DS5500_Homework1

Noah DeMoes

October 4, 2019

Problem 1

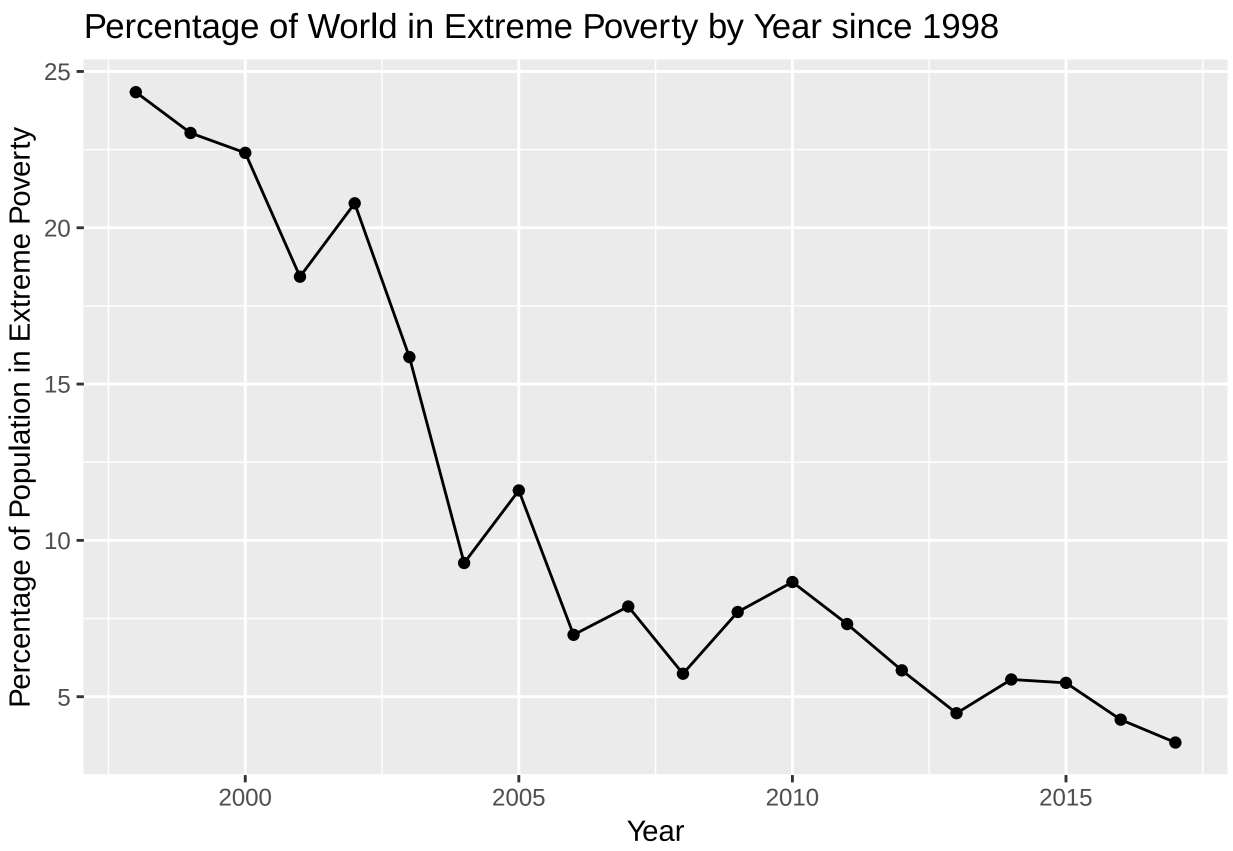
I first scored a 31%, but I took it again and received a 100%.

The question that surprised me the most was about the expected number of children in 2100. The answer is 2 billion which is the same number of children that exist today. I thought that this prediction is interesting given life expectancy is expected to increase. Perhaps, the model is taking into account the Earth's carry capacity. As the number of adults increases, there is prediction that either people will stop having as many children.

The question I selected was problem 2.

In the last 20 years the proportion of people living in extreme poverty worldwide, has…?

```
# Read in files
gapfiles<-list.files()

# Get all csv files
alllfiles<-list.files(gapfiles[1],pattern = '*.csv')

#Get only the disaster files that relate to deaths
Disasterfiles<-c()
i=1
for (f in alllfiles){
  a=grepl('extreme_poverty',f)

  if( a == TRUE){
    Disasterfiles[i]=f
    i=i+1
    }

}


###Read in extreme poberty file
for (f in Disasterfiles){
  df=read.csv(paste0(gapfiles[1],'/',Disasterfiles[1]))
}

df[is.na(df)]<-remove()
dflast20<-df%>%subset(df$time>=1998)
#### Summarize

poverty_summary<-dflast20 %>% group_by(time) %>% summarise(poverty=mean(extreme_poverty_percent_people_


ggplot(data=poverty_summary)+geom_line(aes(x=poverty_summary$time,y=poverty_summary$poverty))+geom_poin
ggtitle('Percentage of World in Extreme Poverty by Year since 1998')+xlab('Year')+ylab("Percentage of P
```



