# post1

*Eden*

*2017.10.26*

# Topic: Graphing using ggplot2()

## Introduction:

This is a "blog post" introducing (or "reintroducing") the function ggplot2() in R language and how we can use it to set up graphs and visualize the data we want to analyze. In this post, I will show the basics of the ggplot function as well as the geom_violin function which has not been discussed during lectures.

## Motivation:

Graphing is a neat way to *explore relationships between data* and it's always been fun for me to do that!

```
#let's load the packages we are going to use first
library(readr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

### Let's prepare some data for the players in Premier Leagues 2017-18 preseason stats from (https://drive.google.com/file/d/0B640e9RWqrjAeHZ5eTVJV2VURE0/view)

### Note that this data only contains players who played more than 600 minutes

My favorite player Eden Hazard!!

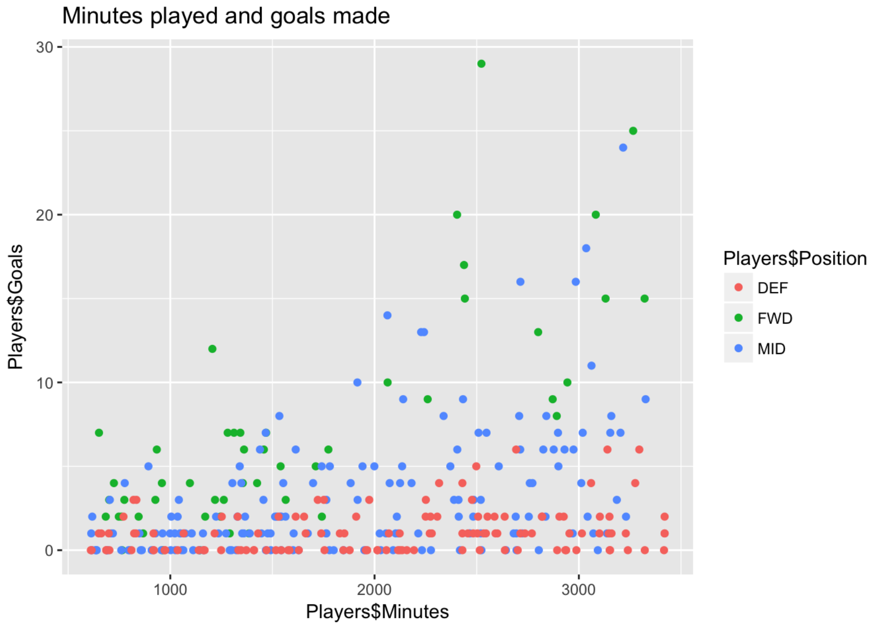
```
#create a vector containing the data we are going to use
library(readxl)
Players <- read_excel("~/stat133/stat133-hws-fall17/post01/data/players.xlsx")
```

# Basics:

## Let’s first review what we’ve learned in class

Here’s how we can use the data we loaded to create a basic scatter plot displaying the relationship between minutes played and goals made.

```
ggplot(Players, aes(x = Players$Minutes, y = Players$Goals, color = Players$Position)) +
  geom_point()+
  ggtitle('Minutes played and goals made')
```



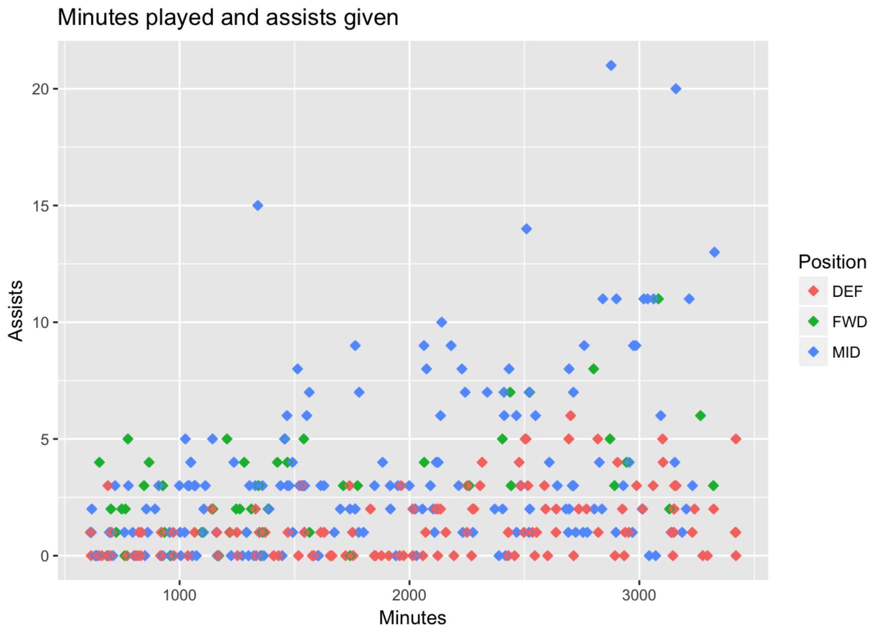
About this graph:

As we can see on the graph, DEF players generally don’t score much however long they played. For MID players, they score more. And FWD strikers scored the most. Some of them made a lot of goals even though they did not played very long.

The result shown actually makes sense since strikers are expected to goal more. And the data correpond to the expectation. This graph helps us easily confirm that.

## Now let’s try the relationship between minutes played and assists

