K-means clustering on a multivariate dataset

Student No: 9914462

The data I will be using for my analysis is on global football team rankings, provided by the data analytics website FiveThirtyEight. The dataset ranks football teams based on their calculated SPI (Soccer Power Index) rating, an estimate of a team's overall strength.
Every team has an offensive rating that represents the number of goals it would be expected to score against an average team on a neutral field, and a defensive rating that represents the number of goals it would be expected to concede. These ratings, in turn, produce an overall SPI rating, which represents the percentage of available points - a win is worth 3 points, a draw worth 1 point, and a loss worth 0 points - the team would be expected to take if that match were played over and over again.[2]

The dataset we use[1] ranks 629 football teams by their SPI rating, as well as lists their offensive and defensive ratings.

We aim to cluster this data to group similarly ranked teams, and explore how large of a gap there is between the "elite" football teams and teams that don't regularly compete for the biggest competitions.

Before we do anything, we have to load our data into R, and load the necessary packages, which will allow us to plot our data using *ggplot*

```
data <- read.csv("spi_global_rankings.csv")
library(tidyverse)
library(cowplot)
```

We need to find an appropriate number k, for which we will perform our clustering. We use the elbow method detailed in the notes:

```
    elbow_data <- NULL
> for (i in 1:20) {
+     clustering <- kmeans(data[5:6], centers=i, nstart=10)
+     elbow_data <- rbind(elbow_data, data.frame(total_ss = clustering$tot.withinss, k=i))
+ }
> ggplot(data = elbow_data, aes(x=k, y=total_ss)) + geom_line() + geom_point() +
+     ylab("Total within cluster sum of squares") + xlab("Number of clusters")
```

Starting with $k = 1$ through to $k = 20$ on the offensive and defensive rating columns, we get the elbow plot in figure 1.

Looking at the plot in figure 1, $k = 3$ seems like a sensible choice, there isn't as sharp of a gradient as we might hope for, so choosing $k = 3$ may not be the best option.

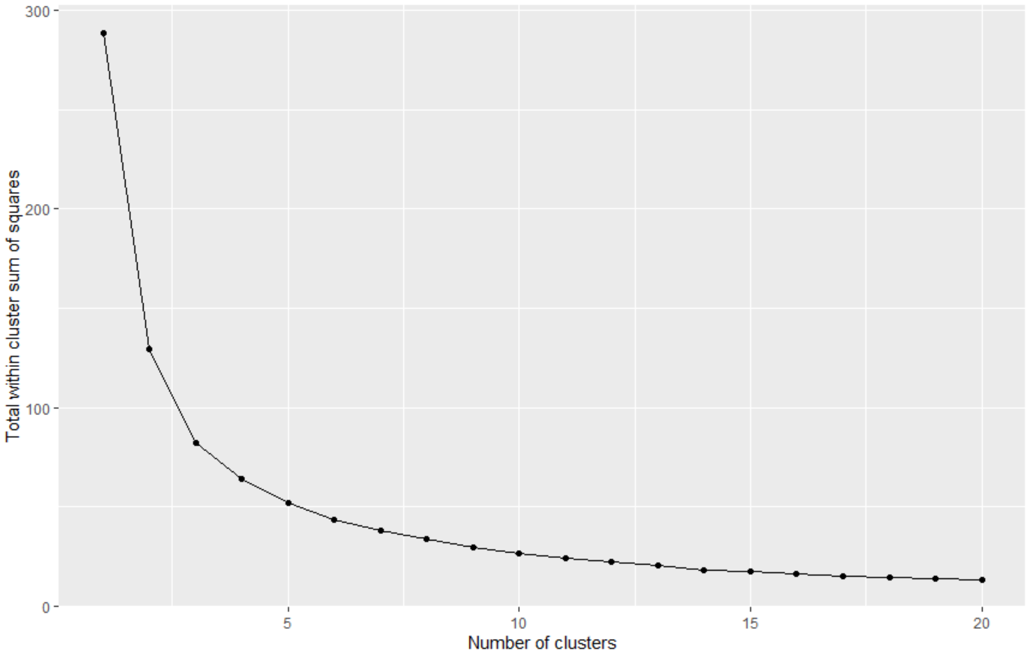

Figure 1: Elbow plot

Now that we have our k value, we can apply the k-means clustering algorithm.

```
k <- 3
clustering <- kmeans(data[,c("off", "def")], centers=k, nstart=10)
data$clusters <- as.factor(clustering$cluster)
ggplot(data = data, aes(x=off, y=def)) + geom_point(aes(colour = clusters), size = 3) + ylab("Defe
guides(colour = FALSE) + geom_text(aes(label = name), nudge_y = 0.08, check_overlap = TRUE)
```