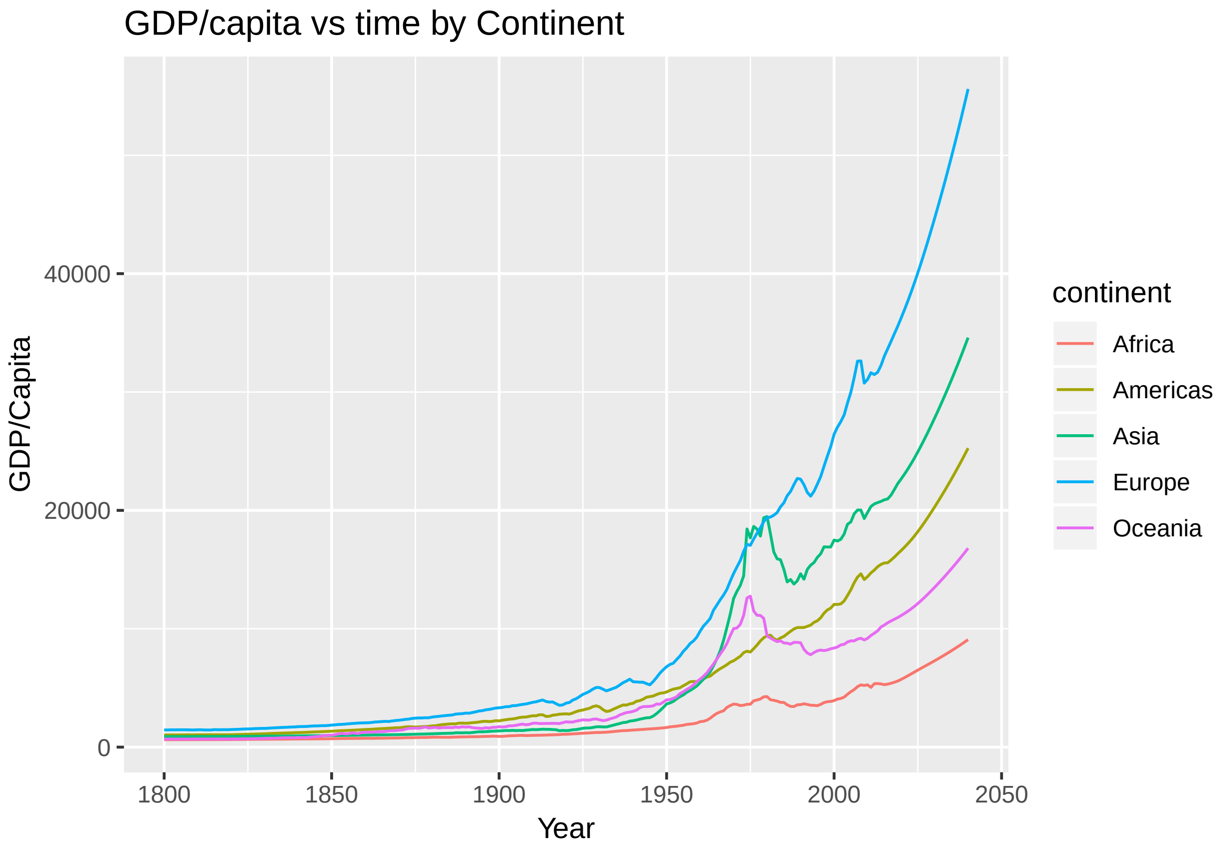
```
difference=poverty_summary$poverty[1]-poverty_summary$poverty[20]
print(difference)
```

```
## [1] 20.8045
```

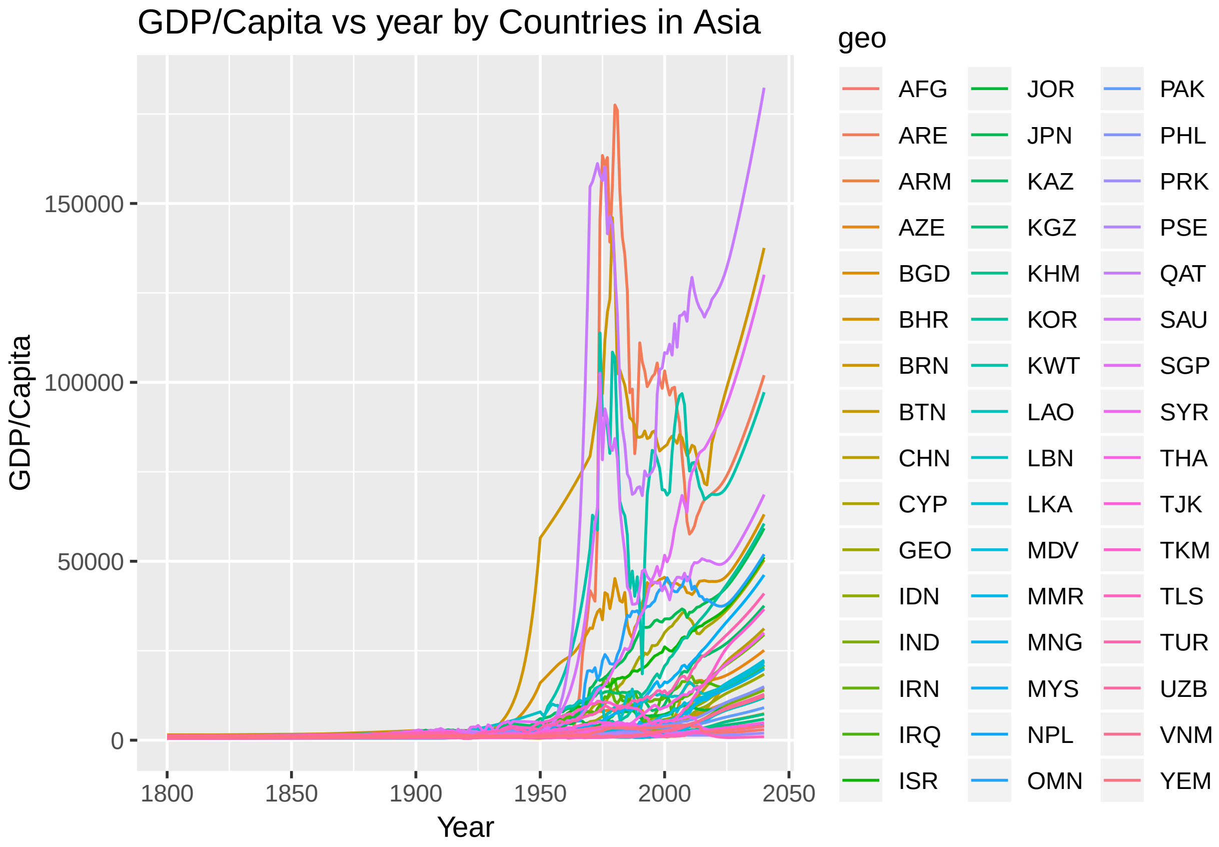
The average percentage of people in the World that live in extreme poverty has decreased by 20.80% from 199 to 2017.

