Term Project Report

DA5020 Collect/Store/Retrieve Data

Harsh Desai
MS in Bioinformatics (Spring 2017)

Premier League 2016-17 Player Database

Motivation: I am a Bioinformatics student, so ideally I should select any biological database for the final term project. However, biological data is big responsibility to work with and I have worked on biological data during my other courses. So, my final choice was to work with sports data and I have decided to work on Premier League 2016-17 player and stats database.

Introduction:

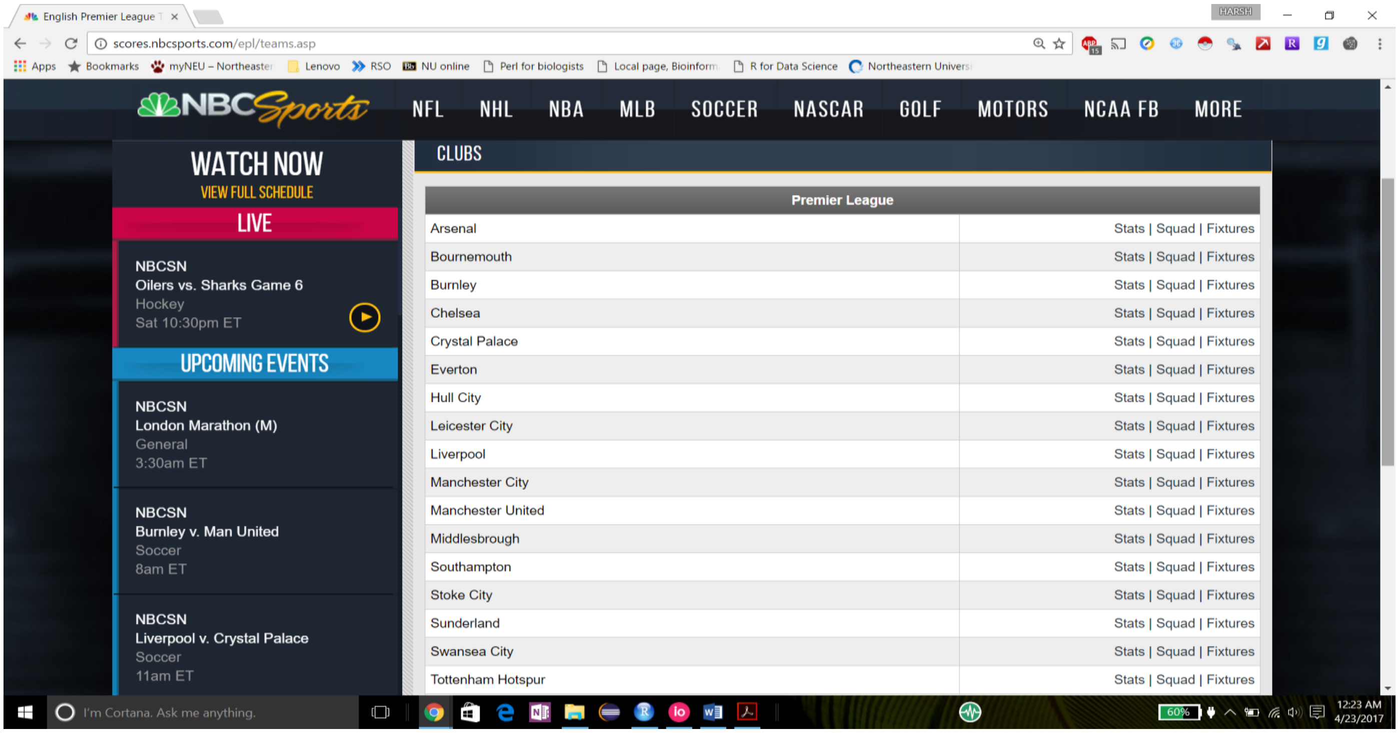
Premier League or English Premier League is a major soccer league contested between top 20 teams in England every year. It is the one of the most followed and major soccer league in the world. For the final term project, I have decided to collect data for the Premier League 2016-17 season, which is also current running season. I have tried to collect most of the possible and significant data for the season.

Work Approach:

1)Collecting 2) Storing 3) Retrieval 4) Graphical Representation

1)Collecting:

For the collection of the data I will be scraping directly from the www.premierleague.com and www.nbcsports.com



WATCH NOW
VIEW FULL SCHEDULE

LIVE

NBCSN
Oilers vs. Sharks Game 6
Hockey
Sat 10:30pm ET

UPCOMING EVENTS

NBCSN
London Marathon (M)
General
3:30am ET

NBCSN
Burnley v. Man United
Soccer
8am ET

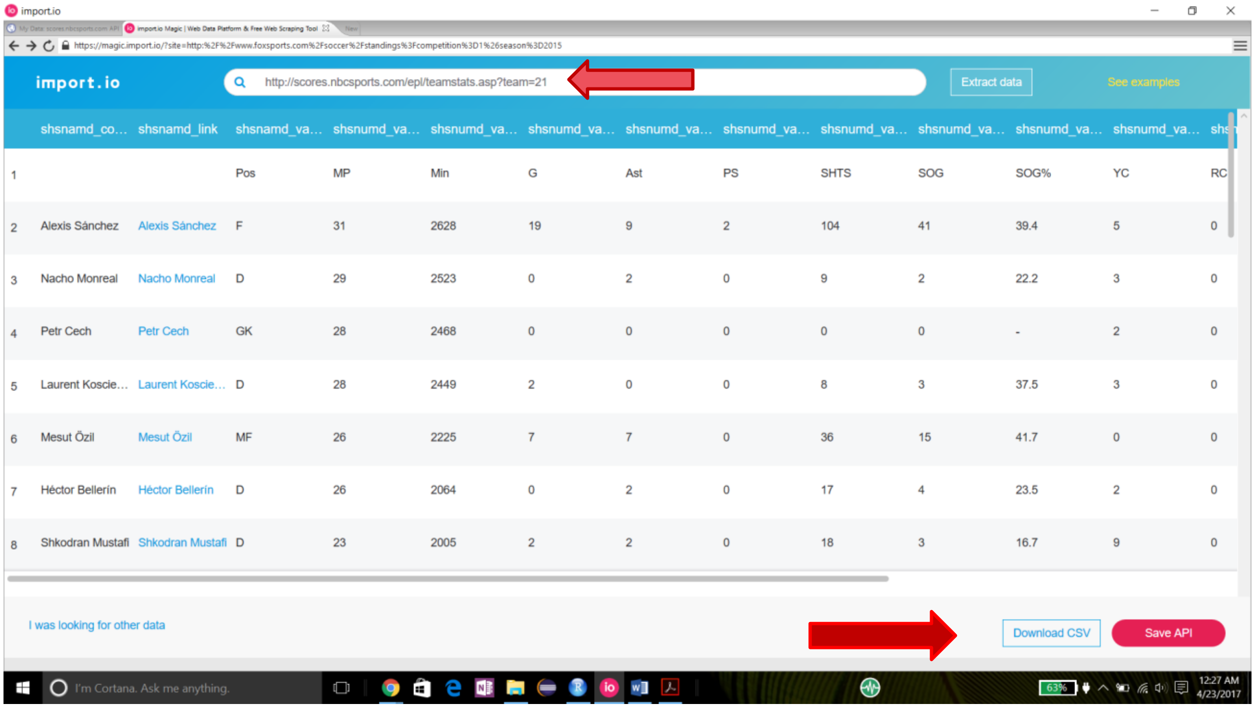
NBCSN
Liverpool v. Crystal Palace
Soccer
11am ET

CLUBS

Premier League	
Arsenal	Stats | Squad | Fixtures
Bournemouth	Stats | Squad | Fixtures
Burnley	Stats | Squad | Fixtures
Chelsea	Stats | Squad | Fixtures
Crystal Palace	Stats | Squad | Fixtures
Everton	Stats | Squad | Fixtures
Hull City	Stats | Squad | Fixtures
Leicester City	Stats | Squad | Fixtures
Liverpool	Stats | Squad | Fixtures
Manchester City	Stats | Squad | Fixtures
Manchester United	Stats | Squad | Fixtures
Middlesbrough	Stats | Squad | Fixtures
Southampton	Stats | Squad | Fixtures
Stoke City	Stats | Squad | Fixtures
Sunderland	Stats | Squad | Fixtures
Swansea City	Stats | Squad | Fixtures
Tottenham Hotspur	Stats | Squad | Fixtures

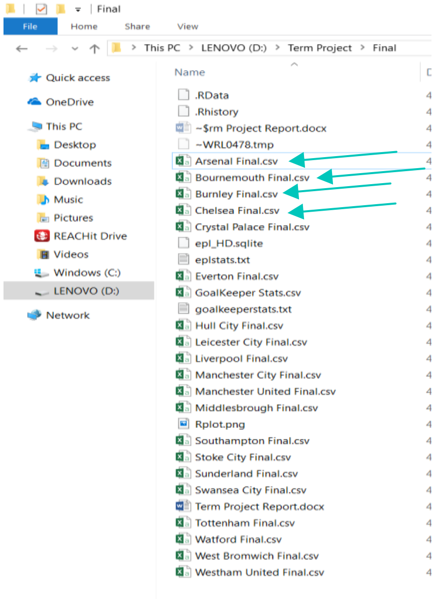
I am using import.io as a scraping tools and URL get query to collect data team wise using team id as URL query.

For example: http://scores.nbcsports.com/epl/teamstats.asp?team=21, here team=21 is a team query for the team called Arsenal. I have extracted data using import.io tool using URL as input.

