**Teoriorienteret Metode**

**Part 2 - Transforming Historical Research through Digital Methods and Approaches**

**Af Claus Vestergaard Olesen**

**Abstract**

With my project I want to tell if there may be more coincidences or connections in professional football in Denmark than imagined. I have looked at season 2016/17 and 2017/18 in the best football league in Denmark and found data to explain statistics such as goals scored, shots on target, red card received and so on. The digital tools I have used is mainly RStudios and ggplot. I have moved the data from an official football website and into RStudios so that the results of the data would become much more transparent and easier to deal with.

**Introduction and background**

The reason why I have chosen this topic is because I have a huge interest for football and spend a lot of time watching it. Therefore, it occurred to me that an investigation of the statistics I had found qua my datasets would be interesting. I hope to find some connections throughout the data that I wouldn´t have found without the use of digital methods. Because then it would help me understanding the need for these digital tools and how they make research much easier and more transparent.

Through my project I have converted the original RStudios script to a version of R called RMarkdown. I did that, because it later on in the process would be much easier to modify my project into word. Because I have used RMarkdown my script looks slightly different from the R scripts we have worked with during classes.

It was my purpose to use WebScraping in the process of transforming the statistics of data to a dataset in RStudios. Because of the tangible datasets I had found it was unnecessary, and Max Odsbjerg Petersen advised me not to use WebScraping.

**Data Acquisition and Processing**

The data used for this assignment was picked up from the official website of the best football league in Denmark (Superligaen). The statistics is compiled by the Danish football association (Det Danske Fodboldforbund) and is Open Data available for everyone. In other words, everyone who want these kinds of information can get access to them by a single click in their web browser and need no appurtenant password or login. The websites for the two seasons are listed below.

<http://www.superliga.dk/performance-centre/2016-17>

<http://www.superliga.dk/performance-centre/2017-18>

**Software framework and Implementation and Empirical Results**

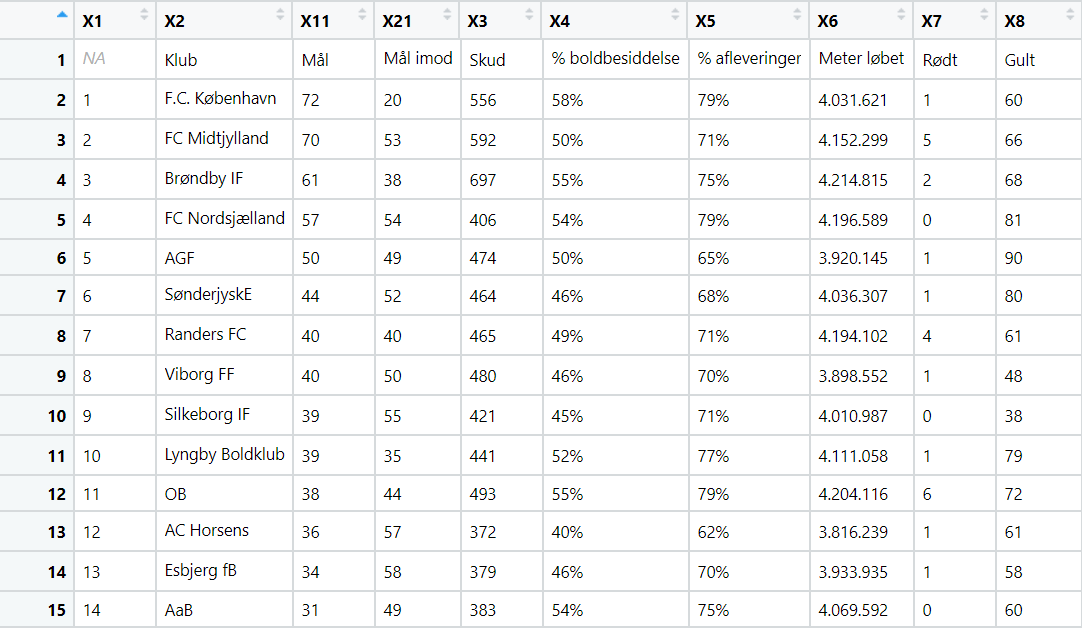
My general idea for this project was to move some data from the webpage of the best national football league in Denmark by Web Scraping. I found out it wasn´t necessary to use Web Scraping, because my datasets isn´t that extensive. Max advised me to do all of my project in RMarkdown. I installed some primarily packages in RStudios. The packages I used was: library(rvest), library(tidyverse), library(lubridate), library(tidytext) and I ran those packages as commands in RStudios.

After that I wanted to tell RMarkdown which HTML code my data from the website had. I ran the command SL1617 <- read\_html("http://www.superliga.dk/performance-centre/2016-17"). By that the data for the season 2016/17 was assembled in a map called SL1617. I did the same thing for the season 2017/18 by using another website and called it 20172018.