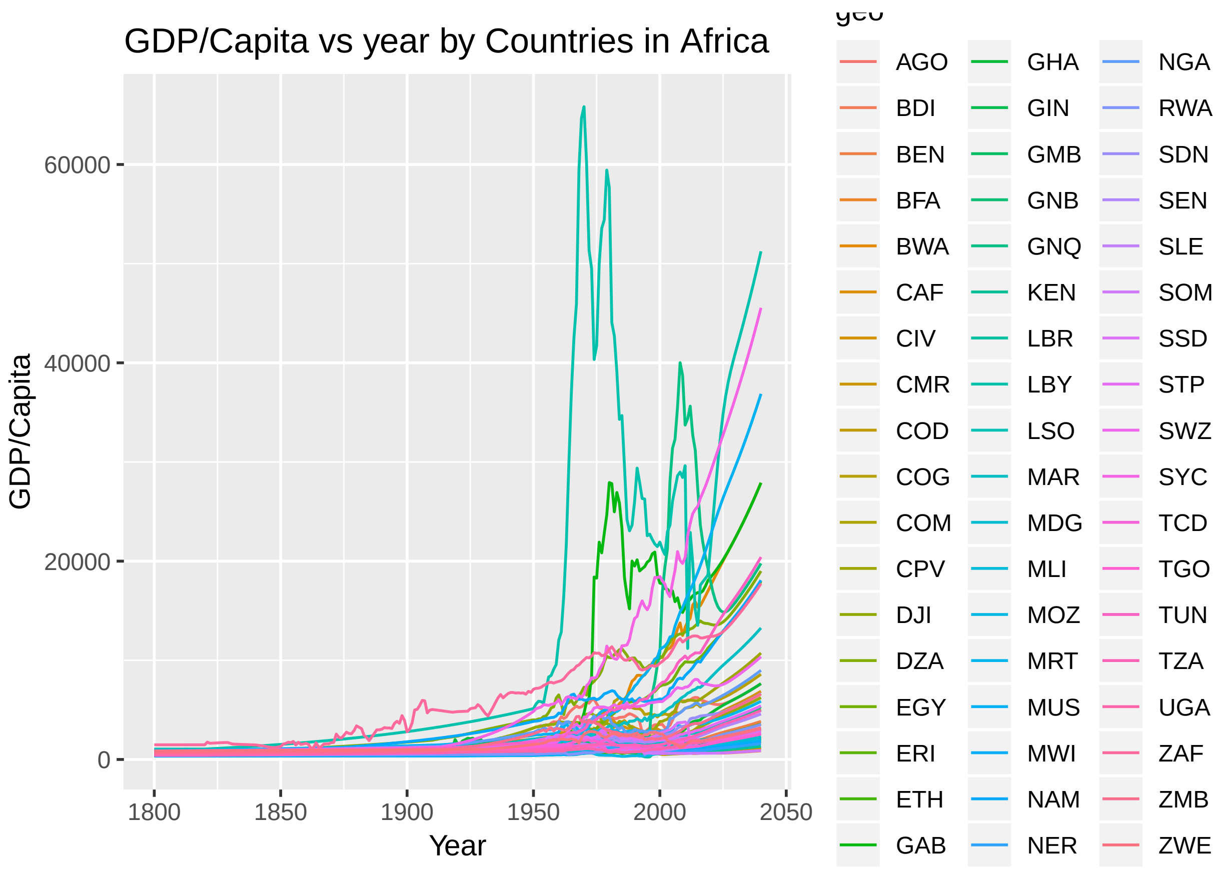
Problem 2

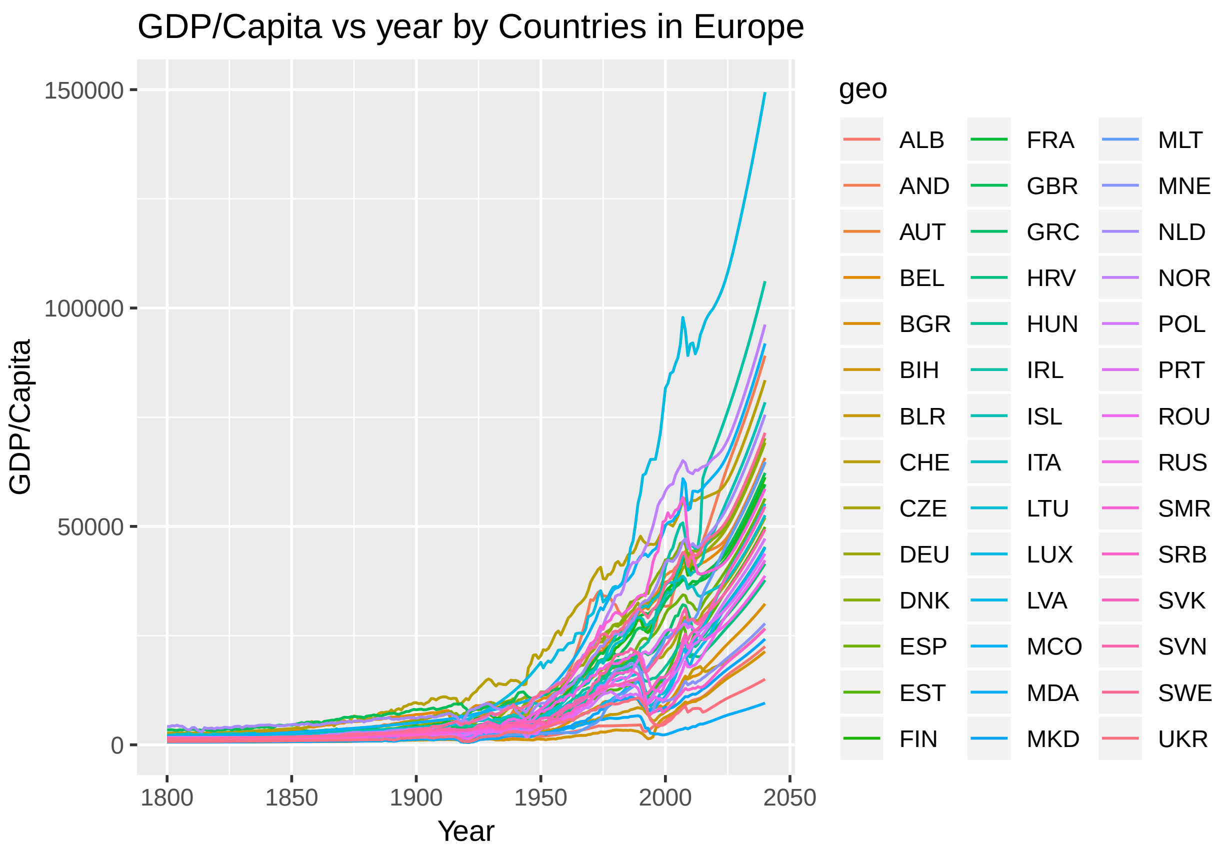
```
library(countrycode)
GDPfiles<-c()
i=1
for (f in alllfiles){
  a=grepl('income_per_person_gdppercapita_ppp_inflation_adjusted',f)
  b=grepl('us',f)

  if( a == TRUE){
    GDPfiles[i]=f
    i=i+1
    }

}

for (f in GDPfiles){
  df=read.csv(paste0(gapfiles[1],'/',GDPfiles[1]))
}

df$geo=as.character(lapply(df$geo,toupper))
df$continent<-countrycode(sourcevar = as.character(df$geo),origin = 'iso3c',destination = 'continent')





gdpsummary<-df%>%group_by(continent,time)%>%summarise(gdp=mean(income_per_person_gdppercapita_ppp_infla



ggplot(data=gdpsummary)+geom_line(aes(x=time,y=gdp,group=continent,colour=continent))+
  ggtitle('GDP/capita vs time by Continent')+xlab('Year')+ylab('GDP/Capita')
```

