615 HW3

Shangchen Han

10/4/2019

Problem 1

```
attach(gapminder)
#How many continents are included in the data set?
data <- gapminder
num_cont <-  data %>% count(continent)
num_cont
```

```
## # A tibble: 5 x 2
##   continent     n
##   <fct>     <int>
## 1 Africa      624
## 2 Americas    300
## 3 Asia        396
## 4 Europe      360
## 5 Oceania      24
```

So, the number of continents is five.

```
#How many countrys are included? How many countries per continent?
num_coun <- data %>% count(country)
num_coun
```

```
## # A tibble: 142 x 2
##    country         n
##    <fct>       <int>
##  1 Afghanistan    12
##  2 Albania        12
##  3 Algeria        12
##  4 Angola         12
##  5 Argentina      12
##  6 Australia      12
##  7 Austria        12
##  8 Bahrain        12
##  9 Bangladesh     12
## 10 Belgium        12
## # ... with 132 more rows
```

```
num_coun_per_cont <- data %>% group_by(continent) %>% summarise(country %>% unique %>% length)
num_coun_per_cont
```

```
## # A tibble: 5 x 2
##   continent `country %>% unique %>% length`
##   <fct>                                <int>
## 1 Africa                                  52
## 2 Americas                                25
## 3 Asia                                    33
## 4 Europe                                  30
## 5 Oceania                                  2
```

There are 142 countries. And there are 52, 25, 33, 30, 2 countries in Africa, Americas, Asia, Europe, and Oceania, respectively.

```
#Using the gapminder data, produce a report showing the continents in the dataset, total population per

Per <- data %>% group_by(continent) %>% summarise(population_million = sum(pop)/1000000, GDP_million = s
kable(cbind(Per), caption = "Total population and total GDP for each continents", align = "c", booktab =
```

Table 1: Total population and total GDP for each continents

continent	population_million	GDP_million
Africa	6187.5860	1.3689029
Americas	7351.4385	2.1408331
Asia	30507.3339	3.1292516
Europe	6181.1153	5.2090112
Oceania	212.9921	0.4469186

```
#Produce a well-labeled table that summarizes GDP per capita for the countries in each continent, contr
Summary_1952 <- data %>% filter(year == 1952)
Summary_2007 <- data %>% filter(year == 2007)
Per_1952 <- Summary_1952 %>% group_by(continent) %>% summarise(Total_GDP_thousand = sum(gdpPercap)/1000
Per_2007 <- Summary_2007 %>% group_by(continent) %>% summarise(Total_GDP_thousand = sum(gdpPercap)/1000
kable(cbind(Per_1952), caption = "Summary GDP per capita for the countries in  each continents in 1952"
```

Table 2: Summary GDP per capita for the countries in each continents in 1952

continent	Total_GDP_thousand	Ave_GDP_thousand	Max_GDP_thousand	Min_GDP_thousand
Africa	65.13377	1.252573	4.725295	0.2988462
Americas	101.97656	4.079063	13.990482	1.3977171
Asia	171.45097	5.195484	108.382353	0.3310000
Europe	169.83172	5.661057	14.734233	0.9735332
Oceania	20.59617	10.298086	10.556576	10.0395956

```
kable(cbind(Per_2007), caption = "Summary GDP per capita for the countries in each continents in 2007",
```

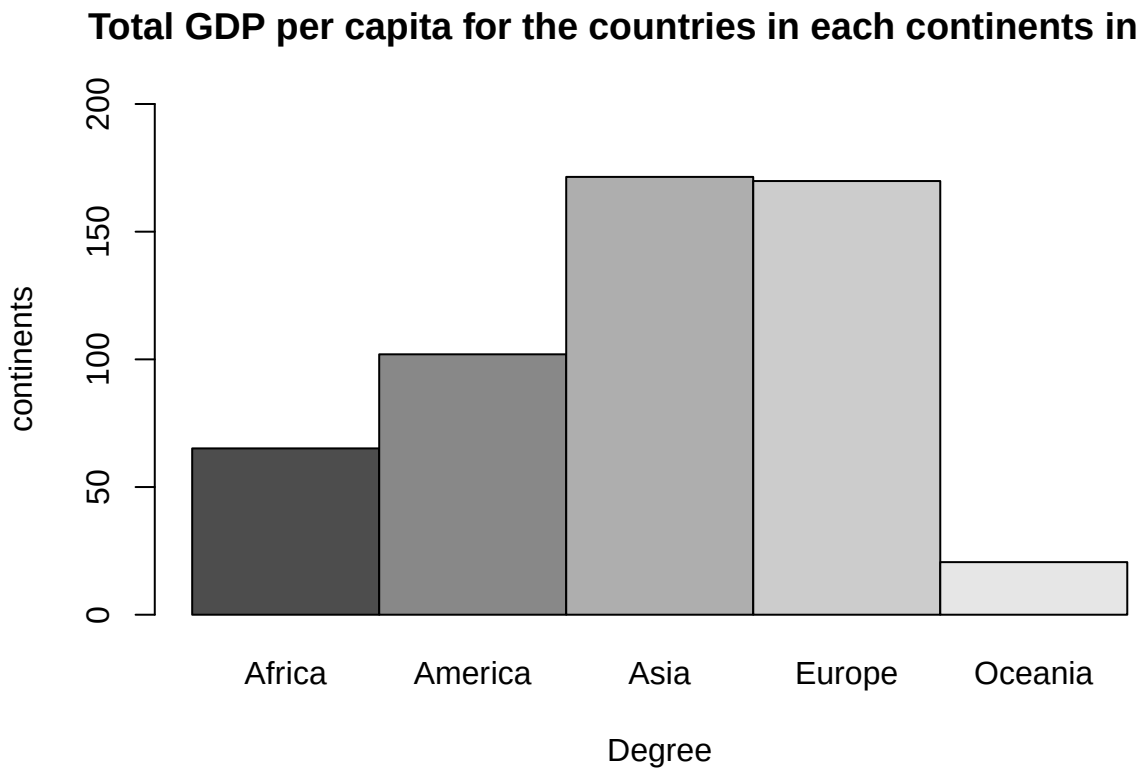
Table 3: Summary GDP per capita for the countries in each continents in 2007