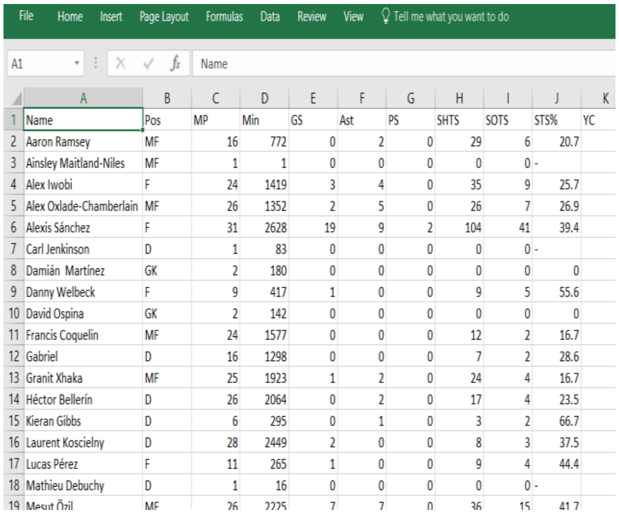

import.io
https://magic.import.io/?site=http:%2F%2Fwww.foxsports.com%2Fsoccer%2Fstandings%3Fcompetition%3D1%26season%3D2015
import.io
http://scores.nbcsports.com/epl/teamstats.asp?team=21
Extract data
See examples
I was looking for other data
Download CSV
Save API
I'm Cortana. Ask me anything.
63%
12:27 AM
4/23/2017

	shsnamd_co…	shsnamd_link	shsnumd_va…	shsnumd_va…	shsnumd_va…	shsnumd_va…	shsnumd_va…	shsnumd_va…	shsnumd_va…	shsnumd_va…	shsnumd_va…	shsnumd_va…	shs
1			Pos	MP	Min	G	Ast	PS	SHTS	SOG	SOG%	YC	RC
2	Alexis Sánchez	Alexis Sánchez	F	31	2628	19	9	2	104	41	39.4	5	0
3	Nacho Monreal	Nacho Monreal	D	29	2523	0	2	0	9	2	22.2	3	0
4	Petr Cech	Petr Cech	GK	28	2468	0	0	0	0	0	-	2	0
5	Laurent Koscie…	Laurent Koscie…	D	28	2449	2	0	0	8	3	37.5	3	0
6	Mesut Özil	Mesut Özil	MF	26	2225	7	7	0	36	15	41.7	0	0
7	Héctor Bellerín	Héctor Bellerín	D	26	2064	0	2	0	17	4	23.5	2	0
8	Shkodran Mustafi	Shkodran Mustafi	D	23	2005	2	2	0	18	3	16.7	9	0

Import.io allow us to store collected data into .csv files. So, I have stored 20 .csv file for each team separately.



.RData
.Rhistory
~$rm Project Report.docx
~WRL0478.tmp
Arsenal Final.csv
Bournemouth Final.csv
Burnley Final.csv
Chelsea Final.csv
Crystal Palace Final.csv
epl_HD.sqlite
eplstats.txt
Everton Final.csv
GoalKeeper Stats.csv
goalkeeperstats.txt
Hull City Final.csv
Leicester City Final.csv
Liverpool Final.csv
Manchester City Final.csv
Manchester United Final.csv
Middlesbrough Final.csv
Rplot.png
Southampton Final.csv
Stoke City Final.csv
Sunderland Final.csv
Swansea City Final.csv
Term Project Report.docx
Tottenham Final.csv
Watford Final.csv
West Bromwich Final.csv
Westham United Final.csv

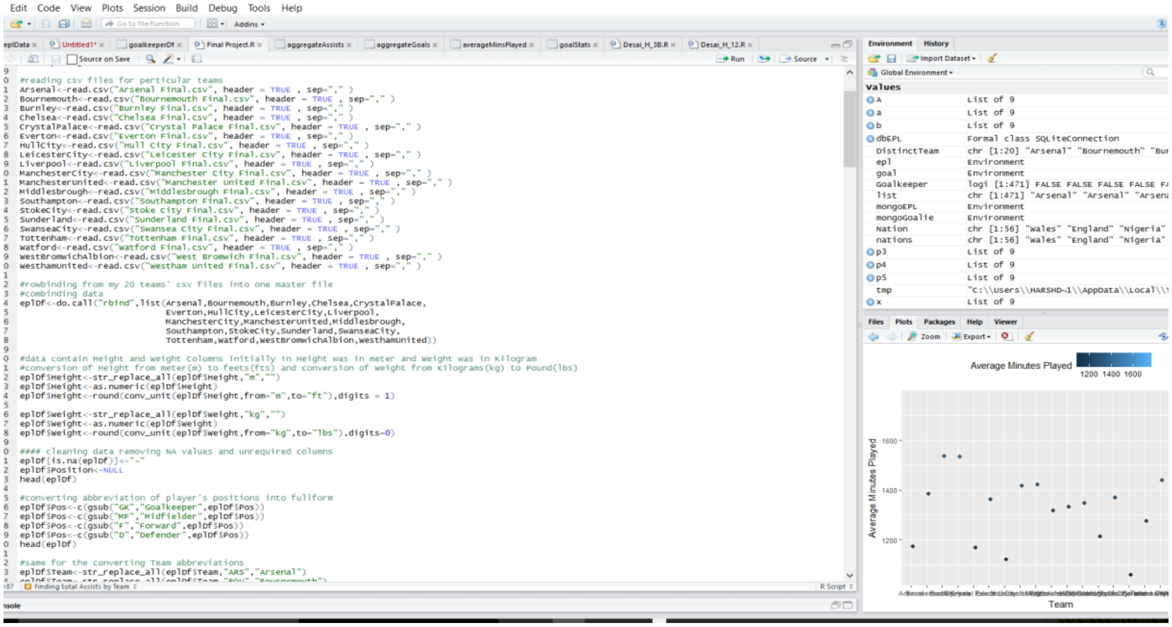


	A	B	C	D	E	F	G	H	I	J	K
1	Name	Pos	MP	Min	GS	Ast	PS	SHTS	SOTS	STS%	YC
2	Aaron Ramsey	MF	16	772	0	2	0	29	6	20.7	
3	Ainsley Maitland-Niles	MF	1	1	0	0	0	0	0	-	
4	Alex Iwobi	F	24	1419	3	4	0	35	9	25.7	
5	Alex Oxlade-Chamberlain	MF	26	1352	2	5	0	26	7	26.9	
6	Alexis Sánchez	F	31	2628	19	9	2	104	41	39.4	
7	Carl Jenkinson	D	1	83	0	0	0	0	0	-	
8	Damián Martínez	GK	2	180	0	0	0	0	0	0	
9	Danny Welbeck	F	9	417	1	0	0	9	5	55.6	
10	David Ospina	GK	2	142	0	0	0	0	0	0	
11	Francis Coquelin	MF	24	1577	0	0	0	12	2	16.7	
12	Gabriel	D	16	1298	0	0	0	7	2	28.6	
13	Granit Xhaka	MF	25	1923	1	2	0	24	4	16.7	
14	Héctor Bellerín	D	26	2064	0	2	0	17	4	23.5	
15	Kieran Gibbs	D	6	295	0	1	0	3	2	66.7	
16	Laurent Koscielny	D	28	2449	2	0	0	8	3	37.5	
17	Lucas Pérez	F	11	265	1	0	0	9	4	44.4	
18	Mathieu Debuchy	D	1	16	0	0	0	0	0	-	
19	Mesut Özil	MF	26	2225	7	7	0	36	15	41.7	

Reading data into R using read.csv() function and combining all 20 files into master file using R functions. I might have combined data manually but I wanted to use R functions that's the main I choose to extract file per teams and not a single master file.

Packages used: stringr ,measurements

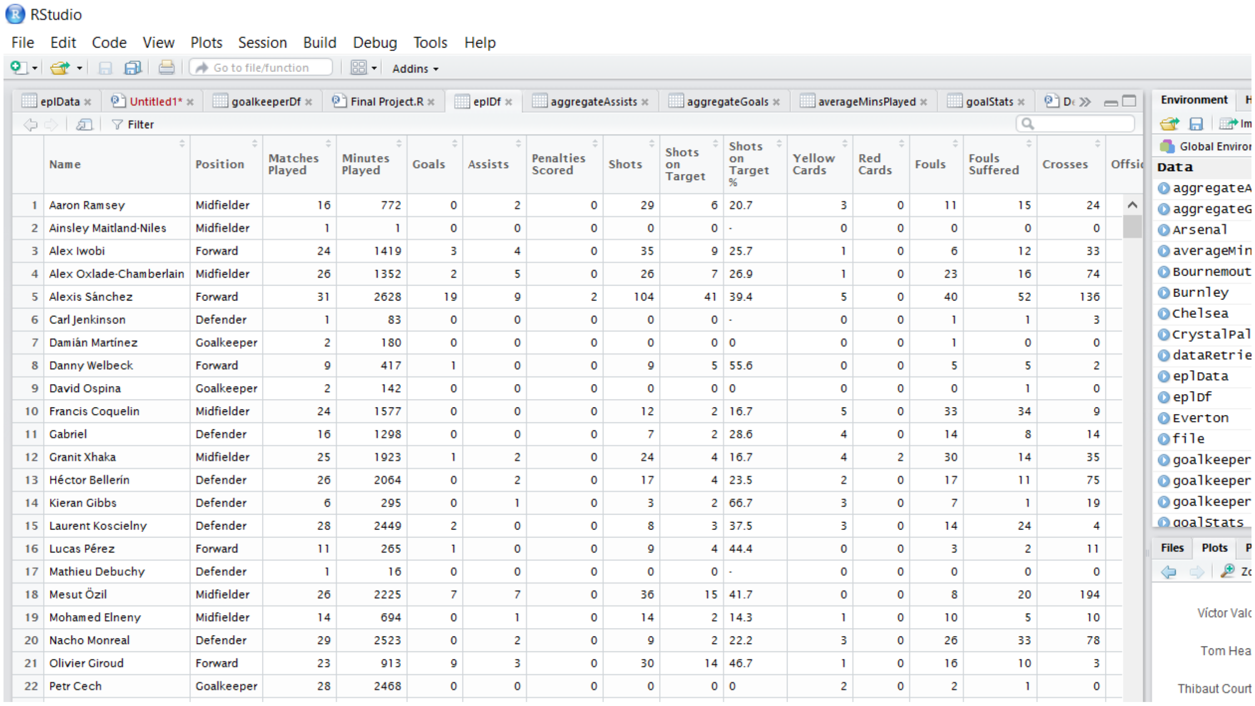
Loading .csv files, combining files and Cleaning data



```
#reading csv files for perticular teams
Arsenal<-read.csv("Arsenal Final.csv", header = TRUE , sep="," )
Bournemouth<-read.csv("Bournemouth Final.csv", header = TRUE , sep="," )
Burnley<-read.csv("Burnley Final.csv", header = TRUE , sep="," )
Chelsea<-read.csv("Chelsea Final.csv", header = TRUE , sep="," )
CrystalPalace<-read.csv("Crystal Palace Final.csv", header = TRUE , sep="," )
Everton<-read.csv("Everton Final.csv", header = TRUE , sep="," )
HullCity<-read.csv("Hull City Final.csv", header = TRUE , sep="," )
LeicesterCity<-read.csv("Leicester City Final.csv", header = TRUE , sep="," )
Liverpool<-read.csv("Liverpool Final.csv", header = TRUE , sep="," )
ManchesterCity<-read.csv("Manchester City Final.csv", header = TRUE , sep="," )
ManchesterUnited<-read.csv("Manchester United Final.csv", header = TRUE , sep="," )
Middlesbrough<-read.csv("Middlesbrough Final.csv", header = TRUE , sep="," )
Southampton<-read.csv("Southampton Final.csv", header = TRUE , sep="," )
StokeCity<-read.csv("Stoke City Final.csv", header = TRUE , sep="," )
Sunderland<-read.csv("Sunderland Final.csv", header = TRUE , sep="," )
SwanseaCity<-read.csv("Swansea City Final.csv", header = TRUE , sep="," )
Tottenham<-read.csv("Tottenham Final.csv", header = TRUE , sep="," )
Watford<-read.csv("Watford Final.csv", header = TRUE , sep="," )
WestBromwichAlbion<-read.csv("West Bromwich Final.csv", header = TRUE , sep="," )
WesthamUnited<-read.csv("Westham United Final.csv", header = TRUE , sep="," )

#rowbinding from my 20 teams' csv files into one master file
#combinding data
eplDf<-do.call("rbind",list(Arsenal,Bournemouth,Burnley,Chelsea,CrystalPalace,
                            Everton,HullCity,LeicesterCity,Liverpool,
                            ManchesterCity,ManchesterUnited,Middlesbrough,
                            Southampton,StokeCity,Sunderland,SwanseaCity,
                            Tottenham,Watford,WestBromwichAlbion,WesthamUnited))

#data contain Height and Weight Columns initially in Height was in meter and Weight was in Kilogram
#conversion of Height from meter(m) to feets(fts) and conversion of Weight from Kilograms(kg) to Pound(lbs)
eplDf$Height<-str_replace_all(eplDf$Height,"m","")
eplDf$Height<-as.numeric(eplDf$Height)
eplDf$Height<-round(conv_unit(eplDf$Height,from="m",to="ft"),digits = 1)

eplDf$Weight<-str_replace_all(eplDf$Weight,"kg","")
eplDf$Weight<-as.numeric(eplDf$Weight)
eplDf$Weight<-round(conv_unit(eplDf$Weight,from="kg",to="lbs"),digits=0)

#### cleaning data removing NA values and unrequired columns
eplDf[is.na(eplDf)]<-"-"
eplDf$Position<-NULL
head(eplDf)

#converting abbreviation of player's positions into fullform
eplDf$Pos<-c(gsub("GK","Goalkeeper",eplDf$Pos))
eplDf$Pos<-c(gsub("MF","Midfielder",eplDf$Pos))
eplDf$Pos<-c(gsub("F","Forward",eplDf$Pos))
eplDf$Pos<-c(gsub("D","Defender",eplDf$Pos))
head(eplDf)

#same for the converting Team abbreviations
eplDf$Team<-str_replace_all(eplDf$Team,"ARS","Arsenal")
eplDf$Team<-str_replace_all(eplDf$Team,"BOU","Bournemouth")
```

