Tidyverse Problem Set

Chuning Yuan

September 29, 2019

The purpose of this problem set is to provide data contexts in which to exercise the capabilitiues of the tidyverse. While some questons require specific answers, other parts of the problems have been written to be purposely ambiguous, requiring you to think through the presentation details of your answer.

HOLD THE PRESSES!
As I was preparing to post these problems yesterday, I noticed that tidyr had been updata in the last few weeks. I was looking for more exercises on gather() and spread() – which are always difficult to master. And I found that they have been superceded!! Why do I love working with R as the tidyversie is on a path of continuous improvement? Because the improvements come from developers who write things like this:

For some time, it's been obvious that there is something fundamentally wrong with the design of spread() and gather(). Many people don't find the names intuitive and find it hard to remember which direction corresponds to spreading and which to gathering. It also seems surprisingly hard to remember the arguments to these functions, meaning that many people (including me!) have to consult the documentation every time. Hadley Wickham, Pivot Vingette

So... before you do anymore tidyverse exercises, Read this tidyr 1.0.0.

Then go to the tidyr cran page and to the examples and exercies in the new vignettes.

In your solutions to the problems below, if you need to use table reshaping functions from TidyR, be sure that you use pivot_longer(), and pivot_wider().

Problem 1

Load the gapminder data from the gapminder package.

How many continents are included in the data set?

```
data <- gapminder
number_cont <- data$continent %>% unique %>% length
number_cont
```

```
## [1] 5
```

How many countrys are included? How many countries per continent?

```
number_coun <- data$country %>% unique %>% length
number_coun
```

```
## [1] 142
```

```
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
```

Table 1: Total population and total GDP per continents

continent	Population	GDP
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

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

Using the gapminder data, produce a report showing the continents in the dataset, total population per continent, and GDP per capita. Be sure that the table is properly labeled and suitable for inclusion in a printed report.

```
Per <- data %>% group_by(continent) %>% summarise(Population = sum(pop)/1000000, GDP = sum(gdpPercap)/1


kable(cbind(Per), caption = "Total population and total GDP per continents", align = "c", booktab =T, l
```

Produce a well-labeled table that summarizes GDP per capita for the countries in each continent, contrasting the years 1952 and 2007.

```
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"

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

Product a plot that summarizes the same data as the table. There should be two plots per continent.

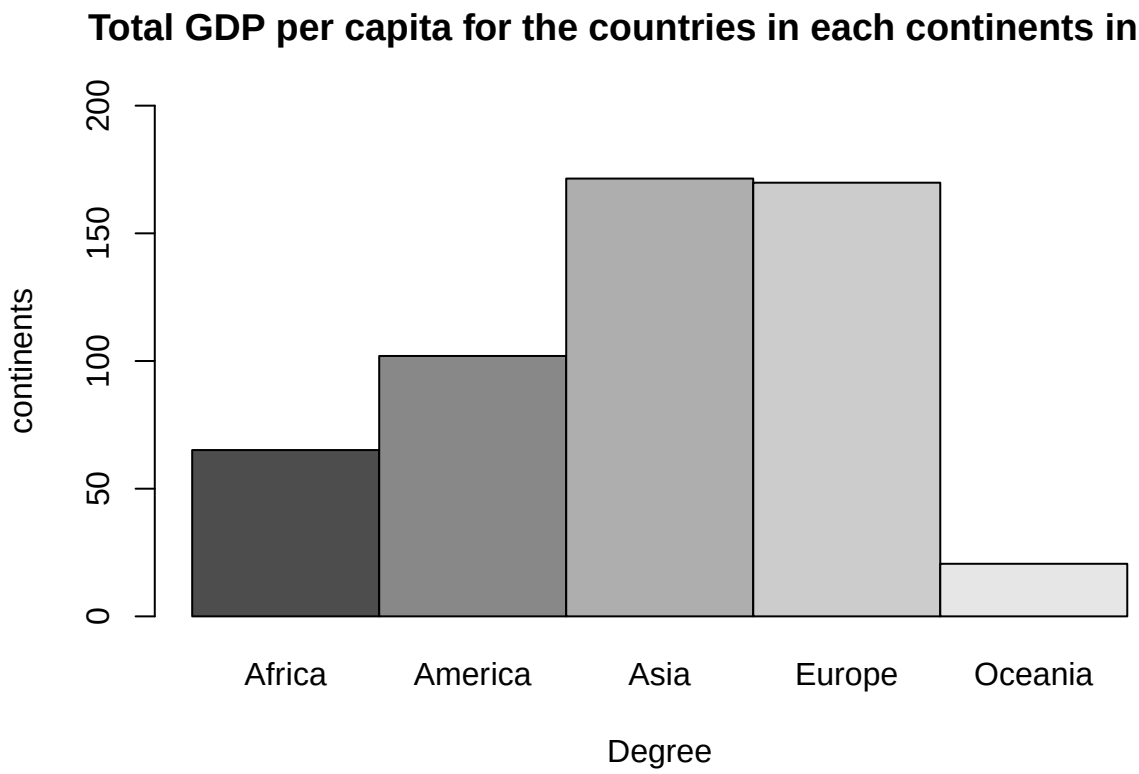
```
Total_1952 <- data %>% filter(year==1952)
Total_1952 <- Total_1952 %>% group_by(continent) %>% summarise(Total_GDP_thousand = sum(gdpPercap)/1000)
Total_1952
```