continent	Total_GDP_thousand	Ave_GDP_thousand	Max_GDP_thousand	Min_GDP_thousand
Africa	160.62970	3.089033	13.20648	0.2775519
Americas	275.07579	11.003032	42.95165	1.2016372
Asia	411.60989	12.473027	47.30699	0.9440000
Europe	751.63445	25.054482	49.35719	5.9370295
Oceania	59.62038	29.810188	34.43537	25.1850091

```
#Product a plot that summarizes the same data as the table. There should be two plots per continent.
Total_1952 <- data %>% filter(year==1952)
Total_1952 <- Total_1952 %>% group_by(continent) %>% summarise(Total_GDP_thousand = sum(gdpPercap)/1000
Total_1952
```

```
## # A tibble: 5 x 2
##   continent Total_GDP_thousand
##   <fct>                  <dbl>
## 1 Africa                  65.1
## 2 Americas               102.
## 3 Asia                   171.
## 4 Europe                 170.
## 5 Oceania                 20.6
```

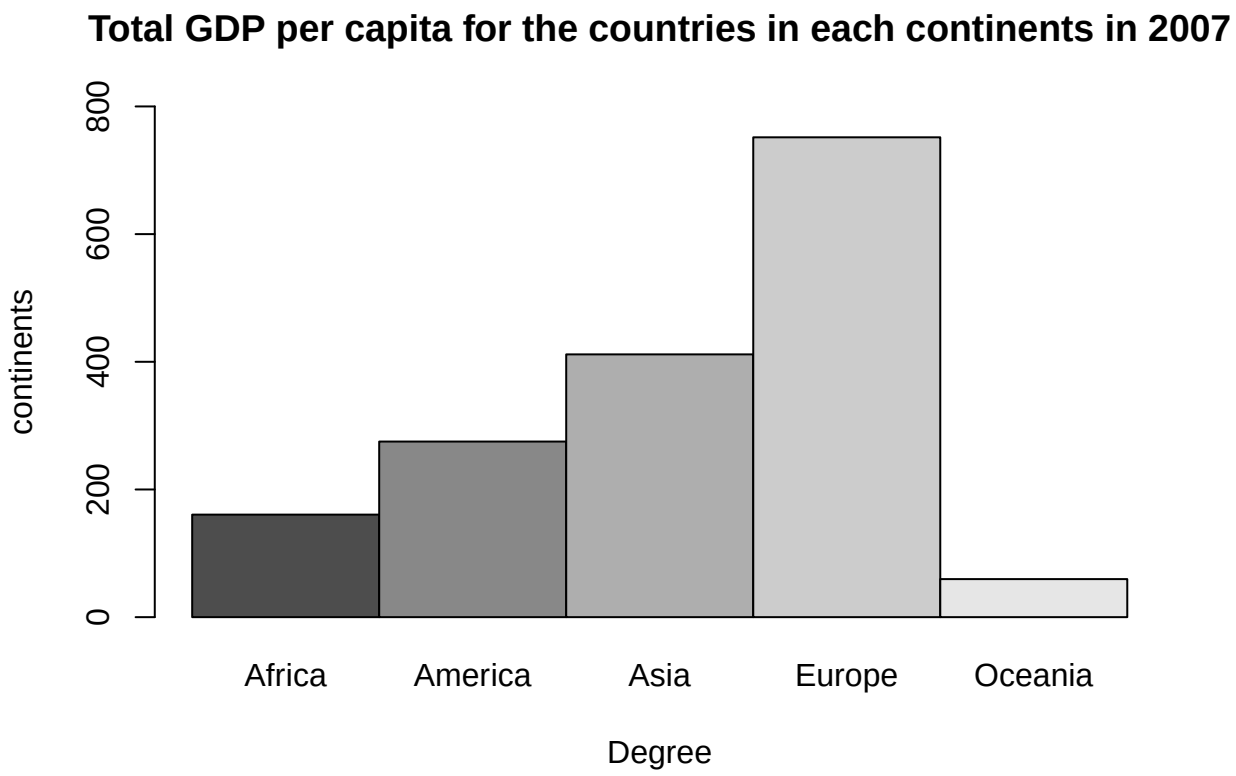
```
barplot(as.matrix(Total_1952[,2]),beside = T,legend.text = T,main = "Total GDP per capita for the counti
```



