|  |
| --- |
|  |
| 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:

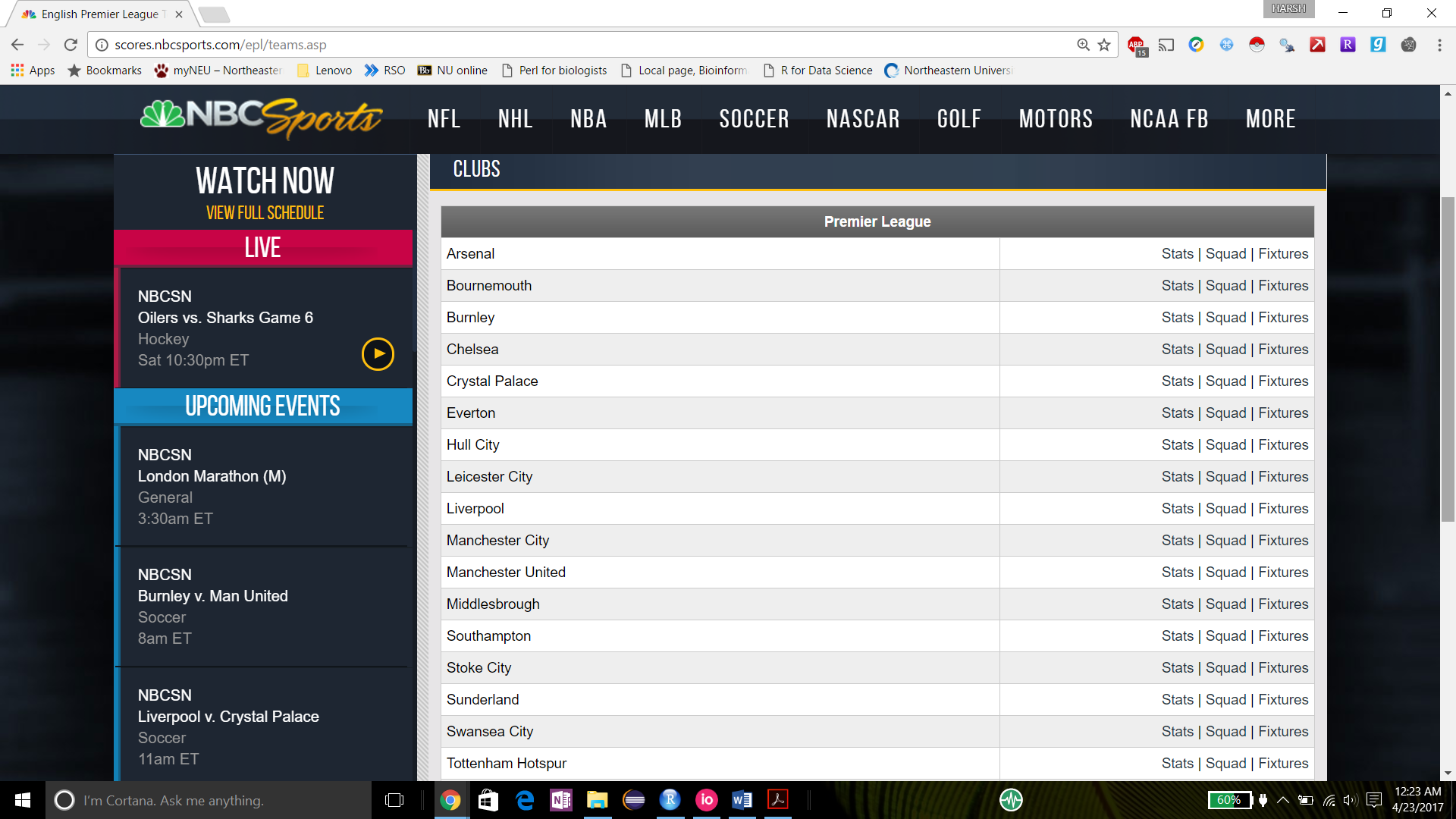
Premier League or English Premier League is a major soccer league contested between top 20 teams in England every year. It is the one of the most followed and major soccer league in the world. For the final term project, I have decided to collect data for the Premier League 2016-17 season, which is also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