```
## # 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
```

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

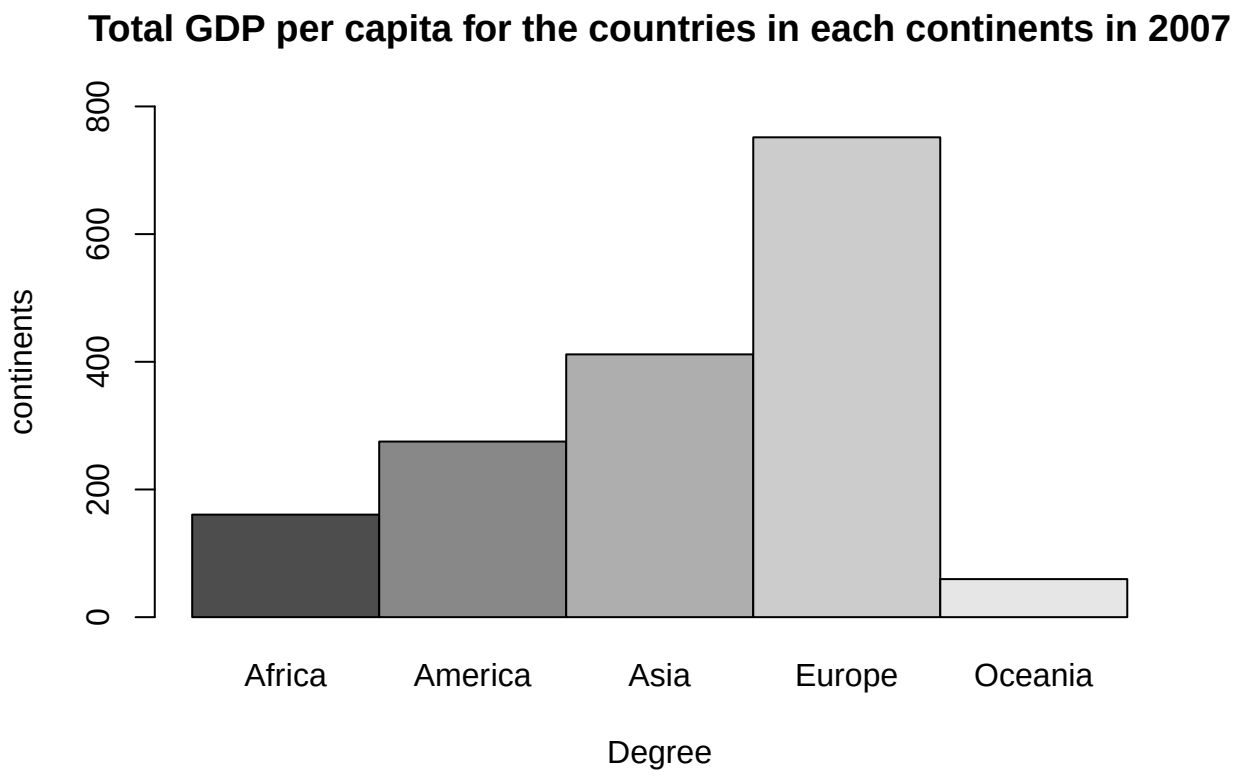
```
barplot(as.matrix(Total_1952[,2]),beside = T,legend.text = T,main = "Total GDP per capita for the counti
```