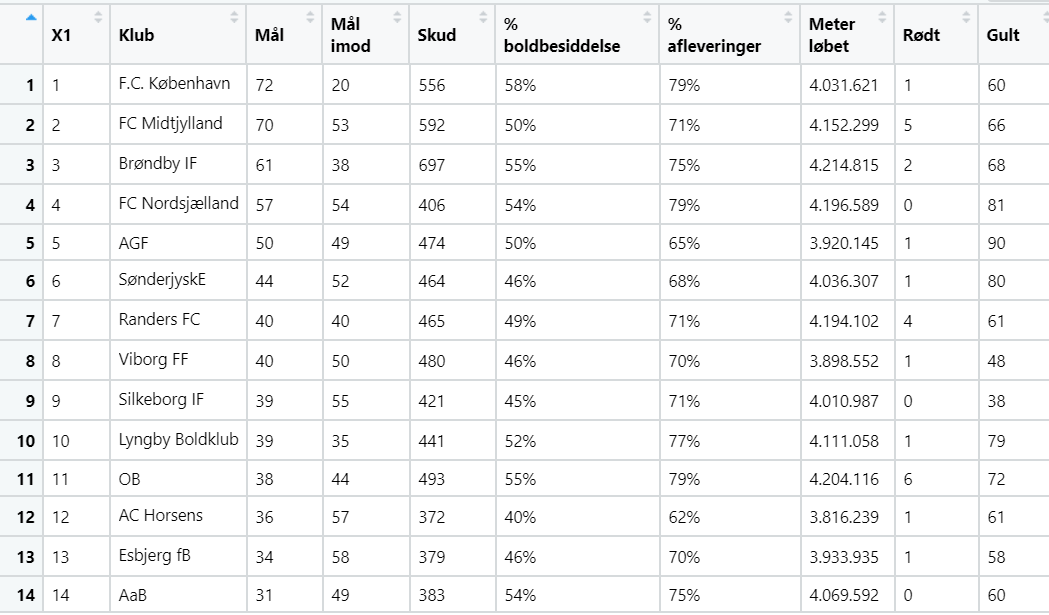
Now I had my data, but I needed to remodel it so that I would end up with the table of the data, so that I could start looking for coincidences and connections. From now on I´m just explaining how I managed to get my final dataset in RStudios for season 2016/17 because the procedure is exactly the same for season 2017/18. The way it worked out was by finding the exact code from the website where the statistics I needed was placed. I did that with the command ```{r}SL1617 %>%html\_table()-> SL1617```.

The HTML from the website had given me four different packages with tables, but now I knew it was package one and two where the information I wanted was stored. To eliminate the data, I had no use of I ran two new commands. ```{r}SL1617[[1]] -> klub2``` and ```{r}SL1617[[2]] -> restenaftabellen```. I called the useful packages klub2 and restenaftabellen, so that I could recognize these packages in global environment in RStudious. I could have called them anything I wanted it really didn´t play a role.

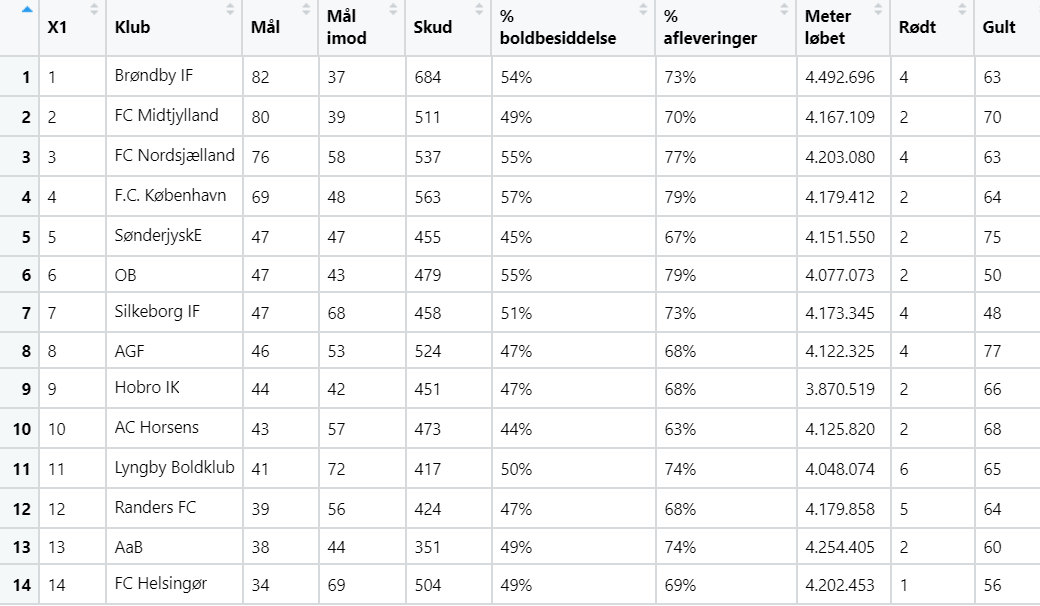
To bind the packages together I used the command: ```{r}bind\_cols(klub2,restenaftabellen) -> SL1617 ```. Now all the information I wanted was gathered in one table. I had a little technical problem, because the table had created an extra column and it didn´t look very good. It looked like this:



To remove the upper line in my table I used two commands. ```{r}write\_csv(SL1617, "SL1617.csv")``` and ```{r}SL1617 <- read\_csv("SL1617.csv", skip = 1)```. The commands created a new table which was much better looking because now all there was, was football statistics and no extra columns to confuse anyone. The table for the two seasons ended up looking like this:



The table for season 2016/17



The table for season 2017/18

Now I was ready to use the data and find the answers for my questions regarding possible coincidences and connections in Danish football. To understand how these new tables in RStudios could be put to best use I learned about ggplot. It is a core member of the tidyverse (one of the packages I installed at the beginning). Ggplot made it easy to access a dataset in RStudios. I learned about the program on a website about data visualization.[[1]](#footnote-1)

Now I had learned about the tools and was ready to search for answers. The first thing that came to mind was if shots on target (Skud) and goal scored (Mål) had a connection. To explore the amount of coincidence in this presumption I ran this command:

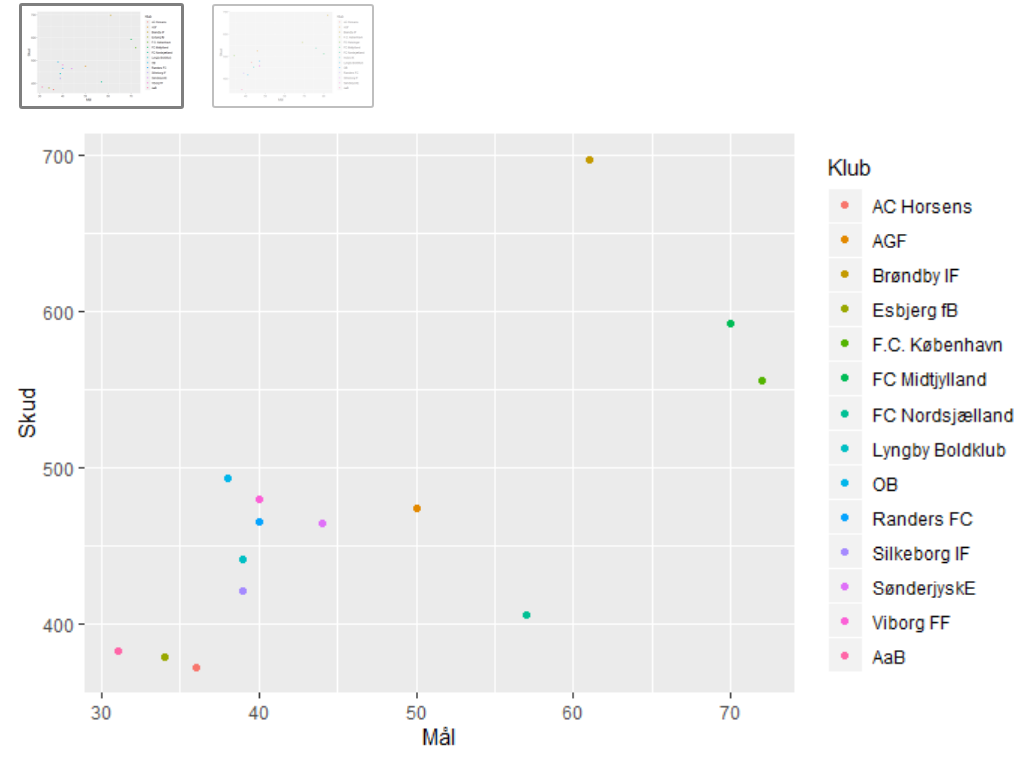
```{r}ggplot(data = SL1617) + geom\_point(mapping = aes(x = Mål, y = Skud, color = Klub))

ggplot(data = SL20172018) + geom\_point(mapping = aes(x = Mål, y = Skud, color = Klub)) ```

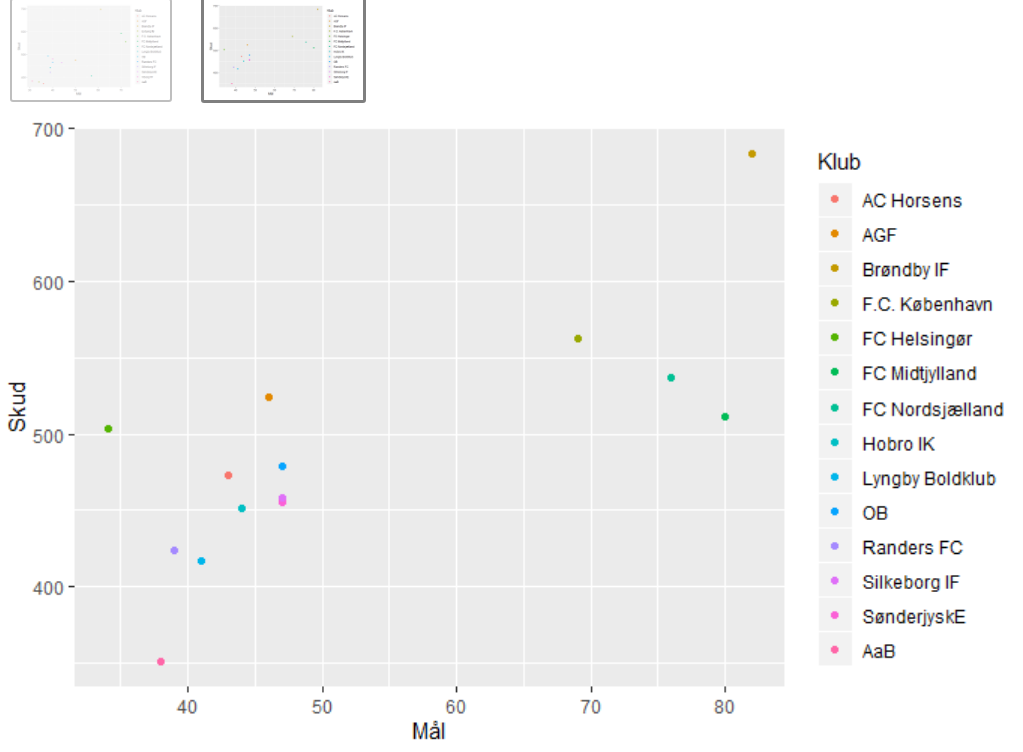
I ran both of the datasets in the same command to save myself some time and to avoid doing it all twice. The fact that I put an extra column in the command called Klub (Club) was to make it perceptible which information that belonged to each of the different clubs.

|  |  |
| --- | --- |
|  |  |

The final coordinate systems ended up looking like this:



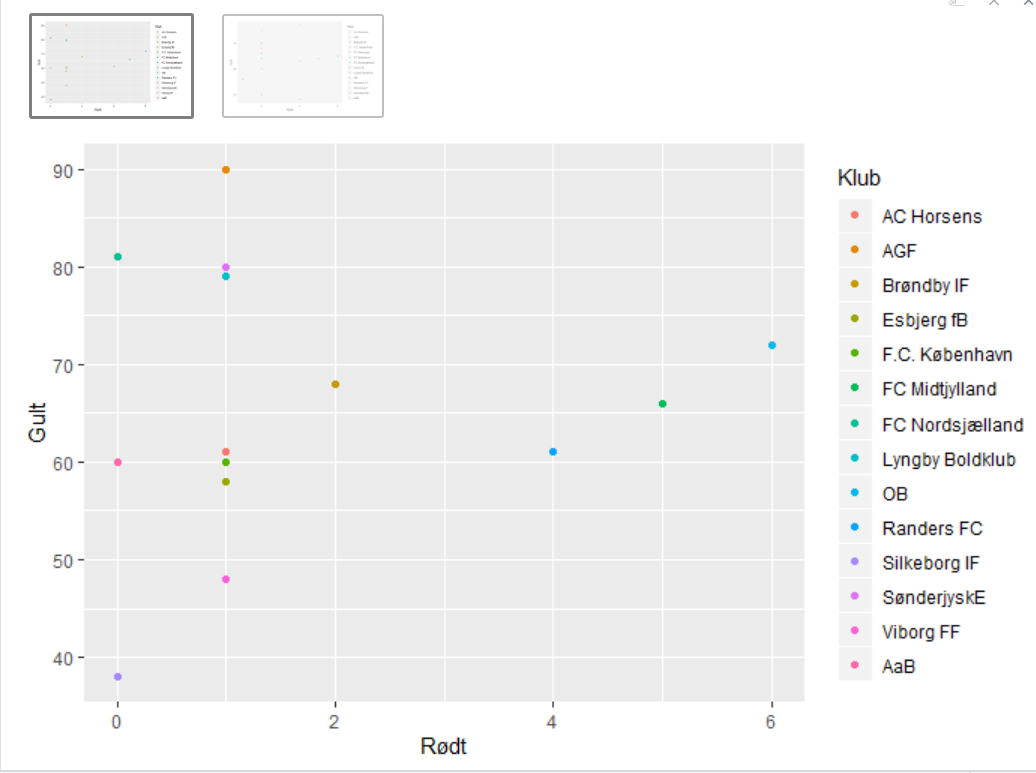
