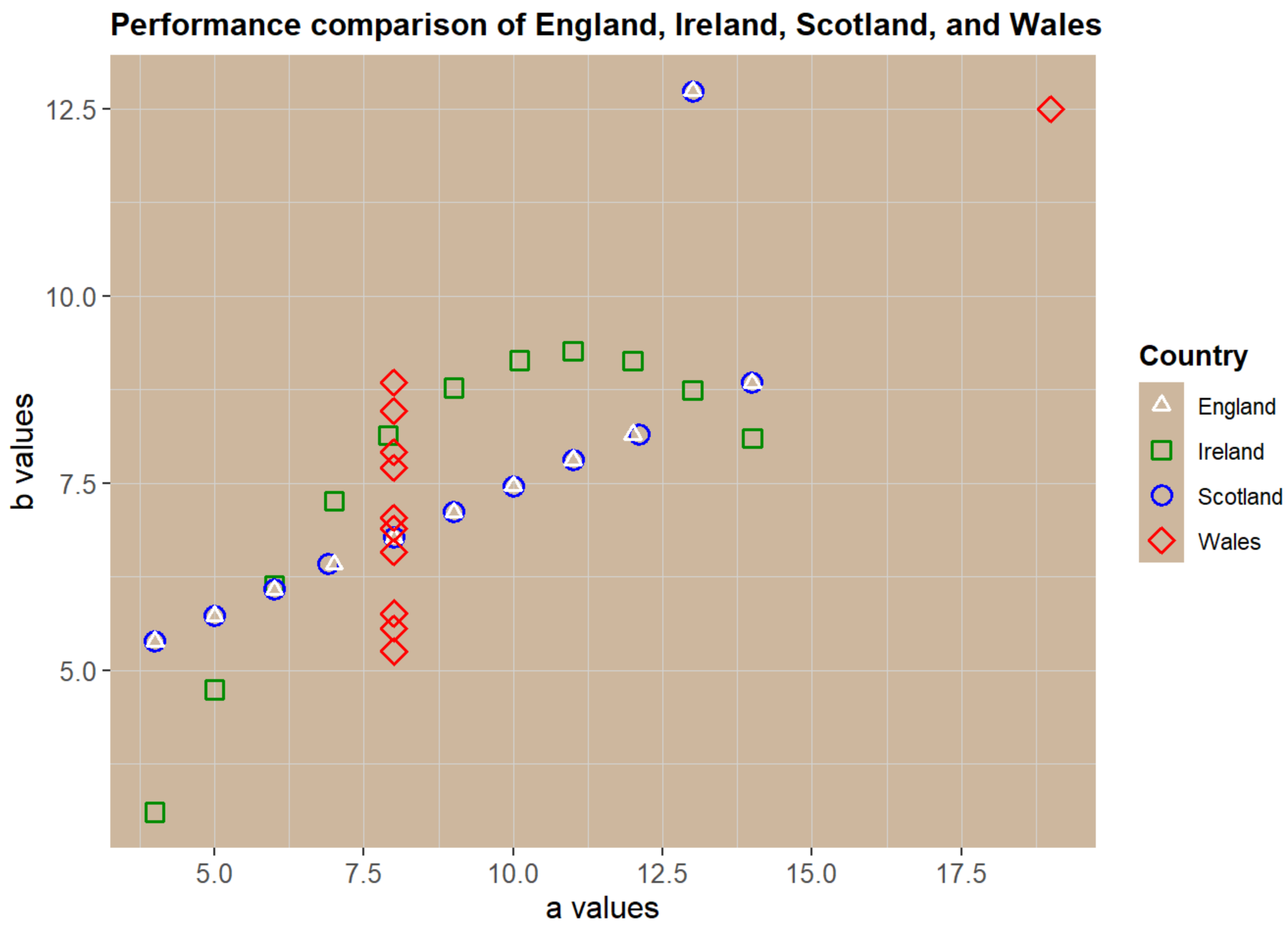
Assignment 1: Scatterplot

Introduction:

This scatterplot visualisation presents a performance comparison of England, Ireland, Scotland, and Wales in relation to variables ‘a’ and ‘b’.The colour of each country has been chosen to be the same as their national rugby t-shirt.



Techniques that have been used:

Aesthetics

Axes

Axis x and Axis y are defined to represent variable ‘a’ and ‘b’ respectively.

Colour

The colour of the datapoints should be representative of each country:

- England: White
- Ireland: Green
- Scotland: Blue
- Wales: Red

White colour does not stand out much in a grey background, it is not easy to find a colour where white, green, blue, and red are visible enough. After trying different colours, I finally chose ‘bisque3’ for the background. I have used the same colour to define the legend in order to facilitate the linking between the legend and the plot.

I tried changing alpha to help visualise all overlapping datapoints but I thought it did not help enough. At the same time, datapoints with lower alpha were more difficult to be seen. I realised that changing colour, shape, and size was enough to show these datapoints.

Geometry

By default, the shape of the countries was set to ‘circle’. There are cases in which datapoints of different countries have the same values. In this cases, they overlapped and it’s just possible to see one datapoint. Using different sizes per country can solve this issue, however it can be misleading and make us think that a country has more importance than others. Because of this, I have decided to change the shape of each country. I have tried few combinations until I was happy on how the overlapping points were being seeing.

