

Visualising the PISA reading scores

Part 1

The key aspects of the visualisation decisions behind the plot given in the assignment are as follows:

- The aesthetic mapping:
The plot has country names on the x axis arranged in increasing order by Value field of Boys and y axis has the Value field which is numeric. The shape, colour and fill of the points are based on Subject field which has three categories (Boy, Girl and Total) but data is represented only for Boys and Girls in the plot.
- The axes:
The axis title and ticks are not present in the plot. The x axis values are tilted to 45 degrees and the size is also small than the default size. Also, two values on x-axis are highlighted in different colours to interpret their values easily (OECD-Average in black and Ireland in Red). The x scale is also expanded a little bit, to make the visualisation look good.
The y axis text has different position than the default position, the text is inside the plot right above the grid line of y axis.
- The gridlines:
The major gridlines for x are blank and minor gridlines for y are blank in the plot instead there is a white line which looks like grid line from x axis till the value of Boy data point and a blue line from Boy data point to Girl data point. This helps in the easier interpretation of the difference between Boy and Girl value. eg. We can clearly see that the difference is least in Colombia.
- The legend:
The legend is at the bottom left corner instead on the right side and the legend labels are placed horizontally. It shows the shape and colour used to represent the points on the plot for Boy and Girl. The legend title is not present.
- The background:
The background is a very light colour (#E2EDF3) with white gridlines.
- Colour:
The colours used to represent the data points are:
For Boy: The circles are filled with a dark blue colour(#406D89) and the boundary of the circle is white which is very lightly visible, but it gives a very good visualisation.
For Girl: The rhombus is filled with the same colour as background(#E2EDF3) and the colour of boundary of rhombus is the same dark blue colour and same white boundary is used outside the rhombus.
Apart from the colours of data points the combination of colours are chosen such that it gives the information required easily just by looking at the graph.
- Title and Subtitle:
The title and subtitle are both present on the same line, title has a size larger than the other part.

The fill in the rhombus is same as the background colour it should not be transparent.

