Programowanie w R i wizualizacja danych- projekt zaliczeniowy- sprawozdanie.

Damian Gortych 402663 IiAD

1. Przygotowanie danych

Dane na temat zaludnienia do pracy nad projektem pobrałem ze strony https://data.world/ .

Ich początkową formę zamieszczam w pliku Gortych_dane_surowe.

Projekt rozpocząłem od wczytania danych.

```
### Wczytanie danych
```{r}
data1<-read_excel("Country-Metadata.xls")
data2<-read_excel("Country-Population.xls")
data3<-read_excel("Fertility-Rate.xls")
data4<-read_excel("Life-Expectancy-At-Birth.xls")
```
```

Kolejnym krokiem było usunięcie niepełnych oraz niepotrzebnych danych z plików.

```
### Usuniecie niepelnych i niepotrzebnych danych
```{r}

data2<-na.omit(data2)
data3<-na.omit(data3)
data4<-na.omit(data4)
data1<-subset(data1, select = -SpecialNotes)
data1<-na.omit(data1)

```
```

Następnie wykonałem statystyki opisowe dla każdego z plików.

```
### Statystyki opisowe
```{r}
class(data1)
str(data1)
typeof(data1)
summary(data1)

class(data2)
str(data2)
typeof(data2)
summary(data2)

class(data3)
str(data3)
typeof(data3)
summary(data3)

class(data4)
str(data4)
typeof(data4)
summary(data4)
```
```

Ostatnim krokiem było zapisanie przekształconych danych.

```
### Zapis danych
```{r}

write.xlsx(data1,file="data1.xls")
write.xlsx(data2,file="data2.xls")
write.xlsx(data3,file="data3.xls")
write.xlsx(data4,file="data4.xls")

```
```

Formę tych danych zamieszczam w pliku Gortych_dane_przekształcone

2. Praca z pakietem tidyverse oraz wizualizacja.

Dane nie wymagały wstępnej pracy, natomiast konieczne było ich przekształcenie w celu wykonania konkretnych wizualizacji.

1. Wykres Stopnia dochodu w zależności od regionu.

Pierwszą wizualizacją jest wykres stopnia dochodu w zależności od regionu. W celu jego wykonania posłużyłem się danymi z pliku data1.xls. Zawiera on miedzy innymi informacje o grupie dochodowej oraz regionie położenia dla każdego kraju.

Posługują się faktorem oraz pakietem tidyverse przygotowałem dane potrzebne do wykonania wizualizacji.

```
###  1) Wykres Stopnia dochodu w zaleznosci od regionu
```{r}

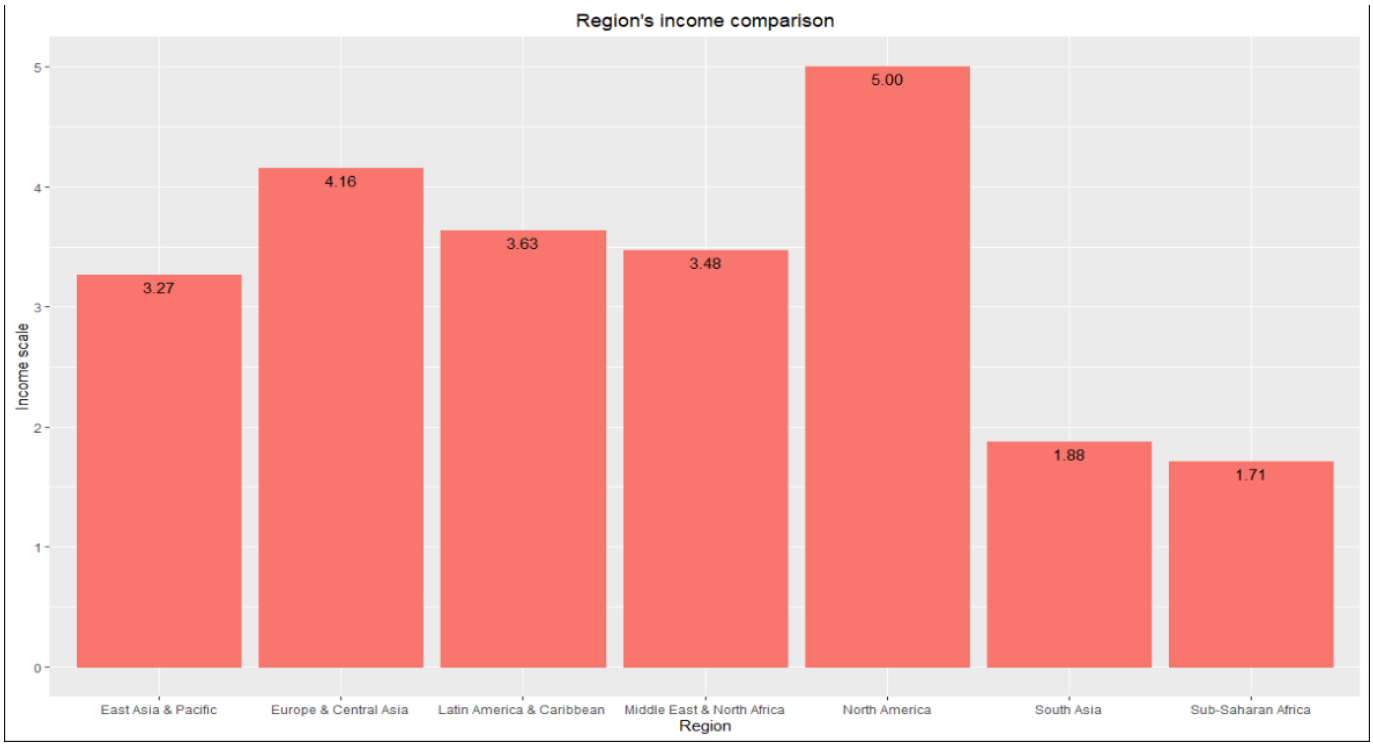
income_faktor<-factor(data1$IncomeGroup)
levels(income_faktor)<-c(5,5,1,2,3)