Values	
A	List of 9
a	List of 9
b	List of 9
dbEPL	Formal class SQLiteConnection
DistinctTeam	chr [1:20] "Arsenal" "Bournemouth" "Bur
epl	Environment
goal	Environment
Goalkeeper	logi [1:471] FALSE FALSE FALSE FALSE FA
list	chr [1:471] "Arsenal" "Arsenal" "Arsena
mongoEPL	Environment
mongoGoalie	Environment
Nation	chr [1:56] "Wales" "England" "Nigeria"
nations	chr [1:56] "Wales" "England" "Nigeria"
p3	List of 9
p4	List of 9
p5	List of 9
tmp	"C:\\Users\\HARSHD~1\\AppData\\Local\\T
x	List of 9

Combined final data frame :



	Name	Position	Matches Played	Minutes Played	Goals	Assists	Penalties Scored	Shots	Shots on Target	Shots on Target %	Yellow Cards	Red Cards	Fouls	Fouls Suffered	Crosses	Offsi
1	Aaron Ramsey	Midfielder	16	772	0	2	0	29	6	20.7	3	0	11	15	24	
2	Ainsley Maitland-Niles	Midfielder	1	1	0	0	0	0	0	-	0	0	0	0	0	
3	Alex Iwobi	Forward	24	1419	3	4	0	35	9	25.7	1	0	6	12	33	
4	Alex Oxlade-Chamberlain	Midfielder	26	1352	2	5	0	26	7	26.9	1	0	23	16	74	
5	Alexis Sánchez	Forward	31	2628	19	9	2	104	41	39.4	5	0	40	52	136	
6	Carl Jenkinson	Defender	1	83	0	0	0	0	0	-	0	0	1	1	3	
7	Damián Martínez	Goalkeeper	2	180	0	0	0	0	0	0	0	0	1	0	0	
8	Danny Welbeck	Forward	9	417	1	0	0	9	5	55.6	0	0	5	5	2	
9	David Ospina	Goalkeeper	2	142	0	0	0	0	0	0	0	0	0	1	0	
10	Francis Coquelin	Midfielder	24	1577	0	0	0	12	2	16.7	5	0	33	34	9	
11	Gabriel	Defender	16	1298	0	0	0	7	2	28.6	4	0	14	8	14	
12	Granit Xhaka	Midfielder	25	1923	1	2	0	24	4	16.7	4	2	30	14	35	
13	Héctor Bellerín	Defender	26	2064	0	2	0	17	4	23.5	2	0	17	11	75	
14	Kieran Gibbs	Defender	6	295	0	1	0	3	2	66.7	3	0	7	1	19	
15	Laurent Koscielny	Defender	28	2449	2	0	0	8	3	37.5	3	0	14	24	4	
16	Lucas Pérez	Forward	11	265	1	0	0	9	4	44.4	0	0	3	2	11	
17	Mathieu Debuchy	Defender	1	16	0	0	0	0	0	-	0	0	0	0	0	
18	Mesut Özil	Midfielder	26	2225	7	7	0	36	15	41.7	0	0	8	20	194	
19	Mohamed Elneny	Midfielder	14	694	0	1	0	14	2	14.3	1	0	10	5	10	
20	Nacho Monreal	Defender	29	2523	0	2	0	9	2	22.2	3	0	26	33	78	
21	Olivier Giroud	Forward	23	913	9	3	0	30	14	46.7	1	0	16	10	3	
22	Petr Cech	Goalkeeper	28	2468	0	0	0	0	0	0	2	0	2	1	0	

Collected Variables:

Basic Player Info: Name, Position, Team, Matches Played, Minutes Played, Nationality etc

Standard Stats: Goals, Assists, Shots, Yellow Cards, Red Cards, Crosses, Offside etc.

Goalkeeper Stats: Saves, Saves, Penalty Kicks Saved, Goals Allowed, Shots faced etc.

2)Storing:

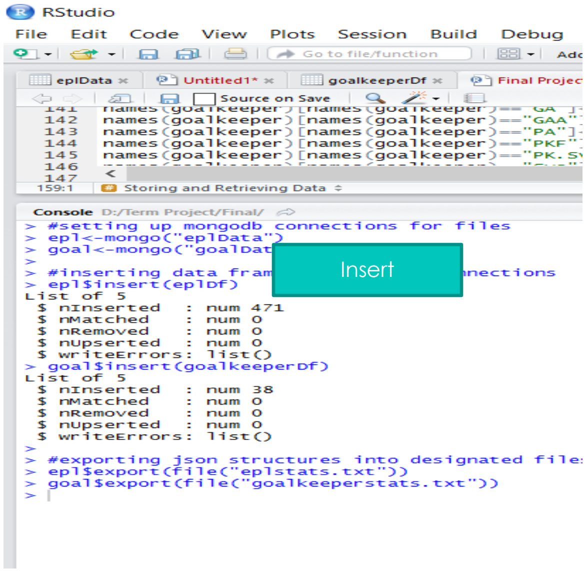
Initially, I was planning to store collected data using Relational Database. After working on relational database schema, I realized that it will be difficult to store data into relational database due to complexity between player positions and their statistics. For example, Goalkeeper statistics attributes are totally different than other player position's attribute for soccer. So, I decided to go with non-relational (NoSQL) database system to store collected data. It is simpler than relational database system, other advantages of NoSQL database system are no unique constrain, massive parallel processing and tolerance for failure. Non-relational system I have used for project is MongoDB.

Packages used: mongolite

Steps for running MongoDB server
1.installing MongoDB from mongodb.com
2.Setting up MongoDB environment
mkdir \data\db
"C:\Program Files\MongoDB\Server\3.4\bin\mongod.exe" --dbpath d:\test\mongodb\data
3. C:\MongoDB\Server\3.2\bin\mongod.exe   Code to start MongoDB server
3. Connect using R

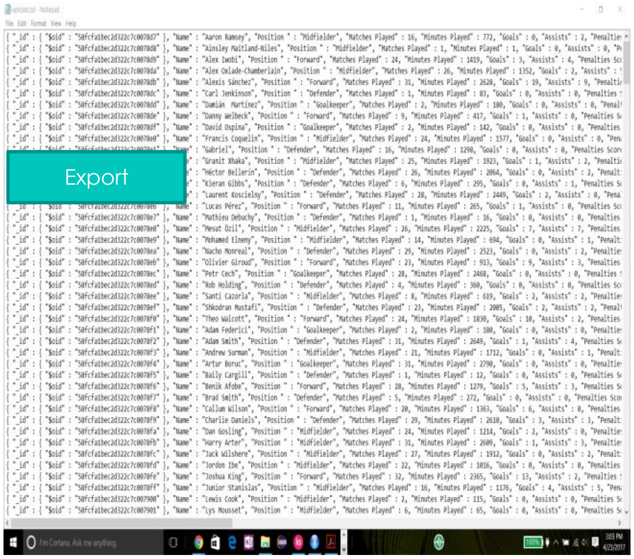
After connecting MongoDB server with R studio using mongolite R package, next step is to insert collected and cleaned data into MongoDB system using insert() function provided my mongolite. Mongolite also provides option of exporting JSON structure to file using export() function.

