Assignment 4

Part 1: Summary of Visualizations:

The aesthetic mappings

- X -axis represents countries, Y- axis represents Reading performance, Points = Gender (BOYS and GIRLS)
- Ireland is represented in red colour, OECD Average values are represented in black color to differentiate from the rest

The axes

- X-axis is discrete scale – i.e., categorical values – Country Names,
- Y-axis is in continuous scale with the Reading performance values
- The data is ordered by scores of boys in ascending order
- The axis titles are removed for both X and Y axes.
- Y axis scales are between 340 and 560.
- The labels are inside the plot rather than being on the outside and sitting on the grid lines
- X-axis label is tilted to 45 degrees angle, to prevent running into each other. The country names are displayed in full.

The gridlines

- Only major grid lines on X – axis.
- The X axis grid lines start from the first Y axis grid line – leaving a space at the bottom
- X-grid lines rather than running all the way, they end at the boy scores.
- Only major Grid lines on Y axis
- Connectors between boy's and girl's values, points paired together to easily connect them.
- Grid lines don't show up in the points. There is a white background surrounding the points

The Legend

- Legend is moved to bottom left and horizontal.
- The GIRL legend rhombus point has a white background
- Fill for the rhombus is light blue on the plot.

The background

- The background for the plot is blue in colour and there is slightly darker blue line on top

Colours Used

Background: #E4EDF2, Boys scores: #4B6C86, Girls scores: Border: #4B6C86, Fill: #E4EDF2, Ireland – Red, OECD Average: Black, Annotation Line at top: #498EBF

Other Visual details

- The title and source for the plot is mentioned left aligned and right aligned respectively
- Data Detail is mentioned as sub-title
- The font for all text has been modified from the default values.

Part 2:
Plot Explanation:
Data Wrangling:

PISA data provided is for multiple years, I have filtered by year 2018, and have used Country code library to create a Country column that contains the full names of countries, and custom matches for few countries and average value to match the given plot. Have created another Column 'BoysScore', which has the BOYS scores for each country.
I have also created a dataframe pisa2018Colour to hold the text colours, grey for all countries, black for average value and red for Ireland.

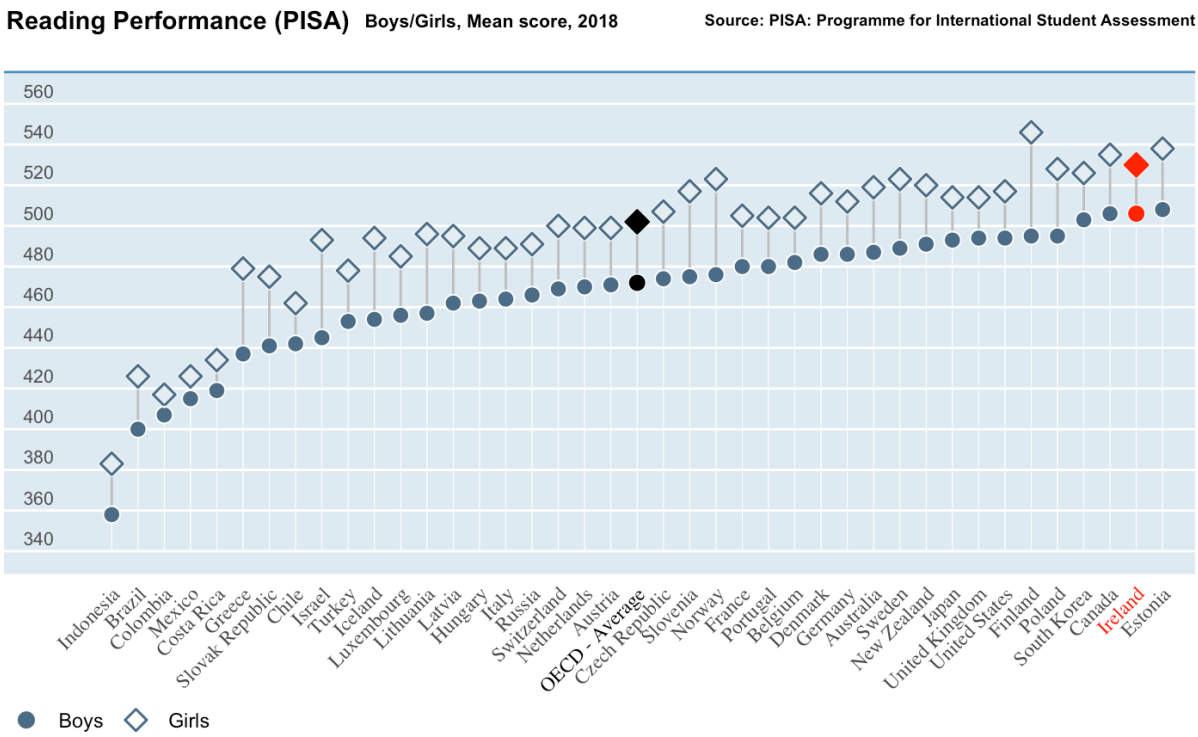
GGPlot Features Used:
Aesthetics: X axis - Reordered by Country, Boys Score. Y axis has value, shape and colour is given by Subject.