income_vector<-as.numeric(as.character(income_faktor))
data1 <- data1 %>% add_column(income_vector)

newdata<-data1 %>% select(income_vector,Region) %>% group_by(Region)  %>% summarise(income=mean(income_vector))
newdata <- arrange(newdata,income)

ggplot(data=newdata,aes(Region,income)) + geom_bar(aes(fill='#A4A4A4', color="darkred"),stat="identity",show.legend = FALSE) + ggtitle("Region's income
comparison") + theme(plot.title = element_text(hjust = 0.5)) + ylab("Income scale") + geom_text(aes(label=sprintf("%0.2f", round(income, digits =
2)),vjust=1.5))


```
```



Jak widać na wykresie zdecydowanie dominuje region Północnej Ameryki. Regionami najbiedniejszymi pod względem ekonomicznym są natomiast Południowa Azja oraz Afryka Subsaharyjska.

2. Wykres zmiany populacji na przestrzeni lat dla 3 krajów z jej największa liczbą w roku 1960 .

W celu wykonania drugiej wizualizacji posłużyłem się danymi na temat liczby populacji dla danego kraju na przestrzeni lat począwszy od roku 1960. Zdecydowałem się na przestawienie zmian dla 3 krajów dla których liczba populacji była największa w tymże roku.

```
###  2) Wykres zmiany populacjji na przestrzeni lat dla 3 krajow z jej najwieksza liczba w roku 1960
```{r}