```
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
```

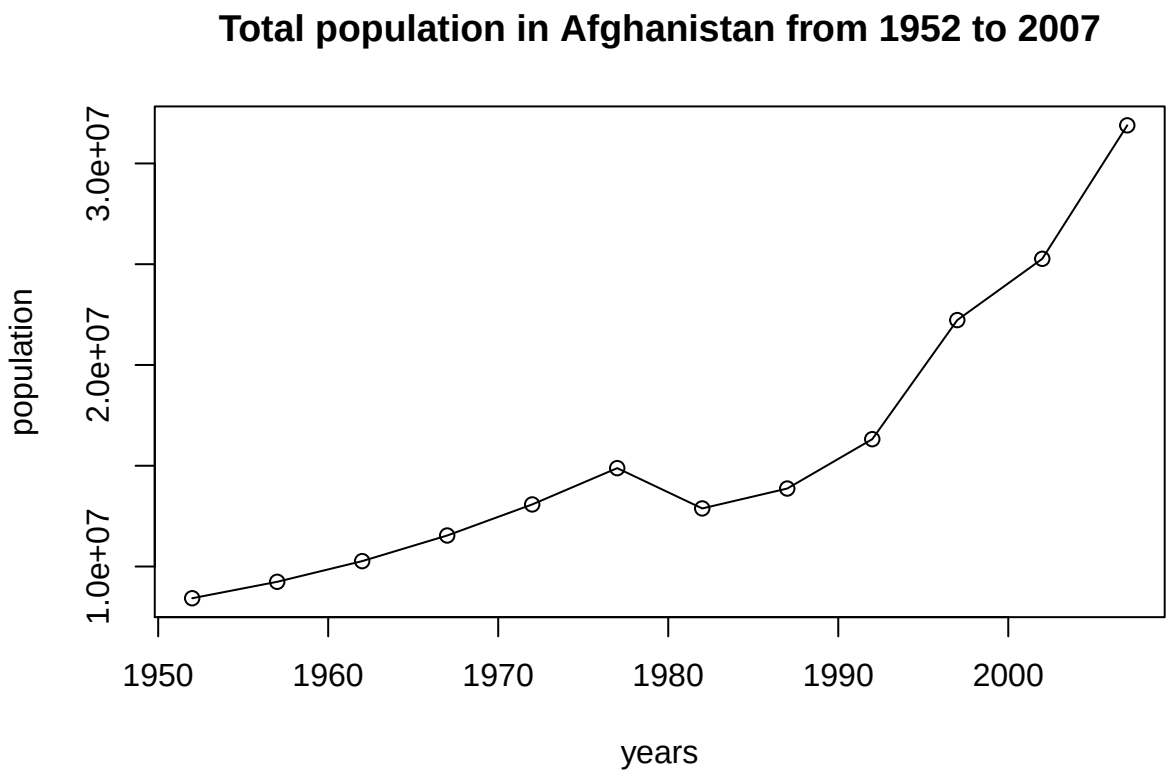
```
barplot(as.matrix(Total_2007[,2]),beside = T,legend.text = T,main = "Total GDP per capita for the counti
```



Which countries in the dataset have had periods of negative population growth?

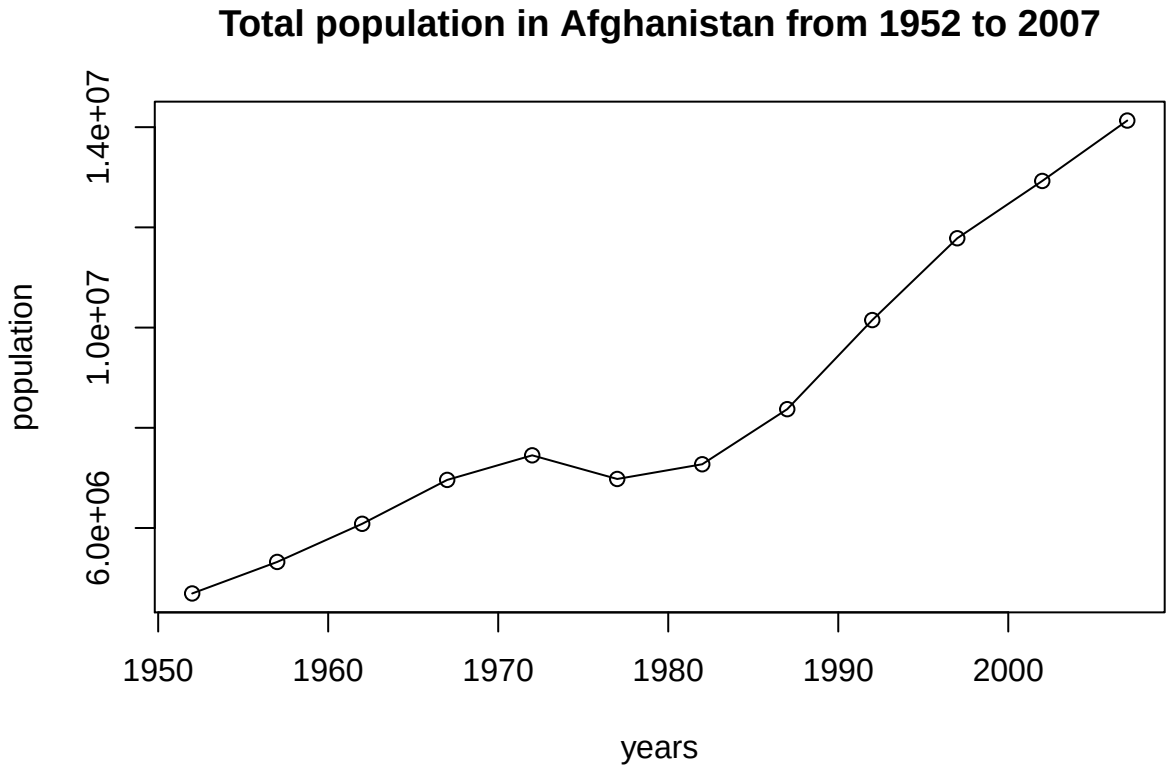
```
Asian_countries <- data %>% filter(continent == "Asia")