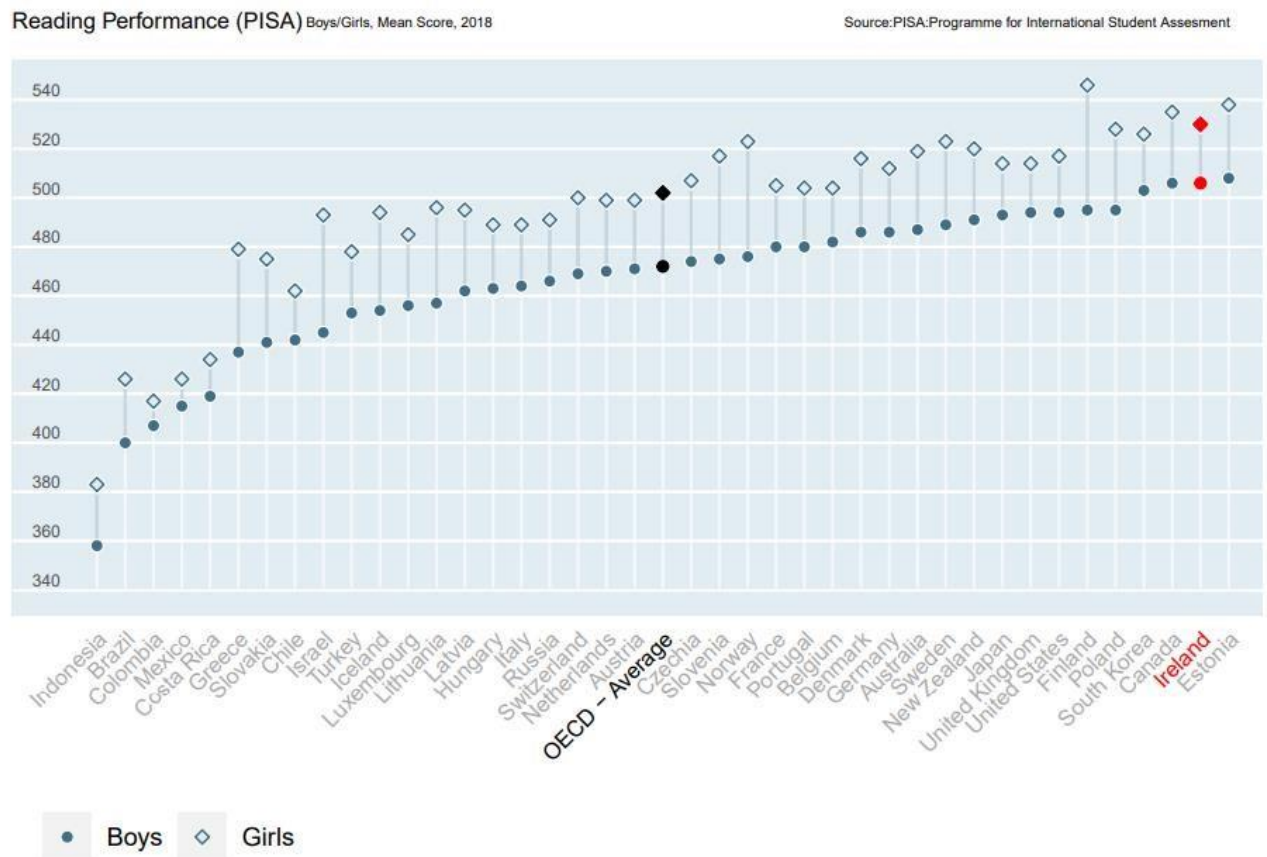
Part 2

The plot is reproduced by using only the ggplot2 features, the features and values used are explained below:

- The aesthetic mapping:
The mapping of x-axis in the increasing order of value by Boy is done by using `reorder(Country, Value)` by filtering only the Boy data and using `fct_inorder(Country_names)` in the `geom_point`. The shape, color and fill are set to the SUBJECT variable.
- The axes:
The axis title and ticks are set to blank and the position, size and angle are adjusted by using the parameters `hjust`, `vjust`, `size` and `angle`. The two values are highlighted by passing a vector of customised colours to the `color` parameter. The x axis scale is expanded using `expand_scale()` function.
The position of y axis text is adjusted using `margin` function and the values are set using `scale_y_continuous()` function.
- The gridlines:
The major x gridlines and minor y gridlines are set to blank. The other lines (white and blue) are plotted using `geom_segment`. `Geom_segment` is used two times, one for the white line (x axis to the Boy data point) and one for blue line (Boy to Girl data point) by passing the required values to `x`, `y`, `xend` and `yend` parameters.
- The legend:
Legend position is adjusted by setting `position` to the bottom, `justification` to left and `box` as horizontal. The title is set to blank and the legend values are changed to Boys and Girls instead of BOY and GIRL. This is done in `scale_manual` functions by passing these values as labels.
- The background:
The background colour is changed to `#E2EDF3`.
- Colour:
The fill colour, colour and shape are set using `scale_fill_manual`, `scale_color_manual` and `scale_shape_manual`. The colours used are:
For Boy: Fill and colour used is `"#406D89"` and the boundary is white. For Girl: The fill is `"#E2EDF3"` and the colour is `"#406D89"` and the boundary is white.
The background is `"#406D89"`, the values and data points of Ireland and OECDAverage are coloured in red and black respectively.
- Title and Subtitle:
The title and subtitle are set in the same line with different sizes by adjusting the `size`, `vjust` and `hjust` parameters for both title and subtitle.

Some key points to keep in mind:

The rhombus for Girl's data point is not transparent it should be filled with light blue colour.
 All the data points have a white boundary to make the visualisation clearer.
 The x axis text is dark grey for all values except OECD-Average(black) and Ireland(Red) to emphasize more on these values.