```
ggplot(Players, aes(x = Players$Minutes, y = Players$Assists, color = Players$Position)) +
  geom_point() +
#we can modify the shape of the points and add title to the graph
  geom_point(shape = 5) +
  ggtitle('Minutes played and assists given') +
  labs(x = 'Minutes', y = 'Assists') +
  scale_color_discrete(name = 'Position')
```



About this graph:

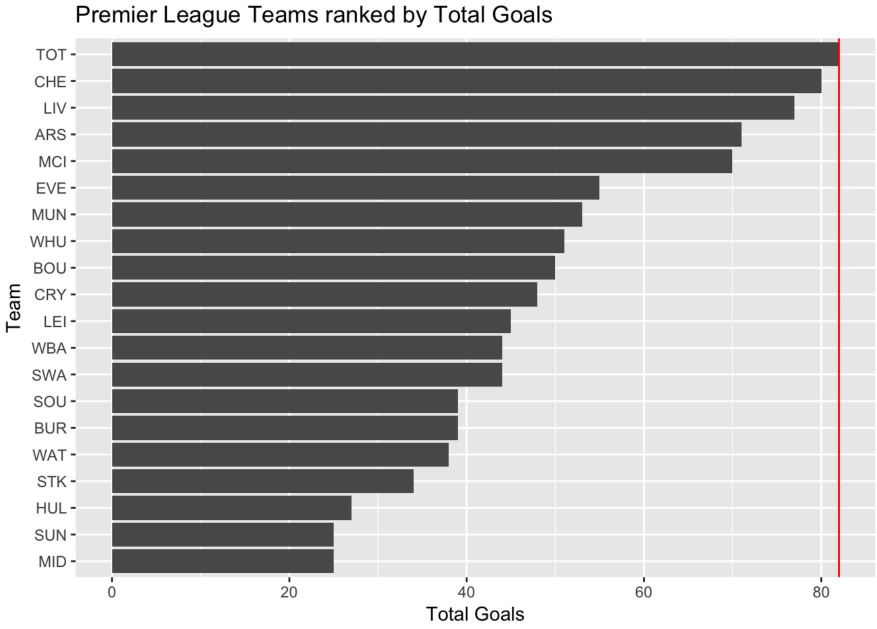
It’s clear that generally, MID plyaers got more assists. This also makes sense since MID players are responsible for giving good passes to FWD to score.

### Next I’ll rank each team based on the goals they made using a bar plot.

```
#firstly, we'll need to create a new vector containing the total goals of each team
Team <- summarise(group_by(Players, team = as.character(Team)),
  goals = sum(Goals))
```

```
## Warning: Mangling the following names: <U+00A3> -> <U+00A3>. Use
## enc2native() to avoid the warning.
```