Coordinate system for Goals (Mål) and shots on target (Skud) for season 2016/17



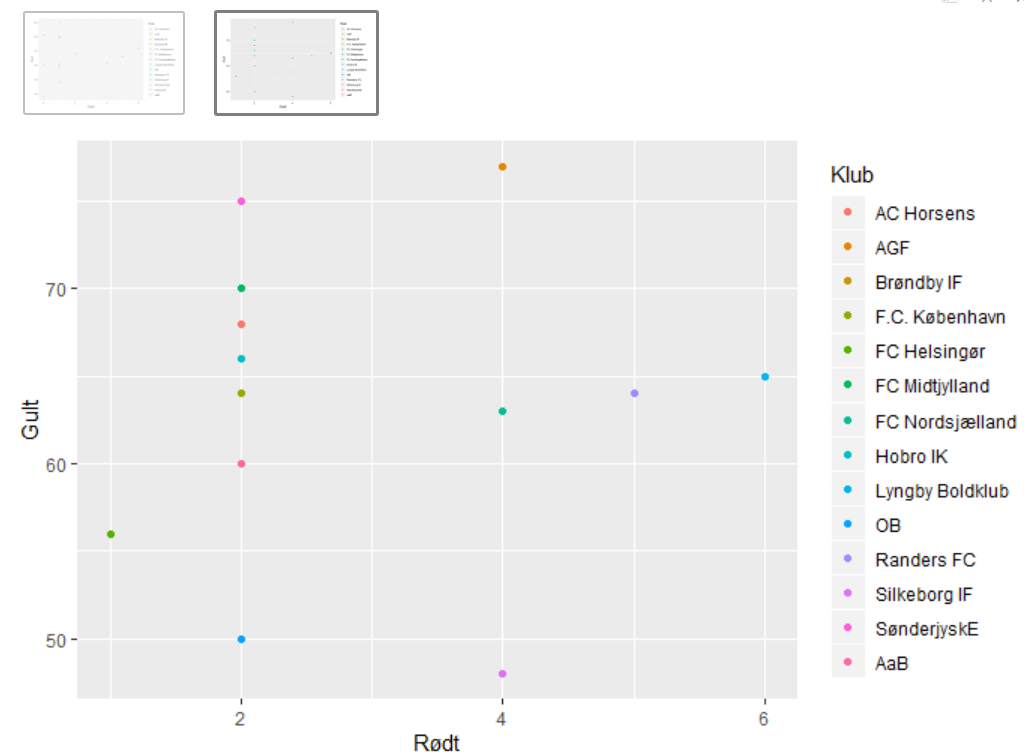
Coordinate system for Goals (Mål) and shots on target (Skud) for season 2017/18

For both of the seasons it is obvious that shots on target and goals have a connection. The team who creates few chances are less likely to score many goals and that is clearly stated in the exploration. Vise versa are teams who score a lot of goals very likely to create a lot of chances and therefore have a lot of shots on target. The easy way to check this is to look for a somewhat linear function from bottom left corner to upper right corner in the coordinate system. If such a line is possible to see, there most certainly is a connection.

Next, I would like to look into the interrelationship between received red and yellow cards. I thought there might be a connection between these two things, but to be sure I ran a function to investigate it in RStudios. The function I ran was as you see below and ended up looking like this: ```{r}ggplot(data = SL1617) + geom\_point(mapping = aes(x = Rødt, y = Gult, color = Klub))

ggplot(data = SL20172018) + geom\_point(mapping = aes(x = Rødt, y = Gult, color = Klub))```

The amount of red (Rødt) and yellow (Gult) cards received by each club in the Superliga for season 2016/17



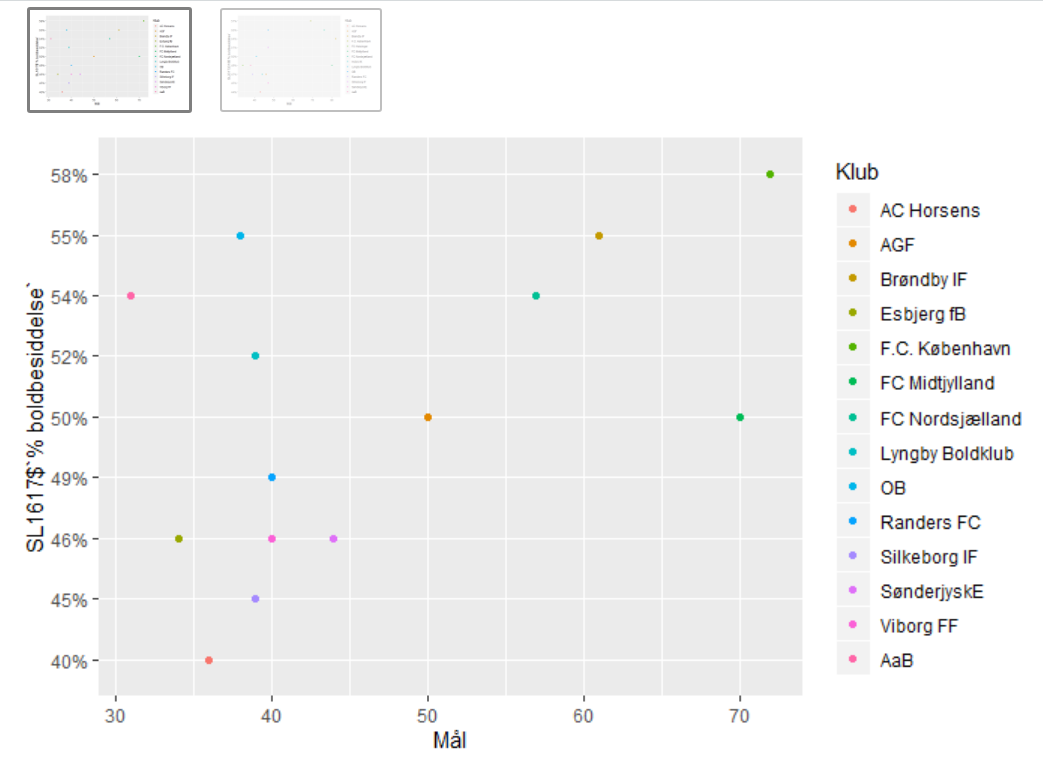
The amount of red (Rødt) and yellow (Gult) cards received by each club in the Superliga for season 2017/18

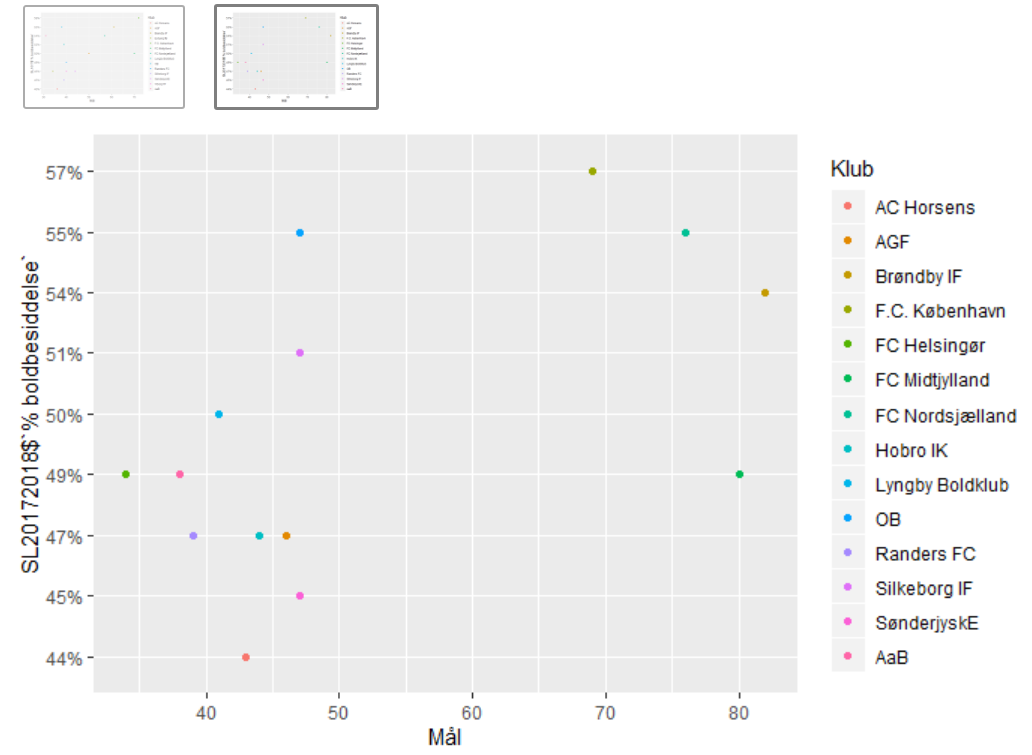