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



There was a decrease from 1977 to 1982 in Afghanistan.

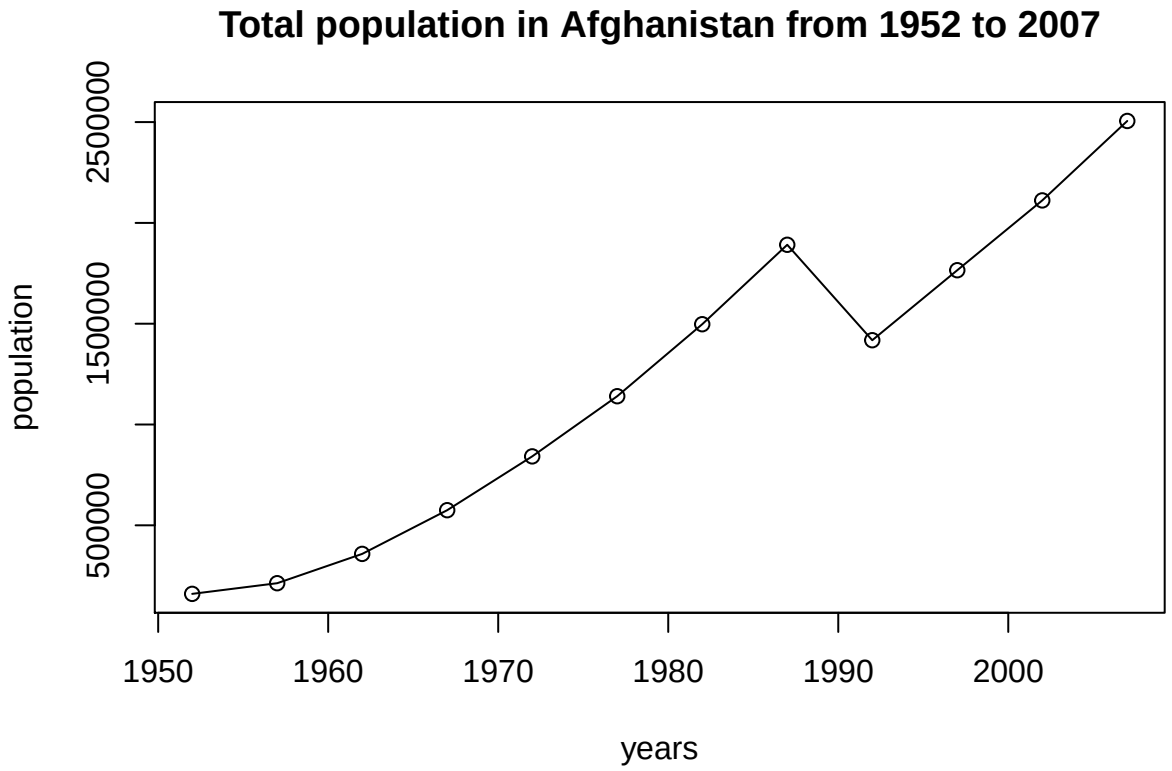
```
#For Cambodia:
Cam <- Asian_countries[37:48,]
plot(y=Cam$pop,x=Cam$year,type = "o",xlab = "years" ,ylab = "population", main = "Total population in A
```