X-gridlines – using Geom_segment – lines xend – country and y=340 and yend = BoysScore
Connector line between Girls and Boys Score – using geom line group by country.
Data points – Using geom_point with aesthetic colour = subject. Scale shape manual and scale colour manual has been used to customize the points, shape 16 and 23 have been chosen for Boys and girls respectively and colours as per the matched colours mentioned above.
Also, another two layers of geom_point has been used for GIRLS point to increase the thickness using Stroke and hide the grid lines passing through the points and to create the white background around the point. The white background has also been created for boys score point.
A thick line at the top of the plot is added using annotate feature.
Scale_Y_continuous has been used to keep the line breaks between 340,560.
Labs feature has been used to create the plot title, subtitle and caption. Vjust and hjust have been used to adjust their positions.
Scale_X_discrete has been used to increase the plot width to accommodate the Y labels being on the inside.

Theme – set axis titles to blank using axis.title, grid lines for x to be blank, minor grid lines for y to be blank, set axis lines to blank, set panel.border to be blank, and panel.background to be chosen colour using element rect, axis.text.y using element text, make it move over the y grid lines using the margin feature, adjust the position using vjust and hjust. Axis text x, make the x labels tilted 45 degrees and adjust the colour of the text using the pisa2018Colour dataframe. Finally set the legend horizontal using legend.direction, set legend.key to be blank, adjust position using legend.position, and set legend.title to be blank.

Visualizations not able to achieve:
Couldn't increase the width of the Y-axis, to make the connecting lines clear between the Girls and Boys points, to match the given graph.
Couldn't match the fonts of the text to the given font, unable to figure out the font.

Plot 1:


Reading Performance (PISA) Boys/Girls, Mean score, 2018
Source: PISA: Programme for International Student Assessment
560
540
520
500
480
460
440
420
400
380
360
340
Indonesia
Brazil
Colombia
Mexico
Costa Rica
Greece
Slovak Republic
Chile
Israel
Turkey
Iceland
Luxembourg
Lithuania
Latvia
Hungary
Italy
Russia
Switzerland
Netherlands
Austria
OECD - Average
Czech Republic
Slovenia
Norway
France
Portugal
Belgium
Denmark
Germany
Australia
Sweden
New Zealand
Japan
United Kingdom
United States
Finland
Poland
South Korea
Canada
Ireland
Estonia
Boys
Girls