Code:

```
#Import the required Libraries
```

```
library(ggplot2  
)
```

```
library(dplyr)
```

```
library(countrycode)
```

```
library(tidyr)
```

```
library(forcats  
)
```

```
#Read and add the Country names column data_original
```

```
<- read.csv('OECD_PISA.csv')
```

```
data_original$Country <- countrycode(data_original$ï..LOCATION,origin =  
'iso3c', destination = 'country.name')
```

```
#Set NA to OECD-Average
```

```
data_original$Country[is.na(data_original$Country)] <- 'OECD - Average'
```

```
#Part 2 plot: Subset the data only for Boy and Girl
```

```
data <- data_original[which(data_original$SUBJECT %in% c('BOY','GIRL') &  
data_original$TIME == '2018'),] data_1 <- data[which(data$SUBJECT ==  
'BOY'),] data_1 <- data_1[order(data_1$Value),] data_2 <-  
data[which(data$SUBJECT == 'GIRL'),]
```

```
#rbind both boy and girl data to get the order of Countries according to  
Boy
```

```
data_3 <- rbind(data_1,data_2) Country_names
```

```
<- data_3$Country
```

```
#use pivot_wider to use it in geom_segment
```

```
data_new <- pivot_wider(data,names_from = SUBJECT,values_from = Value)
```

```
data4 <- subset(data,SUBJECT == 'GIRL')['Value']
```

```
#For theme colors
```

```
col_avg <- subset(data, Country == "OECD - Average")
```

```
col_ire <- subset(data, Country == "Ireland") custom_clr
```

```
<- ifelse(data_1$Country == 'Ireland', "red",
```

```
ifelse(data_1$Country == 'OECD - Average',  
'black',"darkgrey"))
```

```
plot1 <- ggplot(data =subset(data,SUBJECT == 'BOY'),
```

```
aes(x=reorder(Country,Value), y=Value)) +
```

```
geom_segment(data =data_new, #for Boy to Girl blue lines
```

```
aes(x = reorder(Country,BOY),y = BOY,
```

```
= reorder(Country,BOY),yend = GIRL),
```

```
xend
```

```
size =
```

```

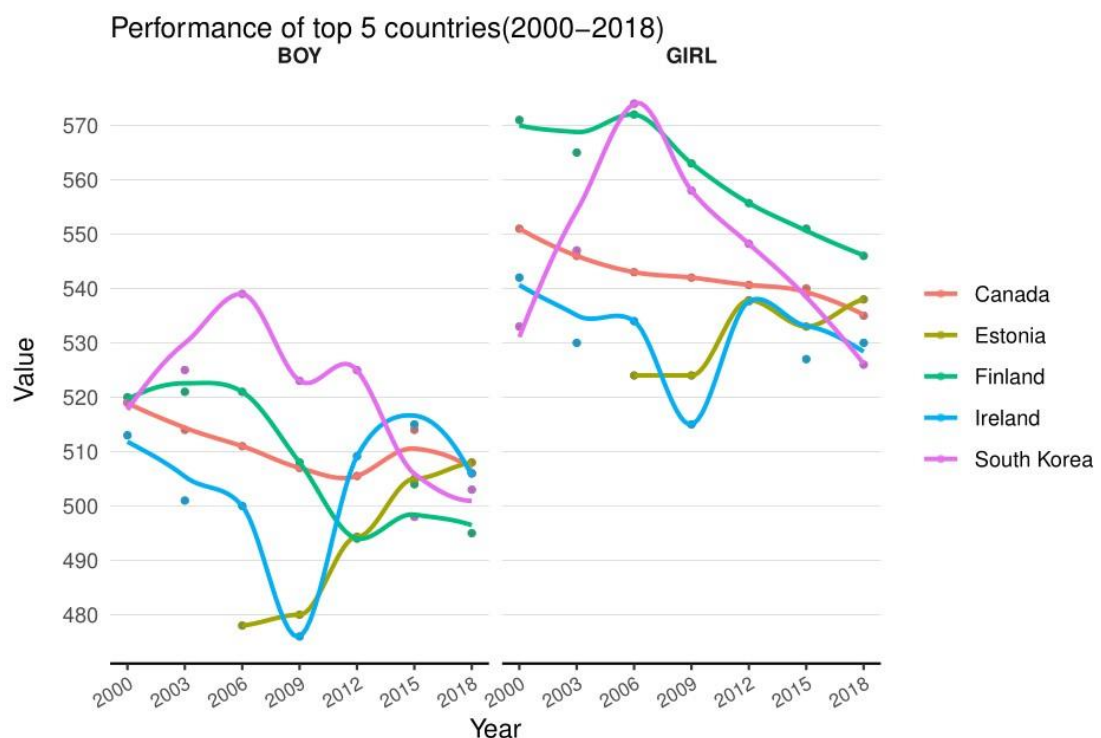
0.5,alpha=0.4,color = '#406D89')+      geom_segment(data =
data_3, #for white gridlines
      aes(x = Country,y = 340,xend = Country,yend = Value),
size = 0.5,alpha=0.4,color = 'white')+      #for the white boundaries
on both boy and girl data points      geom_point(data =
subset(data_3,SUBJECT == 'BOY'), color="white",      shape =
16,size = 2, show.legend = FALSE)+      geom_point(data =
subset(data_3,SUBJECT == 'GIRL'), color="white",      shape
= 23,size = 2, show.legend = FALSE)+
      #for the main data points
      geom_point(data = data_3,aes(x = fct_inorder(Country_names),
y=Value,      color=SUBJECT, shape = SUBJECT,fill =
SUBJECT))+
      scale_fill_manual(values = c("BOY" = "#406D89","GIRL"= "#E2EDF3"),
labels = c("Boys", "Girls"))+
      scale_color_manual(values = c("BOY" = "#406D89","GIRL"= "#406D89"),
labels = c("Boys", "Girls")) +      scale_shape_manual(values = c("BOY"
= 16, "GIRL" = 23),      labels = c("Boys",
"Girls"))+      scale_y_continuous(breaks=seq(340, 560, 20))+
      scale_x_discrete(expand = expand_scale(add = c(3,1.2)))
#Theme settings
plot1+ #highlight data points for Ireland and OECD-Average
      geom_point(data=col_avg,colour="black",shape = c(21,23),fill = 'black')
+
      geom_point(data=col_ire,colour = 'red',shape = c(21,23),fill = 'red')+
labs(title = "Reading Performance (PISA)",
      subtitle = "Boys/Girls, Mean Score, 2018
Source:PISA:Programme for International Student Assesment",
legend )+
      theme(axis.text.x = element_text(angle = 45, hjust=1,size =8,
colour = custom_clr),      axis.text.y = element_text(vjust = -
0.3,
      margin = margin(l = 20, r = -20),size =
6),
      axis.ticks = element_blank(),
axis.title = element_blank(),
      panel.background = element_rect(fill = "#E2EDF3"),
panel.grid.major.x = element_blank(),
panel.grid.minor.y = element_blank(),      legend.title
= element_blank(),      legend.position="bottom",
legend.box = "horizontal",      legend.justification =
"left",      legend.box.background = element_blank(),
plot.title = element_text(size = 8,vjust = -2.5),
      plot.subtitle = element_text(size = 5,vjust = 5.5,hjust = 0.9))