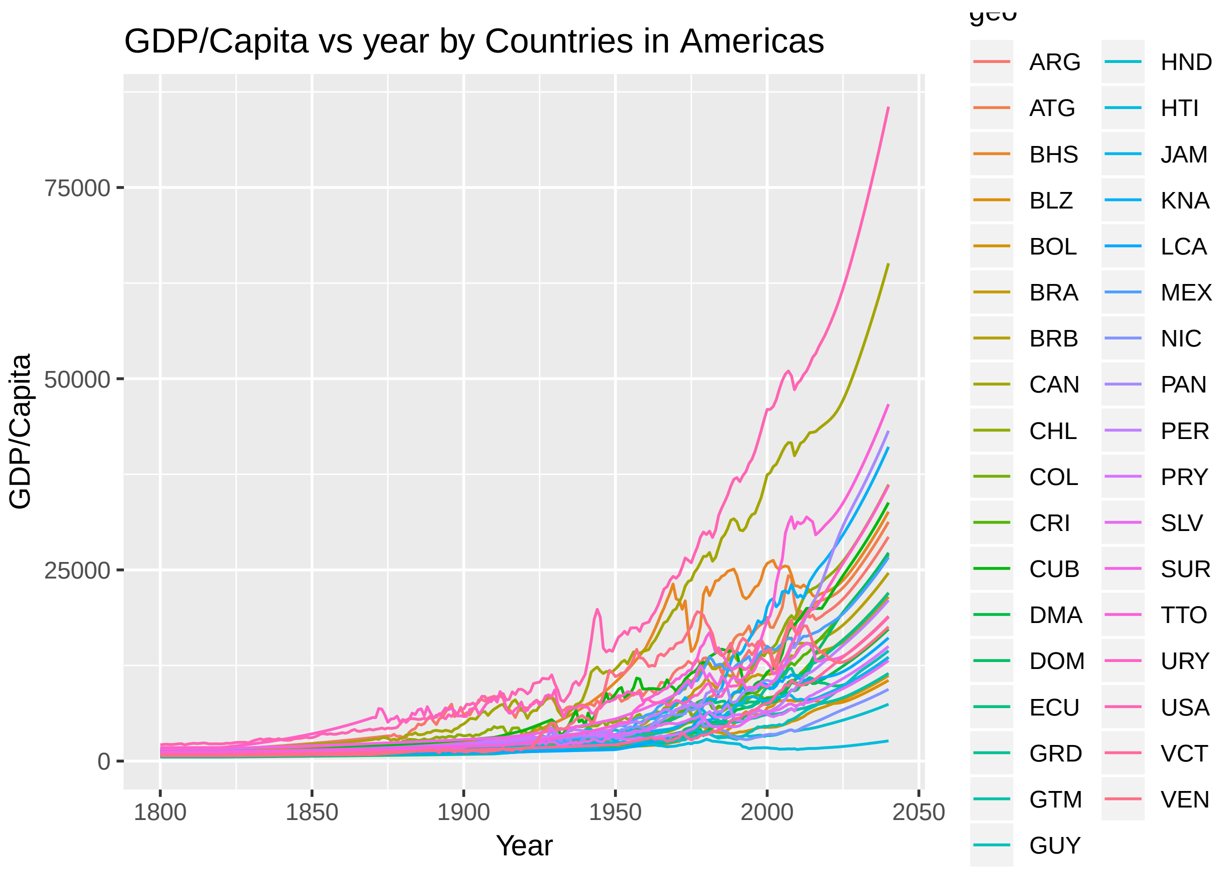


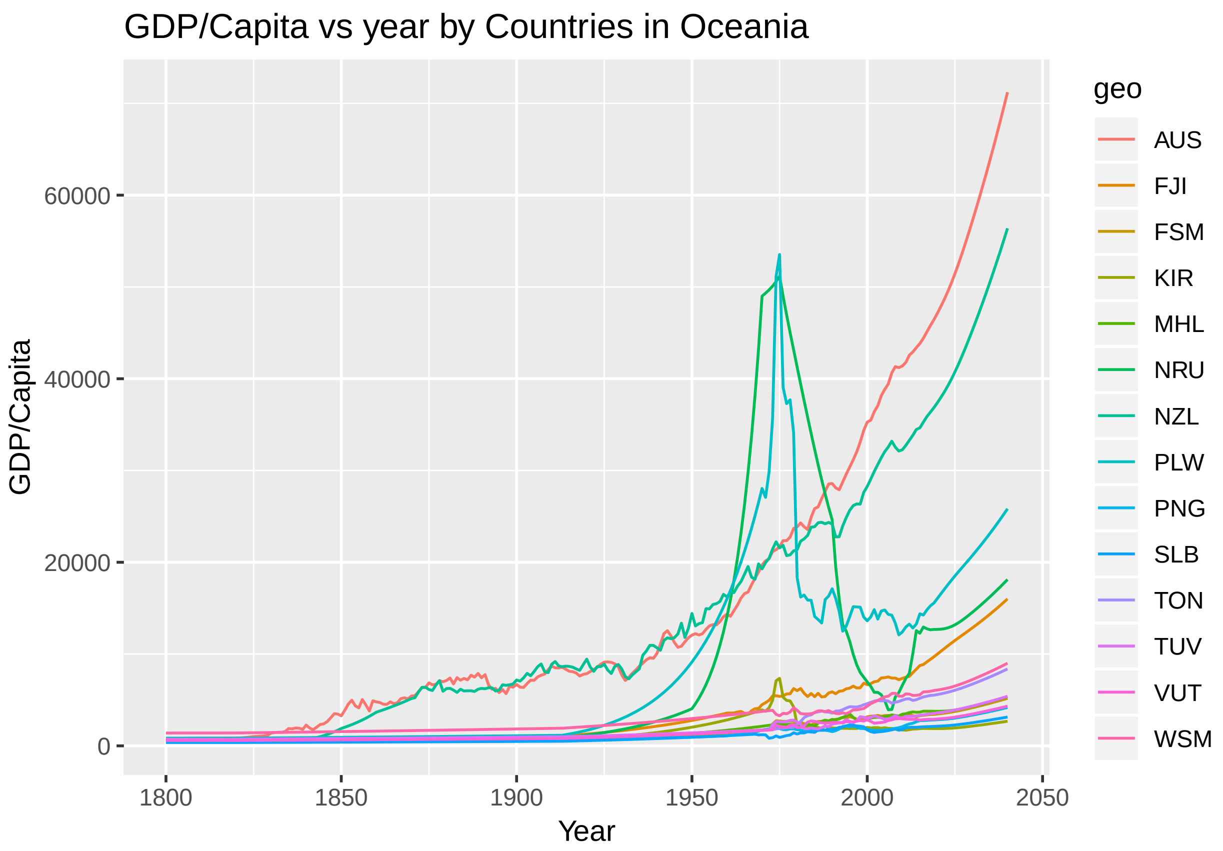
```
for (cont in unique(df$continent)){
dg=df[df$continent==cont,]
print(ggplot(data=dg)+geom_line(aes(x=time,y=income_per_person_gdppercapita_ppp_inflation_adjusted,grou
```