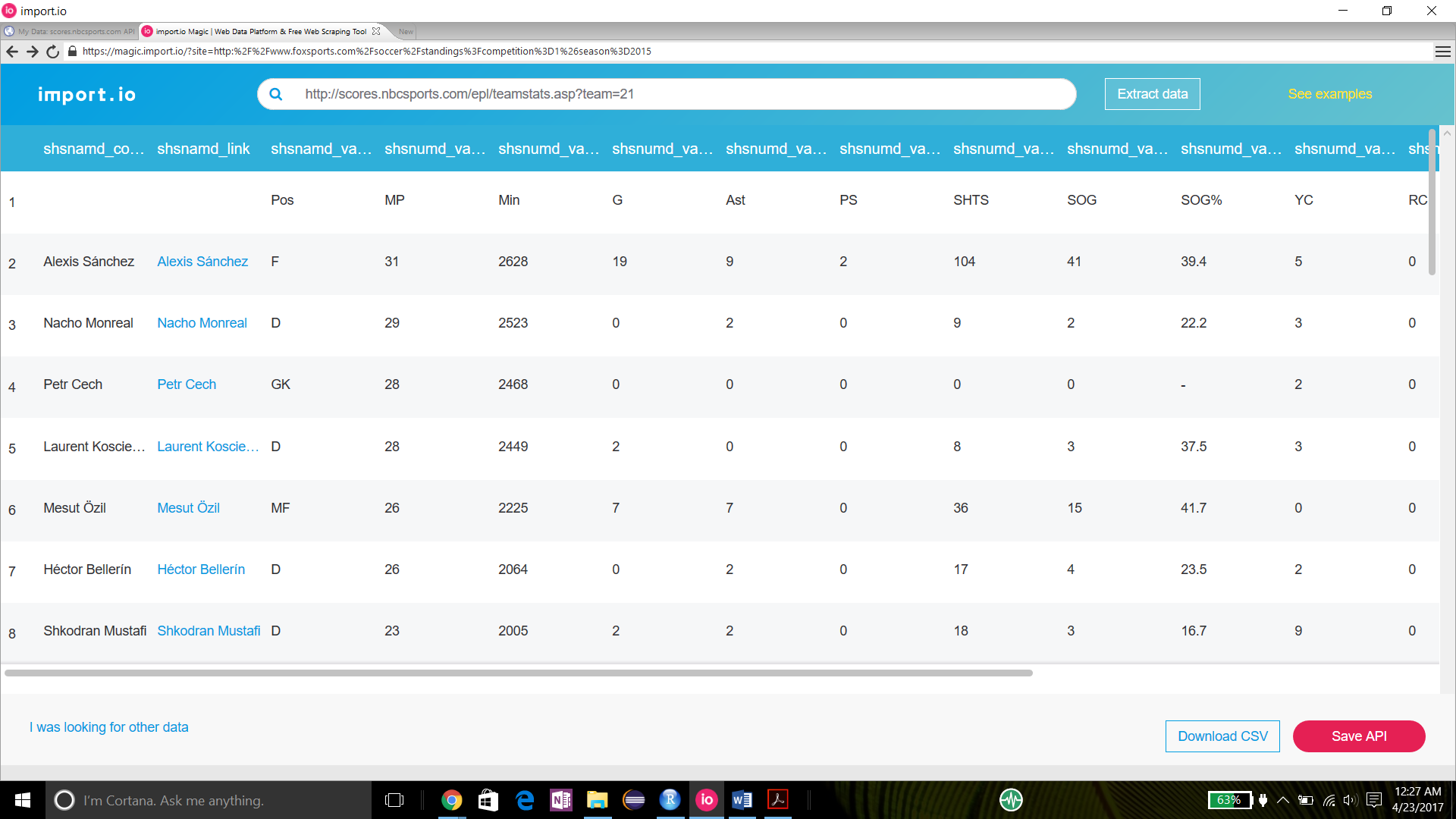
**1)Collecting:**

For the collection of the data I will be scraping directly from the [www.premierleague.com](http://www.premierleague.com) and [www.nbcsports.com](http://www.nbcsports.com)

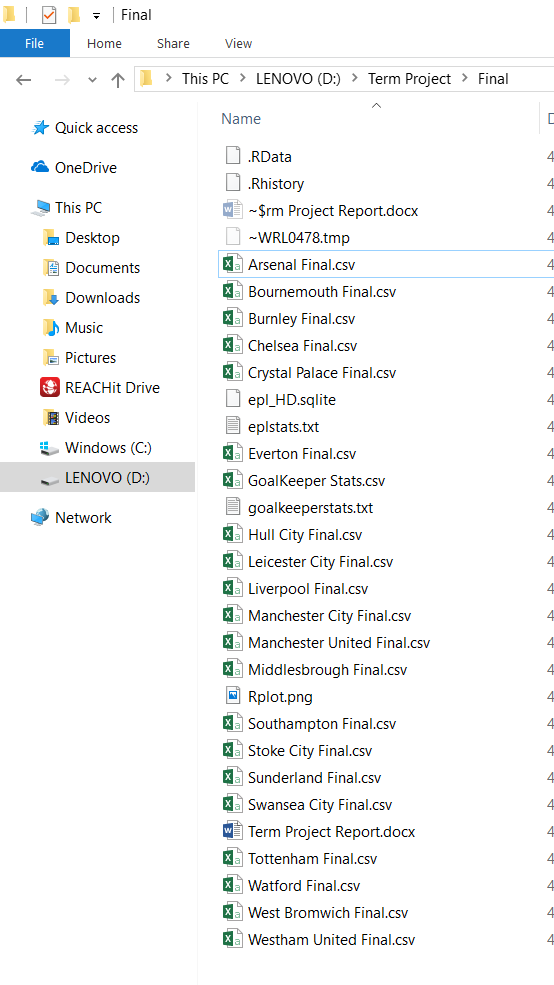
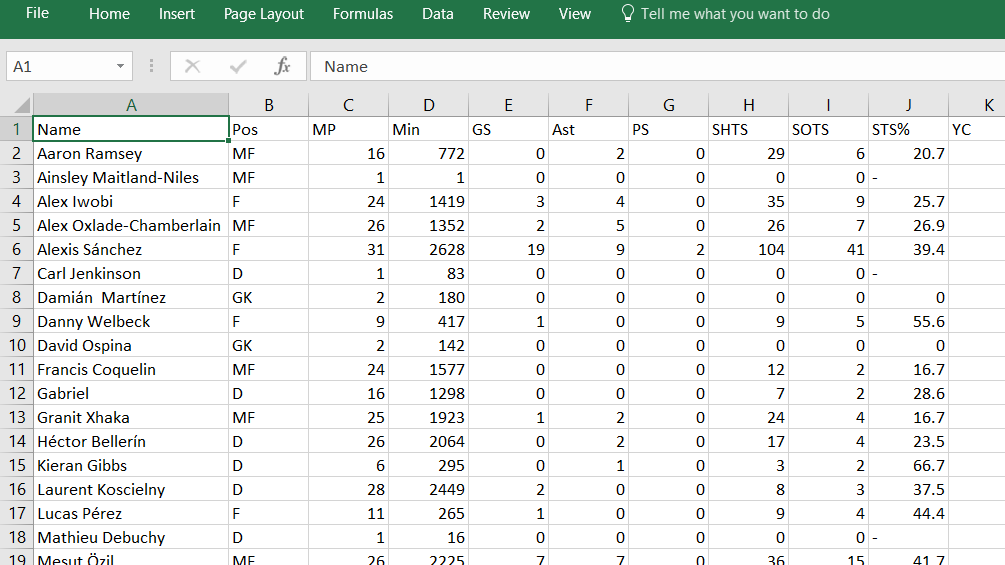


I am using import.io as a scraping tools and URL get query to collect data team wise using team id as URL query.

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.