Setting MongoDB connections, inserting files and exporting JSON structure to designated files.


RStudio
File Edit Code View Plots Session Build Debug
Go to file/function
eplData
Untitled1*
goalkeeperDf
Final Projec
Source on Save
159:1
Storing and Retrieving Data
Console D:/Term Project/Final/
Insert

```
names(goalkeeper)[names(goalkeeper)=="GA"]
names(goalkeeper)[names(goalkeeper)=="GAA"]
names(goalkeeper)[names(goalkeeper)=="PA"]
names(goalkeeper)[names(goalkeeper)=="PKF"]
names(goalkeeper)[names(goalkeeper)=="PK.S
```

```
> #setting up mongodb connections for files
> epl<-mongo("eplData")
> goal<-mongo("goalDat
>
> #inserting data fram                nections
> epl$insert(eplDf)
List of 5
 $ nInserted   : num 471
 $ nMatched    : num 0
 $ nRemoved    : num 0
 $ nUpserted   : num 0
 $ writeErrors: list()
> goal$insert(goalkeeperDf)
List of 5
 $ nInserted   : num 38
 $ nMatched    : num 0
 $ nRemoved    : num 0
 $ nUpserted   : num 0
 $ writeErrors: list()
>
> #exporting json structures into designated file:
> epl$export(file("eplstats.txt"))
> goal$export(file("goalkeeperstats.txt"))
>
```



```
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078d7" }, "Name" : "Aaron Ramsey", "Position " : "Midfielder", "Matches Played" : 16, "Minutes Played" : 772, "Goals" : 0, "Assists" : 2, "Penaltie
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078d8" }, "Name" : "Ainsley Maitland-Niles", "Position " : "Midfielder", "Matches Played" : 1, "Minutes Played" : 1, "Goals" : 0, "Assists" : 0, "Pe
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078d9" }, "Name" : "Alex Iwobi", "Position " : "Forward", "Matches Played" : 24, "Minutes Played" : 1419, "Goals" : 3, "Assists" : 4, "Penalties Sco
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078da" }, "Name" : "Alex Oxlade-Chamberlain", "Position " : "Midfielder", "Matches Played" : 26, "Minutes Played" : 1352, "Goals" : 2, "Assists" : :
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078db" }, "Name" : "Alexis Sánchez", "Position " : "Forward", "Matches Played" : 31, "Minutes Played" : 2628, "Goals" : 19, "Assists" : 9, "Penalti
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078dc" }, "Name" : "Carl Jenkinson", "Position " : "Defender", "Matches Played" : 1, "Minutes Played" : 83, "Goals" : 0, "Assists" : 0, "Penalties :
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078dd" }, "Name" : "Damián  Martínez", "Position " : "Goalkeeper", "Matches Played" : 2, "Minutes Played" : 180, "Goals" : 0, "Assists" : 0, "Penal
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078de" }, "Name" : "Danny Welbeck", "Position " : "Forward", "Matches Played" : 9, "Minutes Played" : 417, "Goals" : 1, "Assists" : 0, "Penalties S
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078df" }, "Name" : "David Ospina", "Position " : "Goalkeeper", "Matches Played" : 2, "Minutes Played" : 142, "Goals" : 0, "Assists" : 0, "Penalties
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078e0" }, "Name" : "Francis Coquelin", "Position " : "Midfielder", "Matches Played" : 24, "Minutes Played" : 1577, "Goals" : 0, "Assists" : 0, "Pen
                                                     "Gabriel", "Position " : "Defender", "Matches Played" : 16, "Minutes Played" : 1298, "Goals" : 0, "Assists" : 0, "Penalties Scor
                                                   e" : "Granit Xhaka", "Position " : "Midfielder", "Matches Played" : 25, "Minutes Played" : 1923, "Goals" : 1, "Assists" : 2, "Penalti
                                                   e" : "Héctor Bellerín", "Position " : "Defender", "Matches Played" : 26, "Minutes Played" : 2064, "Goals" : 0, "Assists" : 2, "Penalt
                                                   e" : "Kieran Gibbs", "Position " : "Defender", "Matches Played" : 6, "Minutes Played" : 295, "Goals" : 0, "Assists" : 1, "Penalties S
                                                   e" : "Laurent Koscielny", "Position " : "Defender", "Matches Played" : 28, "Minutes Played" : 2449, "Goals" : 2, "Assists" : 0, "Pena
                                                     : "Lucas Pérez", "Position " : "Forward", "Matches Played" : 11, "Minutes Played" : 265, "Goals" : 1, "Assists" : 0, "Penalties Sco
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078e7" }, "Name" : "Mathieu Debuchy", "Position " : "Defender", "Matches Played" : 1, "Minutes Played" : 16, "Goals" : 0, "Assists" : 0, "Penalties
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078e8" }, "Name" : "Mesut Özil", "Position " : "Midfielder", "Matches Played" : 26, "Minutes Played" : 2225, "Goals" : 7, "Assists" : 7, "Penalties
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078e9" }, "Name" : "Mohamed Elneny", "Position " : "Midfielder", "Matches Played" : 14, "Minutes Played" : 694, "Goals" : 0, "Assists" : 1, "Penalt
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078ea" }, "Name" : "Nacho Monreal", "Position " : "Defender", "Matches Played" : 29, "Minutes Played" : 2523, "Goals" : 0, "Assists" : 2, "Penaltie
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078eb" }, "Name" : "Olivier Giroud", "Position " : "Forward", "Matches Played" : 23, "Minutes Played" : 913, "Goals" : 9, "Assists" : 3, "Penalties
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078ec" }, "Name" : "Petr Cech", "Position " : "Goalkeeper", "Matches Played" : 28, "Minutes Played" : 2468, "Goals" : 0, "Assists" : 0, "Penalties :
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078ed" }, "Name" : "Rob Holding", "Position " : "Defender", "Matches Played" : 4, "Minutes Played" : 360, "Goals" : 0, "Assists" : 0, "Penalties Sco
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078ee" }, "Name" : "Santi Cazorla", "Position " : "Midfielder", "Matches Played" : 8, "Minutes Played" : 619, "Goals" : 2, "Assists" : 2, "Penaltie
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078ef" }, "Name" : "Shkodran Mustafi", "Position " : "Defender", "Matches Played" : 23, "Minutes Played" : 2005, "Goals" : 2, "Assists" : 2, "Penal
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f0" }, "Name" : "Theo Walcott", "Position " : "Forward", "Matches Played" : 24, "Minutes Played" : 1830, "Goals" : 10, "Assists" : 2, "Penalties
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f1" }, "Name" : "Adam Federici", "Position " : "Goalkeeper", "Matches Played" : 2, "Minutes Played" : 180, "Goals" : 0, "Assists" : 0, "Penaltie
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f2" }, "Name" : "Adam Smith", "Position " : "Defender", "Matches Played" : 31, "Minutes Played" : 2649, "Goals" : 1, "Assists" : 4, "Penalties S
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f3" }, "Name" : "Andrew Surman", "Position " : "Midfielder", "Matches Played" : 21, "Minutes Played" : 1712, "Goals" : 0, "Assists" : 1, "Penalt
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f4" }, "Name" : "Artur Boruc", "Position " : "Goalkeeper", "Matches Played" : 31, "Minutes Played" : 2790, "Goals" : 0, "Assists" : 0, "Penaltie
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f5" }, "Name" : "Baily Cargill", "Position " : "Defender", "Matches Played" : 1, "Minutes Played" : 12, "Goals" : 0, "Assists" : 0, "Penalties S
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f6" }, "Name" : "Benik Afobe", "Position " : "Forward", "Matches Played" : 28, "Minutes Played" : 1279, "Goals" : 5, "Assists" : 3, "Penalties S
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f7" }, "Name" : "Brad Smith", "Position " : "Defender", "Matches Played" : 5, "Minutes Played" : 272, "Goals" : 0, "Assists" : 0, "Penalties Scor
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f8" }, "Name" : "Callum Wilson", "Position " : "Forward", "Matches Played" : 20, "Minutes Played" : 1363, "Goals" : 6, "Assists" : 0, "Penalties
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078f9" }, "Name" : "Charlie Daniels", "Position " : "Defender", "Matches Played" : 29, "Minutes Played" : 2610, "Goals" : 3, "Assists" : 3, "Penalt
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078fa" }, "Name" : "Dan Gosling", "Position " : "Midfielder", "Matches Played" : 24, "Minutes Played" : 1214, "Goals" : 2, "Assists" : 0, "Penaltie
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078fb" }, "Name" : "Harry Arter", "Position " : "Midfielder", "Matches Played" : 31, "Minutes Played" : 2609, "Goals" : 1, "Assists" : 3, "Penaltie
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078fc" }, "Name" : "Jack Wilshere", "Position " : "Midfielder", "Matches Played" : 27, "Minutes Played" : 1912, "Goals" : 0, "Assists" : 2, "Penalt
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078fd" }, "Name" : "Jordon Ibe", "Position " : "Midfielder", "Matches Played" : 22, "Minutes Played" : 1016, "Goals" : 0, "Assists" : 0, "Penalties
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078fe" }, "Name" : "Joshua King", "Position " : "Forward", "Matches Played" : 32, "Minutes Played" : 2365, "Goals" : 13, "Assists" : 2, "Penalties :
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c0078ff" }, "Name" : "Junior Stanislas", "Position " : "Midfielder", "Matches Played" : 16, "Minutes Played" : 1176, "Goals" : 4, "Assists" : 5, "Pen
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c007900" }, "Name" : "Lewis Cook", "Position " : "Midfielder", "Matches Played" : 2, "Minutes Played" : 115, "Goals" : 0, "Assists" : 0, "Penalties S
{ "_id" : { "$oid" : "58fcfa1bec2d322c7c007901" }, "Name" : "Lys Mousset", "Position " : "Midfielder", "Matches Played" : 6, "Minutes Played" : 65, "Goals" : 0, "Assists" : 0, "Penalties S
```