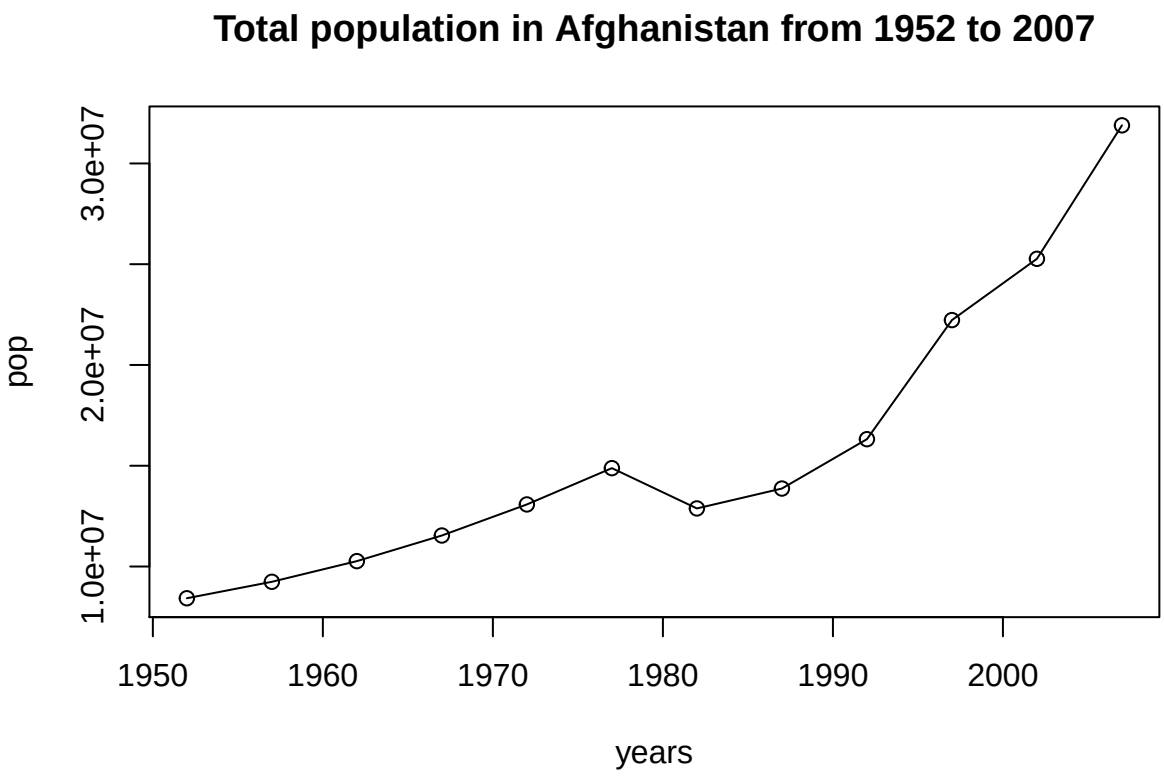

```
Total_2007 <- data %>% filter(year==2007)
Total_2007 <- Total_2007 %>% group_by(continent) %>% summarise(Total_GDP_thousand = sum(gdpPercap)/1000)
Total_2007
```

```
## # A tibble: 5 x 2
##   continent Total_GDP_thousand
##   <fct>                  <dbl>
## 1 Africa                 161.
## 2 Americas               275.
## 3 Asia                   412.
## 4 Europe                 752.
## 5 Oceania                 59.6
```

```
barplot(as.matrix(Total_2007[,2]),beside = T,legend.text = T,main = "Total GDP per capita for the counti
```

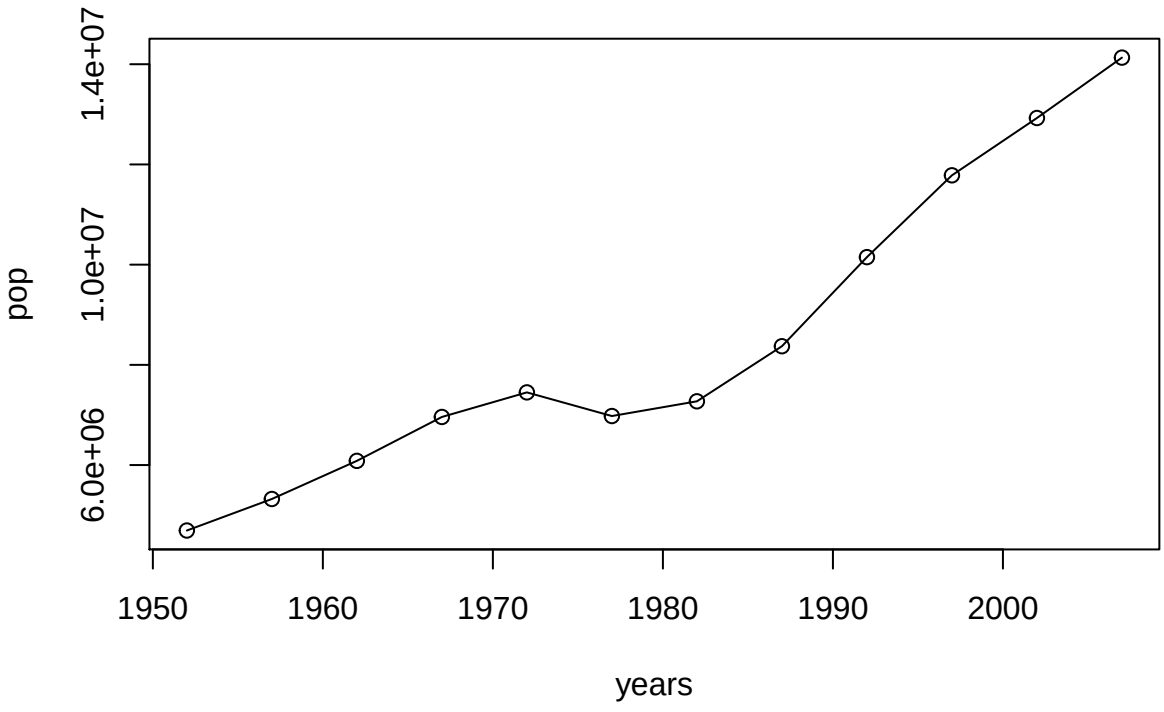



```
#Which countries in the dataset have had periods of negative population growth? Illustrate your answer
Asian_countries <- data %>% filter(continent == "Asia")

#For Afghanistan:
Afg <- Asian_countries[1:12,]
plot(y=Afg$pop,x=Afg$year,type = "o",xlab = "years" ,ylab = "pop", main = "Total population in Afghanis
```



