Data Visualisation Assignment1 Group D

Swaroop(17230755)/Vignesh(17231743)/Vinoop(17231748)

16 February 2018

Overview:

According to the book 'The Rational Optimist: How Prosperity Evolve' by Mat Ridley -

- Fertility rates in richer countries are low.
- 'Developing' countries are showing a decrease in average fertility.

We illustrate Ridley's hypothesis through appropriate visualisations. This involves comparing the world population growth rate over the years and visualising the relation between how the fertility, GDP/Income and Education of the high, middle, and low income countries varies.

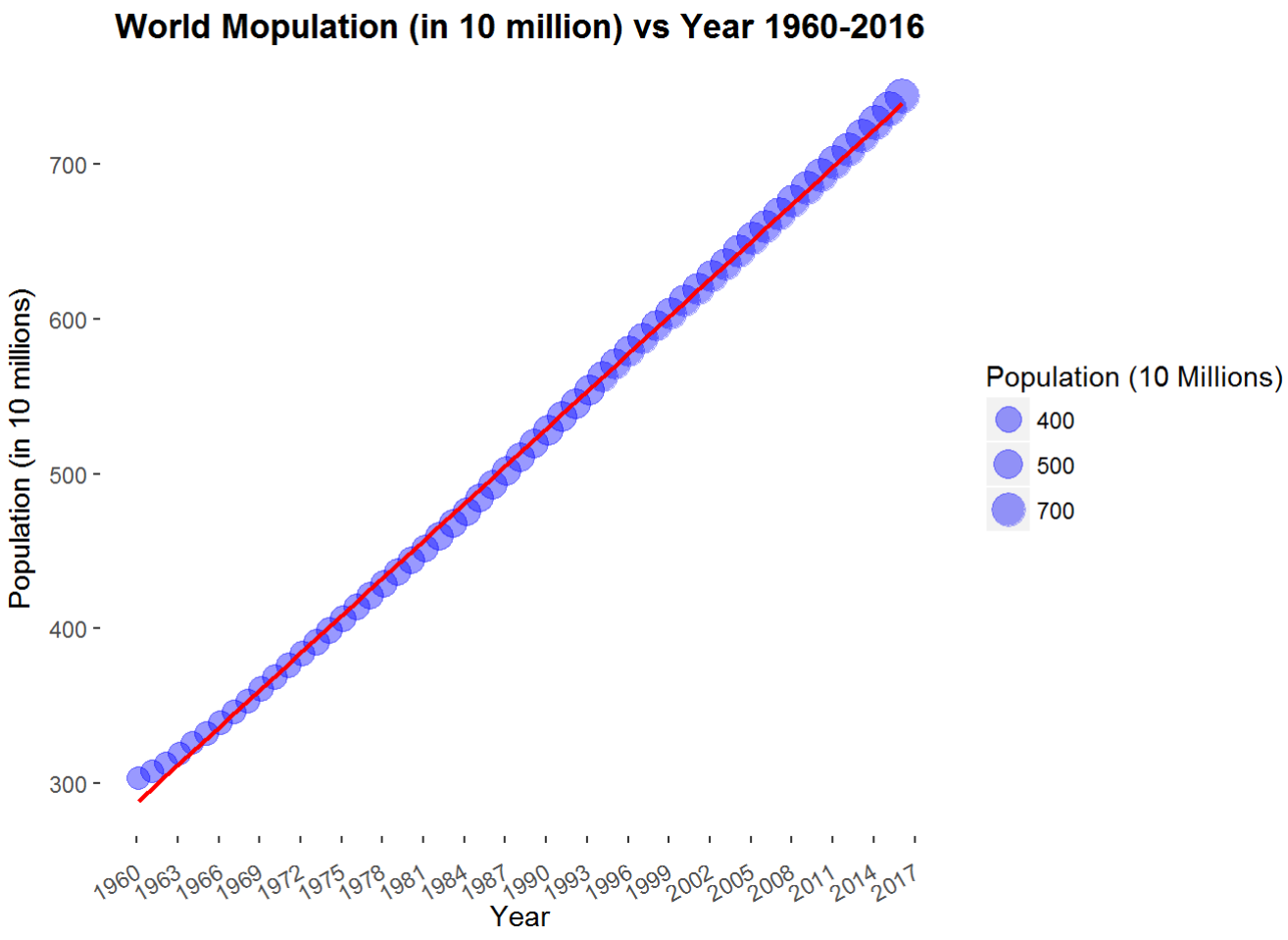
Loading the required packages:

```
library(tidyverse )
library(plyr)
library(scales)
library(dplyr)
library(lubridate)
library(zoo)
library(ggplot2)
if (!require("RColorBrewer")) {
install.packages("RColorBrewer")
library(RColorBrewer)
}
```

Percentage Increase in World Population:

When we were visualising the World Population, we found that the world population tend to increase linearly with time. We can confirm the same by looking into the below graph where the red line represents the World Population and the black line represents the regression line.