```

Part 3

The plot below shows the trend of performance of top 5 countries over the years (2000-2018). These countries are selected based on:

Top 5 countries with the highest values of SUBJECT = "TOT" in 2018 are considered. As these countries showed the best performance in 2018 for both Boys and Girls, we can visualise the trends (if any) for these top 5 countries over the years.



From the plot above we can visualise the trends for the top 5 countries (Canada, Estonia, Finland, Ireland and South Korea). Some trends and comparisons visible from the plot are:

- South Korea, Finland and Canada shows a decrease in the performance from 2000-2018 for both boys and girls.
- Ireland's performance decreased a lot in 2009 for both and now it is same as it was in 2000.
- Estonia's performance has improved a lot since 2006 for both.
- South Korea's trend is unstable over the years for Boys.
- The performance trend for Boys and Girls for all 5 countries is similar as we can see that girls performance is better as compared to boys for all countries.

Some feature used to produce this plot are:

- The year is in x axis and Value in Y axis, the colour is according to 5 countries present in the legend and the facets has 2 values Boy and Girl.
- `Geom_smooth` is used to understand the trend more easily, for the precise values `geom_point` is used with `alpha = 0.6`.
- Theme is like the plot in Part 2 with very few minor changes like background colour, axis ticks and legend position.

Code:

#Part3: Top 5 countries with the highest values of SUBJECT = "TOT" in 2018 are considered. As these countries showed the best performance in 2018 for both Boys and Girls, we can visualise the trends(if any) for these top 5 countries over the years.

```
#Select the top 5 countries with highest Value of Total in latest year available(2018) and compare the performance for Boy and Girl across all the years. data_tot <- data_original[which(data_original$SUBJECT == 'TOT' & data_original$TIME == '2018'),]  
data_tot <- tail(data_tot[order(data_tot$Value),],n=5)
```

#Subset these 5 countries for Boy and Girl

```
data_plot2 <- data_original[which(data_original$Country %in% data_tot$Country),]  
data_plot2 <- data_plot2[which(data_plot2$SUBJECT %in% c('BOY','GIRL')),]
```

```
plot2 <- ggplot(data_plot2,aes(x=TIME,y=Value, color = Country,fill=Country))+  
  geom_point(color="black",  
             shape = 16,size = 1.3, show.legend = FALSE,alpha=0.6)+  
  geom_point(alpha=0.6,size = 1)+ geom_smooth( method = "loess",  
size = 0.9,se = FALSE)+  
  scale_x_continuous(breaks=seq(2000, 2018, 3))+  
  scale_y_continuous(breaks=seq(480, 580, 10))+ facet_grid(cols  
= vars(SUBJECT))+  
  labs(title = "Performance of top 5 countries(2010-2018)", x="Year", y="Value")+  
  
  theme_classic()+  
  theme(axis.text.x = element_text(angle = 30, hjust=1,size =8),  
axis.ticks.y = element_blank(),  
axis.title.y=element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)),  
axis.line.y = element_blank(),  
legend.title = element_blank(),  
panel.grid.major.x = element_blank(),  
panel.grid.minor.y = element_blank(),  
panel.grid.major.y = element_line(colour="grey", size = 0.1),  
plot.title = element_text(size = 12,vjust = -2.5),  
plot.subtitle = element_text(size = 8,vjust = 5.5,hjust = 1.2),  
strip.text.x = element_text(size=9, face="bold"),  
strip.background = element_blank()  
)  
plot2
```