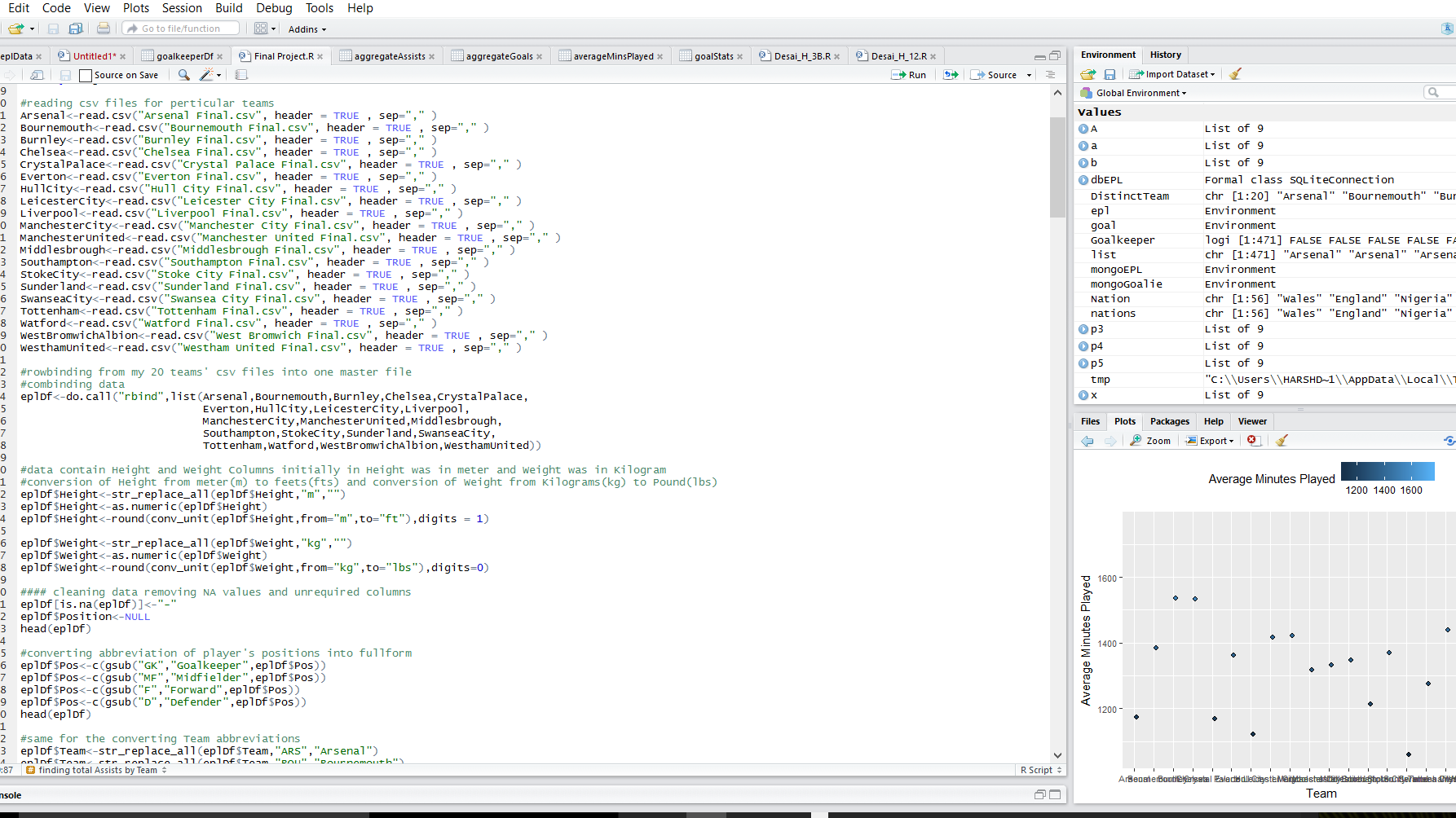
Import.io allow us to store collected data into .csv files. So, I have stored 20 .csv file for each team separately.

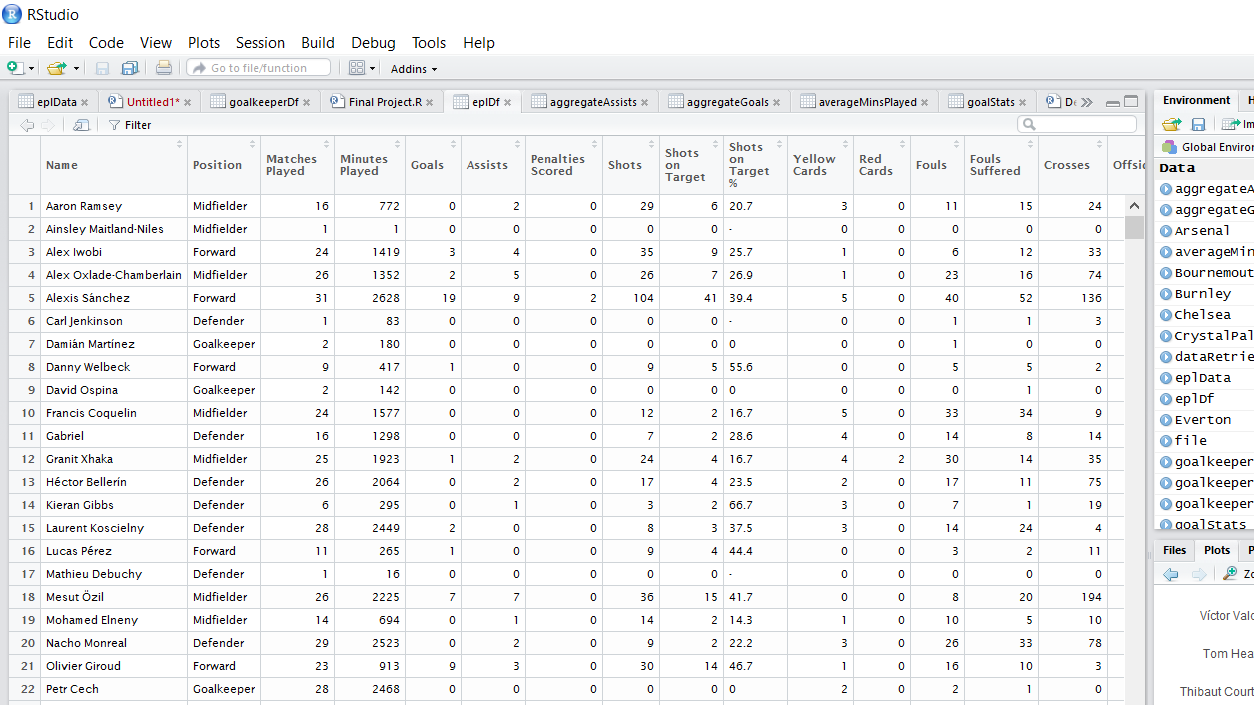
Reading data into R using read.csv() function and combining all 20 files into master file using R functions. I might have combined data manually but I wanted to use R functions that’s the main I choose to extract file per teams and not a single master file.