GDP/Capita vs year by Countries in Asia
GDP/Capita
150000
100000
50000
0
1800
1850
1900
1950
2000
2050
Year
geo
AFG
ARE
ARM
AZE
BGD
BHR
BRN
BTN
CHN
CYP
GEO
IDN
IND
IRN
IRQ
ISR
JOR
JPN
KAZ
KGZ
KHM
KOR
KWT
LAO
LBN
LKA
MDV
MMR
MNG
MYS
NPL
OMN
PAK
PHL
PRK
PSE
QAT
SAU
SGP
SYR
THA
TJK
TKM
TLS
TUR
UZB
VNM
YEM


GDP/Capita vs year by Countries in Africa
GDP/Capita
60000
40000
20000
0
1800
1850
1900
1950
2000
2050
Year
geo
AGO
BDI
BEN
BFA
BWA
CAF
CIV
CMR
COD
COG
COM
CPV
DJI
DZA
EGY
ERI
ETH
GAB
GHA
GIN
GMB
GNB
GNQ
KEN
LBR
LBY
LSO
MAR
MDG
MLI
MOZ
MRT
MUS
MWI
NAM
NER
NGA
RWA
SDN
SEN
SLE
SOM
SSD
STP
SWZ
SYC
TCD
TGO
TUN
TZA
UGA
ZAF
ZMB
ZWE


GDP/Capita vs year by Countries in Europe
GDP/Capita
150000
100000
50000
0
1800
1850
1900
1950
2000
2050
Year
geo
ALB
AND
AUT
BEL
BGR
BIH
BLR
CHE
CZE
DEU
DNK
ESP
EST
FIN
FRA
GBR
GRC
HRV
HUN
IRL
ISL
ITA
LTU
LUX
LVA
MCO
MDA
MKD
MLT
MNE
NLD
NOR
POL
PRT
ROU
RUS
SMR
SRB
SVK
SVN
SWE
UKR


GDP/Capita vs year by Countries in Americas
GDP/Capita
75000
50000
25000
0
1800
1850
1900
1950
2000
2050
Year
geo
ARG
ATG
BHS
BLZ
BOL
BRA
BRB
CAN
CHL
COL
CRI
CUB
DMA
DOM
ECU
GRD
GTM
GUY
HND
HTI
JAM
KNA
LCA
MEX
NIC
PAN
PER
PRY
SLV
SUR
TTO
URY
USA
VCT
VEN



```
GDPdf=df
```

I first looked at the GDP/Capita for each Continent as a function of time from 1800 to 2050. It appears as though for the first 150 years each continent was on the same tradjectory with only a small variance between continents. After 1975, however, each continent's GDP/capita seperates, with Europe leading, followed by Asia, Americas, Oceania, and Africa. Perhaps the baby boom contributed to the seperation of GDP.

Additionally, I provided graphics of the GDP/Capita by country seperated by continent. Asia, Oceania, and Africa all have one or two countries that spike in GDP/Capita during 1975 while the countries in Europe and Americas seem to follow more stable trends. Perhaps, this means that countries in Europe and the Americas are more dependent on each other and work together more than countries on other continents. However, USA seems to do well despite variations in other countries.

Problem 3

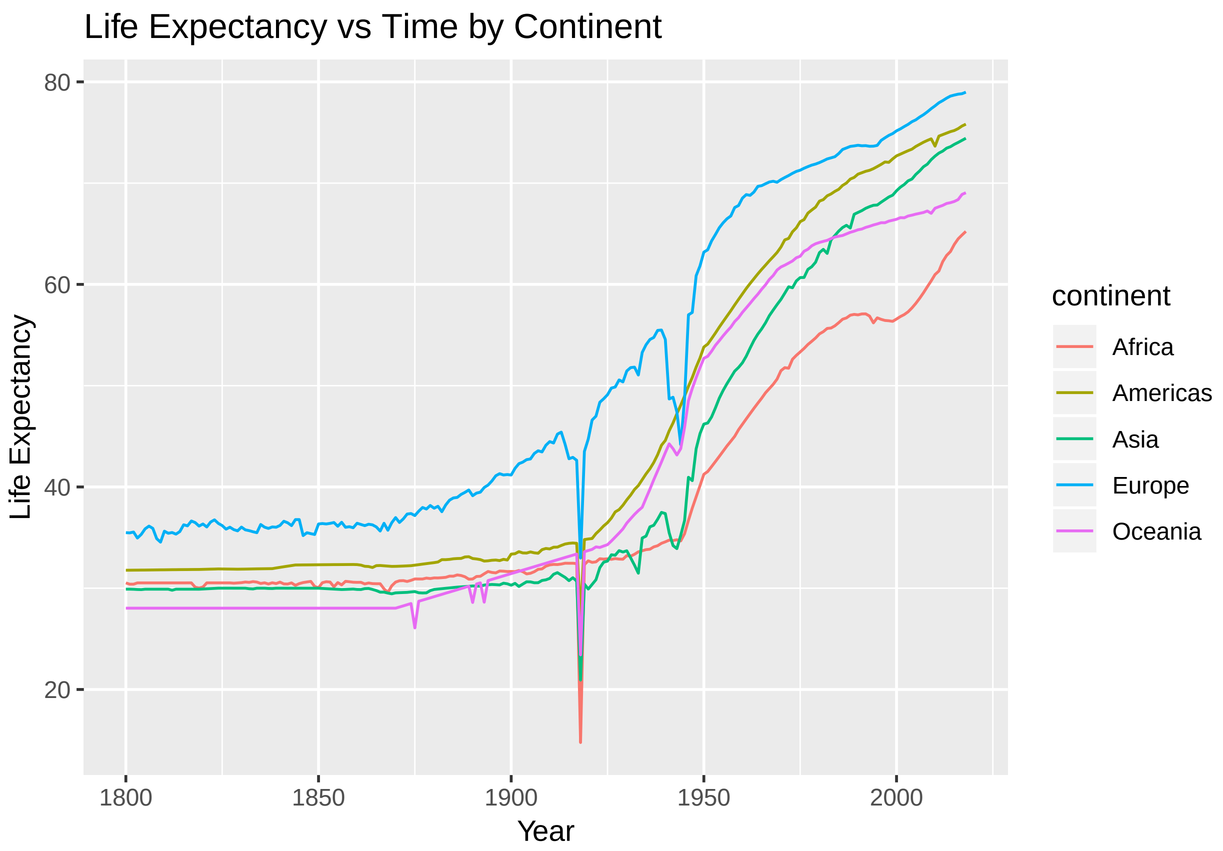
```
lifemortalfiles<-c()
i=1
for (f in alllfiles){
  a=grepl('child_mortality',f)
  b=grepl('us',f)

  if( a == TRUE){
    lifemortalfiles[i]=f
    i=i+1
    }
}

for (f in lifemortalfiles){
  child_df=read.csv(paste0(gapfiles[1],'/',lifemortalfiles[1]))
}


child_df$geo=as.character(lapply(child_df$geo,toupper))


lifemortalfiles<-c()
i=1

for (f in alllfiles){
  a=grepl('life_expectancy_years',f)
  b=grepl('us',f)

  if( a == TRUE){
    lifemortalfiles[i]=f
    i=i+1
    }

}

for (f in lifemortalfiles){
  lifeexpectancy_df=read.csv(paste0(gapfiles[1],'/',lifemortalfiles[1]))
}
lifeexpectancy_df$geo=as.character(lapply(lifeexpectancy_df$geo,toupper))


p3df=merge(GDPdf,child_df,by=c('geo','time'),all=T)

p3df=remove_missing(p3df)
```

```
## Warning: Removed 7056 rows containing missing values.
```