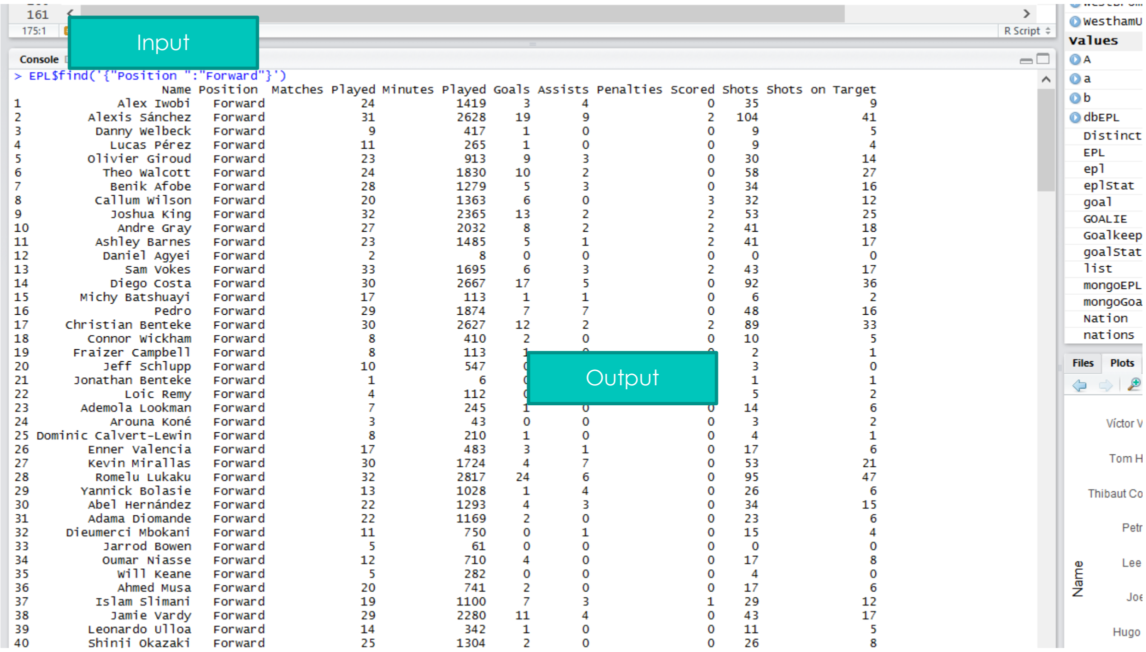
3.Retrieval:

Functions used : sort(),find(),distinct()

Packages used: dplyr,ggplot2,mongolite

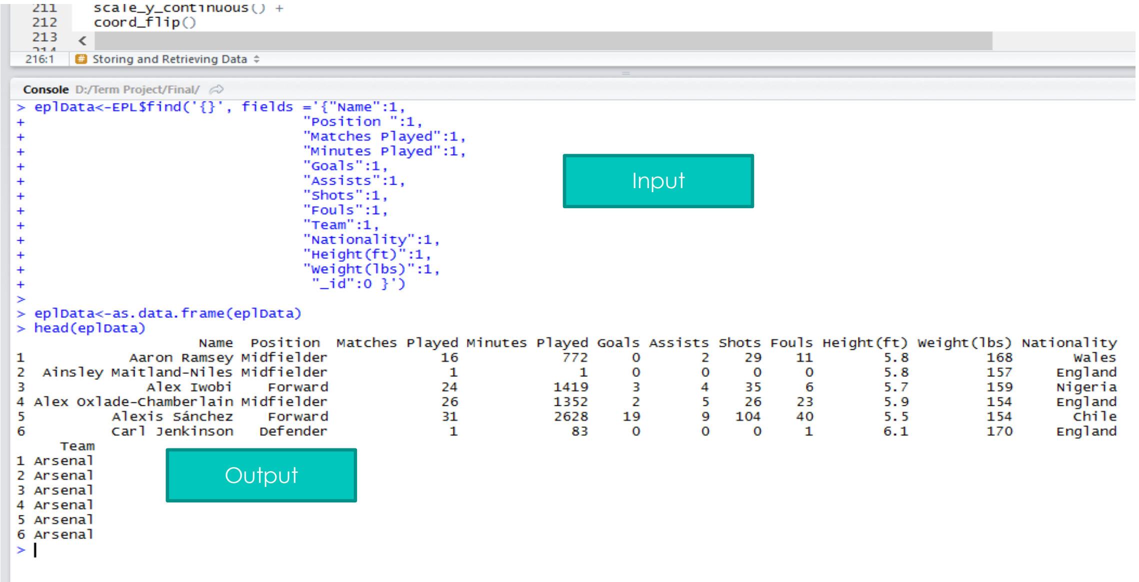
find() function is used to find specific value from the stored data.

Firstly using find() function to search all the forward position players from the database.



```
> EPL$find('{"Position ":"Forward"}')
                    Name Position  Matches Played Minutes Played Goals Assists Penalties Scored Shots Shots on Target
1             Alex Iwobi   Forward             24           1419     3       4                0    35               9
2         Alexis Sánchez   Forward             31           2628    19       9                2   104              41
3          Danny Welbeck   Forward              9            417     1       0                0     9               5
4            Lucas Pérez   Forward             11            265     1       0                0     9               4
5         Olivier Giroud   Forward             23            913     9       3                0    30              14
6           Theo Walcott   Forward             24           1830    10       2                0    58              27
7            Benik Afobe   Forward             28           1279     5       3                0    34              16
8          Callum Wilson   Forward             20           1363     6       0                3    32              12
9            Joshua King   Forward             32           2365    13       2                2    53              25
10            Andre Gray   Forward             27           2032     8       2                2    41              18
11         Ashley Barnes   Forward             23           1485     5       1                2    41              17
12          Daniel Agyei   Forward              2              8     0       0                0     0               0
13             Sam Vokes   Forward             33           1695     6       3                2    43              17
14           Diego Costa   Forward             30           2667    17       5                0    92              36
15       Michy Batshuayi   Forward             17            113     1       1                0     6               2
16                 Pedro   Forward             29           1874     7       7                0    48              16
17     Christian Benteke   Forward             30           2627    12       2                2    89              33
18        Connor Wickham   Forward              8            410     2       0                0    10               5
19      Fraizer Campbell   Forward              8            113     1       0                0     2               1
20          Jeff Schlupp   Forward             10            547     0                              3               0
21      Jonathan Benteke   Forward              1              6     0                              1               1
22             Loic Remy   Forward              4            112     0                              5               2
23       Ademola Lookman   Forward              7            245     1       0                0    14               6
24           Arouna Koné   Forward              3             43     0       0                0     3               2
25 Dominic Calvert-Lewin   Forward              8            210     1       0                0     4               1
26        Enner Valencia   Forward             17            483     3       1                0    17               6
27        Kevin Mirallas   Forward             30           1724     4       7                0    53              21
28         Romelu Lukaku   Forward             32           2817    24       6                0    95              47
29       Yannick Bolasie   Forward             13           1028     1       4                0    26               6
30        Abel Hernández   Forward             22           1293     4       3                0    34              15
31        Adama Diomande   Forward             22           1169     2       0                0    23               6
32     Dieumerci Mbokani   Forward             11            750     0       1                0    15               4
33          Jarrod Bowen   Forward              5             61     0       0                0     0               0
34          Oumar Niasse   Forward             12            710     4       0                0    17               8
35            Will Keane   Forward              5            282     0       0                0     4               0
36            Ahmed Musa   Forward             20            741     2       0                0    17               6
37         Islam Slimani   Forward             19           1100     7       3                1    29              12
38           Jamie Vardy   Forward             29           2280    11       4                0    43              17
39        Leonardo Ulloa   Forward             14            342     1       0                0    11               5
40        Shinji Okazaki   Forward             25           1304     2       0                0    26               8
```

Using find() function to fetch multiple columns from the database and to make data frame

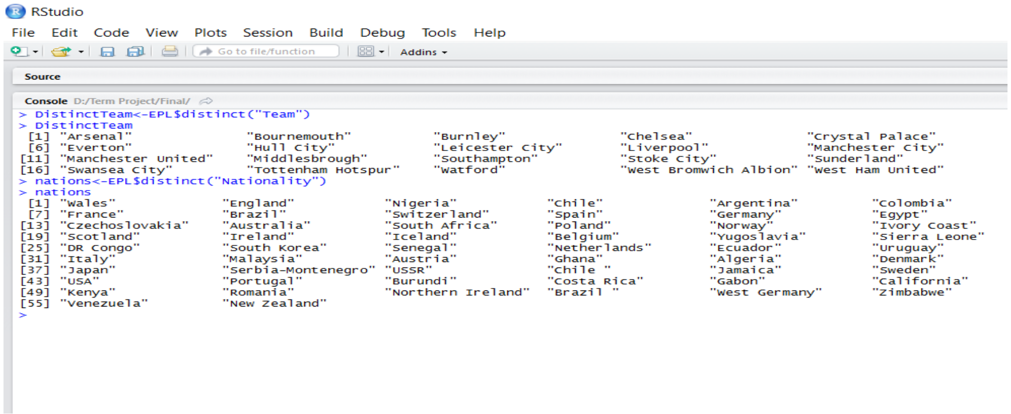


```
scale_y_continuous() +
coord_flip()
```

```
> eplData<-EPL$find('{}', fields ='{"Name":1,
+                                  "Position ":1,
+                                  "Matches Played":1,
+                                  "Minutes Played":1,
+                                  "Goals":1,
+                                  "Assists":1,
+                                  "Shots":1,
+                                  "Fouls":1,
+                                  "Team":1,
+                                  "Nationality":1,
+                                  "Height(ft)":1,
+                                  "Weight(lbs)":1,
+                                   "_id":0 }')
>
> eplData<-as.data.frame(eplData)
> head(eplData)
                     Name  Position  Matches Played Minutes Played Goals Assists Shots Fouls Height(ft) Weight(lbs) Nationality
1            Aaron Ramsey Midfielder             16            772     0       2    29    11        5.8         168       Wales
2   Ainsley Maitland-Niles Midfielder              1              1     0       0     0     0        5.8         157     England
3               Alex Iwobi    Forward             24           1419     3       4    35     6        5.7         159     Nigeria
4 Alex Oxlade-Chamberlain Midfielder             26           1352     2       5    26    23        5.9         154     England
5           Alexis Sánchez    Forward             31           2628    19       9   104    40        5.5         154       Chile
6           Carl Jenkinson   Defender              1             83     0       0     0     1        6.1         170     England
     Team
1 Arsenal
2 Arsenal
3 Arsenal
4 Arsenal
5 Arsenal
6 Arsenal
> |
```

The other function used for retrieval was distinct(),which will find distinct values from particular columns.