Packages used: stringr ,measurements

Loading .csv files, combining files and Cleaning data



Combined final data frame :



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

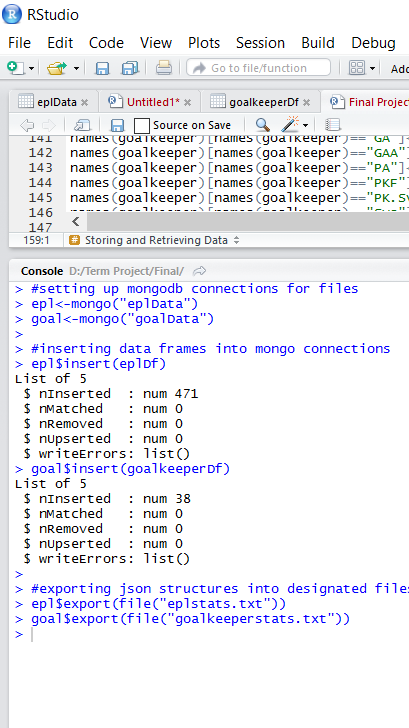
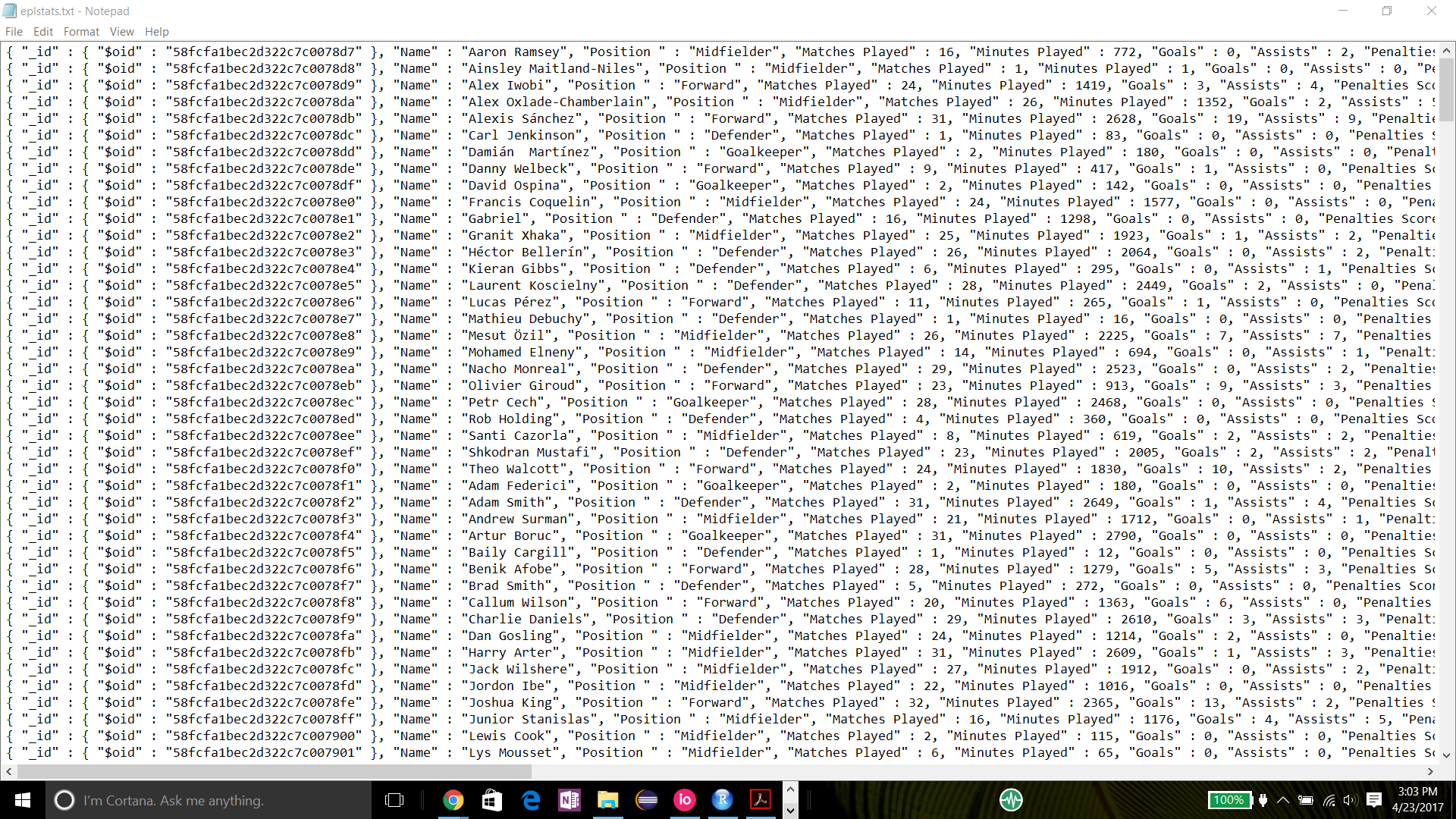
"C:\Program Files\MongoDB\Server\3.4\bin\mongod.exe" --dbpath d:\test\mongodb\data

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.