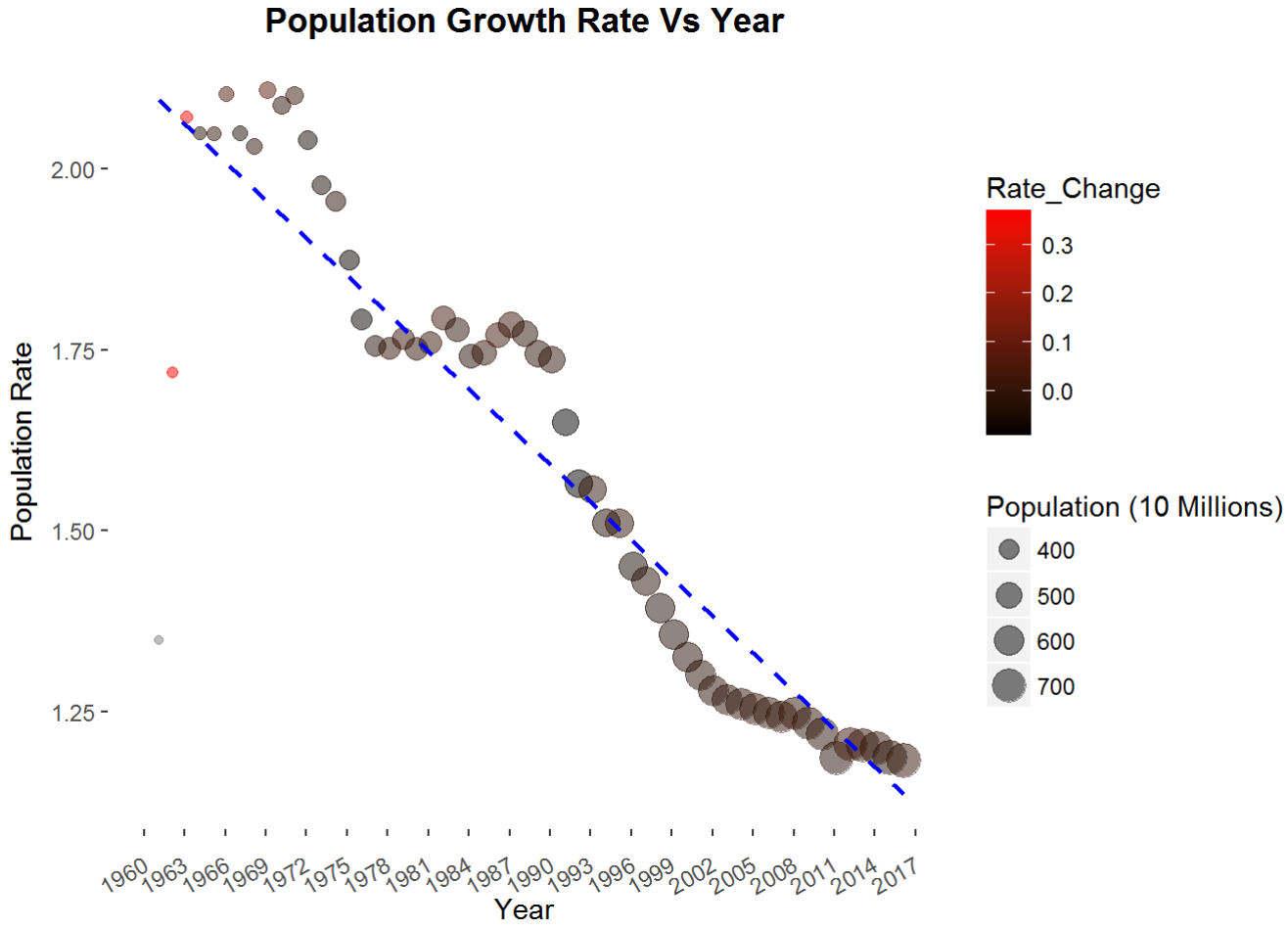
```
#Reading the WorldPopulation.csv file from the directory
world_population <- read_csv("WorldPopulation.csv")
#Preprocessinng the Data using the  gather function
world_population <- gather(world_population, key=Year, value = Population, `1960`:`2016`)
#Considering the Year and Population column from the above preproceesed data
world_population <- world_population[,c("Year","Population")]
#Changing Year Colun to the date format
world_population$Year <- as.Date(as.character(world_population$Year), format="%Y")
#Dividing the world popluation by 10000000 so as to get the numbers in 10 million range
world_population$Population <- world_population$Population / 10000000
#Renaming the Population column to Population (10 Millions)
colnames(world_population)[2] <- "Population (10 Millions)"

p<- ggplot(data=world_population, aes(x=Year, y=`Population (10 Millions)`, size=`Population
 (10 Millions)`))
p + geom_point(alpha=0.4, color="blue") +
  scale_size_area(max_size = 6, breaks=c(400, 500, 700))+
  ylab("Population (in 10 millions)")+
  scale_x_date(date_breaks = "3 year", date_labels = "%Y")+
  ggtitle("World Mopulation (in 10 million) vs Year 1960-2016") +
  stat_smooth(method=lm, se= FALSE, colour = "red", size = 0.8)+
  theme(axis.text.x = element_text(angle = 30, hjust=1, vjust = .5),
        plot.title = element_text(hjust = 0.5, face="bold"),
        panel.background = element_blank())
```



We need to modify the data in order to check for the growth rate of world population over a period of time. With the help of the dplyr package, we are add new columns - “Previous_YearPopulation”, “Change” and “Percent_Change” columns. We then plot a graph of Year Vs Population Growth Rate change.

We see that, the world witnessed a very high population growth of around 2 percent in the 1960s. The 1970s witnessed a change in population growth and started a downward trend that has continued to this day (as shown by the regression line). The world population is now growing at 1.25 percent.

```
# Adding the new columns PreviousYear_Population, Change, Percent_change and Rate_change usin
g Mutate function
world_population <- world_population %>%
  mutate(Previous_YearPopulation = lag(`Population (10 Millions)`, 1),
         Change = `Population (10 Millions)` - Previous_YearPopulation ,
         Percent_Change = Change/Previous_YearPopulation*100,
         Rate_Change = Percent_Change - lag(Percent_Change,1))
#Renaming the Change column to Percentage change
colnames(world_population)[4] <- "Population_Change"

p<- ggplot(world_population, aes(x=Year, y=Percent_Change, size=`Population (10 Millions)`))

p + geom_point(alpha=0.5, aes(color=Rate_Change), na.rm=T) +
  ylab("Population Rate") +
  scale_colour_gradient(low = "black", high = "red")+
  scale_x_date(date_breaks = "3 year", date_labels = "%Y") +
  ggtitle("Population Growth Rate Vs Year") +
  stat_smooth(method=lm, se= FALSE, colour = "blue", size = 0.8, linetype="dashed", alpha=0.7
) +
  theme(axis.text.x = element_text(angle = 30, hjust=1, vjust = .5),
        plot.title = element_text(hjust = 0.5, face="bold"),
        panel.background = element_blank())
```



Change in Fertility Rates:

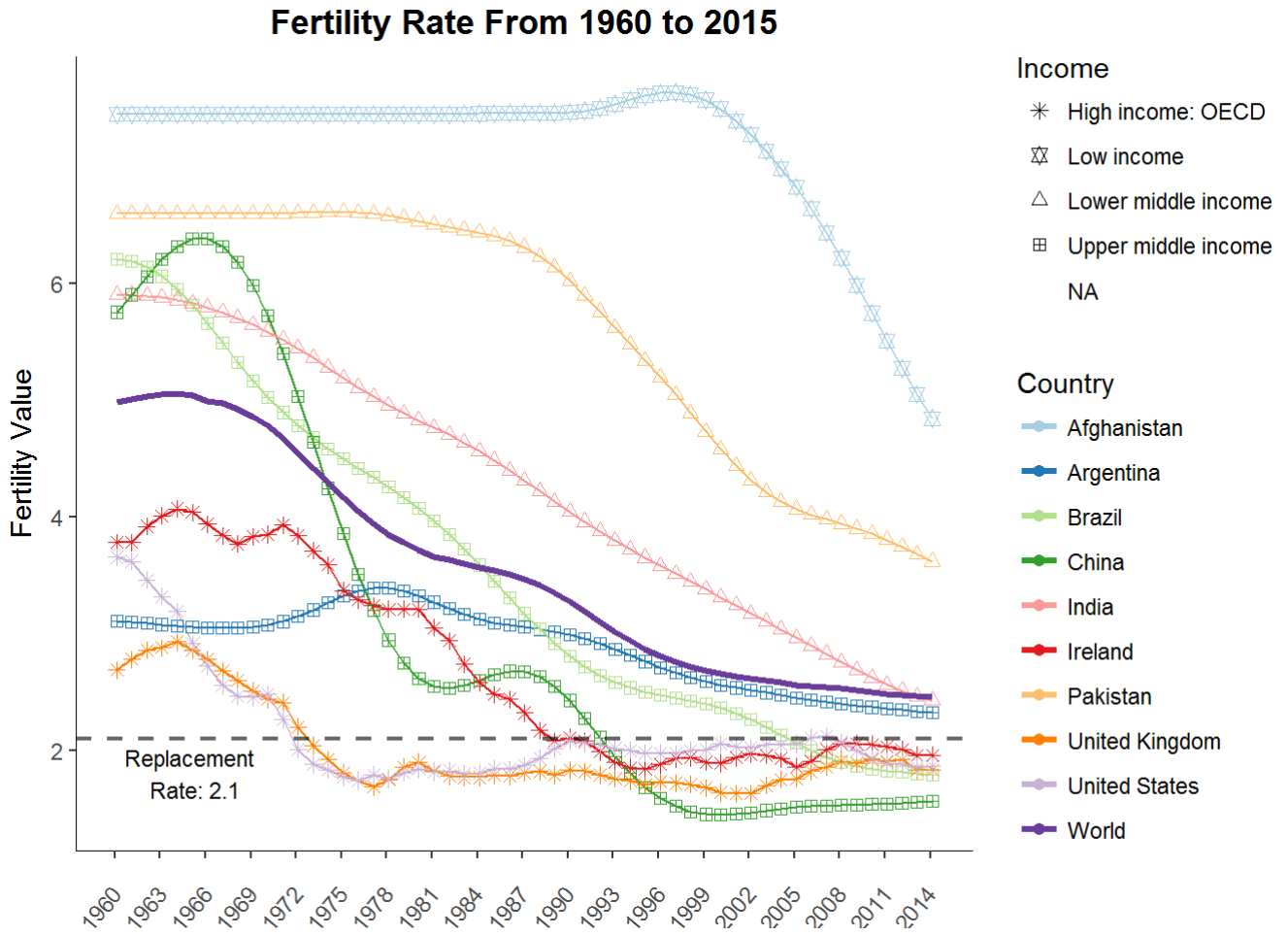
According to Ridley's hypothesis, the average fertility rate of the world is falling. We select a few countries belonging to different income groups and different regions and visualise their fertility rate for different years. The plot confirms Ridley's argument that higher income countries such as the United Kingdom and Ireland have a lower fertility rate. Lower income countries such as Pakistan and Afghanistan have a very high fertility rate.

One general trend that we have observed is that world's fertility rate is decreasing and heading towards Replacement Rate (2.1). This in actual is not a bad thing since there are limited resources available in the world.

Pre Processing Data:

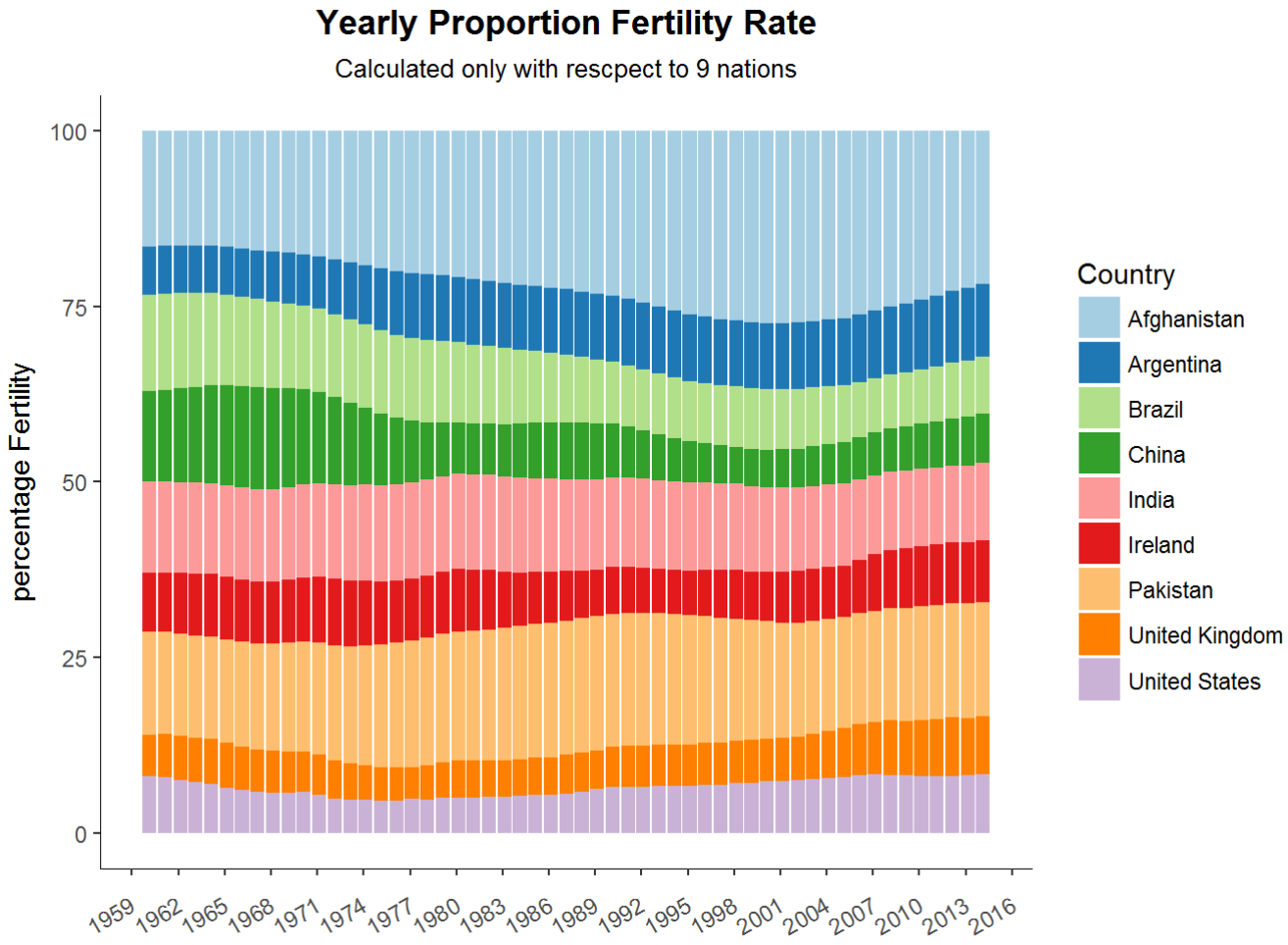
```
# Loading the nations data and fertility data from home directory.
fert_data <- read_csv("Fertility.csv")
nat_data <- read_csv("nations.csv", col_types = cols(population=col_double()))
#Appending the World data to fert_data
fert_world <- fert_data[fert_data$`Country or Area` == 'World', ]
#Converting Date - string column to Date type
fert_world$Year <- as.Date(as.character(fert_world$Year), format="%Y")

# Selectiing few coutries which belongs to Rich, middle and Poor categories.
# Selected countries are Brazil, United States, Cambodia, Argentina, India, Pakistan, German
y, Ireland, China, United Kingdom and United States
selc_data = fert_data[fert_data$`Country or Area` %in% c('Brazil', 'United States', 'Afghanis
tan', 'Argentina', 'India', 'Pakistan', 'Ireland', 'China', 'United Kingdom', 'World'), ]
selc_data$Year <- as.Date(as.character(selc_data$Year), format="%Y")
selc_data['Income'] <- nat_data[match(selc_data$`Country or Area`, nat_data$country), 11]
```

```
p<- ggplot(selc_data, aes(x = Year, y=Value, color=`Country or Area`), show.legend=F)
p + geom_line() +
  geom_point(size = 1.8, aes(shape=Income), na.rm = T) +
  scale_shape_manual(values = c(8, 11, 2, 12))+
  geom_line(data=fert_world, size=1.2)+
  scale_colour_brewer(palette = "Paired", name = "Country",
                      labels=c('Afghanistan', 'Argentina', 'Brazil',
                               'China', 'India', 'Ireland', 'Pakistan',
                               'United Kingdom', 'United States', 'World')) +
  ylab("Fertility Value") +
  ggtitle("Fertility Rate From 1960 to 2015") +
  scale_x_date(date_breaks = "3 year", date_labels = "%Y")  +
  #labs(caption = "SWAROOP: 17230755")+
  theme(panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black", size = 0.25),
        axis.title.x=element_blank(),
        axis.text.x = element_text(angle = 50, hjust=1, vjust = .5),
        legend.key = element_rect(fill = NA, colour = NA, size = 0.25),
        plot.title = element_text(hjust = 0.5, face="bold"))+
  geom_hline(yintercept = 2.1, size=0.8, linetype = 2, alpha=0.6)+
  annotate("text", x = as.Date("1965", format="%Y"), y = 1.8,
           label = c("Replacement \n Rate: 2.1"), fontface=1, size=3)
```



The below graph shows us that low income countries have higher fertility rates. Even though the fertility rates around the world has decreased, the low income countries still have a much higher fertility rate than high income countries. This may lead to strain on their resources leading to lower GDP per capita.