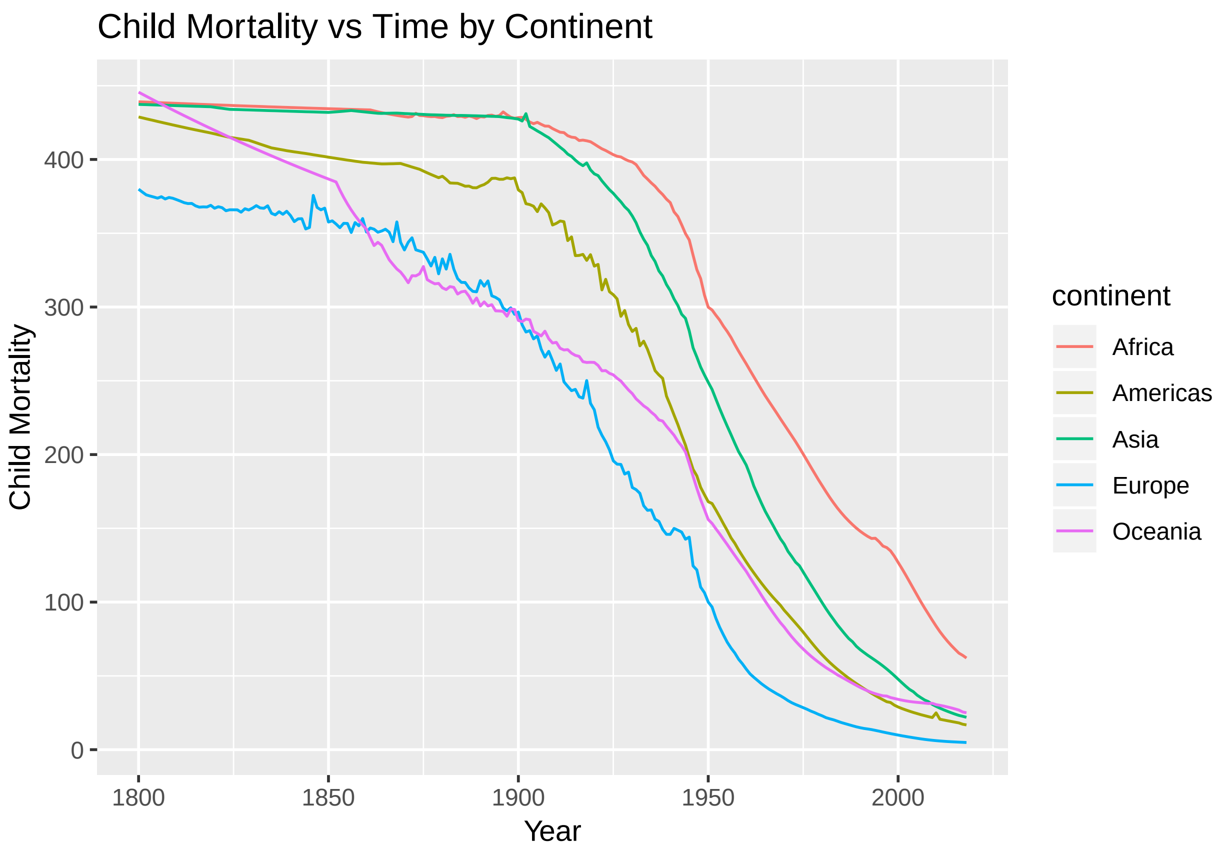
```
p3df=merge(p3df,lifeexpectancy_df,by=c('geo','time'),all=T)
p3df=remove_missing(p3df)
```

```
## Warning: Removed 4189 rows containing missing values.
```