Export

Insert

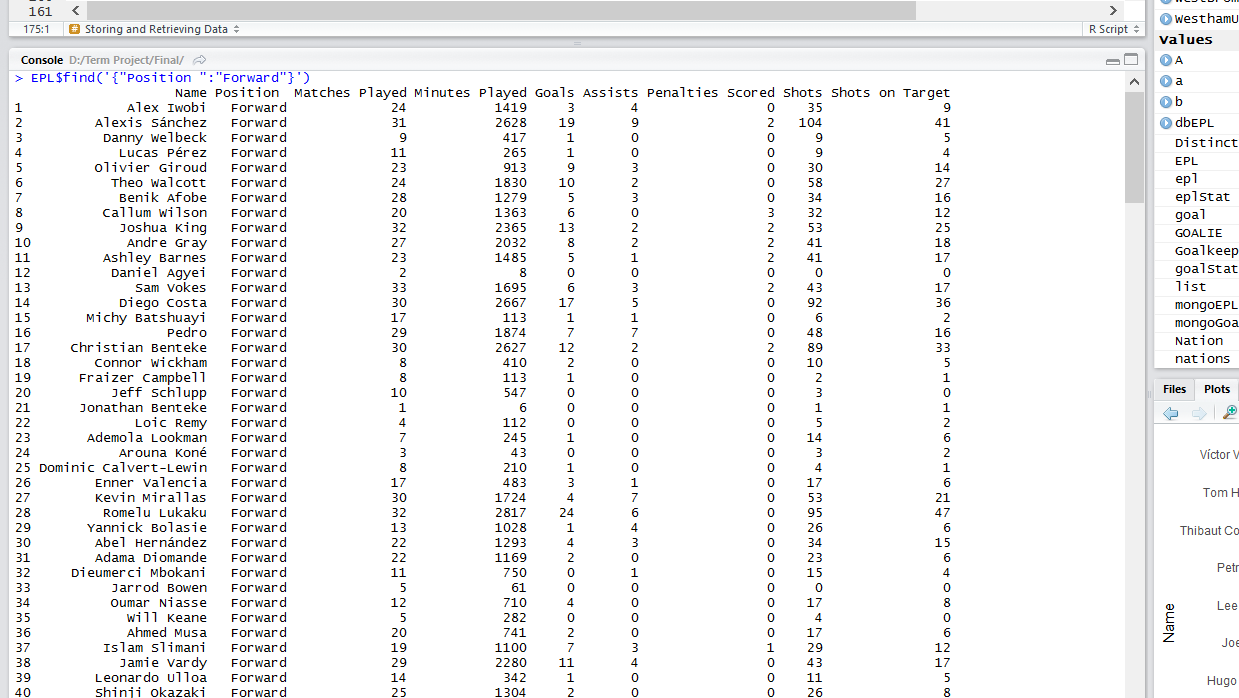
**3.Retrieval:**

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

Packages used: dplyr,ggplot2,mongolite

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

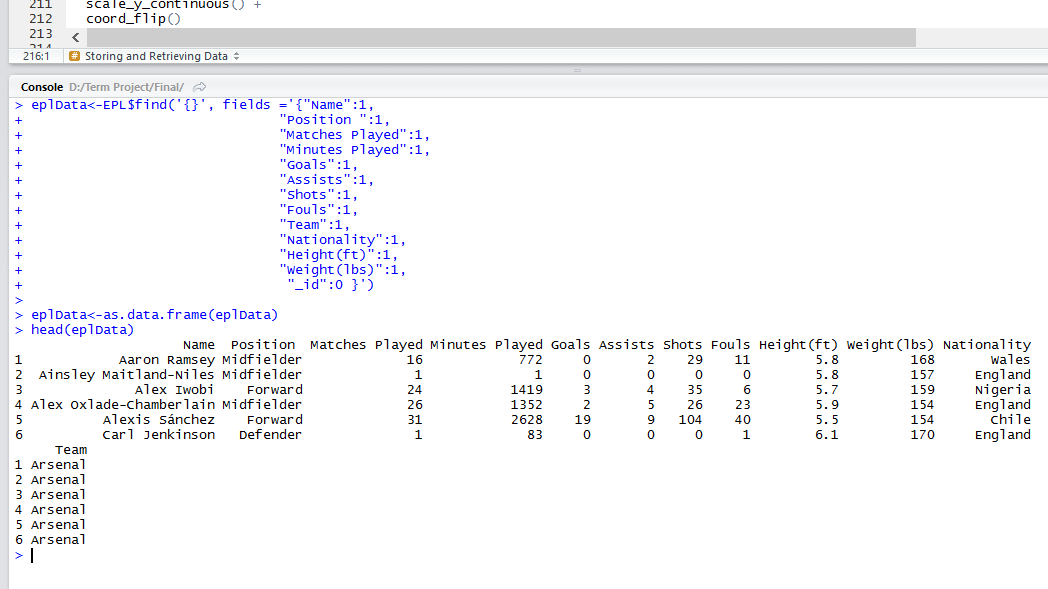
Firstly using find() function to search all the forward position players from the database.