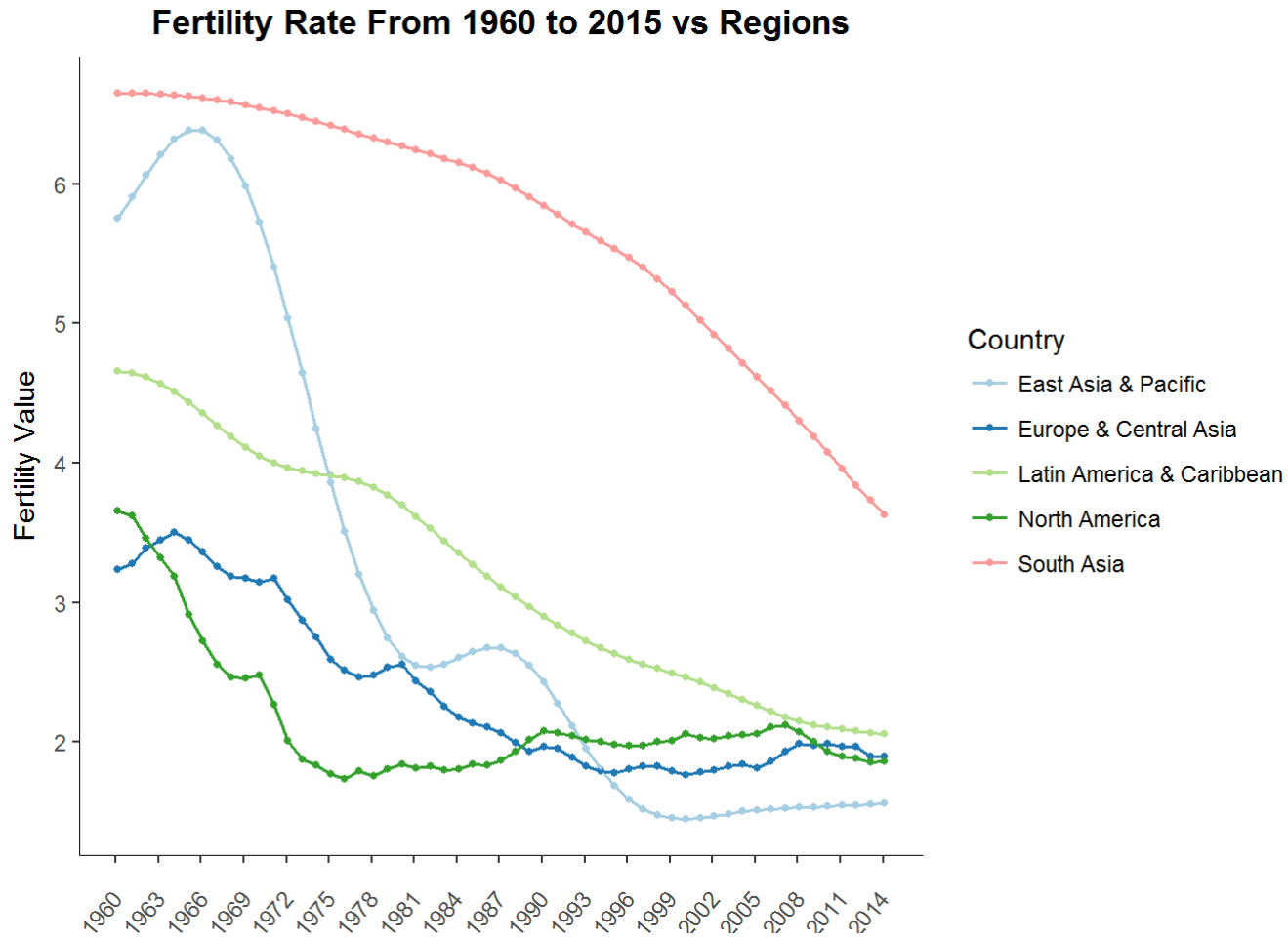
```
#Converting all Fertility Value column to percent to plot the proportion graph.
per_fertility =
  ddply(selc_data[selc_data$`Country or Area` != 'World', ], "Year", transform, percent_ferti
lity = Value/sum(Value)* 100)



ggplot(per_fertility, aes(x = Year, y=percent_fertility, fill=Country.or.Area))+
  geom_bar( stat="identity")+
  scale_fill_brewer(palette = "Paired", name = "Country",
                    labels=c('Afghanistan', 'Argentina', 'Brazil', 'China', 'India',
                             'Ireland', 'Pakistan', 'United Kingdom', 'United States')) +
  ylab("percentage Fertility") +
  labs(title='Yearly Proportion Fertility Rate',
       subtitle='Calculated only with rescpect to 9 nations') +
  scale_x_date(date_breaks = "3 year", labels = date_format("%Y"))  +
  theme(panel.grid.major = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black", size = 0.25),
        axis.title.x=element_blank(), axis.text.x = element_text(angle = 30, hjust=1, vjust =
 .5),
        legend.key = element_rect(fill = NA, colour = NA, size = 0.25))+
  theme(plot.title = element_text(hjust = 0.5, face="bold"),
        plot.subtitle = element_text(hjust = 0.5))
```



Regional view of the fertility rates around the world:

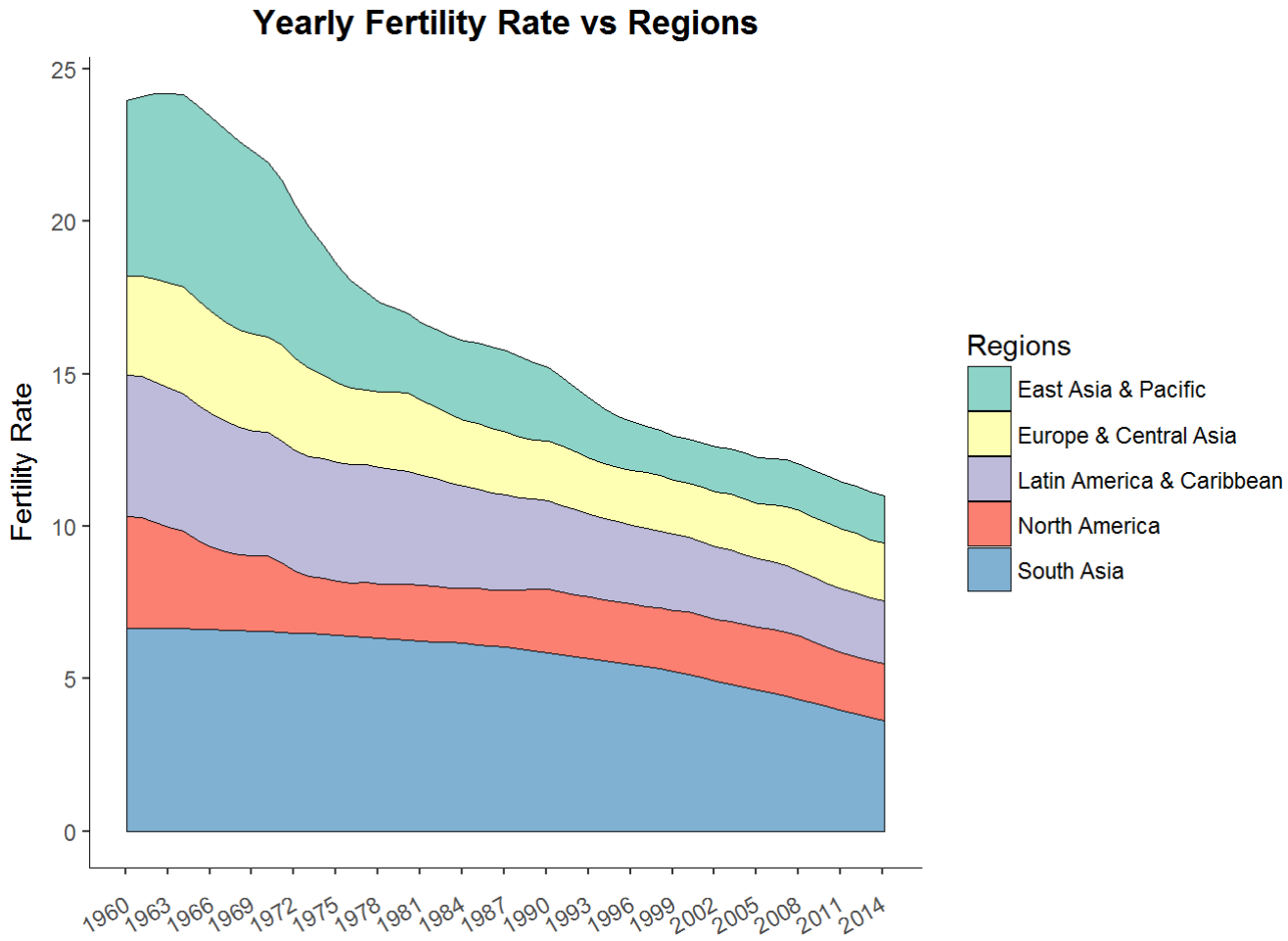
The countries mentioned above are grouped according to their regions. We can clearly see that Europe, North America and East Asia have a much lower fertility rate when compared to South Asia or Latin America. This can be attributed to the fact that a majority of the high-income countries are located in either Europe, North America or East Asia.

```
Reg_data <- selc_data[selc_data$`Country or Area` != 'World', ]
colnames(Reg_data)[1] <- 'country'
Reg_data['Region'] <- nat_data[match(Reg_data$country, nat_data$country), 10]
Reg_data <- Reg_data %>% group_by(Year, Region) %>% dplyr::summarise(Value=mean(Value, na.rm=
TRUE))
Reg_data$Year <- as.Date(as.character(Reg_data$Year), format="%Y")

d<- ggplot(Reg_data, aes(x = Year, y=Value, color= Region))

d + geom_point(size = 1)+
  geom_line(size=0.6)+
  scale_colour_brewer(palette = "Paired", name = "Country")+
  ylab("Fertility Value") +
  ggtitle("Fertility Rate From 1960 to 2015 vs Regions") +
  scale_x_date(date_breaks = "3 year", date_labels = "%Y")+
  theme(panel.grid.major = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black", size = 0.25),
        axis.title.x=element_blank(),
        axis.text.x = element_text(angle = 50, hjust=1, vjust = .5),
        legend.key = element_rect(fill = NA, colour = NA, size = 0.25),
        plot.title = element_text(hjust = 0.5, face="bold"))
```



The graph below clearly shows us that East Asia, Europe & Central Asia and Latin American region have witnessed a sharp decrease in fertility rate over the years. North America and South Asia have shown a much smaller decrease in fertility rate. Majority of the population still lives in South Asia.

```
ggplot(Reg_data, aes(x=Year, y=Value, fill=Region)) +
  geom_area(stat="identity", colour="black", size = .2)  +
  scale_fill_brewer(palette = "Set3", name = "Regions",
                    labels=c("East Asia & Pacific", "Europe & Central Asia",
                             "Latin America & Caribbean", "North America",
                             "South Asia")) +
  ylab("Fertility Rate") +
  labs(title='Yearly Fertility Rate vs Regions') +
  scale_x_date(date_breaks = "3 year", labels = date_format("%Y"))  +
  theme(panel.grid.major = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black", size = 0.25),
        axis.title.x=element_blank(), axis.text.x = element_text(angle = 30, hjust=1, vjust =
 .5),
        legend.key = element_rect(fill = NA, colour = NA, size = 0.25))+
  theme(plot.title = element_text(hjust = 0.5, face="bold"),
        plot.subtitle = element_text(hjust = 0.5))
```