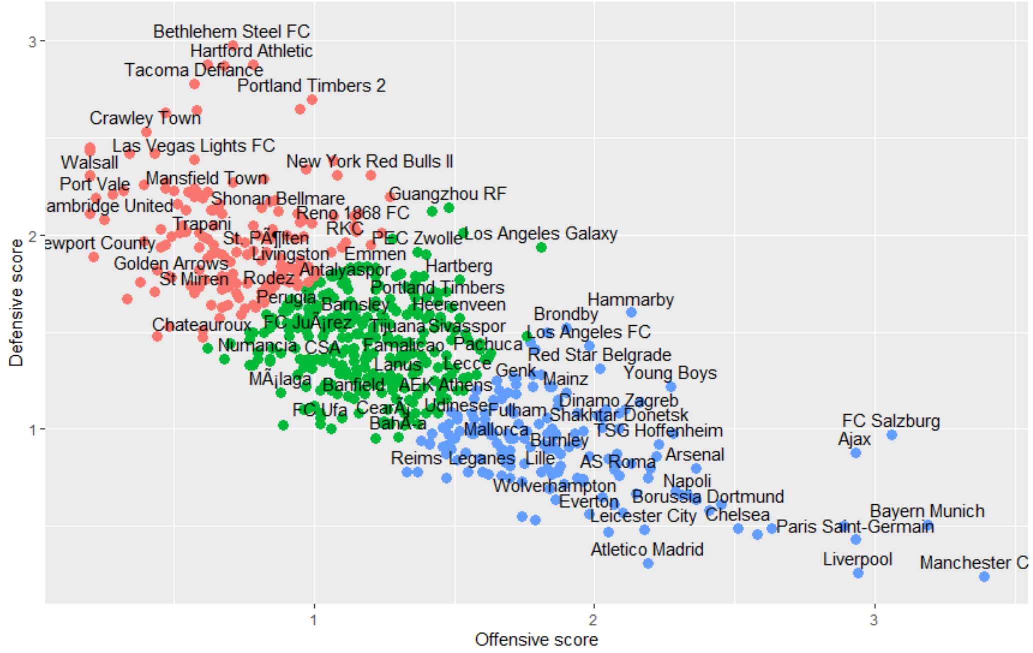

Figure 2: K-means clustering, k=3

Running the k-means code previously described, we get Figure 2. The plot shows out of the 269 teams, we can split them in to three clear categories:

- Teams that compete in European club competitions and/or one of the top European leagues.
- Teams that compete in the top leagues of the weaker footballing countries/second tier of the stronger footballing countries.
- Teams that compete in the lower leagues of weaker and stronger footballing countries.

Let's say we want to further analyse the elite teams, the teams coloured in blue in Figure 2. The following code:

```
nrow(data[data$clusters==3,])
```

gives us how many teams fall in to the blue category, which is subsequently calculated to be 161.

We save these teams in to a new dataset:

```
elite <- data[data$clusters==3,]
```

Now let's perform our k-means clustering on this new dataset. First we run another elbow plot using the new dataset. As we see in Figure 3, we have a very similar plot to before, and again $k = 3$ seems like a sensible choice.

Now we run our k-means code again, just replacing *data* for *elite*. The results are plotted in 4.

So within our our elite teams dataset, we can break it down in to 3 groups. The data points labelled in red correspond to the best teams in the world. They are:

Manchester City, Liverpool, Bayern Munich, Paris Saint-Germain, Barcelona, Juventus, Real Madrid, Chelsea, Ajax, Salzburg.