Output

Input

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

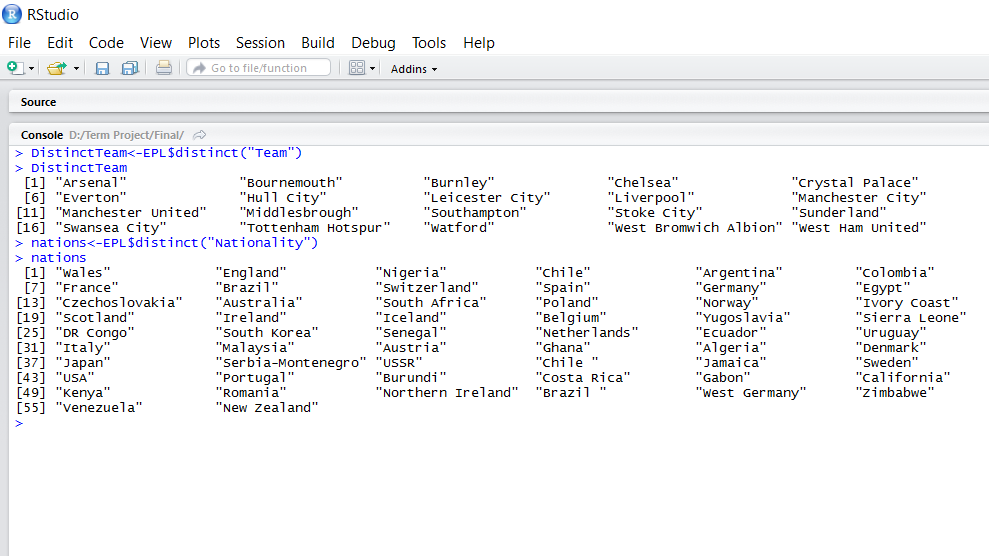


Output

Input

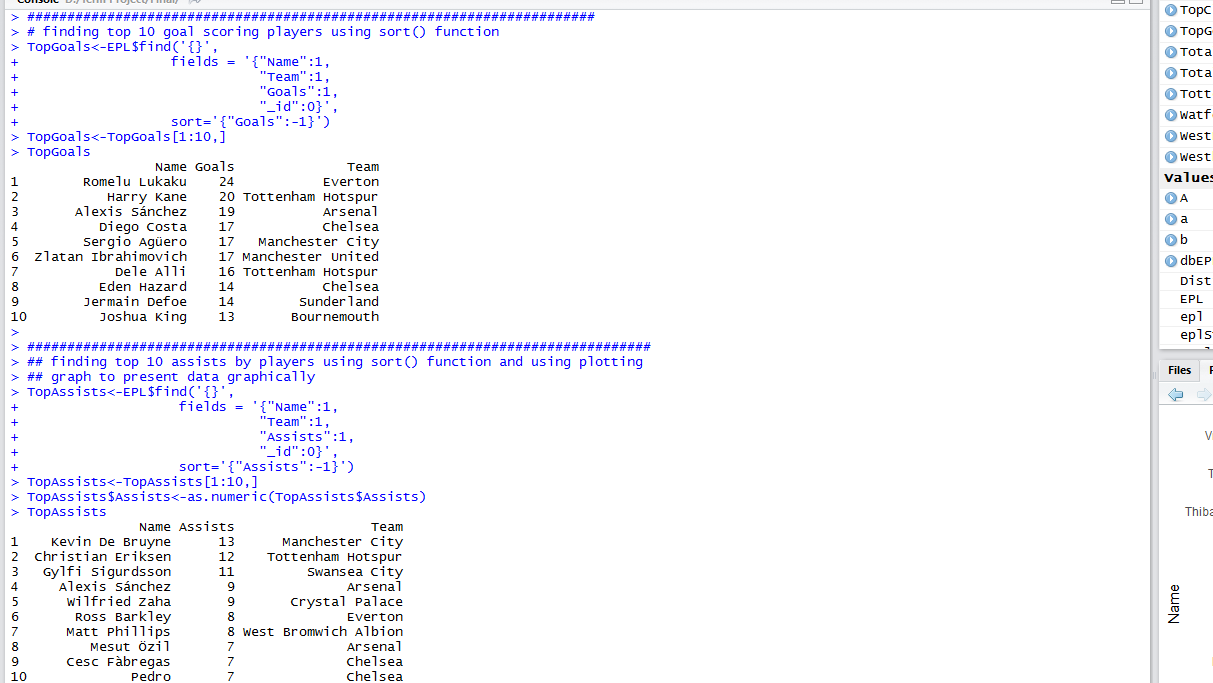
The other function used for retrieval was distinct(),which will find distinct values from particular columns.

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



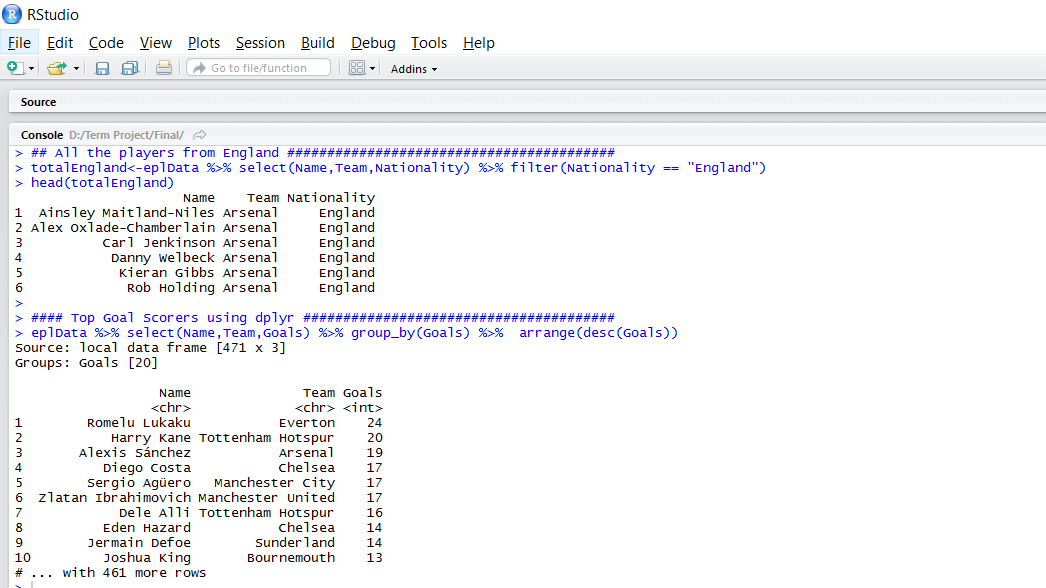
We can see that there are 20 distinct teams in a database and players are from 56 distinct countries.

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.



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

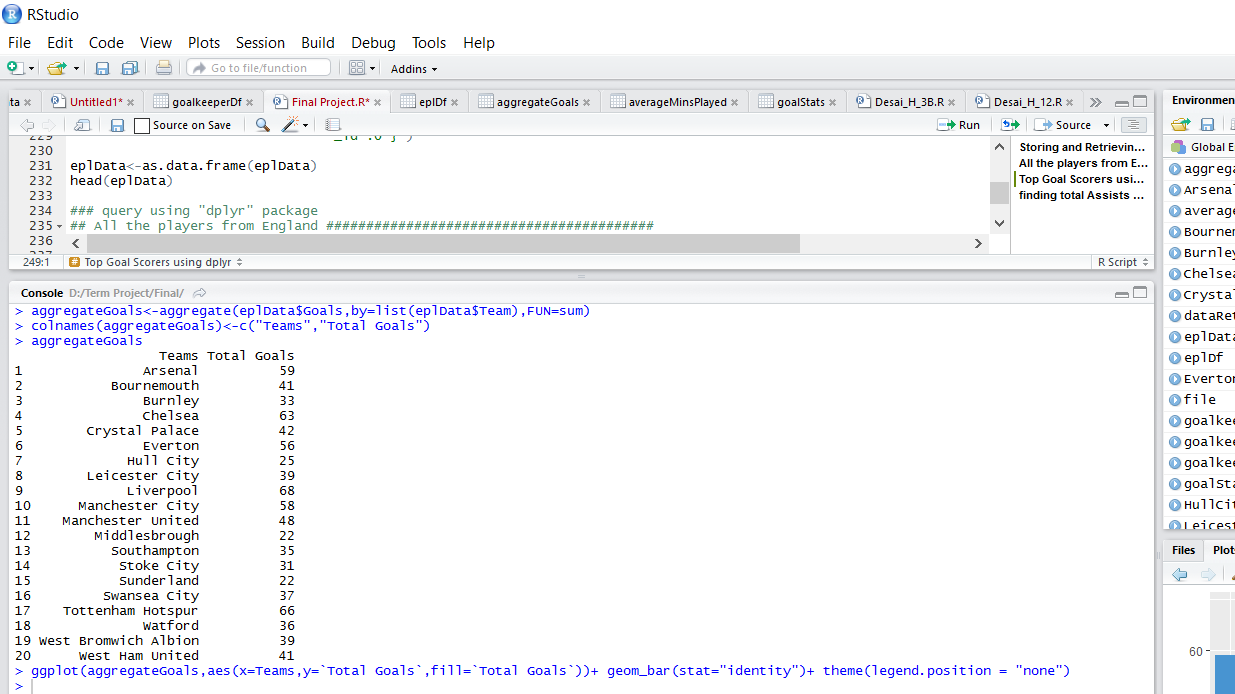
I have used dplyr package to find all the player from England and Top Goal Scorer for the season.