The coordinate systems created by the ggplot functions in RMarkdown makes the transparency of the data much easier to manage. It seems there are no specific interrelationship for the clubs according to receiving red and yellow card. Some clubs get many yellow cards but no red card at all after an entire season, and other clubs get up to six red cards but only receive the average amount of yellow cards. I most conclude that there is no connection at all between red and yellow cards received for the seasons 2016/17 and 2017/18 in the Danish Superliga.

I had a huge problem with my dataset in RStudios. It was easy to use the data which was in a headcolumn consisting of only one word like Mål (Goals), Rødt (Red), Gult (Yellow), Klub (Club) and so on. I couldn´t get the headcolumns consisting of two words working. It was the columns: Mål imod (goals against), % boldbesiddelse (% possession of the ball), % aflevering (% Passing) and Meter Løbet (Meters Run). I was asking my fellow students on this course, but no one could help me. I finally found a solution on a support webpage for RStudios.[[2]](#footnote-2) The thing that needed to be changed was the turf in which the two words were placed and instead of just filling in the turf I had to reference to the specific dataset by using the $ sign and bind the two words inside quotes. An example of how the function would look like now is: ```{r}ggplot(data = SL1617) + geom\_point(mapping = aes(x = Skud, y = SL1617$`Meter løbet`, color = Klub)) + labs(x = "SL1617")

ggplot(data = SL20172018) + geom\_point(mapping = aes(x = Skud, y = SL20172018$`Meter løbet`, color = Klub)) + labs(x = "SL20172018")```

I could now explore the data regarding all of my columns. So now I would like to look for a connection between Goals (Mål) and Possession in percent (% Boldbesiddelse). The command is: ```{r}ggplot(data = SL1617) + geom\_point(mapping = aes(x = Mål, y = SL1617$`% boldbesiddelse`, color = Klub))"

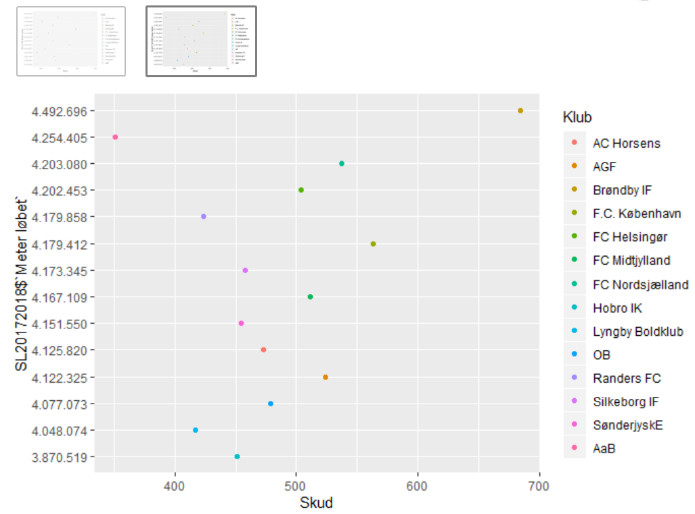