Code Appendix for Plot1:

```
#packages for fonts Not used
# install.packages('extrafont')
# library(extrafont)
# font_import()

library(dplyr)

##
## 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)
library(countrycode)
library(tidyr)
library(ggthemes)

pisa<-read.csv("pisa.csv",fileEncoding = "UTF-8-BOM")

#filter by year 2018
pisa2018<-pisa%>%filter(TIME=='2018')

#Populate country codes with country names, also included custom match for
certain countries
pisa2018$Country = countrycode(pisa2018$LOCATION,origin = "iso3c",destinat
ion =  'country.name', custom_match = c('OAVG' = 'OECD - Average','CZE'='C
zech Republic','SVK'='Slovak Republic'))

#Create a Boys Score column group by country and get min of value as Boys
score is consistently lower than girls for all countries
pisa2018<-pisa2018%>% group_by(Country)%>%mutate(BoysScore=min(Value))
pisa2018<-pisa2018%>% filter(SUBJECT!='TOT')

#Create a dataframe to hold colour for the x axis text, contains one colum
n text color based on the condition, red for ireland, black for oecd avera
ge, grey for others
pisa2018Colour<- pisa2018%>%arrange(BoysScore)%>%select(Country) %>%
  distinct() %>%mutate(textColour=if_else(Country == "OECD - Average", "bl
ack",if_else(Country == "Ireland","red","#565252")))

#Plot
ggplot(pisa2018 , aes(x = reorder(Country,BoysScore), y= Value,shape = SUB
JECT, colour = SUBJECT)) +
  geom_segment(aes(xend = Country,yend=BoysScore),y=340, colour="white", s
ize=0.15, linetype = "solid")+#segment for x grid lines, adding these befo
re the geom points and line
  geom_line(aes(group = Country), colour = "grey", size=0.5) + # connectin
g line between boys and girls score
```

```
  geom_point( size = 3.0, aes(colour = SUBJECT),stroke=0.80)+ # points
  scale_shape_manual(name = "",labels=c("Boys","Girls"),values = c(16, 23)
) +#custom shapes
  scale_colour_manual(name = "",labels=c("Boys","Girls"),values = c( "#4B6
C86","#4B6C86"))+ #custom colour
  geom_point(data = subset(pisa2018, SUBJECT == 'GIRL'), shape=5,color='wh
ite', size=3.4)+#white background for the girl points
  geom_point(data = subset(pisa2018, SUBJECT == 'GIRL'), shape=23,color='#
4B6C86',fill='#E4EDF2',stroke=0.80, size=3.1)+#superimposing on the origin
al point hide grid lines, stroke to increase width
  geom_point(data = subset(pisa2018, SUBJECT == 'BOY'), shape=1,color='whi
te', size=3.1)+#white background for boy points
  geom_point(data = subset(pisa2018, Country == 'Ireland'),mapping =  aes(
x=Country,y=Value), shape=c(16,23),color='red', fill = "red", size = c(3,3
.6))+#red points for Ireland
  geom_point(data = subset(pisa2018, Country == 'OECD - Average'),mapping
=  aes(x=Country,y=Value), shape=c(16,23),color='black', fill = "black", s
ize = c(3,3.6))+#black points for average
  annotate(geom = 'segment', y = Inf, yend = Inf, color = '#498EBF', x = -
Inf, xend = Inf, size = 1) +#blue line at top of plot
  scale_y_continuous(limits = c(340,565), breaks = seq(340,560,by=20))+#y
scale

  labs(title= "Reading Performance (PISA)",subtitle = "Boys/Girls, Mean sc
ore, 2018",caption = "Source: PISA: Programme for International Student As
sessment") +#titles, subtitle and caption
  scale_x_discrete(expand = c(0,5,0,1.25))+#adjust poisition

  theme(axis.title.y = element_blank(),
        axis.title.x = element_blank(),
        panel.grid.major.x =element_blank(),
        panel.grid.minor.x = element_blank(),
        #panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank(),
        axis.line.y = element_blank(),
        axis.line.x = element_blank(),
        axis.ticks = element_blank(),
        panel.background = element_rect(fill = "#deeaf1"),#background colo
ur
        panel.border = element_blank(),
        axis.text.y = element_text(hjust = 2,vjust = -0.5, size=8,margin=m
argin(-1,-35,0,3)),#adjusted margins to get labels into the plot
        axis.text.x = element_text(family="serif",angle = 45, vjust = 1, h
just =1,color = pisa2018Colour$textColour),#colours from created dataframe
        plot.title = element_text(color="black", face='bold',size=11,vjust
= 0,hjust =0.03),
        plot.subtitle = element_text(color = "black", face='bold',size=8,h
just =0.40, vjust = 8),
        plot.caption=element_text(size=7,hjust = 1, face='bold',vjust = 20
7),
        legend.direction = "horizontal",
        legend.key=element_blank(),
        legend.position= c(0.1,-0.29),#adjust legend position
        legend.title = element_blank())
```