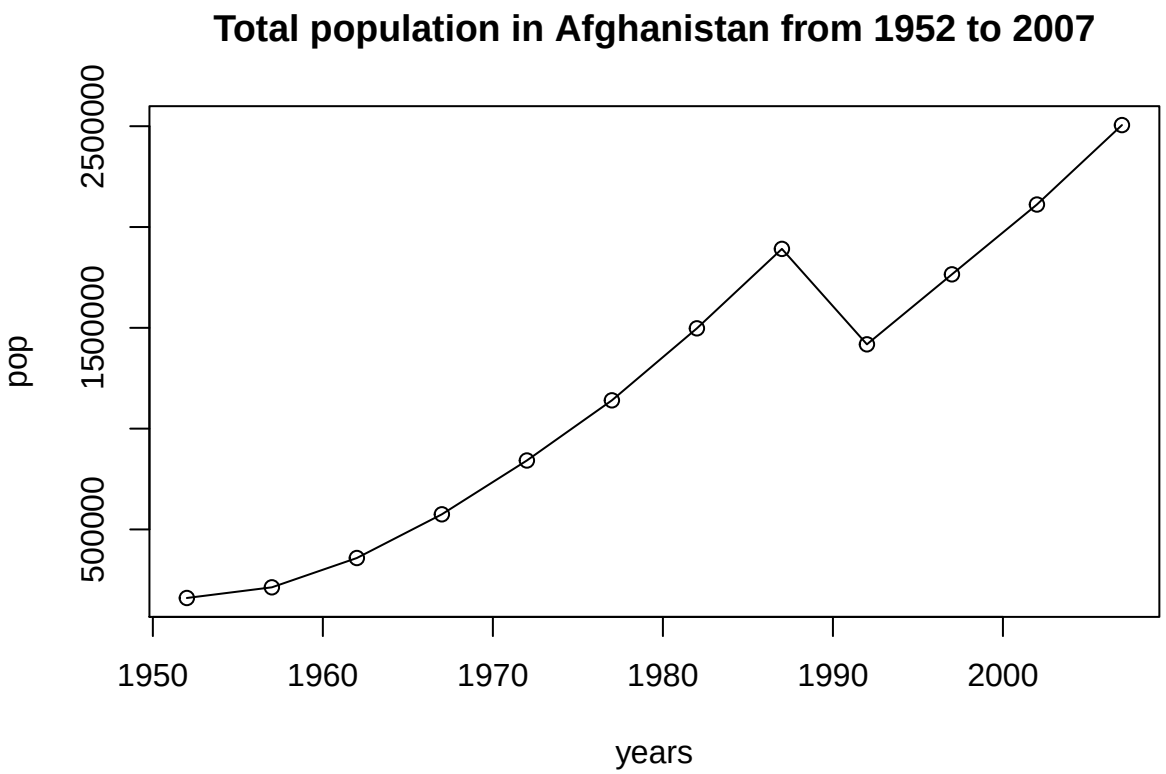
There was a decrease from 1977 to 1982 in Afghanistan.

```
#For Cambodia:
Cam <- Asian_countries[37:48,]
plot(y=Cam$pop,x=Cam$year,type = "o",xlab = "years" ,ylab = "pop", main = "Total population in Afghanis
```



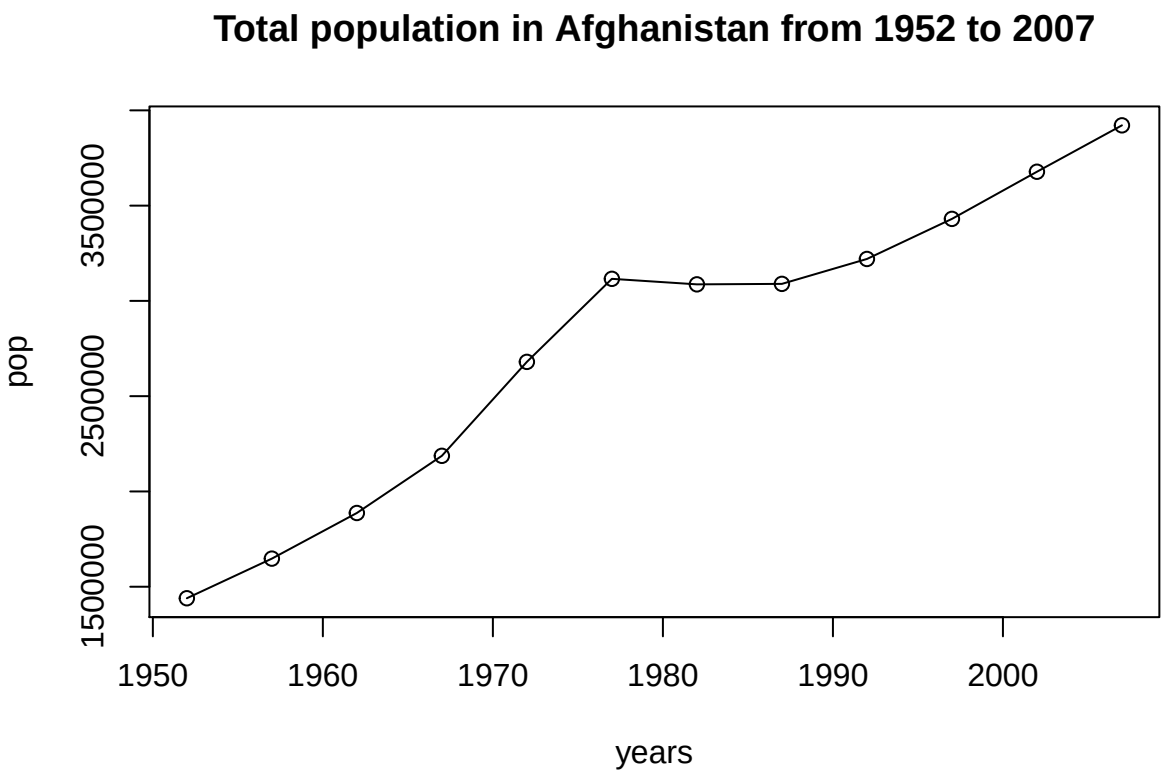
There was a decrease from 1972 to 1977 in Cambodia.

```r
#For Kuwait:
Kuw <- Asian_countries[181:192,]
plot(y=Kuw$pop,x=Kuw$year,type = "o",xlab = "years" ,ylab = "pop", main = "Total population in Afghanis
```