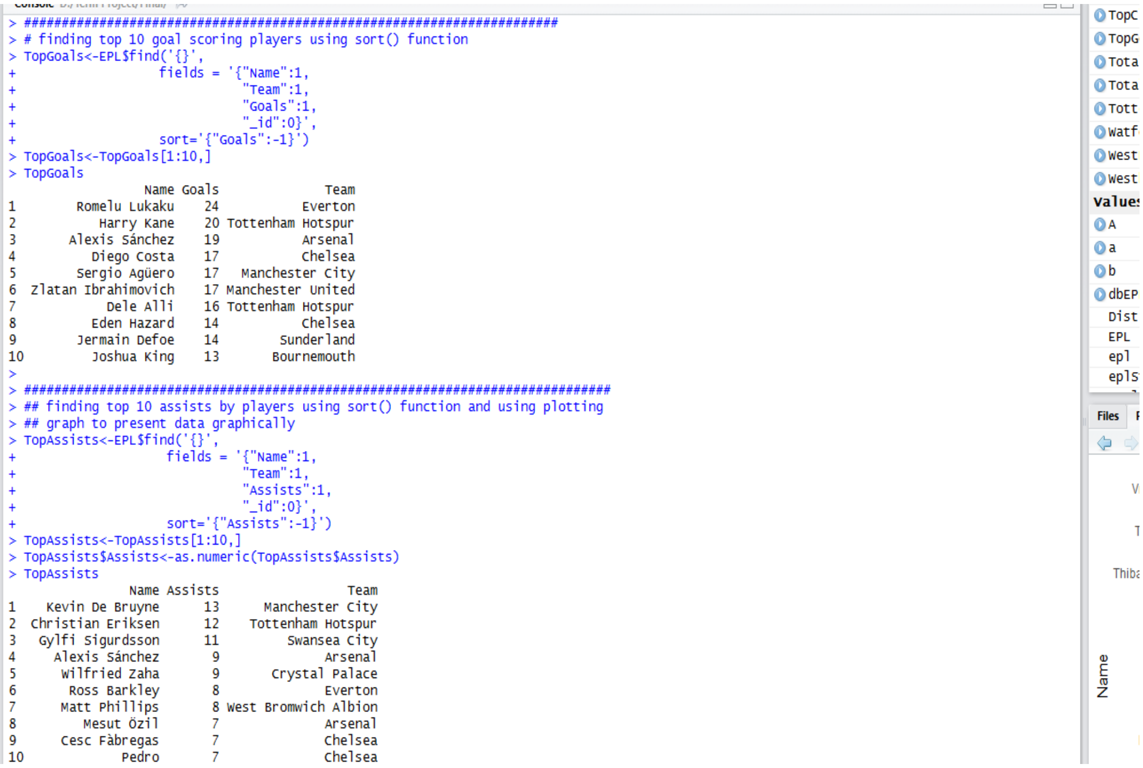
Use of distinct() function to find distinct teams and players from different nations.


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
Source
Console D:/Term Project/Final/

```
> DistinctTeam<-EPL$distinct("Team")
> DistinctTeam
 [1] "Arsenal"              "Bournemouth"          "Burnley"              "Chelsea"              "Crystal Palace"
 [6] "Everton"              "Hull City"            "Leicester City"       "Liverpool"            "Manchester City"
[11] "Manchester United"    "Middlesbrough"        "Southampton"          "Stoke City"           "Sunderland"
[16] "Swansea City"         "Tottenham Hotspur"    "Watford"              "West Bromwich Albion" "West Ham United"
> nations<-EPL$distinct("Nationality")
> nations
 [1] "Wales"              "England"            "Nigeria"            "Chile"              "Argentina"          "Colombia"
 [7] "France"             "Brazil"             "Switzerland"        "Spain"              "Germany"            "Egypt"
[13] "Czechoslovakia"     "Australia"          "South Africa"       "Poland"             "Norway"             "Ivory Coast"
[19] "Scotland"           "Ireland"            "Iceland"            "Belgium"            "Yugoslavia"         "Sierra Leone"
[25] "DR Congo"           "South Korea"        "Senegal"            "Netherlands"        "Ecuador"            "Uruguay"
[31] "Italy"              "Malaysia"           "Austria"            "Ghana"              "Algeria"            "Denmark"
[37] "Japan"              "Serbia-Montenegro"  "USSR"               "Chile "             "Jamaica"            "Sweden"
[43] "USA"                "Portugal"           "Burundi"            "Costa Rica"         "Gabon"              "California"
[49] "Kenya"              "Romania"            "Northern Ireland"   "Brazil "            "West Germany"       "Zimbabwe"
[55] "Venezuela"          "New Zealand"
>
```

We can see that there are 20 distinct teams in a database and players are from 56 distinct countries.

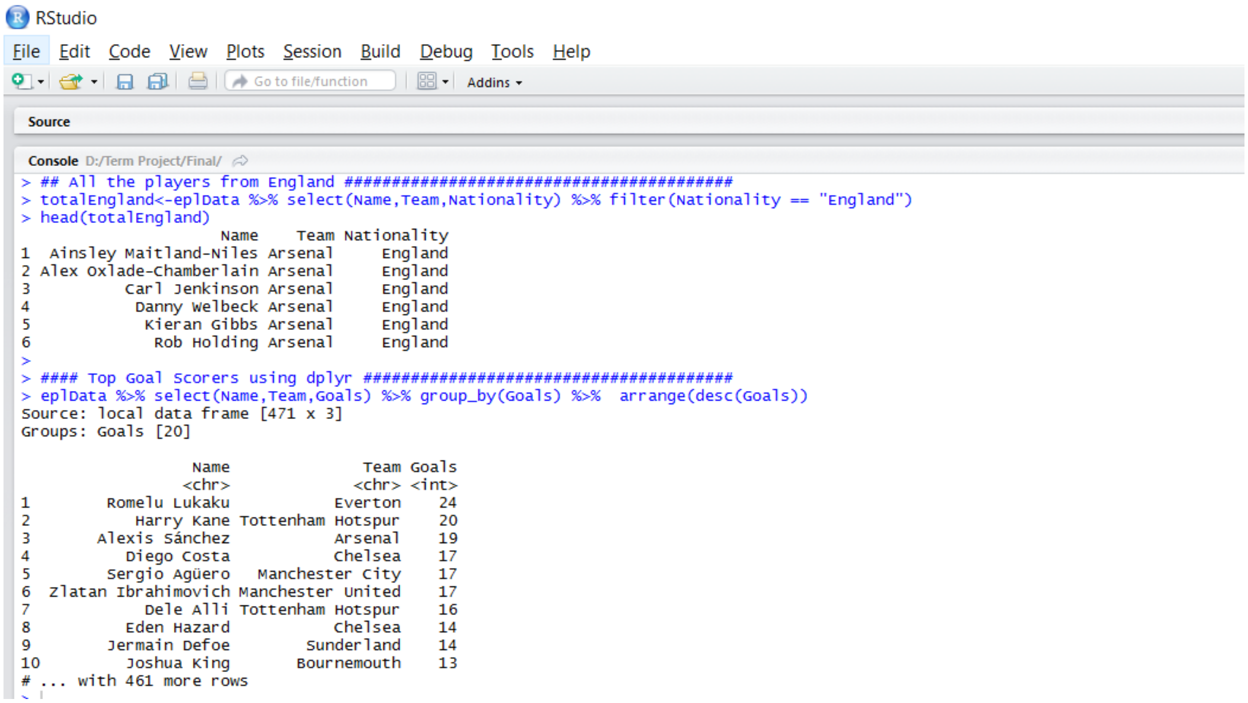
Next function I have used for the retrieval is sort()function. I have used sort()function to find top 10 Goal scorers and top 10 Assists by players.



```
> ########################################################################
> # finding top 10 goal scoring players using sort() function
> TopGoals<-EPL$find('{}',
+                    fields = '{"Name":1,
+                               "Team":1,
+                               "Goals":1,
+                               "_id":0}',
+                    sort='{"Goals":-1}')
> TopGoals<-TopGoals[1:10,]
> TopGoals
                Name Goals              Team
1      Romelu Lukaku    24           Everton
2         Harry Kane    20 Tottenham Hotspur
3     Alexis Sánchez    19           Arsenal
4        Diego Costa    17           Chelsea
5      Sergio Agüero    17   Manchester City
6  Zlatan Ibrahimovich  17 Manchester United
7          Dele Alli    16 Tottenham Hotspur
8        Eden Hazard    14           Chelsea
9      Jermain Defoe    14        Sunderland
10       Joshua King    13       Bournemouth
>
> ###############################################################################
> ## finding top 10 assists by players using sort() function and using plotting
> ## graph to present data graphically
> TopAssists<-EPL$find('{}',
+                     fields = '{"Name":1,
+                               "Team":1,
+                               "Assists":1,
+                               "_id":0}',
+                     sort='{"Assists":-1}')
> TopAssists<-TopAssists[1:10,]
> TopAssists$Assists<-as.numeric(TopAssists$Assists)
> TopAssists
                 Name Assists                 Team
1      Kevin De Bruyne      13      Manchester City
2   Christian Eriksen      12    Tottenham Hotspur
3    Gylfi Sigurdsson      11         Swansea City
4      Alexis Sánchez       9              Arsenal
5      Wilfried Zaha        9       Crystal Palace
6       Ross Barkley        8              Everton
7       Matt Phillips       8 West Bromwich Albion
8          Mesut Özil       7              Arsenal
9       Cesc Fàbregas       7              Chelsea
10              Pedro       7              Chelsea
```

I have also used **dplyr package** to work on my retrieved data. I have used filter, group_by, select, arrange etc.

I have used dplyr package to find all the player from England and Top Goal Scorer for the season.