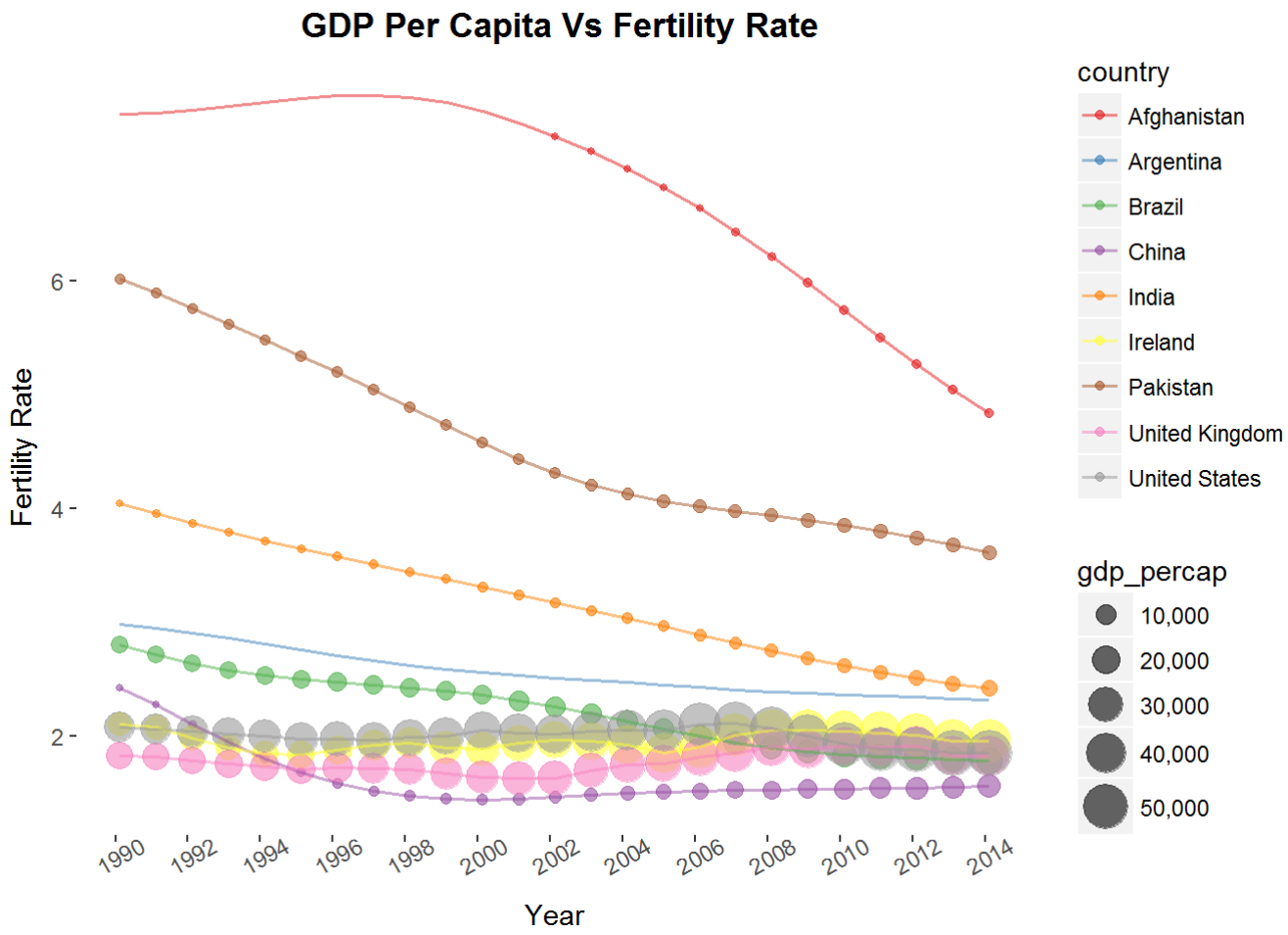


GDP Per Capita vs Fertility Rate:

We plot GDP per capita against fertility rate of certain countries of the world over the years.

We can clearly see that GDP per capita and Fertility rates are closely related. Countries having a low fertility rate usually have a higher per capita GDP as indicated by the size of the bubble. Countries having a higher fertility rate have a significantly lower GDP due to the fact that resources are to be shared among a much larger population. China seems to have realised the issue and have followed a one-child-policy. This seems to have borne fruit as we can see that their per capita GDP has increased significantly over the past 15 years.

```
#Reading the Fertility.csv from the directory
fertility_info <- read.csv("Fertility.csv")
#Reading the nations.csv from the directory
nations_info <- read.csv("nations.csv")
#Reading the Enrollment Percentage.csv from the directory
education_info <- read_csv("Enrollment Percentage.csv")

#Renaming the Country column to country so as to match with other dataset
colnames(fertility_info)[1] <- "country"
#Renaming the Year column to year so as to match with other dataset
colnames(fertility_info)[2] <- "year"

#Converting the year column to factors
fertility_info$year=as.factor(fertility_info$year)
#Converting the year cloumn to factors
nations_info$year=as.factor(nations_info$year)
#Joining the 2 dataframes fertility info and nations info
fertility_Nations_Info <- left_join(x = fertility_info, y = nations_info, by = c("country","y
ear"), all = TRUE)

#Filtering the data as we are considering only specific countries
fertility_Nations_Info <- fertility_Nations_Info %>%
  select(country, population, gdp_percap, region, income, Value, year) %>%
filter((country=="Afghanistan" | country=="Brazil" | country=="India" |
       country=="China" | country=="Argentina" | country=="Ireland" |
       country=="Pakistan" | country=="United States" | country=="United Kingdom")
       & year > 1989)

#Changing the year column to Date format
fertility_Nations_Info$year <- as.Date(as.character(fertility_Nations_Info$year), format = "%
Y")

n <-ggplot(fertility_Nations_Info, aes(x=year, y=Value, size=gdp_percap, color=country))
n + geom_point(alpha = 0.6, na.rm = TRUE) +
  geom_line(size =0.7, alpha=0.5) +
  scale_x_date(date_breaks = "2 years", labels = date_format("%Y")) +
  ggtitle(" GDP Per Capita Vs Fertility Rate") +
  xlab("Year") +
  ylab("Fertility Rate") +
  scale_color_brewer(palette='Set1') +
  scale_size_area(max_size=8, labels  = comma) +
  theme(plot.title = element_text(hjust = 0.5, face="bold"),
  axis.text.x = element_text(angle=30),
  panel.grid.major = element_blank(),
  panel.background = element_blank())
```