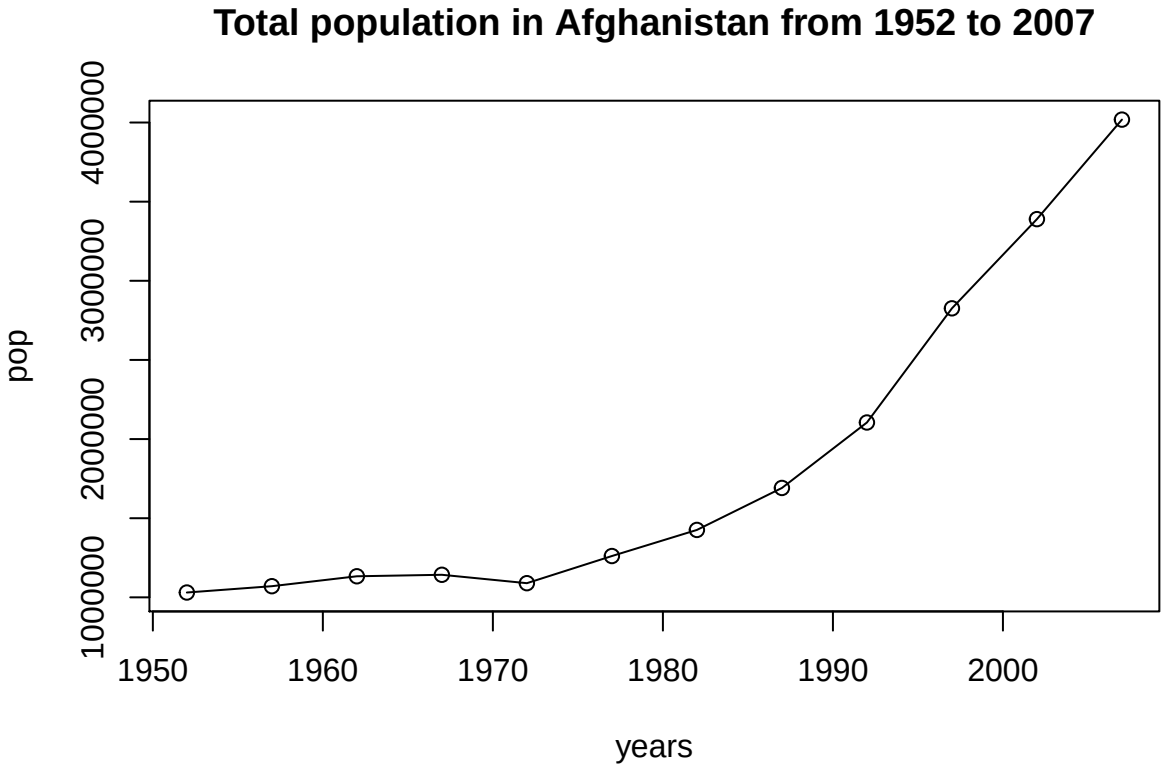
There was a decrease from 1987 to 1992 in Kuwait.

```r
#For Lebanon:
Leb <- Asian_countries[193:204,]
plot(y=Leb$pop,x=Leb$year,type = "o",xlab = "years" ,ylab = "pop", main = "Total population in Afghanis
```



There was a decrease from 1977 to 1987 in Lebanon.

```
#For West Bank and Gaza:
WBG <- Asian_countries[373:384,]
plot(y=WBG$pop,x=WBG$year,type = "o",xlab = "years" ,ylab = "pop", main = "Total population in Afghanis
```



There was a decrease from 1967 to 1972 in Wesr Bank and Gaza.

