset functions.

avg.team.sal=dtSal[,list(salary), by="teamID"]



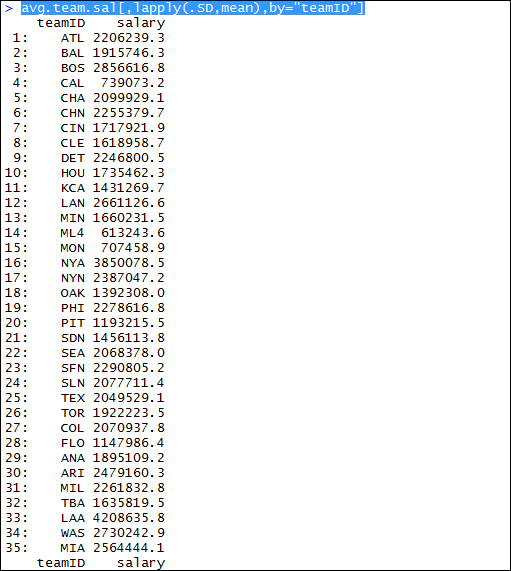
Now let us apply lapply function to get the max salary for the team.

avg.team.sal[,lapply(.SD,max),by="teamID"]



Now let us get the Average salary for each teamID using lapply.

avg.team.sal[,lapply(.SD,mean),by="teamID"]



[BACK TO MENU](#_Introduction)