At the end, I changed some of the sizes. England (triangle) has a smaller size. As we are using different shapes, I think that this is not obvious and helps to visualise overlapping points.

Theme

Finally, I have done some changes on the main title and the legend such as using bold to make the title stand out from other parts of the pdf. Additionally, axis x was showing by default values 5, 10 and 15. As Wales has most of it’s values between 5 and 10, I added extra values: 5, 7.5, 10, 12.5, 15 and 17.5 to make it easier for the reader to identify them.

Table of statistical summary values

In the assignment, we were suggested to calculate the mean, standard deviation and correlation of variable_a and variable_b. I have added an additional table with with the maximum, the minimum, and the quantiles 25%, 50% (median), and 75%:

country	quantile_25_a	quantile_25_b	median_a	median_b	quantile_75_a	quantile_75_b	cor_ab
england	6.50	6.250	9	7.11	11.50	7.98	0.8162867
ireland	6.50	6.695	9	8.14	11.50	8.95	0.8161639
scotland	6.45	6.250	9	7.11	11.55	7.98	0.8150854
wales	8.00	6.170	8	7.04	8.00	8.19	0.8165214

country	min_a	min_b	max_a	max_b	mean_a	mean_b	sd_a	sd_b
england	4	5.39	14	12.74	9	7.500000	3.316625	2.030424
ireland	4	3.10	14	9.26	9	7.500909	3.322951	2.031657
scotland	4	5.39	14	12.74	9	7.500000	3.331966	2.030424
wales	8	5.25	19	12.50	9	7.500909	3.316625	2.030578

This table allows us to see some additional information:

- England, Ireland, and Scotland have the same minimum (4) and maximum (14) for variable_a, while Wales presents higher values in both cases.
- England, Scotland, and Wales have similar minimum and maximum for variable_b, while Ireland presents lower values in both cases.
- By looking at the quantiles, it seems that England, Ireland and Scotland have a similar distribution of variable_a.
- By looking at the quantiles, it seems that all countries have a similar distribution of variable_b.
- All countries have similar mean and standard deviation for variable_a and variable_b

Now, there are things that I would have not been able to realise by just looking at this table:

- England are Scotland have a very similar performance. It is almost identical.
- Wales behaves in a completely different way than the rest of the countries.
- Ireland datapoints seems to follow a quadratic correlation.
- England, Scotland and Wales have outliers. Only three datapoints have variable_a > 15 or variable_b > 10.

Finally, I would like to mention that I did not expect this discrepancy between the correlation presented in the table what is seen in the plot. I would have expected Ireland to have a lower linear correlation in comparison to the other countries.

Appendix

```r
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
# Loading data in dataframe
data1 <- read.csv("week2_data_4cat.csv")
# Creating Scatterplot
g <- ggplot(data1, aes(x=var_a, y = var_b,colour=country,shape = country,alpha=country))+
  geom_point(aes(size=country),stroke=1)+
  scale_colour_manual(values = c("white","green4","blue","red"), labels = c("England","Ireland","Scotland","Wale
s"),name = "Country")+
  scale_shape_manual(values = c(2,22,1,23), labels = c("England","Ireland","Scotland","Wales"),name = "Country")+
  scale_size_manual(values = c(1.5,3,3,3), labels = c("England","Ireland","Scotland","Wales"),name = "Country")+
  scale_alpha_manual(values = c(1,1,1,1), labels = c("England","Ireland","Scotland","Wales"),name = "Country")+
  ggtitle("Performance comparison of England, Ireland, Scotland, and Wales")+
  xlab("a values")+
  ylab("b values")+
  scale_x_continuous(breaks = c(5,7.5,10,12.5,15,17.5,20))+
  theme(panel.background = element_rect(fill = "bisque3"),
        panel.grid.major=element_line(size=0.25,linetype = 'solid',colour = "lightgrey"),
        panel.grid.minor =element_line(size = 0.1,linetype ='solid',colour = "lightgrey"),
        legend.title = element_text(face = "bold",size = 11),
        legend.background = element_rect(fill = "transparent"),
        legend.key = element_rect(fill = "bisque3"),
        axis.text = element_text(size=10),
        axis.title = element_text(size = 12),
        plot.title = element_text(face = "bold",size=12))
g
library(dplyr)
library(knitr)
# Creating table 1
data1_stats<- data1 %>%
  group_by(country) %>%
  summarize(quantile_25_a = quantile(var_a,probs=c(0.25)), quantile_25_b = quantile(var_b,probs=c(0.25)), median_
a = median(var_a), median_b = median(var_b),quantile_75_a = quantile(var_a,probs=c(0.75)), quantile_75_b = quanti
le(var_b,probs=c(0.75)),cor_ab=cor(var_a,var_b))
kable(data1_stats)
# Creating table 2
data1_stats<- data1 %>%
  group_by(country) %>%
  summarize(min_a=min(var_a),min_b=min(var_b),max_a=max(var_a),max_b=max(var_b),mean_a = mean(var_a), mean_b = me
an(var_b), sd_a = sd(var_a), sd_b = sd(var_b))
kable(data1_stats)
```