ggplot(data = SL20172018) + geom\_point(mapping = aes(x = Mål, y = SL20172018$`% boldbesiddelse`, color = Klub))```.



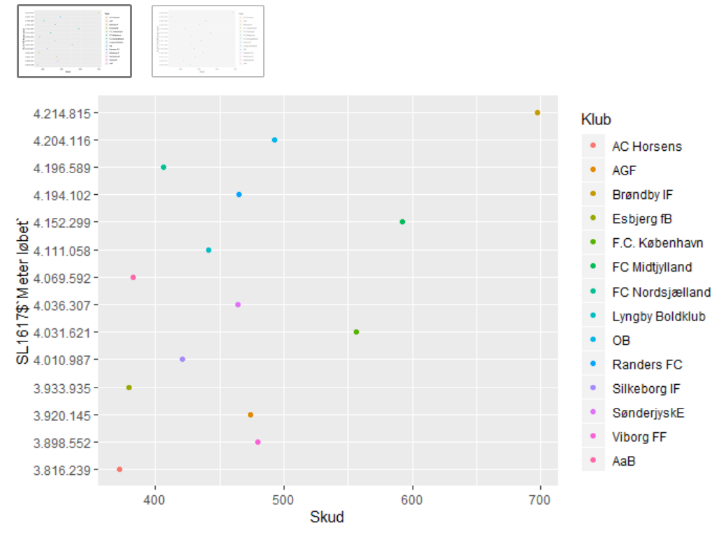
Coordinate system explaining Possession in percent (% Boldbesiddelse) and Goals (Mål) for season 2016/17

Coordinate system explaining Possession in percent (% Boldbesiddelse) and Goals (Mål) for season 2017/18

I can from the coordinate systems conclude that clubs with small amount of possession not score a lot of goals. It is only FC Midtjylland who for both seasons manage to have low possession combined with a lot of goals, otherwise any other team support the statement, that small amount of possession equals few goals.

 Next, I would like to look into the consistency of shorts on target (Skud) and Meters run (Meter løbet). Before I did any research on this topic I had absolutely no idea if the two parameters had any consistency. I ran the same command as I did in the survey on the previous page. The result turned out this way:

A graph explaining the amount of Shots on target (Skud) and Meters run (Meter løbet) for both of the seasons



The survey gives a kind of ambiguous result, because for the season 2016/17 season it seems like there are kind of a connection between the two words I have searched for. It is in this season obvious that the team with the fewest shots on target (AC Horsens) also are the team who have run the shortest distance in the entire league. The team with most shots (Brøndby IF) is also the team who have run fare more than any other team. For season 2017/18 the results are more strayed. There are again one team (Brøndby IF) who have run the most and have the most shots on target. The other teams that year make it harder to determine a conclusion, so in the end I can´t be sure there are any coherence in this screening.