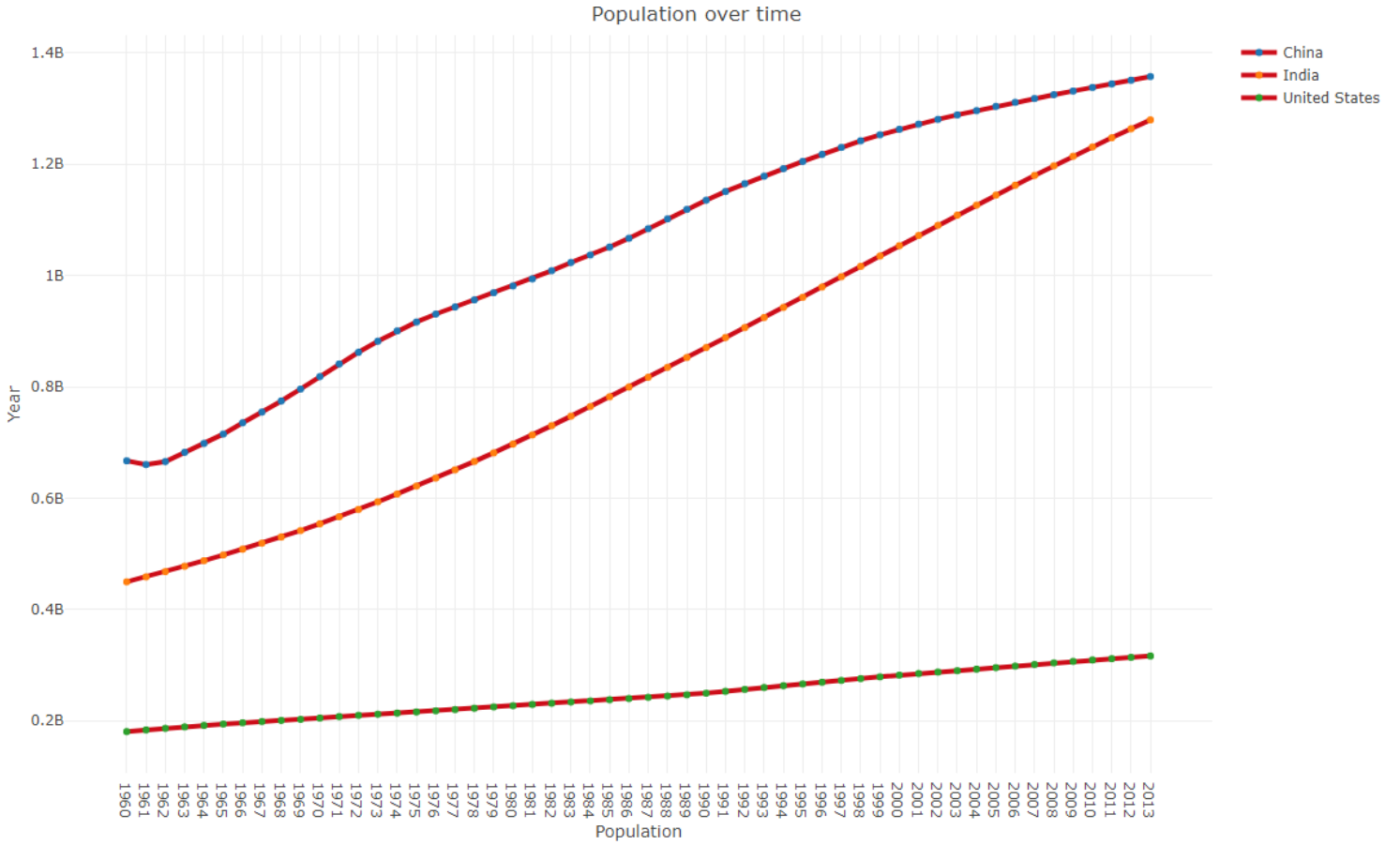
newdata2<-data2[order(-data2[,5]),]
newdata2<-slice(newdata2,1:3)

newdata2 <- newdata2[, -c(1:4)]
newdata2 <- t(newdata2)

newdata2 <- data.frame(years = row.names(newdata2), newdata2)


plot <- plot_ly()%>%
         layout(title = "Population over time",
         xaxis = list(title = "Population"),
         yaxis = list (title = "Year") )
plot <- plot %>% add_trace(x=newdata2[["years"]],y = newdata2[["X1"]], type = 'scatter',
                      mode = 'line+markers', name="China",
                      line = list(color = 'rgb(205, 12, 24)', width = 4))
plot <- plot %>% add_trace(x=newdata2[["years"]],y = newdata2[["X2"]], type = 'scatter',
                      mode = 'line+markers',name="India",
                      line = list(color = 'rgb(205, 12, 24)', width = 4))
plot <- plot %>% add_trace(x=newdata2[["years"]],y = newdata2[["X3"]], type = 'scatter',
                      mode = 'line+markers',name="United States",
                      line = list(color = 'rgb(205, 12, 24)', width = 4))

plot
```
```



Z wykresu można wyczytać, że zdecydowanie najszybszy wzrost populacji wystąpił w Indiach, natomiast w przypadku USA był on bardzo łagodny.

3. Wykres zamiany średniego współczynnika dzietności na przestrzeni lat.

Trzecia wizualizacja ma na celu przedstawienie zamiany średniego współczynnika dzietności na przestrzeni lat. Bardzo ważną informacją jest sposób w jaki został on obliczony, ponieważ z tego powodu można dojść do błędnych wniosków.