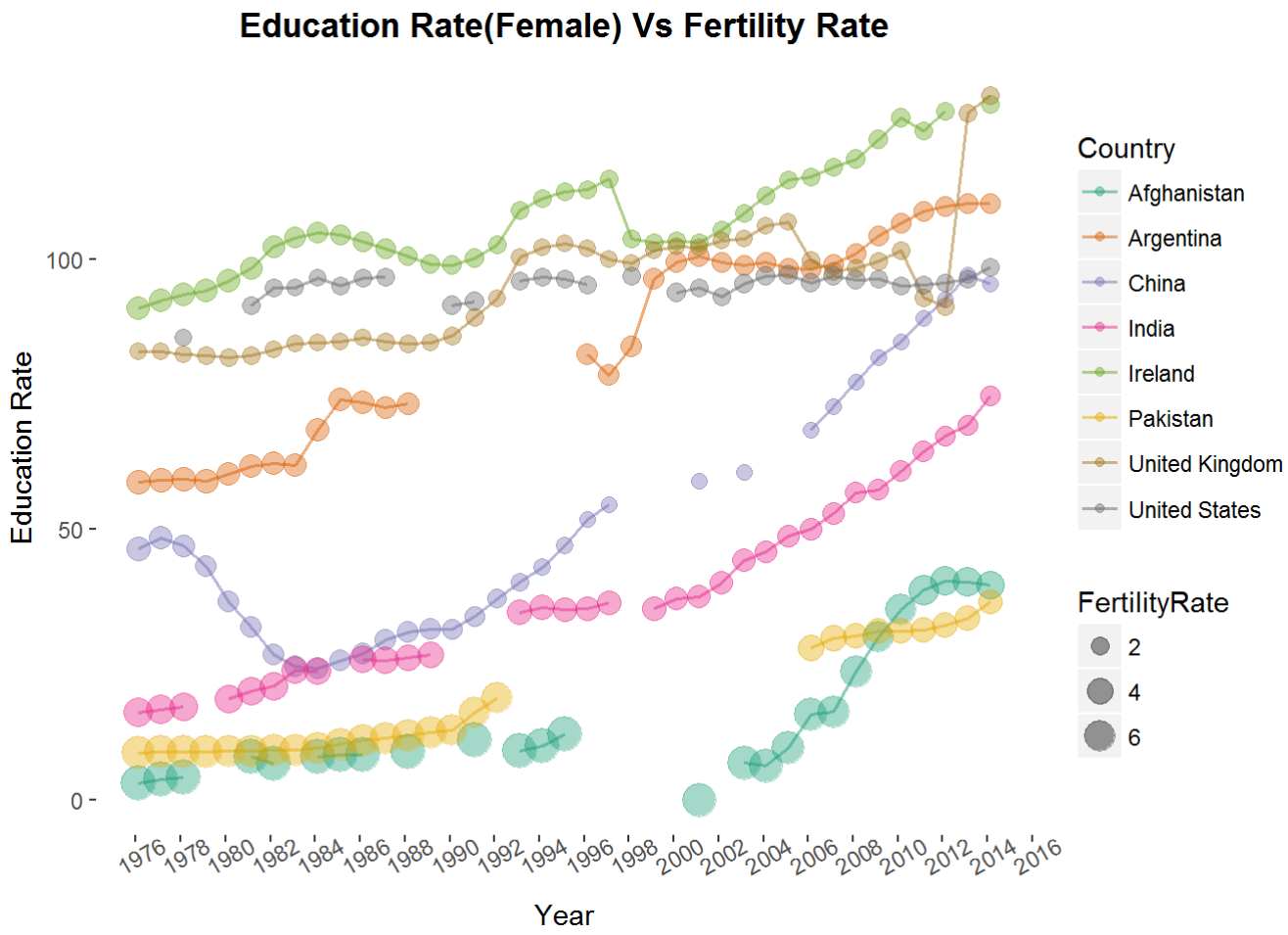


Endication Rate vs Fertility Rate:

We plot Education Enrollment rate against fertility rate of certain countries of the world over the years.

We can clearly see that Enrollment rate of females and Fertility rates are closely related. Countries havin ag low fertility rate usually have a higher enrollment education rate. We can visalise the same in the below graph. Poor countries like Afghanisthan have very low enrollment rate. But the rate is incrreasing rapidly in the developing countries like India. And we can see very high enrollmet rates in the develpoed countries like United Kingdom and United states.

```
#Reading the Enrollment Percentage csv from the directory
education_info <- read_csv("Enrollment Percentage.csv")
#Reading the fertility csv from the directory
fertility_info <- read.csv("Fertility.csv")
#Preprocessing the data using the gather function
education_info <- gather(education_info, key=Year, value = GrossEducationRate,`1970`:`2015`)
#Considering the Country Name grossEducationRate and Year column from education_info df
education_info <- education_info[,c("Country Name","GrossEducationRate", "Year")]

#Renaming the columsn
colnames(fertility_info)[1] <- "country"
colnames(fertility_info)[2] <- "year"
colnames(education_info)[1] <- "country"
colnames(education_info)[3] <- "year"
#Converting the column to the char format
fertility_info$year <- as.character(fertility_info$year)
#Joining the fertility_info and education_info datframes
fertility_education_info <- left_join(x = fertility_info, y = education_info, by =c("country"
,"year"), all = TRUE)
#Converting the year column to the date format
fertility_education_info$year <- as.Date(as.character(fertility_education_info$year), format
 = "%Y")

#Filtering the data as we are considering only specific countries
fertility_education_info <- fertility_education_info %>%
  select(country, GrossEducationRate, Value, year) %>%
  filter((country=="Afghanistan" | country=="India" |
            country=="China" | country=="Argentina" | country=="Ireland" |
            country=="Pakistan" | country=="United States" | country=="United Kingdom")
         & year > 1970)

#Renaming the columns
colnames(fertility_education_info)[1] <- "Country"
colnames(fertility_education_info)[3] <- "FertilityRate"

q <-ggplot(fertility_education_info, aes(x=year, y=GrossEducationRate,
                                         size=FertilityRate, color=Country))
q + geom_line(size =0.7, alpha=0.5)+
  geom_point(alpha = 0.4, na.rm = TRUE)+
  scale_x_date(date_breaks = "2 years", labels = date_format("%Y")) +
  ggtitle("Education Rate(Female) Vs Fertility Rate") +
  xlab("Year") +
  ylab("Education Rate") +
  scale_color_brewer(palette='Dark2') +
  scale_size_area(max_size = 6) +
  theme(plot.title = element_text(hjust = 0.5, face="bold"),
        axis.text.x = element_text(angle=30),
        panel.grid.major = element_blank(),
        panel.background = element_blank())
```



Conclusion:

From the above plots, there is a strong evidence towards Ridley's hypothesis i.e. Countries with high income tends to have low fertility rate like United States, Ireland and China. Moreover, there is also strong evidence that eduction rate has an impact on Income of the countries which inreturn have a relation on fertility rate. In, addition deleveloping countries like India shows a gradual decreasing trend in fertility rate, whose education rate and GDP is increasing over the year. Furthermore, china shows sudden decrease in fertility rate in mid 80's, may due to 1 child policy followed by china.

Contributions:

- Vignesh: Question 1, Question 4 and Report
- Swaroop: Question 2, Question 3 and Report
- Vinoop: Question 3, Question 4 and Report