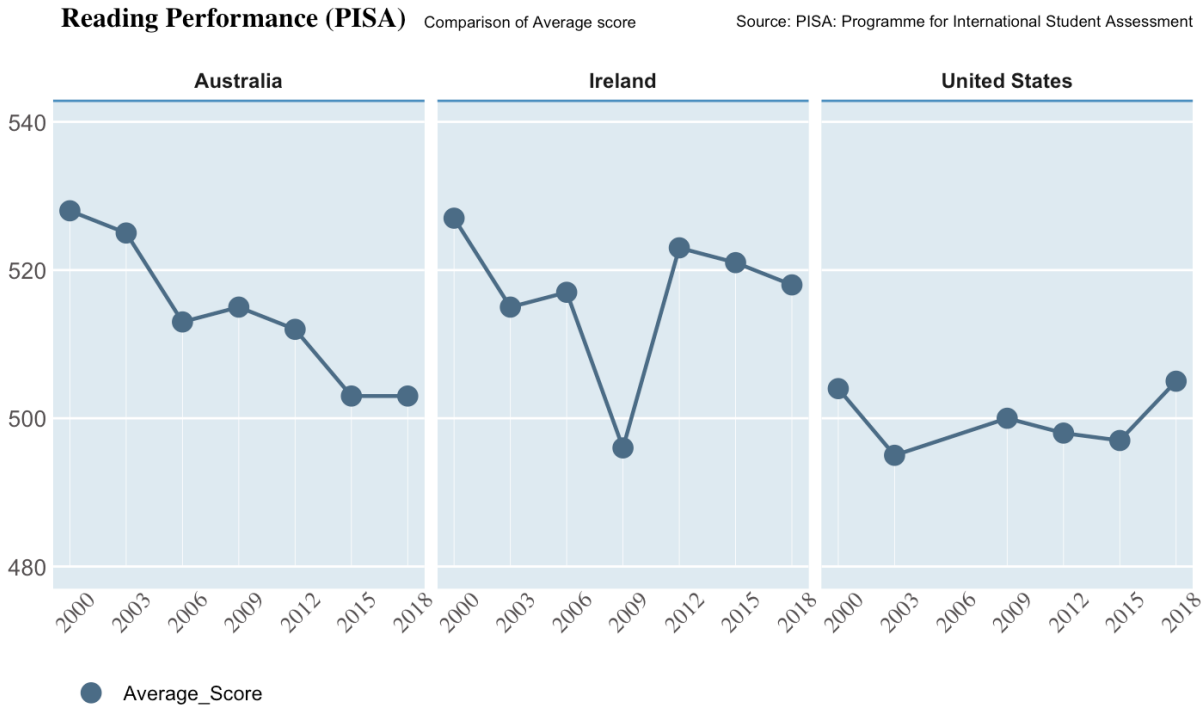
Part 3:

Data Wrangling:

For this plot I have selected three English speaking countries from three continents to make a comparison between them. Hence, I have filtered by LOCATION with values AUS, IRL, USA. Then I have used country codes library to populate country names. I have filtered by subject=TOT, as this is the average of boys and girls score for each country. I have changed TOT to Average Score.