Mianowicie, wykorzystałem pakiet tidyverse do obliczenia średniej arytmetycznej współczynnika dla wszystkich krajów bez uwzględnienia ich populacji. Zatem nie można z niego wyczytać zmiany tego współczynnika dla ogółu ludzkości, jednakże można dojść do innych ciekawych wniosków.

```
###  3) Wykres zamiany sredniegowspolczynnika dzietnosci na przestrzeni lat
```{r}

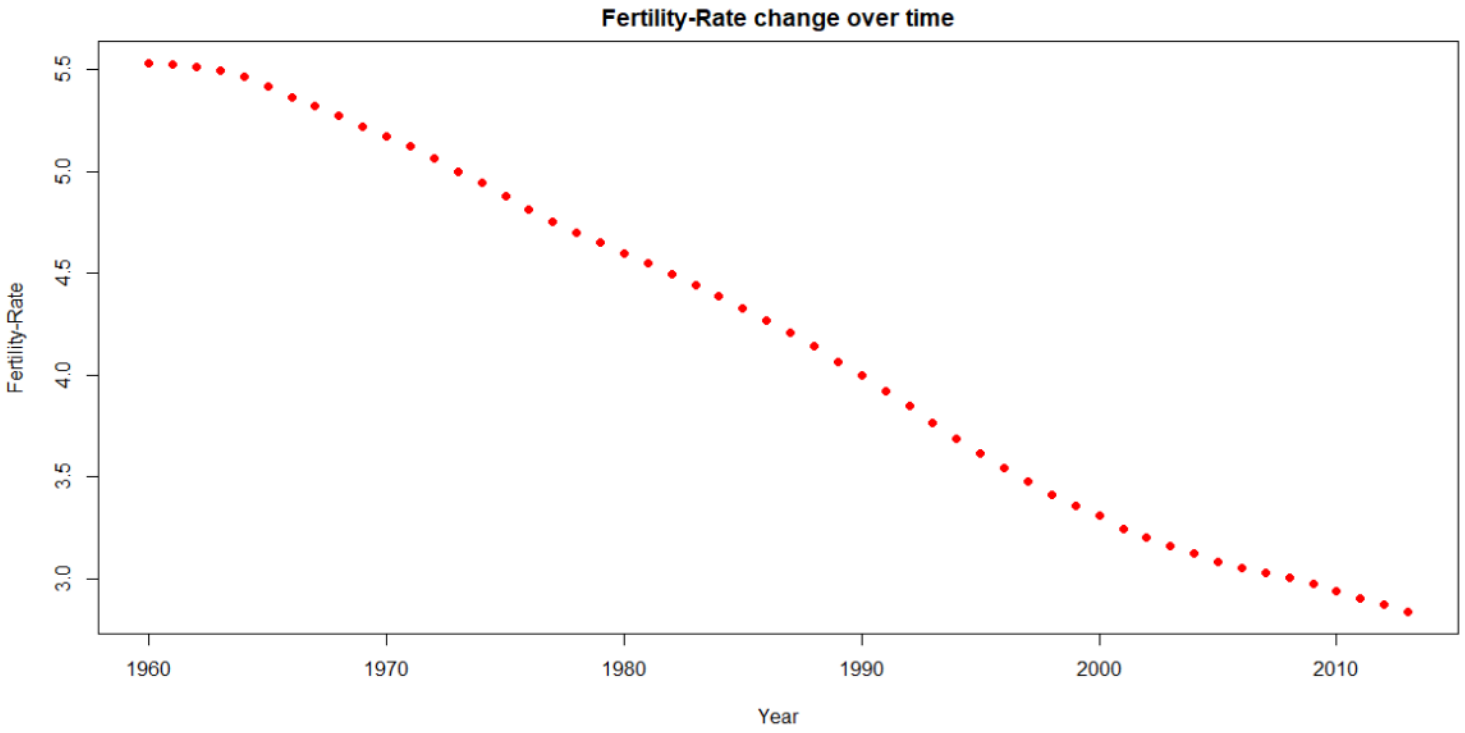
newdata3 <- data3[ , -c(1:4)]

means <- data.frame(colMeans(newdata3))
means <- data.frame(years = row.names(means), means)

plot(y=means$colMeans.newdata3.,x=means$years,xlab="Year",ylab="Fertility-Rate",type = "p", pch=16, col="red")
title("Fertility-Rate change over time")




```
```



Jak widać wskaźnik znacząco spada na przestrzeni lat. Znając jego sposób obliczenia, możemy stwierdzić, że spada on dla większości krajów świata, natomiast rośnie w przypadku krajów o dużej liczbie populacji. Dodatkowe wnioski na ten temat przytoczę przy okazji wizualizacji nr 5.