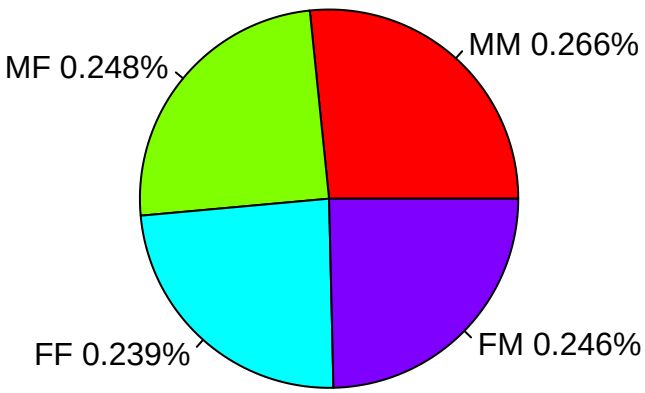
```
#Which countries in the dataset have had the highest rate of growth in per capita GDP? Illustrate your
Highest_rate = Summary_1952 %>% mutate(rate = (Summary_2007$gdpPercap-Summary_1952$gdpPercap)/Summary_19
Highest_rate = Highest_rate %>% filter(rate == max(rate))
Highest_rate
```

```
## # A tibble: 1 x 7
##   country           continent  year lifeExp    pop gdpPercap  rate
##   <fct>             <fct>     <int>   <dbl>  <int>     <dbl> <dbl>
## 1 Equatorial Guinea Africa     1952    34.5 216964      376.  31.4
```

The highest rate of growth country in per capita GDP is Equatorial Guinea with 375.6431%.

Problem 2

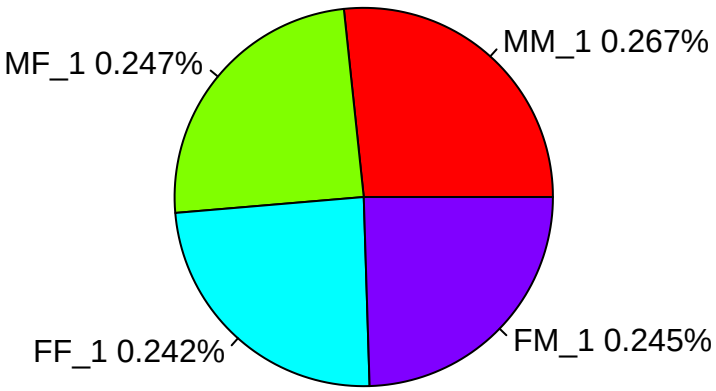
```
#Product a plot the contracts the frequency of these four combinations.
data("Fertility")
MM <- Fertility %>% filter(gender1=="male"& gender2=="male")
MF <- Fertility %>% filter(gender1=="male" & gender2=="female")
FF <- Fertility %>% filter(gender1=="female" & gender2=="female")
FM <- Fertility %>% filter(gender1=="female" & gender2=="male")

slices <- c(67799, 63185,60946,62724)
lbls <- c("MM","MF","FF","FM")
pct <- round(slices/sum(slices),3)
lbls <- paste(lbls,pct)
lbls <- paste(lbls,"%",sep = "")
pie(slices,labels = lbls,col = rainbow(length(lbls)),
    main = "Frequency of these four combinations")
```




```
# Are the frequencies different for women in their 20s and wemen who are older than 29?

Fertility_1 <- Fertility %>% filter(age<30)
Fertility_2 <- Fertility %>% filter(age>29)
MM_1 <- Fertility_1 %>% filter(gender1=="male"& gender2=="male")
MF_1 <- Fertility_1 %>% filter(gender1=="male" & gender2=="female")
FF_1 <- Fertility_1 %>% filter(gender1=="female" & gender2=="female")
FM_1 <- Fertility_1 %>% filter(gender1=="female" & gender2=="male")
```

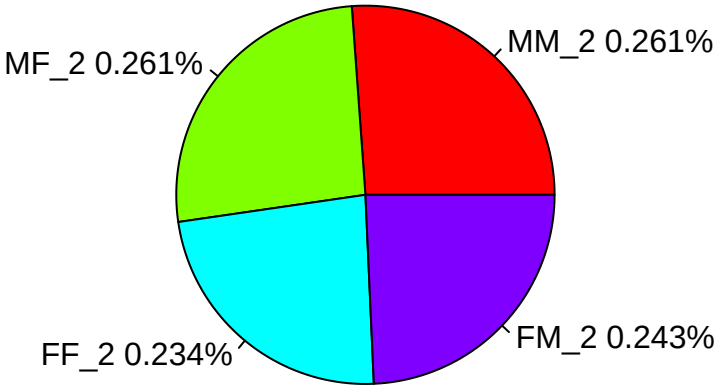
```
MM_2 <- Fertility_2 %>% filter(gender1=="male"& gender2=="male")
MF_2 <- Fertility_2 %>% filter(gender1=="male" & gender2=="female")
FF_2 <- Fertility_2 %>% filter(gender1=="female" & gender2=="female")
FM_2 <- Fertility_2 %>% filter(gender1=="female" & gender2=="male")

slices <- c(24505, 22653,22183,22508)
lbls <- c("MM_1","MF_1","FF_1","FM_1")
pct <- round(slices/sum(slices),3)
lbls <- paste(lbls,pct)
lbls <- paste(lbls,"%",sep = "")
pie(slices,labels = lbls,col = rainbow(length(lbls)),
    main = "Frequency of these four combinations with age under 30")
```

Frequency of these four combinations with age under 30




```
slices <- c(43294, 43294,38763,40216)
lbls <- c("MM_2","MF_2","FF_2","FM_2")
pct <- round(slices/sum(slices),3)
lbls <- paste(lbls,pct)
lbls <- paste(lbls,"%",sep = "")
pie(slices,labels = lbls,col = rainbow(length(lbls)),
    main = "Frequency of these four combinations with age over 30")
```