I could keep presenting new graphs for surveys for a long time, but I have presented the once where I believed was a chance of finding either a connection or a somewhat close to a connection. Many more examinations are in the RStudios file for my surveys which is uploaded to my GitHub account[[3]](#footnote-3), so feel free to take a closer look, on some of the surveys I have left out of this assignment.

**Conclusions**

One thing I found out for sure is that Goals and Shots on targets are related. Both of the graphs created by ggplot in RMarkdown shows clear evidence for the interrelationship between these factors. Honestly it is not a tremendous surprise, but now I have the evidence to support my assumption.

Another thing I found out was the lack of connection between red and yellow cards received by each club in the Superliga. Apparently these two things are independent factors, and that was quite a surprise.

When it comes to the interrelationship between Possession in percent and Goals, I found it a bit surprising that the compiled coordinate system made by ggplot showed such a result. The interrelationship the two factors regarded was easy to see and it was far more perceptible than I had thought.

The digital tool I have been using for this assignment have been of great aid. I see how conjectures are much easier to deal with when I can use digital tools not only because the transparency of the data gets easier but because it really has saved me a lot of time.

**Acknowledgement**

I have to thank Max from Hovedbiblioteket in Aarhus for showing me how to compile my digital dataset and for teaching me how to find the exact HTML code I needed. It was also him who introduced me to ggplot. He has been of great use, and I´m not sure I could have done it without him.

**References**

Websites for my datasets