4. Wykres różnicy średniej długości życia miedzy Polską a Niemcami.

Czwartą wizualizacją jest porównanie zmiany różnicy miedzy średnią długością życia dla Polski i Niemiec na przestrzeni lat.

W celu jej utworzenia wybrałem dane dla obu krajów i obliczyłem różnice miedzy wartościami.

```
###  4) Wykres roznicy sredniej dlugosci zycia miedzy Polska a Niemcami
```{r}

newdata4 <- filter(data4,data4$`Country Name` == "Poland" | data4$`Country Name` == "Germany")
newdata4 <- newdata4[ , -c(1:4)]
newdata4 <- t(newdata4)
newdata4 <- data.frame(years = row.names(newdata4), newdata4)

newdata4 <- mutate(newdata4,diff=X1 - X2)


plot(y=newdata4$diff,x=newdata4$years,xlab="Year",ylab="Life-expectancy-difference",type = "p", pch=16, col="blue")
title("Life-expectancy-difference between Poland and Germany")
lines(newdata4$years,newdata4$diff,type="l")
```
```

Jak widać na wykresie miedzy rokiem 1974 a 1990 wystąpił duży skok różnicy który w przypadku kolejnych lat delikatnie wygasał.

5. Wykresy porównawcze zmiany populacji i współczynnika dzietności.

Ostatnią wizualizacją są dwa wykresy służące do porównania zmian liczby populacji i średniego współczynnika dzietności o którym była mowa w przypadku wizualizacji nr 3. W celu wykonania połączyłem dwie data frames i wykonałem wykresy.

```
###  5) wykresy porownawcze zmiany poppulacji i wspolczynnika dzietnosci
```{r}

newdata5 <- data2[ , -c(1:4)]
newdata5 <- data.frame(Population = colSums(newdata5))
newdata5 <- bind_cols(newdata5,means)