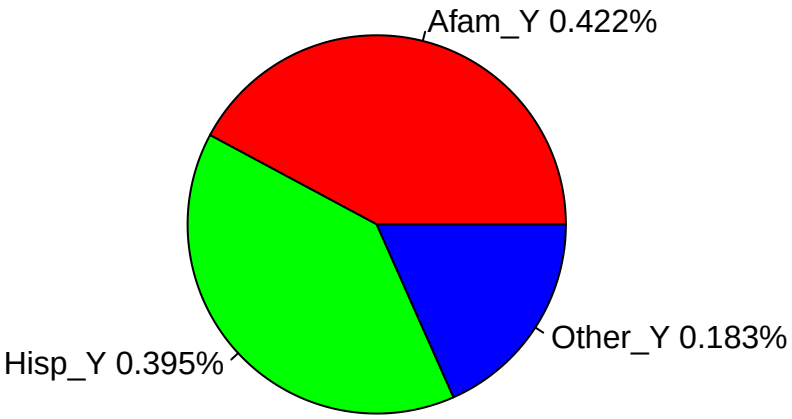
Frequency of these four combinations with age over 30



The percentage of MM has been decreased, compared under 30 to over 30. The other 3 parts increased.

```
#Produce a plot that contrasts the frequency of having more than two children by race and ethnicity.

New_Fertility <- Fertility %>% filter(morekids == "yes")

Afam_Y <- New_Fertility %>% filter(afam == "yes" & hispanic == "no" & other == "no")

Hisp_Y <- New_Fertility %>% filter(afam == "no" & hispanic == "yes" & other == "no")

Other_Y <- New_Fertility %>% filter(afam == "no" & hispanic == "no" & other == "yes")

slices <- c(5933, 5555,2581)
lbls <- c("Afam_Y","Hisp_Y","Other_Y")
pct <- round(slices/sum(slices),3)
lbls <- paste(lbls,pct)
lbls <- paste(lbls,"%",sep = "")
pie(slices,labels = lbls,col = rainbow(length(lbls)),
    main = "Percentage of race and ethnicity")
```

Percentage of race and ethnicity



Problem 3

```
#Use the mtcars and mpg datasets.
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
#How many times does the letter "e" occur in mtcars rownames?
row_n <- row.names(mtcars)
letter = sapply(letters, function(x) x<-sum(x==unlist(strsplit(row_n,""))))
letter
```

```
##  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  v  w  x  y
## 32  2 12  8 25  0  3  3 17  0  0 14  2 16 28  3  0 30  6 18  5  4  1  0  3
##  z
##  2
```

There are 25 letter es which occur in mtcars rownames.

```
#How many cars in mtcars have the brand Merc?
row_n
```

```
##  [1] "Mazda RX4"           "Mazda RX4 Wag"       "Datsun 710"
##  [4] "Hornet 4 Drive"      "Hornet Sportabout"   "Valiant"
##  [7] "Duster 360"          "Merc 240D"           "Merc 230"
## [10] "Merc 280"            "Merc 280C"           "Merc 450SE"
## [13] "Merc 450SL"          "Merc 450SLC"         "Cadillac Fleetwood"
## [16] "Lincoln Continental" "Chrysler Imperial"   "Fiat 128"
## [19] "Honda Civic"         "Toyota Corolla"      "Toyota Corona"
## [22] "Dodge Challenger"    "AMC Javelin"         "Camaro Z28"
## [25] "Pontiac Firebird"    "Fiat X1-9"           "Porsche 914-2"
## [28] "Lotus Europa"        "Ford Pantera L"      "Ferrari Dino"
```

Table 4: Mileage data for Merc cars in mtcars

manufacturer	mpg
Merc 240D	24.4
Merc 230	22.8
Merc 280	19.2
Merc 280C	17.8
Merc 450SE	16.4
Merc 450SL	17.3
Merc 450SLC	15.2

```
## [31] "Maserati Bora"       "Volvo 142E"
```

There are 7 cars which have the brand Merc.

```
#How many cars in mpg have the brand("manufacturer" in mpg) Merc?
Merc = mpg %>% count(manufacturer)
Merc
```

```
## # A tibble: 15 x 2
##    manufacturer     n
##    <chr>        <int>
##  1 audi            18
##  2 chevrolet       19
##  3 dodge           37
##  4 ford            25
##  5 honda            9
##  6 hyundai         14
##  7 jeep             8
##  8 land rover       4
##  9 lincoln          3
## 10 mercury          4
## 11 nissan          13
## 12 pontiac          5
## 13 subaru          14
## 14 toyota          34
## 15 volkswagen      27
```

There are 4 mars in mpg that have the brand Merc.

```
#Contrast the mileage data for Merc cars as reported in mtcars and mpg. Use tables, plots, and a short
MPG_1 = mpg %>% filter(manufacturer == "mercury")
MTCARS_1 = mtcars[8:14,]
NAME_mtcars = row.names(MTCARS_1)
tbl_mtcars = cbind(NAME_mtcars, MTCARS_1$mpg)
tbl_mpg = cbind(MPG_1$manufacturer, MPG_1$cty, MPG_1$hwy)

kable(tbl_mtcars, digits = 2, align = "c", format = "latex", booktabs=TRUE, ,caption = "Mileage data fo
```

```
kable(tbl_mpg, digits = 2, align = "c", format = "latex", booktabs=TRUE, ,caption = "Mileage data for M
```

Problem 4

```
#Draw a sample of 500,000 rows from the babynames data.
library(babynames)
```