* <http://www.superliga.dk/performance-centre/2016-17>
* <http://www.superliga.dk/performance-centre/2017-18>

Introduction for ggplot in Tidyverse

* <https://r4ds.had.co.nz/data-visualisation.html#introduction-1>

Helping with my technical problems in RStudios

|  |  |
| --- | --- |
|  |  |

* <https://support.rstudio.com/hc/en-us/articles/200552336-Getting-Help-with-R>.

Installing Tidyverse for my project

* <https://tidyverse.tidyverse.org/>

Introduction to WebScraping

* <https://librarycarpentry.org/lc-webscraping/>

# B- Required Metadata

# *Table 1 – Software metadata*

|  |  |  |
| --- | --- | --- |
| **Nr** | **(executable) Software metadata description** | ***Please fill in this column*** |
| S1 | Current software version | RStudios 3.6.2 |
| S2 | Permanent link to executables of this version | [https://cran.rproject.org/bin/windows/base/release.htm](https://cran.rproject.org/bin/windows/base/release.htm?fbclid=IwAR1bKuutzMO3xWXdWMPKbzwWAZT54sBwwNDj_59HRs3wUHrd9hf_t-uHx5Q) |
| S3 | Legal Software License | Free for all |
| S4 | Computing platform / Operating System | *Microsoft Windows 64bit* |
| S5 | Installation requirements & dependencies | *Non* |
| S6 | If available Link to user manual - if formally published include a reference to the publication in the reference list | [https://cran.r-project.org/index.html](https://cran.r-project.org/index.html?fbclid=IwAR3lctv3aTLRzCgyR5KyLy-w2BuJByNJBgP7zk9tm7-6nq0e5jH4KPuddrI) |
| S6 | Support email for questions | Non |

1. <https://r4ds.had.co.nz/data-visualisation.html#introduction-1> [↑](#footnote-ref-1)
2. <https://support.rstudio.com/hc/en-us/articles/200552336-Getting-Help-with-R>. [↑](#footnote-ref-2)
3. <https://github.com/Digital-Methods-HASS/AU592093_Olesen_Claus> [↑](#footnote-ref-3)