```
ggplot(data = Team, aes(x = reorder(team , goals), y = goals)) +
  #Bar plot!
  geom_bar(stat = "identity") + labs(y = "Total Goals") +
  labs(x = "Team") + coord_flip() + geom_hline(aes(yintercept = 82), col = "red") + ggtitle("Premier League Teams
ranked by Total Goals")
```



About this graph:

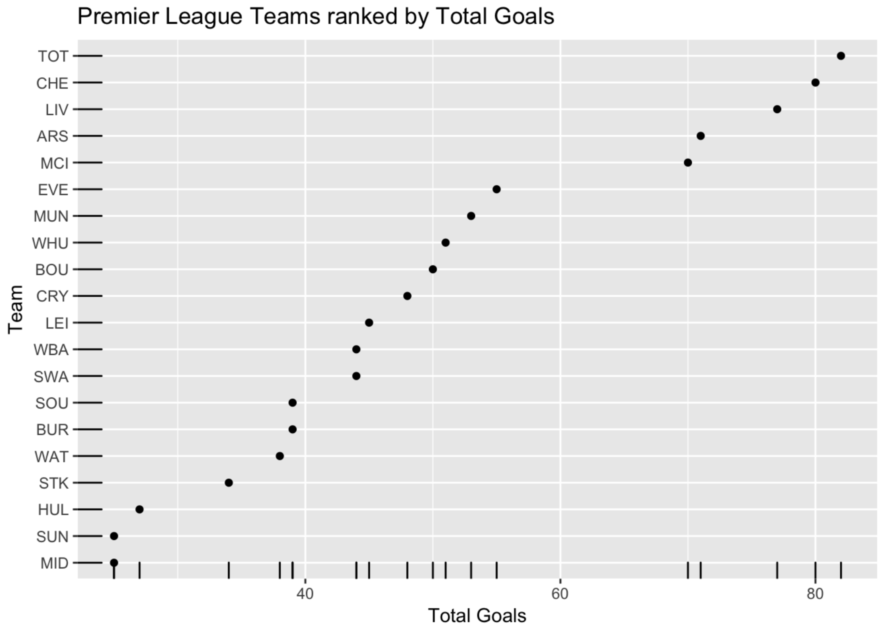
We can easily tell that for the 16/17 season, TOT made most goals so they were really on top of the league on the attack side. As confirmed by the final league table, the top five scoring teams reanked top five. (https://www.google.com/search?q=16%2F17+premier+league+table&oq=16%2F17+premier&aqs=chrome.0.0j69i57j0l4.5337j0j4&sourceid=chrome&ie=UTF-8)

## Time for things that weren’t covered!

### 1.Rug plots in the margins. (http://ggplot2.tidyverse.org/reference/geom_rug.html)

Using a rug plot, we can visualise what number each point on the graph correspond to on the x-axis and y-axis. Let’s try an example about the team ranking of total goals during 16/17 season.

```
ggplot(data = Team, aes(y = reorder(team , goals), x = goals)) +
  #Bar plot!
  geom_point() + labs(x = "Total Goals") +
  labs(y = "Team") + ggtitle("Premier League Teams ranked by Total Goals") +
  #Simply add another geom function geom_rug to get the extra margins which will help you better see the correspon
ding numbers!
  geom_rug()
```



About the graph

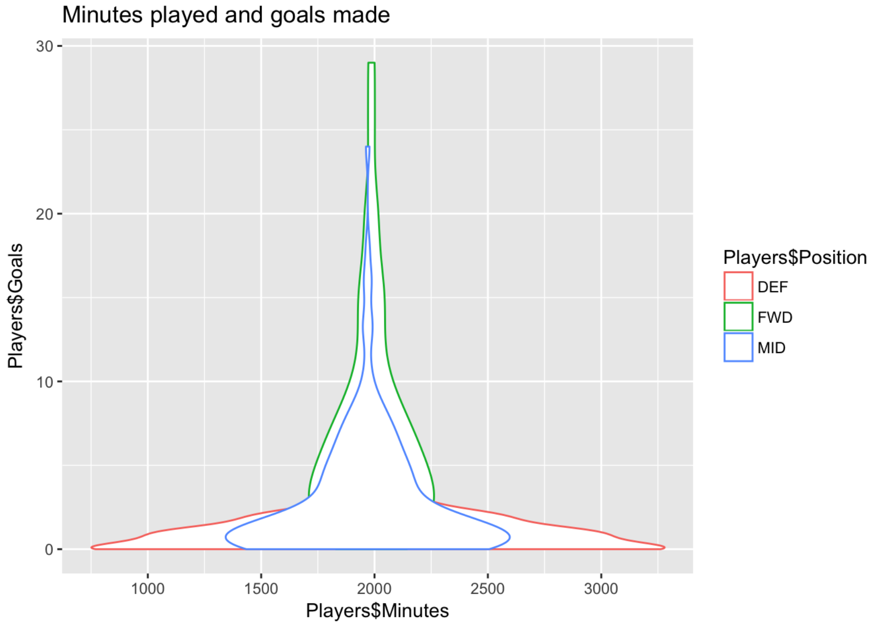
Now we can see the corresponding numbers of the points on the graph. It might not be super useful for most of the time, but it’s still an interssting function.

## 2.Violin plot. (http://ggplot2.tidyverse.org/reference/geom_violin.html)

A violin plot displays a continuous distribution. It is just like a boxplot, different in the way that it’s also a mirrored density plot. We can use the geom_violin function to display the spread of a data set like what we’ll do right now.