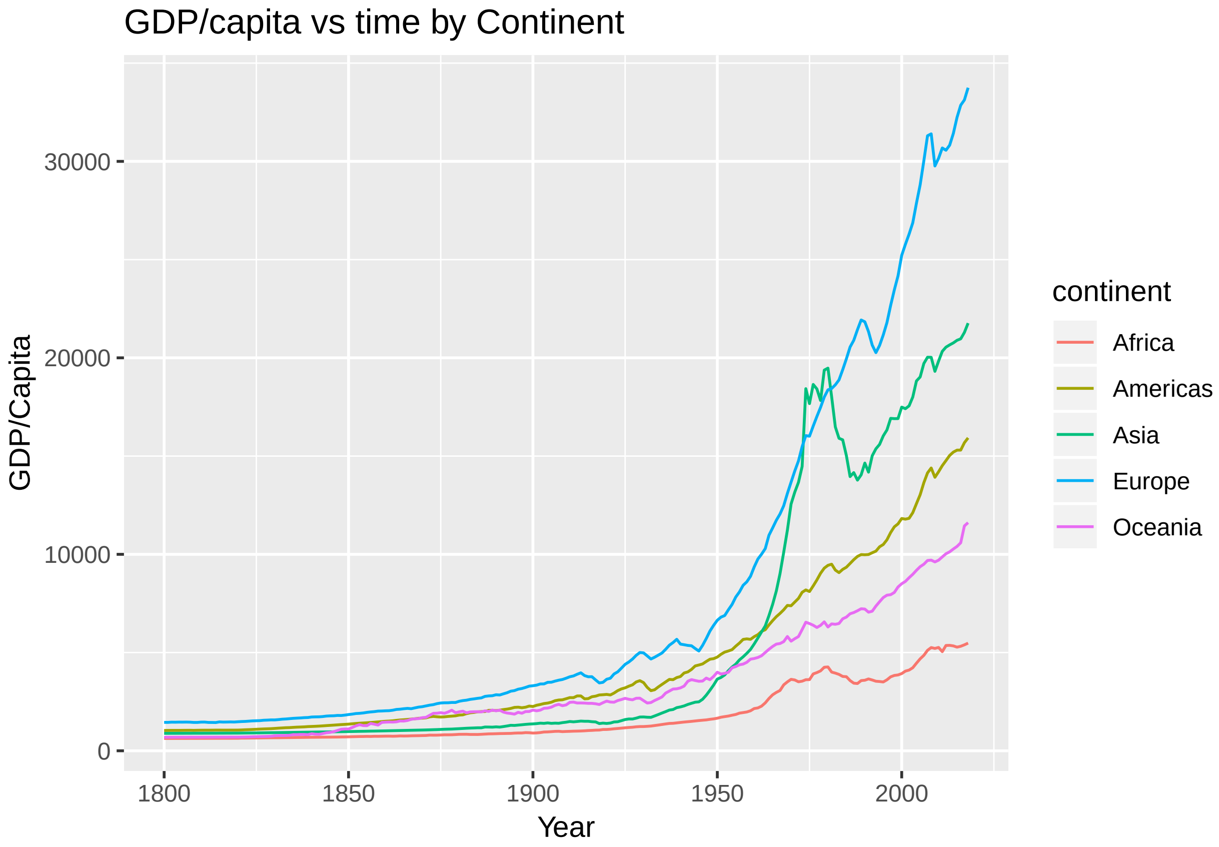
```
gdpsummary<-p3df%>%group_by(continent,time)%>%summarise(gdp=mean(income_per_person_gdppercapita_ppp_infl



ggplot(data=gdpsummary)+geom_line(aes(x=time,y=life_exp,group=continent,colour=continent))+
  ggtitle('Life Expectancy vs Time by Continent')+xlab('Year')+ylab('Life Expectancy')
```



```
ggplot(data=gdpsummary)+geom_line(aes(x=time,y=child_mortality,group=continent,colour=continent))+
  ggtitle('Child Mortality vs Time by Continent')+xlab('Year')+ylab('Child Mortality')
```



```
ggplot(data=gdpsummary)+geom_line(aes(x=time,y=gdp,group=continent,colour=continent))+
  ggtitle('GDP/capita vs time by Continent')+xlab('Year')+ylab('GDP/Capita')
```



The main outlier for life expectance occurs during 1914-1919 and 1939-1944, during world war world war I and II. All countries are effected. Besides that period, life expectance has trended higher as time goes increased from 1800 to 2017. There does appear to be a sharp increase in Life Expectancy after 1950 that trends higher. This can be attributed to advancements in healthcare and modern medicine. This conclusion is supported by the trends in child mortality. Child mortality rates year over year changed moderately between 1800-1900. It wasn't until 1900 did they decrease at a higher rate. It should be noted that modern medicine did not arrive at each continent at the same time which attributes to the slower decrease in mortaility rates for AFrica and Asia compared to Europe, Oceania, and the Americas.

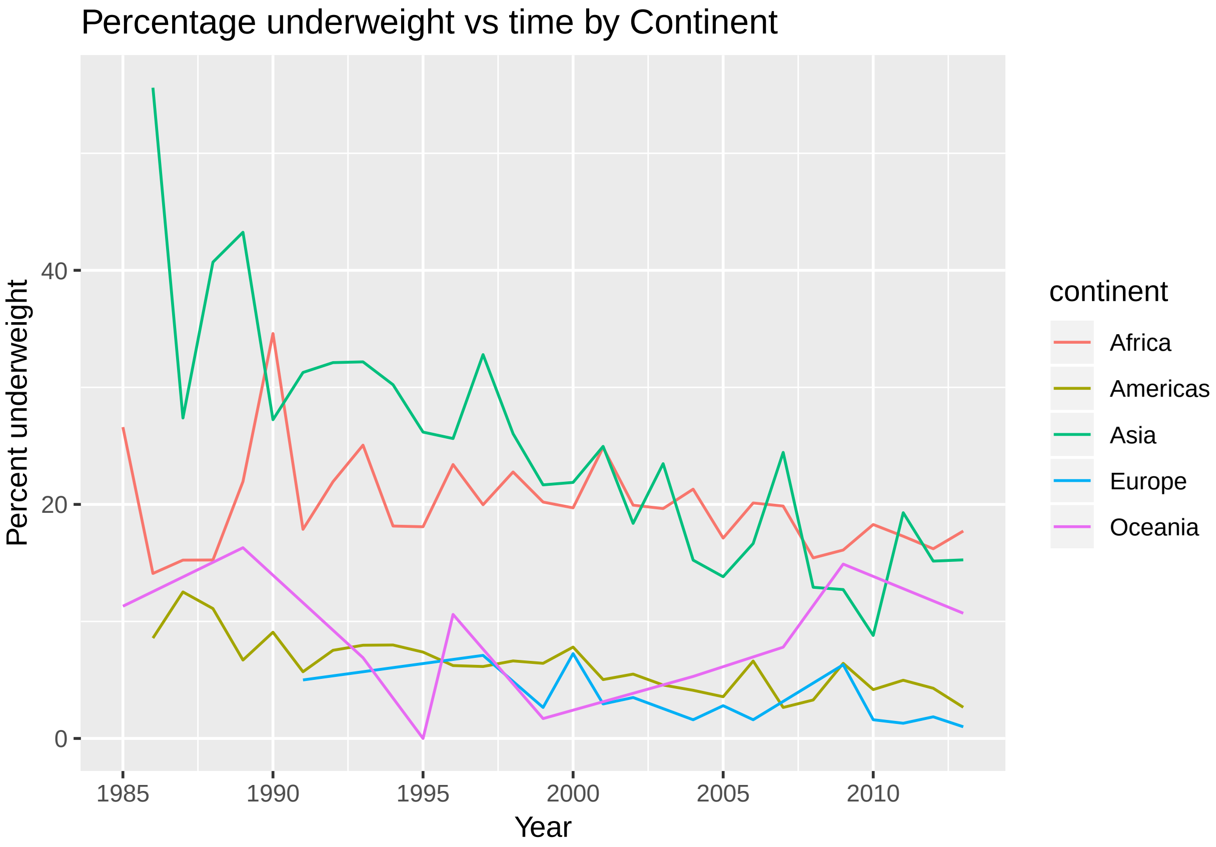
Additionally, rate of change for GDP/captia follows relatively the same reciprocal of child mortality rates. It is more robust to life expectance as shown by the lack of change in GDP/capita during the world wars.