Plot 2:

The aesthetics for the plot has been changed in the X- axis – Time and Y-axis is average values for that year, shape by Subject.
For the plot I have used the same features and theme as before, in addition I have used **facet grid** to display three countries side by side on a similar scale in Y axis, so as to provide the audience with simpler representation to make out the comparison between the three countries easily. I have used a common colour for all the points and the line as the countries are clearly differentiated by the facet grid and the colours will not add any value, the geom line indicates the trend and that is also in the same colour. For USA the data for 2006 is missing. Also, it is not possible to bring the Y labels in the graph, as facet grid alignment collapses.



Code Appendix for Plot 2:

```
#filter by location to get dataframe
plot2<-pisa%>%filter(LOCATION=='AUS'|LOCATION=='IRL'|LOCATION=='USA')

#populate country names from Location code
plot2$Country = countrycode(plot2$LOCATION,origin = "iso3c",destination =
'country.name')

#Filter by TOT, and modify it as average score
plot2<-plot2%>% filter(SUBJECT=='TOT')
plot2$SUBJECT[plot2$SUBJECT == "TOT"] <- "Average_Score"

#Plot 2
gplot_2<-ggplot(plot2, aes(x = TIME, y=Value, shape = SUBJECT, color =
Country))+ #x axis time, y axis value, shape by subject- same for all,
colour by country
  geom_line(size = 0.8) +#line to indicate trend
  geom_segment(aes(xend = TIME), yend = 480, colour="white", size=0.15,
linetype = "solid") +#x axis grid lines

  geom_point(color = "#4B6C86", fill = "#4B6C86", size = 4) +# point with
same color as above

  labs(title= "Reading Performance (PISA)",subtitle = "Comparison of
Average score", caption = "Source: PISA: Programme for International
Student Assessment") +

  scale_shape_manual(values = c(16)) + #specifying custom shape
  scale_fill_manual(values=c("#4B6C86")) +#specifying custom color
  annotate(geom = 'segment', y = Inf, yend = Inf, color = '#498EBF', x = -
Inf, xend = Inf, size = 1) +#blue line at top to maintain theme

  scale_y_continuous(limits = c(480,540),breaks = seq(480,540,by=20))
+#modified breaks
  scale_x_continuous(limits = c(2000,2018), breaks = seq(2000,2018,by=3))
+#specified breaks per year

  scale_color_manual(values = c("#4B6C86", "#4B6C86", "#4B6C86")) +#same
color for all countries as depicting average data for comparison

  theme(axis.title.y = element_blank(),
        axis.title.x = element_blank(),
        panel.grid.major.x =element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.minor.y = element_blank(),
        axis.line.y = element_blank(),
        axis.line.x = element_blank(),
        axis.ticks = element_blank(),
        panel.background = element_rect(fill = "#deeaf1"),
        plot.margin = margin(14, 7, 7, 1.5),
        plot.title = element_text(face = "bold",color="black",
```

```
size=12,vjust = -1,hjust = 0.01, family = "serif"),
        plot.subtitle = element_text(color = "black", size=7,hjust =0.40,
vjust = 6),
        plot.caption=element_text(size=7,hjust = 1,color = "black", vjust
= 197),
        axis.ticks.y = element_blank(),
        axis.text.y= element_text(margin = margin(0, 0, 0, 0) ,size = 10,
color = "#565252"),
        axis.text.x= element_text(family="serif",angle=45, size=10, color
= "#565252", margin = margin(5, 0, 5, 5) ),
        strip.text.x = element_text(size=9, face="bold"),
        strip.background = element_blank(),
        legend.key=element_blank(),
        legend.position= c(0.1,-0.20),
        legend.title = element_blank()#legend at bottom as before

  ) +
  facet_grid(cols = vars(Country))#using facet grid to display countries
side by side

gplot_2<-gplot_2+guides(color=FALSE)#hides geomline legend

gplot_2
```