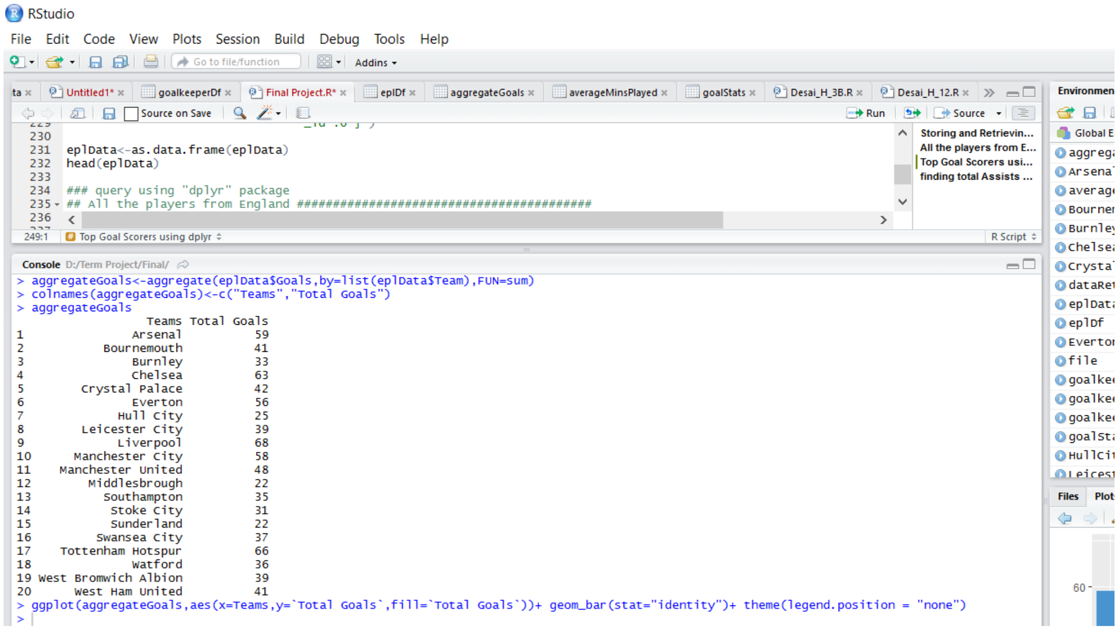

RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
Source
Console D:/Term Project/Final/

```
> ## All the players from England ########################################
> totalEngland<-eplData %>% select(Name,Team,Nationality) %>% filter(Nationality == "England")
> head(totalEngland)
                   Name    Team Nationality
1  Ainsley Maitland-Niles Arsenal     England
2 Alex Oxlade-Chamberlain Arsenal     England
3          Carl Jenkinson Arsenal     England
4           Danny Welbeck Arsenal     England
5            Kieran Gibbs Arsenal     England
6             Rob Holding Arsenal     England
>
> #### Top Goal Scorers using dplyr #######################################
> eplData %>% select(Name,Team,Goals) %>% group_by(Goals) %>%  arrange(desc(Goals))
Source: local data frame [471 x 3]
Groups: Goals [20]

                 Name              Team Goals
                <chr>             <chr> <int>
1       Romelu Lukaku           Everton    24
2          Harry Kane Tottenham Hotspur    20
3      Alexis Sánchez           Arsenal    19
4         Diego Costa           Chelsea    17
5       Sergio Agüero   Manchester City    17
6  Zlatan Ibrahimovich Manchester United    17
7           Dele Alli Tottenham Hotspur    16
8         Eden Hazard           Chelsea    14
9        Jermain Defoe        Sunderland    14
10        Joshua King       Bournemouth    13
# ... with 461 more rows
```

4.Graphical Representation:

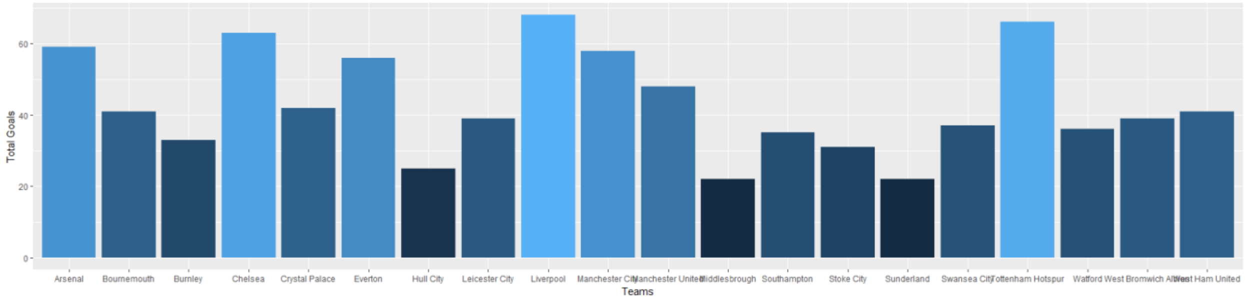
Last but not the list, I have used ggplot2 package to present my data and results graphically.

Use of aggregate function to calculate total number of goals by each team and present data graphically using ggplot.

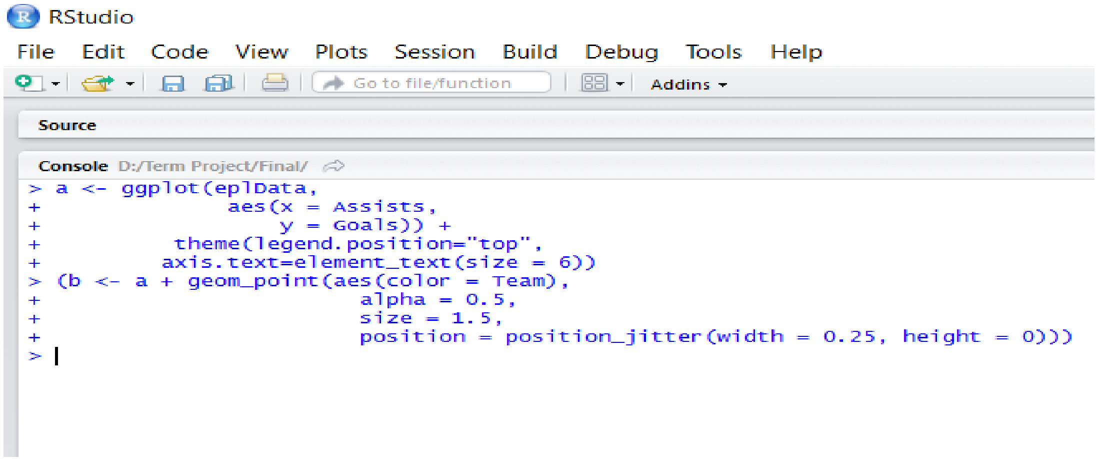


```
eplData<-as.data.frame(eplData)
head(eplData)

### query using "dplyr" package
## All the players from England ########################################
```

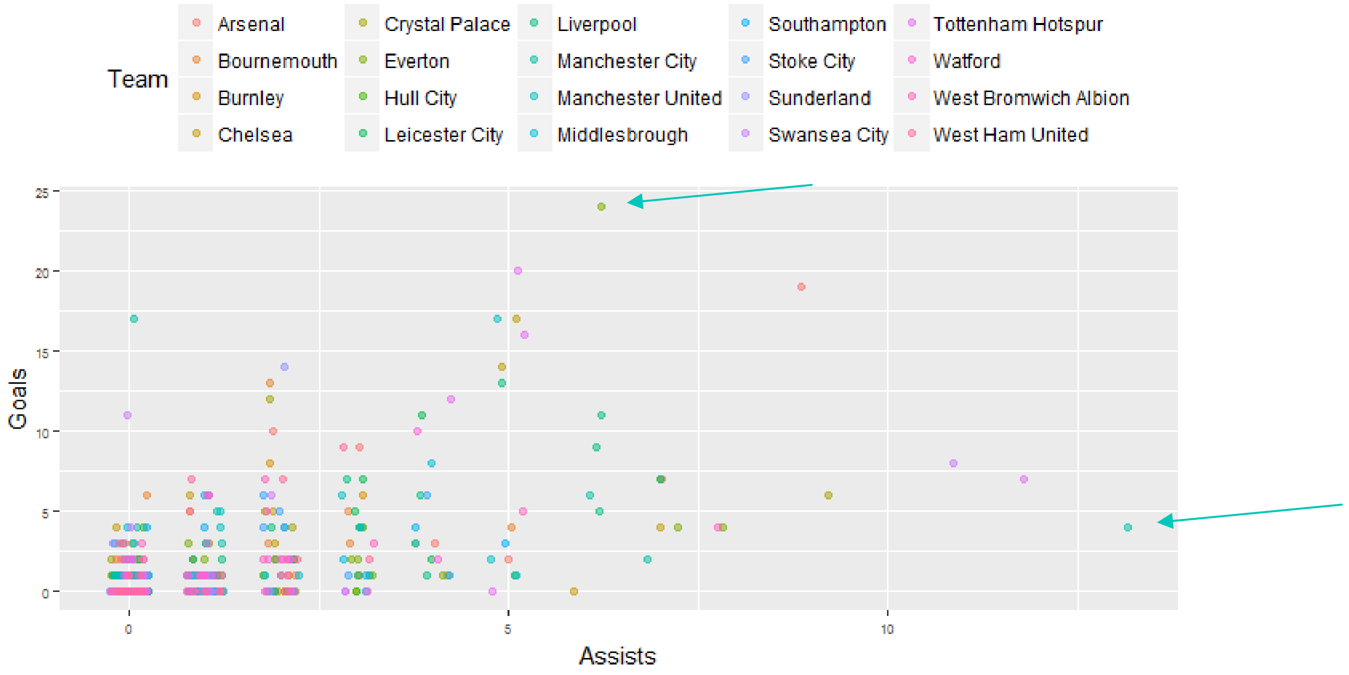
```
> aggregateGoals<-aggregate(eplData$Goals,by=list(eplData$Team),FUN=sum)
> colnames(aggregateGoals)<-c("Teams","Total Goals")
> aggregateGoals
                  Teams Total Goals
1               Arsenal          59
2           Bournemouth          41
3               Burnley          33
4               Chelsea          63
5        Crystal Palace          42
6               Everton          56
7             Hull City          25
8        Leicester City          39
9             Liverpool          68
10      Manchester City          58
11    Manchester United          48
12        Middlesbrough          22
13          Southampton          35
14           Stoke City          31
15           Sunderland          22
16         Swansea City          37
17    Tottenham Hotspur          66
18              Watford          36
19 West Bromwich Albion          39
20      West Ham United          41
> ggplot(aggregateGoals,aes(x=Teams,y=`Total Goals`,fill=`Total Goals`))+ geom_bar(stat="identity")+ theme(legend.position = "none")
> 
```



Then I plotted abundance graph for goals and assists for by the team.