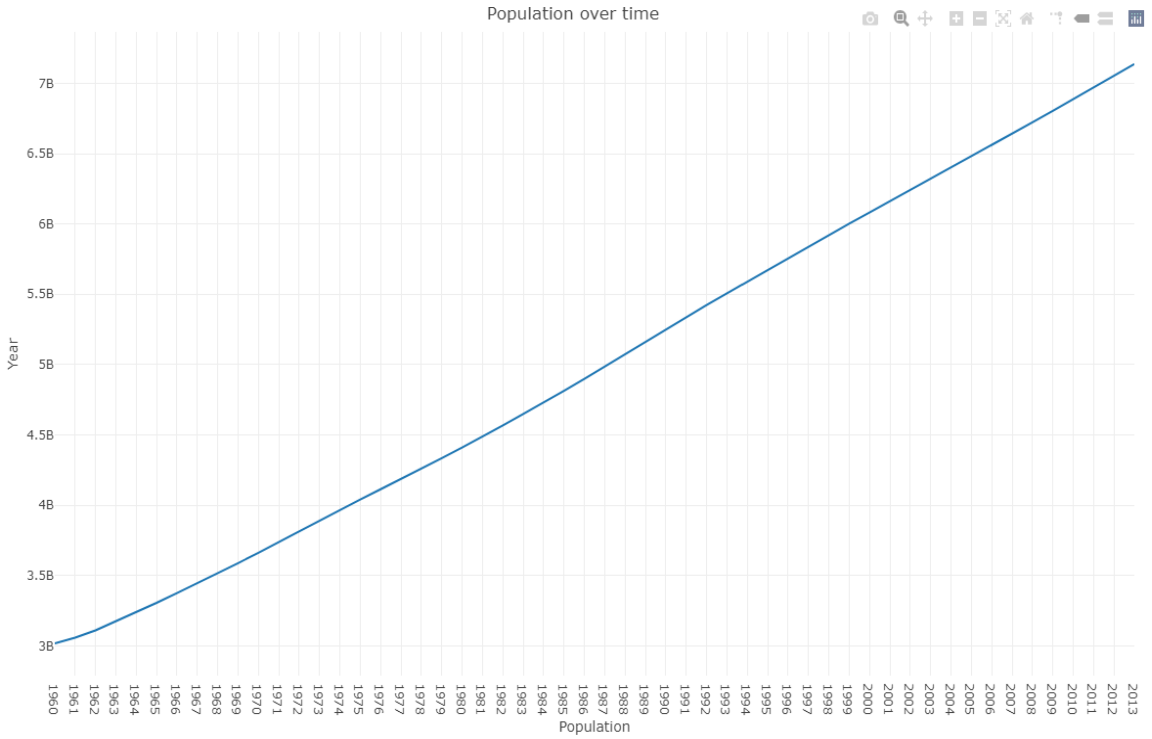

plot1 <- plot_ly(newdata5, x = ~years, y = ~Population, type = 'scatter', mode = 'lines')%>%
         layout(title = "Population over time",
         xaxis = list(title = "Population"),
         yaxis = list (title = "Year") )

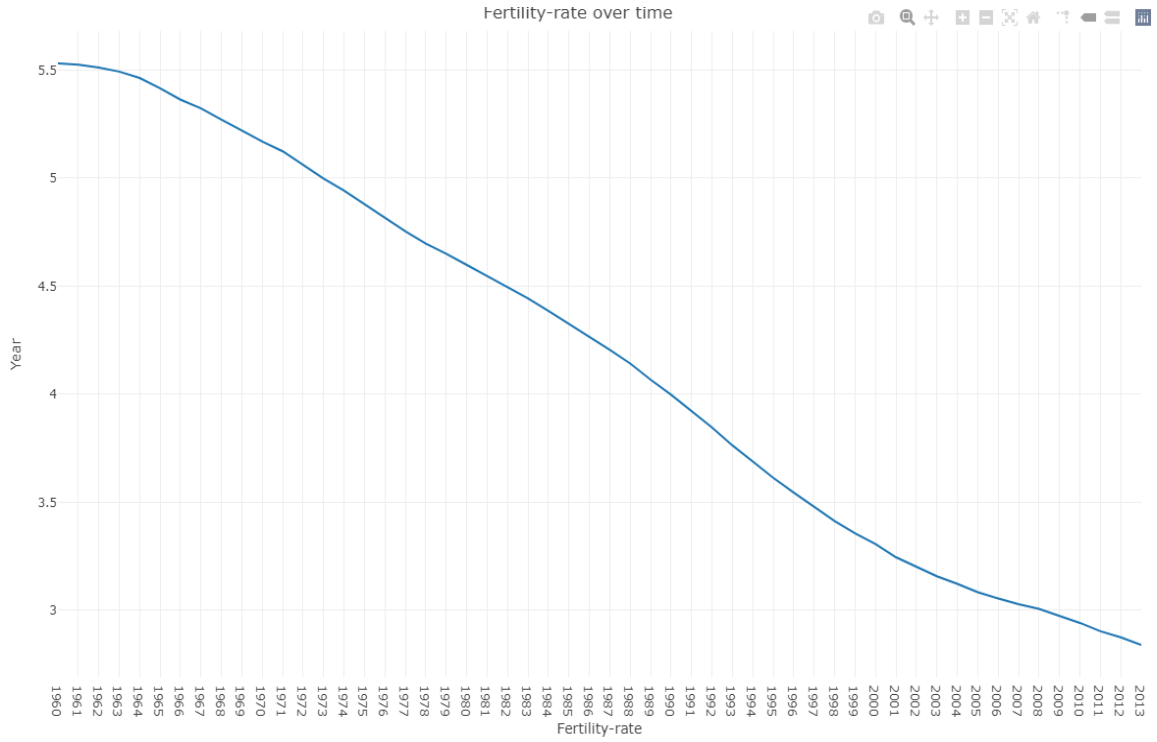

plot2 <- plot_ly(newdata5, x = ~years, y = ~colMeans.newdata3., type = 'scatter', mode = 'lines')%>%
         layout(title = "Fertility-rate over time",
         xaxis = list(title = "Fertility-rate"),
         yaxis = list (title = "Year") )



par(mfrow=c(2,1))
plot1
plot2

```
```





Jak widać pomimo znaczącego spadku współczynnika dzietności, liczba populacji stale rośnie, a nawet zauważalna jest lekka tendencja wzrostowa. Pokazuje to błąd w sposobie liczenia współczynnika, jednakże wiedząc jak zmienia się liczba populacji, można stwierdzić, że dla większości krajów o małej liczbie zaludnienia, współczynnik spada, natomiast rośnie dla krajów o dużej liczbie populacji.