Problem 4

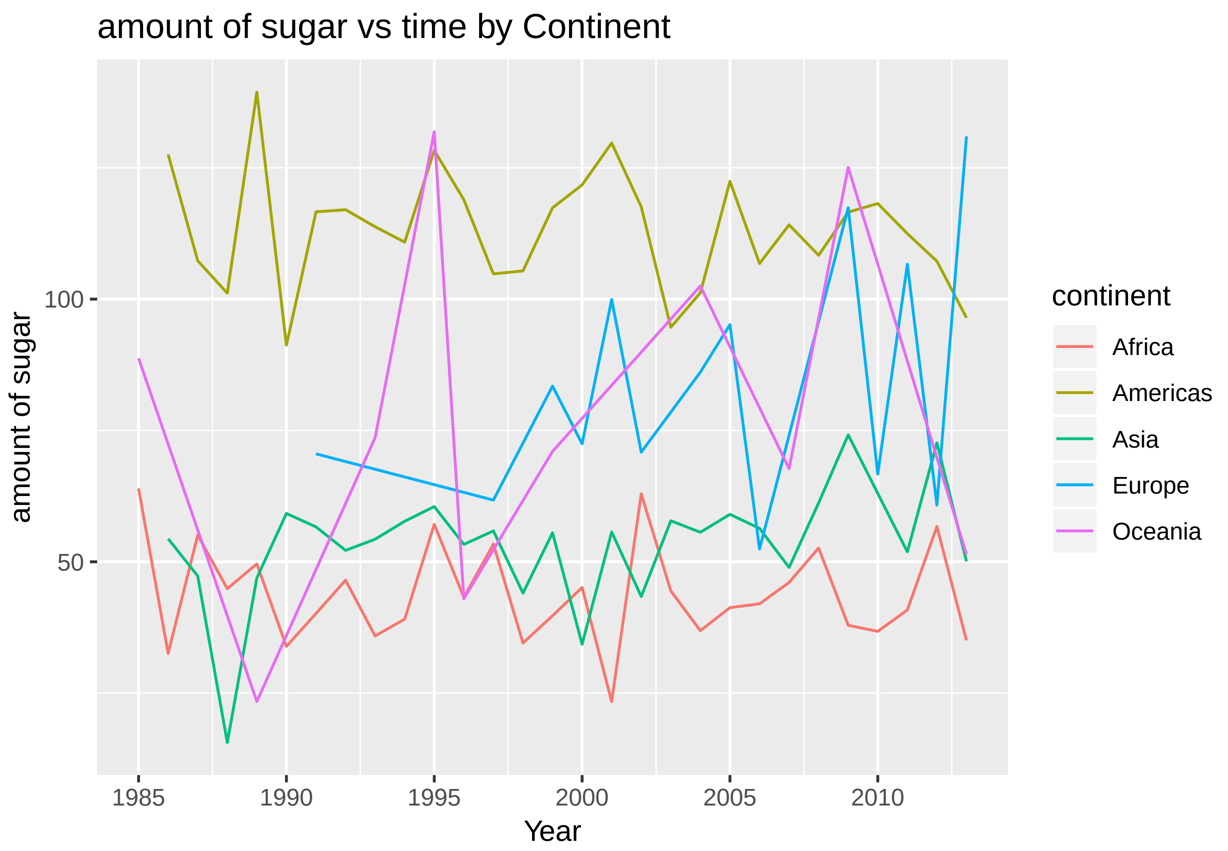
```
cancer<-c()
i=1
for (f in alllfiles){
  a=grepl('malnutr',f)
  b=grepl('us',f)

  if( a == TRUE){
    cancer[i]=f
    i=i+1
    }
}

for (f in cancer){
  malnutr=read.csv(paste0(gapfiles[1],'/',cancer[1]))
}
malnutr$geo=as.character(lapply(malnutr$geo,toupper))



sugar<-c()
i=1
for (f in alllfiles){
  a=grepl('sugar',f)
  b=grepl('us',f)

  if( a == TRUE){
    sugar[i]=f
    i=i+1
    }

}

for (f in sugar){
  sugar_df=read.csv(paste0(gapfiles[1],'/',sugar[1]))
}

sugar_df$geo=as.character(lapply(sugar_df$geo,toupper))

p4df=merge(malnutr,sugar_df,by=c('geo','time'),all=T)
p4df<-remove_missing(p4df)
```

```
## Warning: Removed 7943 rows containing missing values.
```

```
p4df$continent<-countrycode(sourcevar = as.character(p4df$geo),origin = 'iso3c',destination = 'continen

p4summary<-p4df%>%group_by(continent,time)%>%summarise(sugar=mean(sugar_per_person_g_per_day),malnut=mea


ggplot(data=p4summary)+geom_line(aes(x=time,y=malnut,group=continent,colour=continent))+
  ggtitle('Percentage underweight vs time by Continent')+xlab('Year')+ylab('Percent underweight')
```



```
ggplot(data=p4summary)+geom_line(aes(x=time,y=sugar,group=continent,colour=continent))+
  ggtitle('amount of sugar vs time by Continent')+xlab('Year')+ylab('amount of sugar')
```



In problem 4 I looked at the relationship between the amount of sugar consumed by continent and the level of malnurishment in the continent. The result of the amount of sugar by continent is hard to distinguish. Every continent experiences an oscillating pattern for the amount of sugar consumed year over year. Because of the oscillation it is hard to draw any meaningful relationship between the amount of sugar and thelevel of malnutrition.

Problem 5

I utilized static plots.

The advantage of utilizing a static plot is you can select the percise information you would like to present for a given topic. This ability limits the likelihood that an audience will get lost or overwhelmed by too much information. The disadvantage is that it can be hard to convey changes with time for mutliple variables. Additionally, it limits your ability to adapt to a variety of questions, which is one of the advantages of having dynamic plots. Dynamic visualizations enable a research to capture a much wider range of information. Additonally, you can capture changes in two or more variables as a function of another and still maintain understanding.