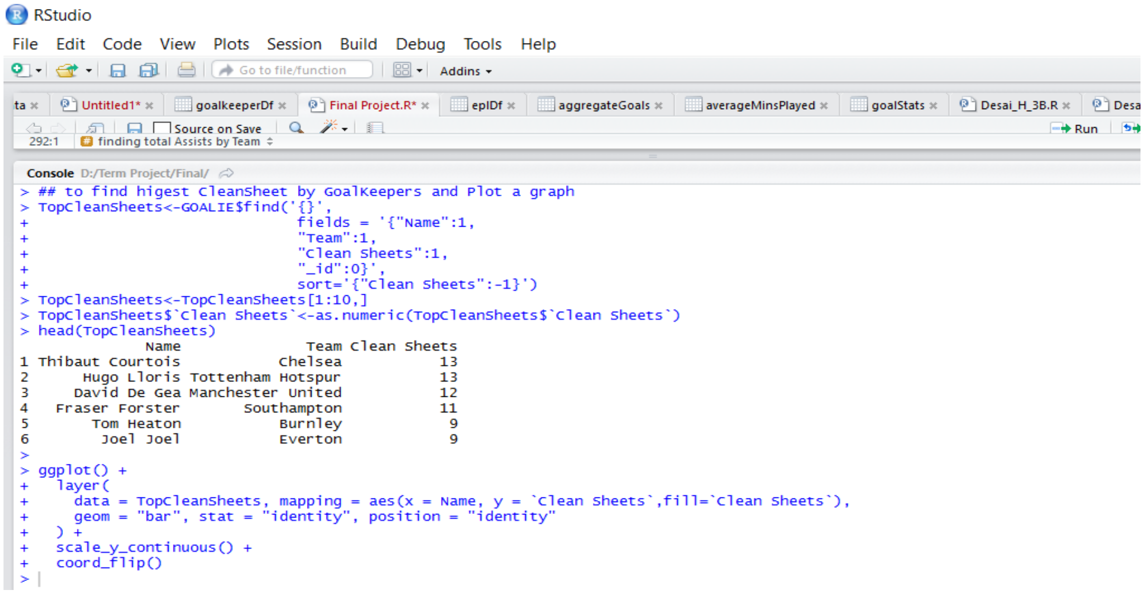


```
> a <- ggplot(eplData,
+             aes(x = Assists,
+                 y = Goals)) +
+         theme(legend.position="top",
+         axis.text=element_text(size = 6))
> (b <- a + geom_point(aes(color = Team),
+                       alpha = 0.5,
+                       size = 1.5,
+                       position = position_jitter(width = 0.25, height = 0)))
> 
```

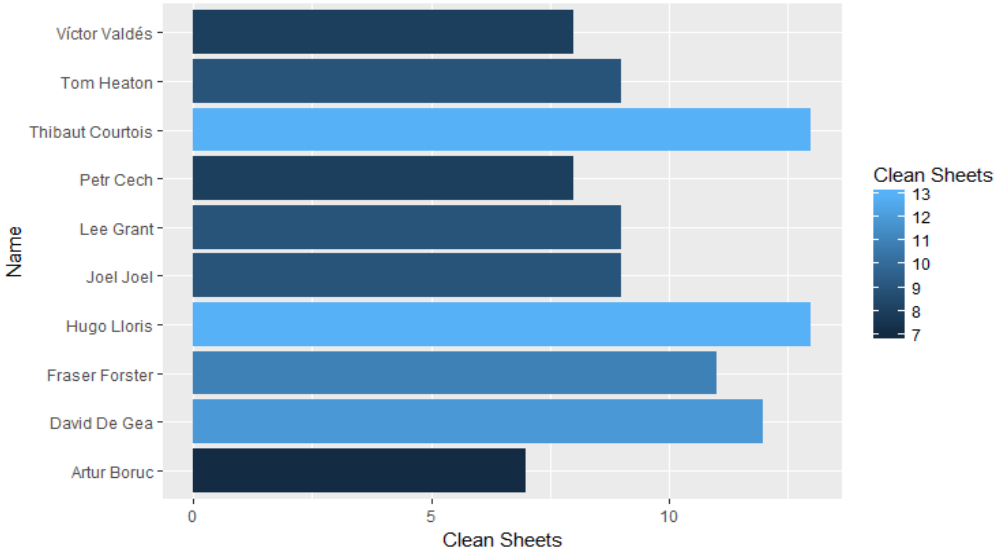


Presented two points in graph shows highest goals and highest assists and color for the team matches with our data. Top goal scorer Romelu Lukaku is from Everton and Top Assists player Kevin De Bruyne is from Manchester City.

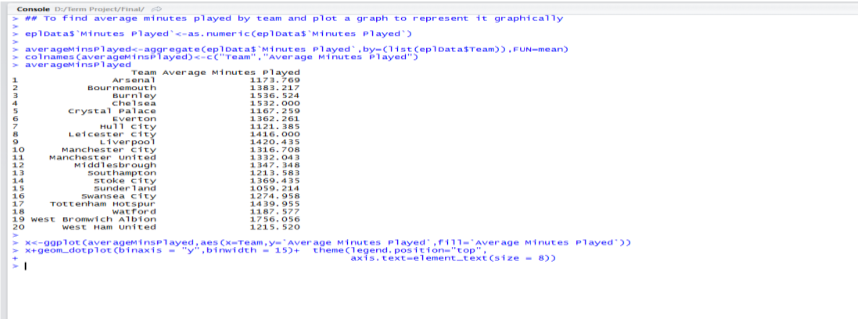
Other graph would be to find Top 10 Cleansheets by Goalkeeper and to plot it.



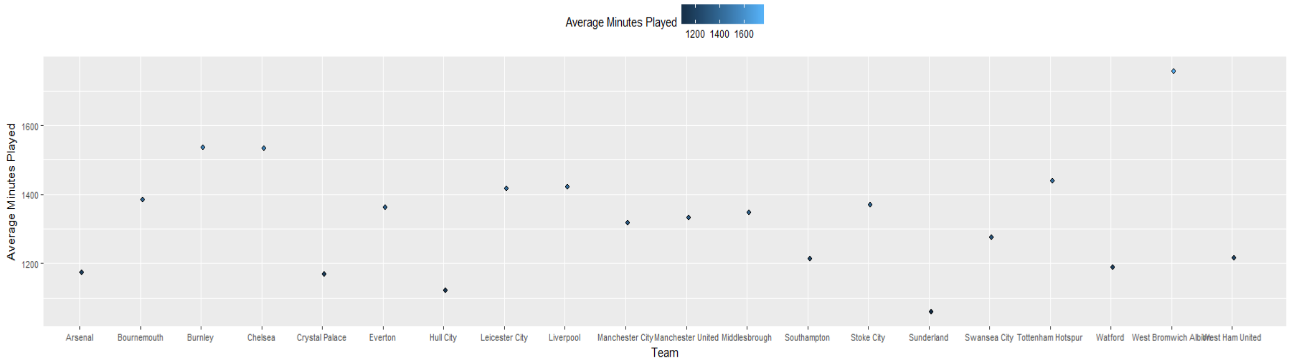
```
> ## to find higest Cleansheet by GoalKeepers and Plot a graph
> TopCleanSheets<-GOALIE$find('{}',
+                             fields = '{"Name":1,
+                             "Team":1,
+                             "Clean Sheets":1,
+                             "_id":0}',
+                             sort='{"Clean Sheets":-1}')
> TopCleanSheets<-TopCleanSheets[1:10,]
> TopCleanSheets$`Clean Sheets`<-as.numeric(TopCleanSheets$`Clean Sheets`)
> head(TopCleanSheets)
             Name              Team Clean Sheets
1 Thibaut Courtois           Chelsea           13
2      Hugo Lloris Tottenham Hotspur           13
3     David De Gea Manchester United           12
4   Fraser Forster       Southampton           11
5       Tom Heaton           Burnley            9
6        Joel Joel           Everton            9
>
> ggplot() +
+   layer(
+     data = TopCleanSheets, mapping = aes(x = Name, y = `Clean Sheets`,fill=`Clean Sheets`),
+     geom = "bar", stat = "identity", position = "identity"
+   ) +
+   scale_y_continuous() +
+   coord_flip()
> |
```



In the end I calculated average minutes played by every team and plotted a graph for it.



```
> ## To find average minutes played by team and plot a graph to represent it graphically
>
> eplData$`Minutes Played`<-as.numeric(eplData$`Minutes Played`)
>
> averageMinsPlayed<-aggregate(eplData$`Minutes Played`,by=(list(eplData$Team)),FUN=mean)
> colnames(averageMinsPlayed)<-c("Team","Average Minutes Played")
> averageMinsPlayed
                   Team Average Minutes Played
1               Arsenal               1173.769
2           Bournemouth               1383.217
3               Burnley               1536.524
4               Chelsea               1532.000
5        Crystal Palace               1167.259
6               Everton               1362.261
7             Hull City               1121.385
8        Leicester City               1416.000
9             Liverpool               1420.435
10      Manchester City               1316.708
11    Manchester United               1332.043
12        Middlesbrough               1347.348
13          Southampton               1213.583
14           Stoke City               1369.435
15           Sunderland               1059.214
16         Swansea City               1274.958
17    Tottenham Hotspur               1439.955
18              Watford               1187.577
19 West Bromwich Albion               1756.056
20      West Ham United               1215.520
>
> x<-ggplot(averageMinsPlayed,aes(x=Team,y=`Average Minutes Played`,fill=`Average Minutes Played`))
> x+geom_dotplot(binaxis = "y",binwidth = 15)+  theme(legend.position="top",
+                                                    axis.text=element_text(size = 8))
> |
```



Future Work:

- To collect more stats for the same seasons and previous seasons
- To use better and dynamic scraping technique
- To collect data for different soccer leagues like La Liga, Champions League etc.
- After collecting other Premier League Seasons data, produce comparative statistics.
- Better graphical representation and will try to create soccer statistics application.

References:

- Collecting, Storing and Retrieving Data by Yatish Jain and Martin Schedlbauer
- Data Manipulation with R by Jaynal Abedin
- Seven Database in Seven Weeks by Eric Redmond and Jim R. Wilson
- www.premierleague.com
- www.nbcsports.com
- http://ggplot2.org/

Acknowledgment:

I would like to thank ,Teaching Assistants for the course DA5020 Jianchao (Jesse) Yang and Japan Mehta for their continuous support throughout the course. My special thanks for the Dr. Durant for continuous guidance and knowledge during course.