Table 5: Mileage data for Merc cars in mpg

manufacturer	cty	hwy
mercury	14	17
mercury	13	19
mercury	13	19
mercury	13	17

```
data = babynames
sub_set <- sample(1:1924655,500000,replace = F)
sub_set <- babynames[sub_set,]
sub_set
```

```
## # A tibble: 500,000 x 5
##     year sex   name          n       prop
##    <dbl> <chr> <chr>     <int>      <dbl>
##  1  1951 M     Shelly       40 0.0000209
##  2  1946 F     Donalee       6 0.00000372
##  3  2014 M     Holsten       5 0.00000245
##  4  2006 M     Kempton      12 0.00000548
##  5  1914 F     Cathleen     28 0.0000352
##  6  1973 F     Sakina       17 0.0000109
##  7  2012 F     Sarabella     7 0.00000362
##  8  1943 F     Ernesteen     6 0.00000418
##  9  1960 F     Martha     5831 0.00280
## 10  1962 F     Karin       825 0.000407
## # ... with 499,990 more rows
```

```
#Produce a tabble that displays the five most popular boy names and girl names in the years 1880,1920,
M_name <- sub_set %>% filter(sex=="M")
F_name <- sub_set %>% filter(sex=="F")

M_name_1880 <- M_name %>% filter(year==1880)
F_name_1880 <- F_name %>% filter(year==1880)
M_name_1920 <- M_name %>% filter(year==1920)
F_name_1920 <- F_name %>% filter(year==1920)
M_name_1960 <- M_name %>% filter(year==1960)
F_name_1960 <- F_name %>% filter(year==1960)
M_name_2000 <- M_name %>% filter(year==2000)
F_name_2000 <- F_name %>% filter(year==2000)


M_name_1880 <- M_name_1880[with(M_name_1880,order(n)),]
F_name_1880 <- F_name_1880[with(F_name_1880,order(n)),]
M_name_1920 <- M_name_1920[with(M_name_1920,order(n)),]
F_name_1920 <- F_name_1920[with(F_name_1920,order(n)),]
M_name_1960 <- M_name_1960[with(M_name_1960,order(n)),]
F_name_1960 <- F_name_1960[with(F_name_1960,order(n)),]
M_name_2000 <- M_name_2000[with(M_name_2000,order(n)),]
F_name_2000 <- F_name_2000[with(F_name_2000,order(n)),]

M_name_1880 <- tail(M_name_1880,n=5)
```

Table 6: The five most popular boy names and girl names in the years 1880,1920, 1960, 2000

1880		1920		1960		2000	
Male	Female	Male	Female	Male	Female	Male	Female
Arthur	Florence	Louis	Marie	Larry	Laura	Jonathan	Destiny
Walter	Clara	Carl	Evelyn	Edward	Elizabeth	David	Grace
Harry	Alice	George	Elizabeth	Gary	Brenda	Ryan	Brianna
George	Minnie	James	Mildred	Michael	Lisa	Brandon	Alyssa
Charles	Mary	Robert	Mary	David	Susan	William	Ashley

```
F_name_1880 <- tail(F_name_1880,n=5)
M_name_1920 <- tail(M_name_1920,n=5)
F_name_1920 <- tail(F_name_1920,n=5)
M_name_1960 <- tail(M_name_1960,n=5)
F_name_1960 <- tail(F_name_1960,n=5)
M_name_2000 <- tail(M_name_2000,n=5)
F_name_2000 <- tail(F_name_2000,n=5)

M_1880 <- M_name_1880[3]
F_1880 <- F_name_1880[3]
M_1920 <- M_name_1920[3]
F_1920 <- F_name_1920[3]
M_1960 <- M_name_1960[3]
F_1960 <- F_name_1960[3]
M_2000 <- M_name_2000[3]
F_2000 <- F_name_2000[3]


tbl = cbind(M_1880,F_1880,
            M_1920,F_1920,
            M_1960,F_1960,
            M_2000,F_2000)
colnames(tbl) <-   c('Male', 'Female',
                     "Male", "Female",
                     'Male', 'Female',
                     "Male", "Female")
kable(tbl, digits = 2,align = "c", format = "latex", booktabs=TRUE, ,caption = "The five most popular b
  add_header_above(c("1880"=2,
                     "1920"=2,
                     "1960"=2,
                    "2000"=2))
```

```
#What names overlap boys and girls?
names = sub_set %>% group_by(name) %>% summarise(lap = length(sex)) %>% filter(lap>1)
names
```

```
## # A tibble: 51,510 x 2
##    name       lap
##    <chr>    <int>
##  1 Aabha        2
##  2 Aadam        8
##  3 Aadan        3
```

```
##  4 Aadarsh      7
##  5 Aaden        5
##  6 Aadhav       2
##  7 Aadhavan     2
##  8 Aadhya       2
##  9 Aadhyan      3
## 10 Aadi         3
## # ... with 51,500 more rows
```

There are 51367 names that are overlapped.

```
#What names were used in the 19th century but have not been used in the 21sth century?
name_19 <- sub_set %>% filter(year<1900)
name_21 <- sub_set %>% filter(year>1999)
name_19 <- name_19 %>% count(name)
name_21 <- name_21 %>% count(name)
name_dif <- subset(name_19, !(name %in% name_21))
```

There were 3612 names used in the 19th century but not in 21th.

```
#Produce a chart that shows the relative frequency of the names "Donald", "Hilary", "Hillary", "Joe","B

Frm_1880_2017 = sub_set %>% filter(year >1879 & year <2018)
Name_1880_2017 = Frm_1880_2017 %>%filter(name == c("Donald", "Hilary", "Hillary", "Joe", "Barrack"))
y = Name_1880_2017 %>% group_by(name) %>% summarise(n = sum(n))
data =  y %>% mutate(frequency = c(84238/sum(n),847/sum(n),2681/sum(n),21876/sum(n)))
Graph = ggplot(data, aes(x = name,y = frequency)) +
geom_bar(stat = "identity")
print(Graph + ggtitle("Frequency of the names -- Donald, Hilary, Hillary, Joe, Barrack"))
```