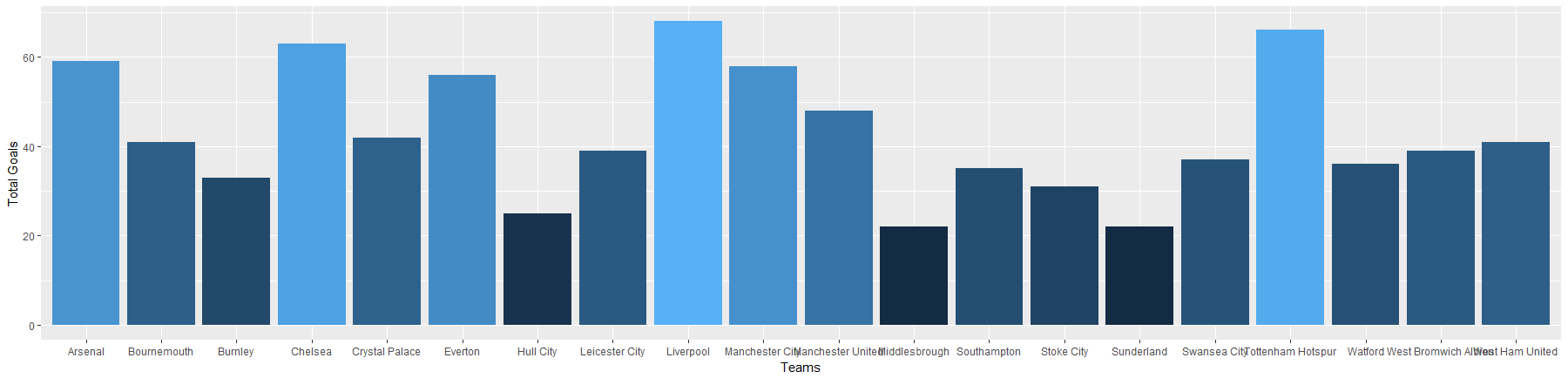


**4.Graphical Representation:**

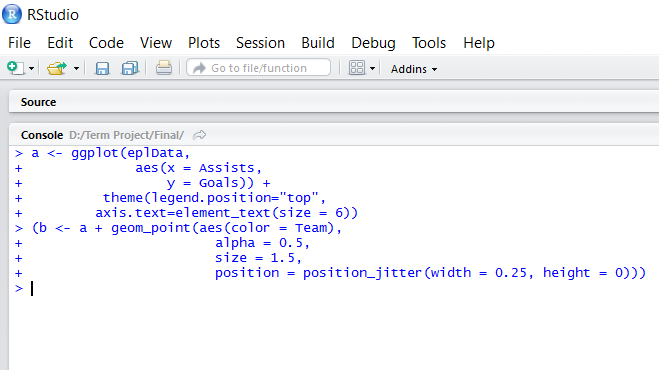
Last but not the list, I have used ggplot2 package to present my data and results graphically.

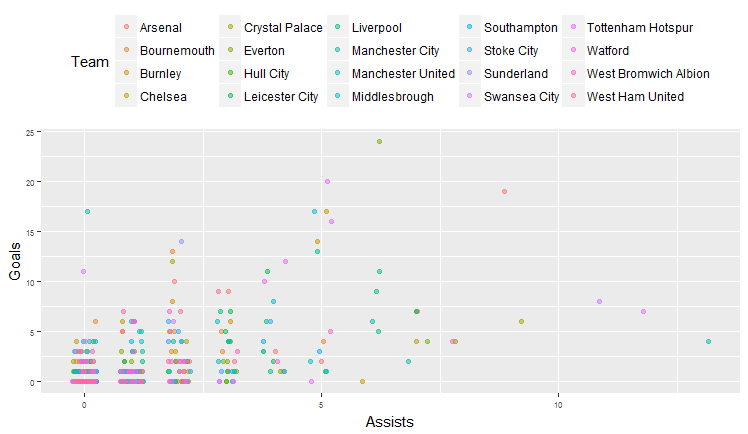
Use of aggregate function to calculate total number of goals by each team and present data graphically using ggplot.





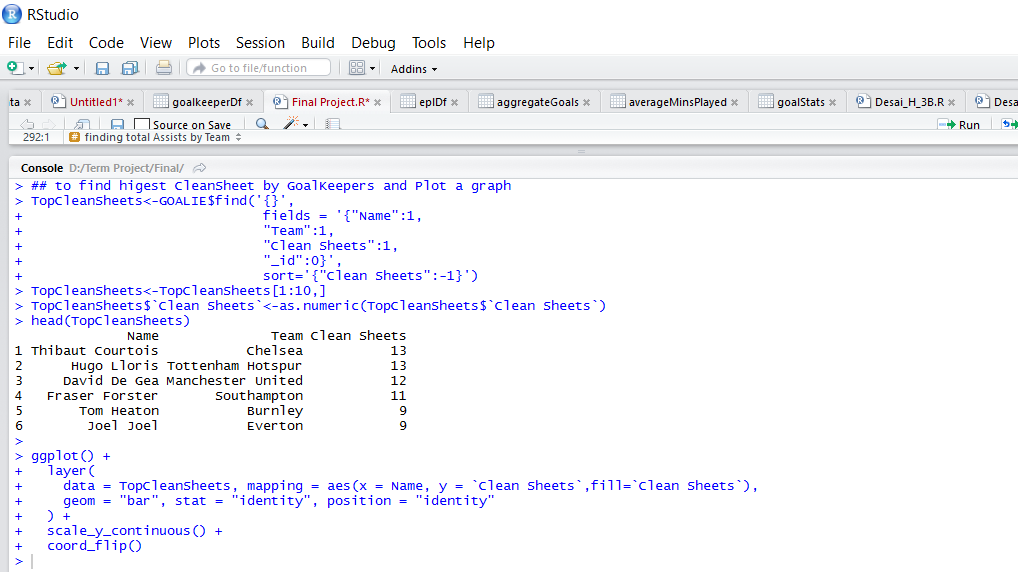
Then I plotted abundance graph for goals and assists for by the team.