There

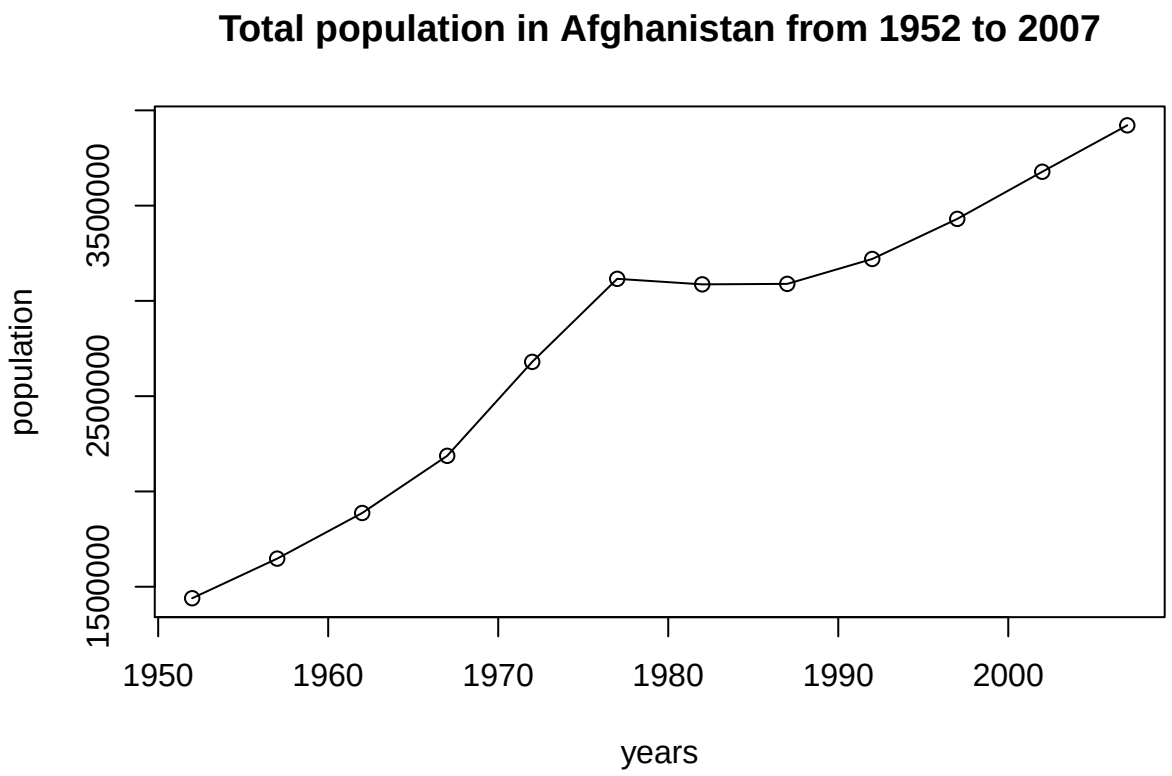
was a decrease from 1972 to 1977 in Cambodia.