This top bracket of teams seems to be correlated with domestic and European success. The 10 teams span 7 different leagues, and indeed the winner of each of these 7 leagues falls within this top 10, with the other 3 teams (Liverpool, Real Madrid, Chelsea) finishing 2nd, 2nd, and 3rd respectively in their leagues. Furthermore, the winners of both European club competitions were won by teams in this top 10.

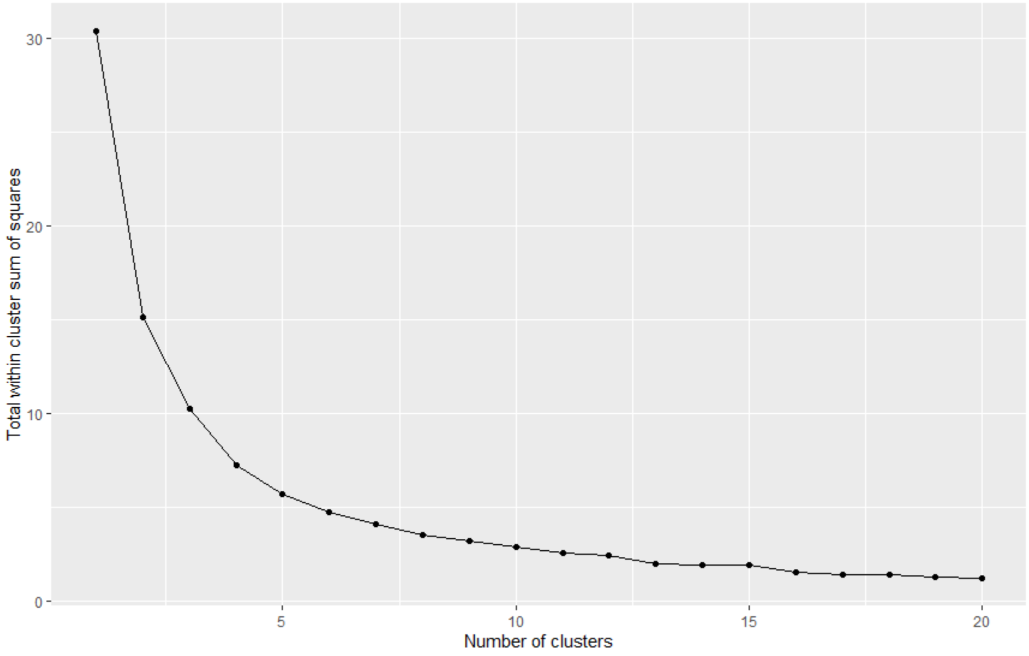

Figure 3: Elbow plot for elite teams

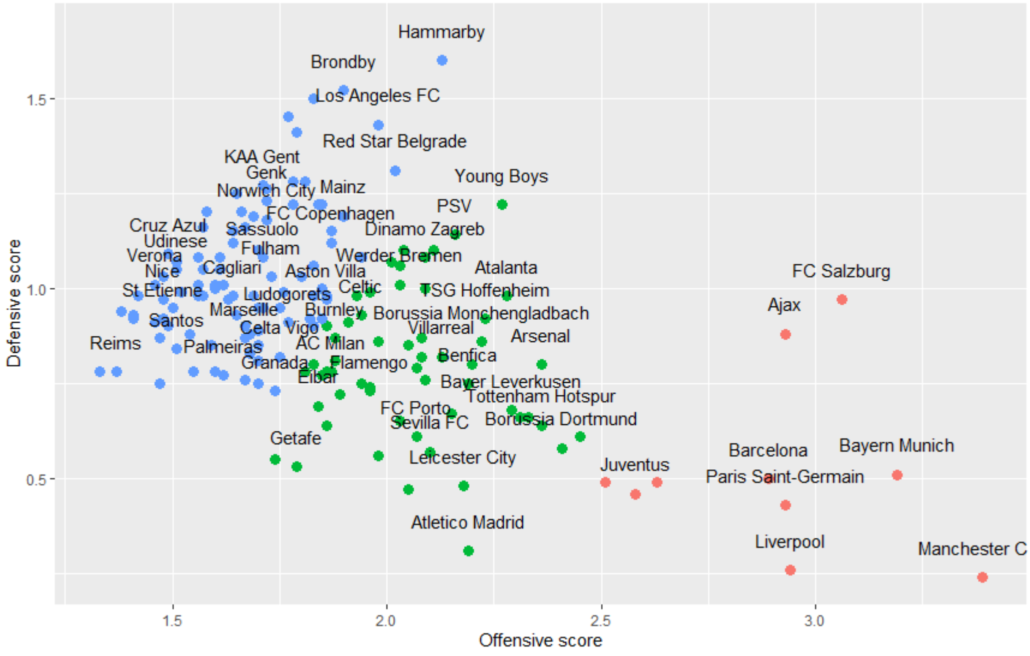

Figure 4: K-means clustering for elite teams, k=3

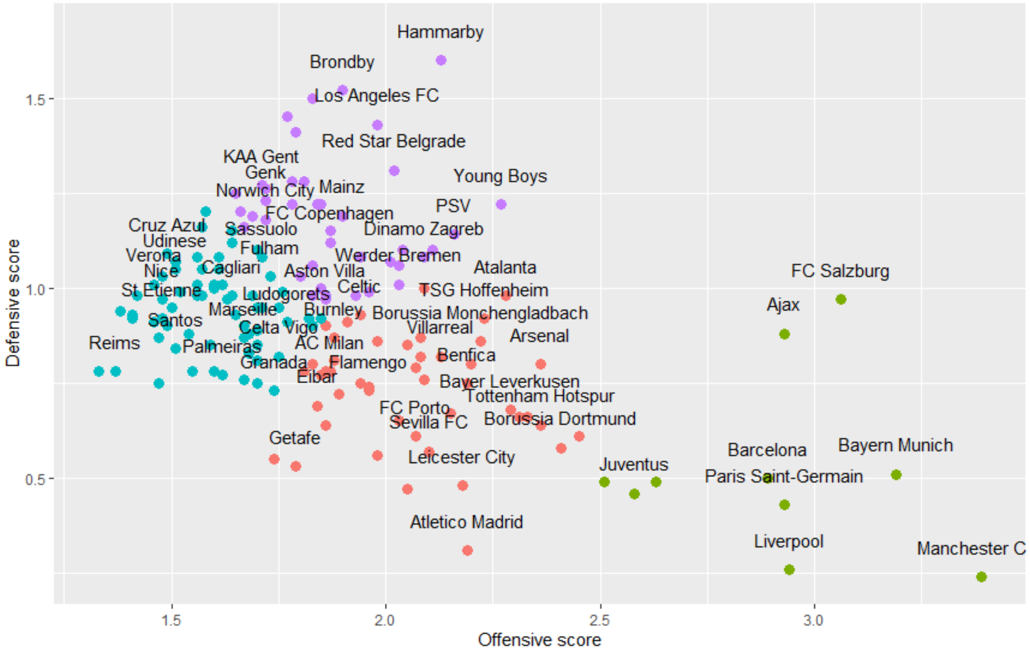

Figure 5: K-means cluster for elite teams, k=4

In conclusion we can be happy with the results produced by the k-means clustering algorithm, it has successfully identified the best performing football teams given their calculated offensive and defensive score. For the elite teams, the best performing group is clearly defined, there are 10 teams that stand out from the rest in 4. The other two groups are less clearly defined, the lowest scoring teams in the middle group are right next to the highest scoring teams in the lowest group. As we discussed before, our elbow-plot didn't have a particularly sharp 'L-shape', so $k = 3$ isn't necessarily the best choice. If we set $k = 4$, we get Figure 5. Choosing $k = 4$ mostly just splits up the lowest group, in to a group that is better defensively but worse offensively (Blue), and a team that is better offensively but worse defensively (Purple).

References

[1] Jay Boice. *FiveThirtyEight Data*. Sept. 2018. URL: `https://github.com/fivethirtyeight/data/tree/master/soccer-spi`.

[2] Jay Boice. *How Our Club Soccer Predictions Work*. Aug. 2018. URL: `https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/`.