```
ggplot(Players, aes(x = Players$Minutes, y = Players$Goals, color = Players$Position)) +
  geom_violin()+
  ggtitle('Minutes played and goals made')
```

```
## Warning: position_dodge requires non-overlapping x intervals
```



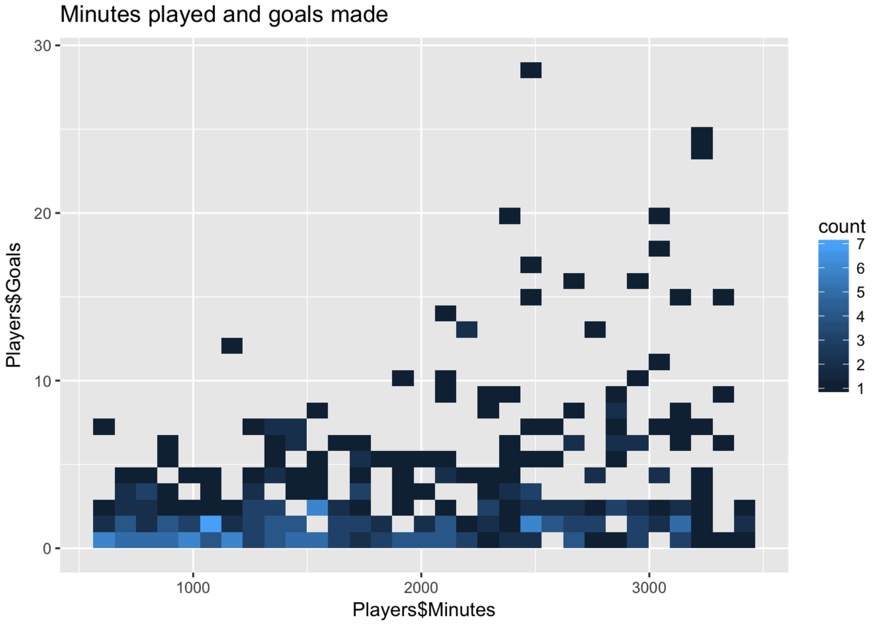
About the graph

In this graph, the density is displayed in a way we can easily compare the goals from players of different positions.

## 3.Heatmap of 2d bin counts. (http://ggplot2.tidyverse.org/reference/geom_bin2d.html)

This geom_bin2d function divid the whole graph into rectangles. In each rectangles, the different color reflect the number of repeats. Another way to put it – we can see how many repeats are there for one point on the graph. Like the following praph.

```
ggplot(Players, aes(x = Players$Minutes, y = Players$Goals)) +
  geom_bin2d()+
  ggtitle('Minutes played and goals made')
```



About the graph

There are less people (darker colors) on the top half of the graph! Not many people can make that many goals. Those outstanding black rectangles represent attackers who are really on top of the games!

## Take-home message:

The ggplot2() is a very comprehensive package of all sorts of useful tools. There are tons more functions to be explored and learnt in the ggplot2() package and we can really do some exciting data analysis by ourselves!

## Reference:

1.http://ggplot2.tidyverse.org/reference/#section-plot-basics

2.https://drive.google.com/file/d/0B640e9RWqrjAeHZ5eTVJV2VURE0/view

3.https://images.cdn.fourfourtwo.com/sites/fourfourtwo.com/files/eden_hazard_1.jpg

4.https://www.google.com/search?q=16%2F17+premier+league+table&oq=16%2F17+premier&aqs=chrome.0.0j69i57j0l4.5337j0j4&sourceid=chrome&ie=UTF-8

5.http://ggplot2.tidyverse.org/reference/geom_rug.html

6.http://ggplot2.tidyverse.org/reference/geom_violin.html

7.http://ggplot2.tidyverse.org/reference/geom_bin2d.html