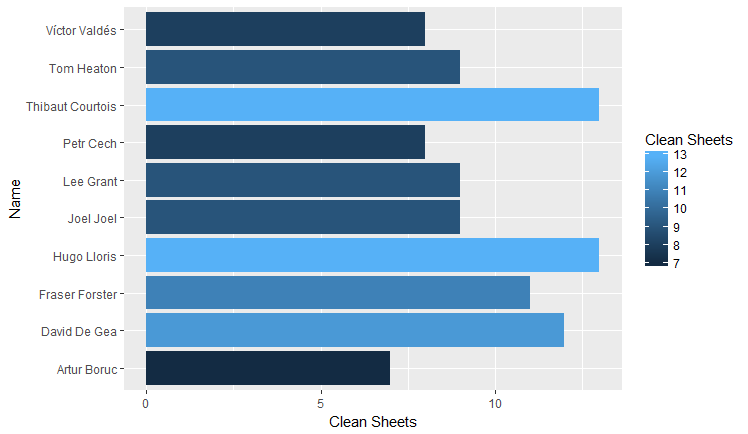




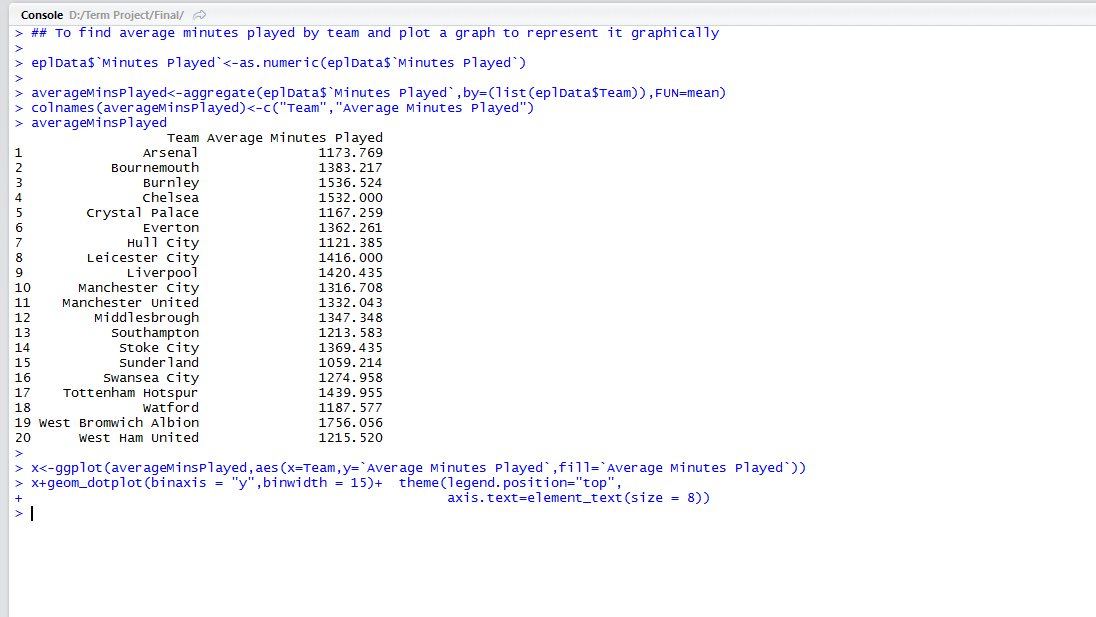
Presented two points in graph shows highest goals and highest assists and color for the team matches with our data. Top goal scorer Romelu Lukaku is from Everton and Top Assists player Kevin De Bruyne is from Manchester City.

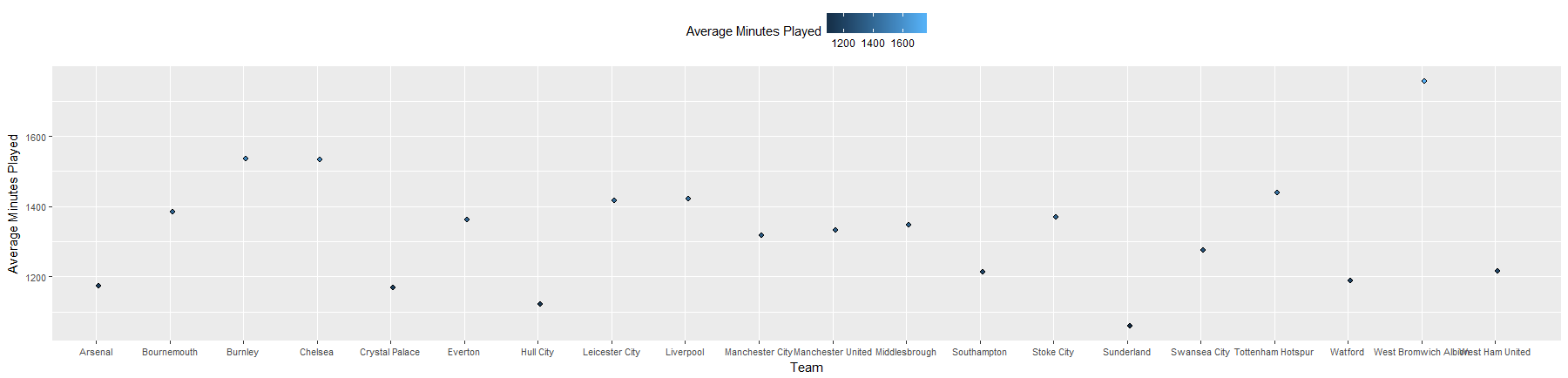
Other graph would be to find Top 10 Cleansheets by Goalkeeper and to plot it.





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





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](http://www.premierleague.com)
* [www.nbcsports.com](http://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.