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



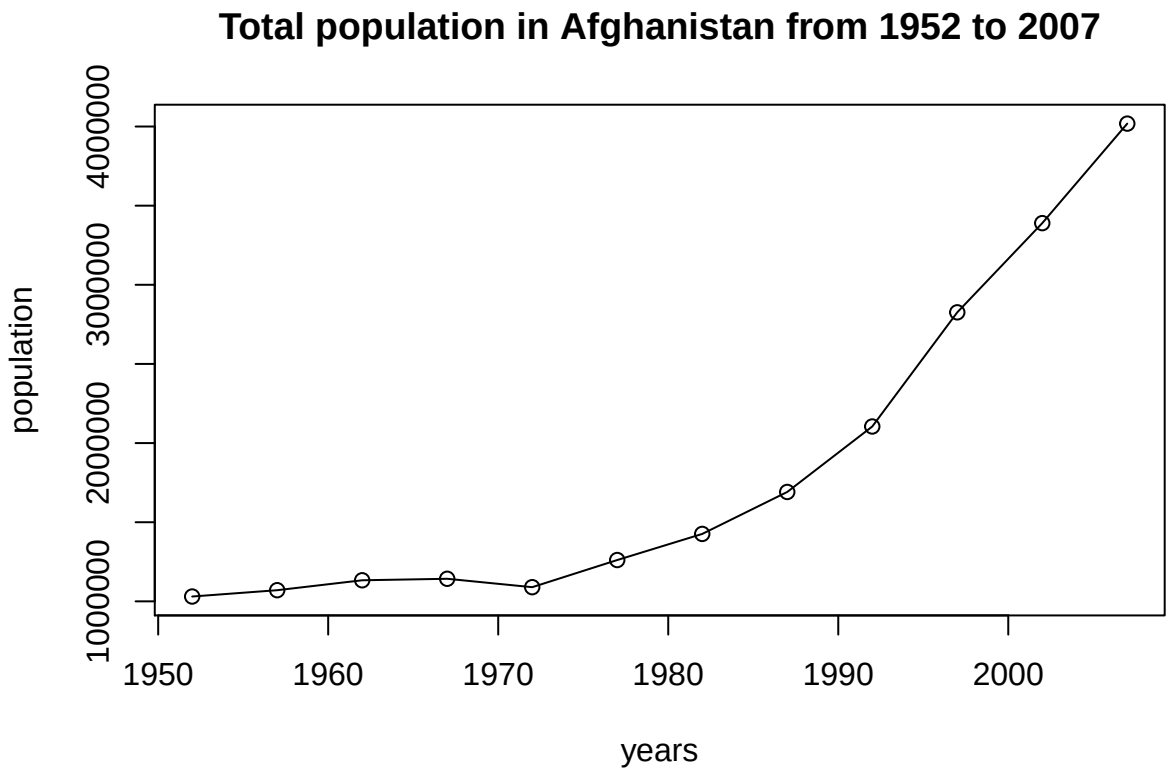
There was a decrease from 1987 to 1992 in Kuwait.

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



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 = "population", main = "Total population in A
```



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?

```
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%. Illustrate your answer with a table or plot.

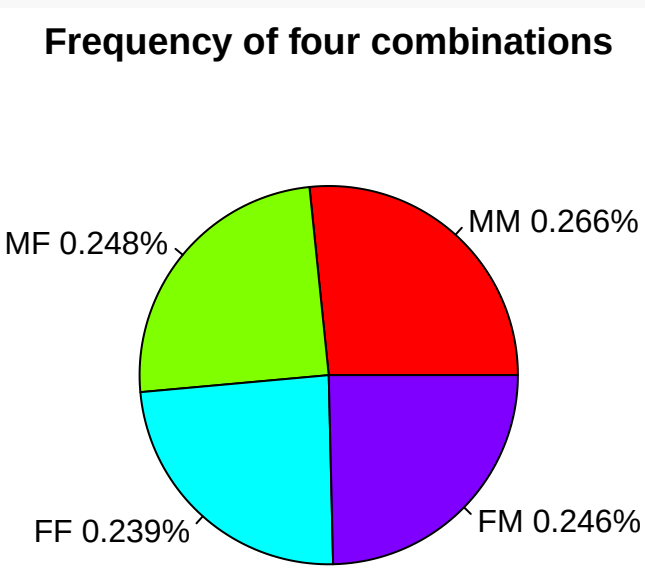
Problem 2

The data for Problem 2 is the Fertility data in the AER package. This data is from the 1980 US Census and is comprised of date on married women aged 21-35 with two or more children. The data report the gender of each woman's first and second child, the woman's race, age, number of weeks worked in 1979, and whether the woman had more than two children.

There are four possible gender combinations for the first two Children. 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 four combinations")
```



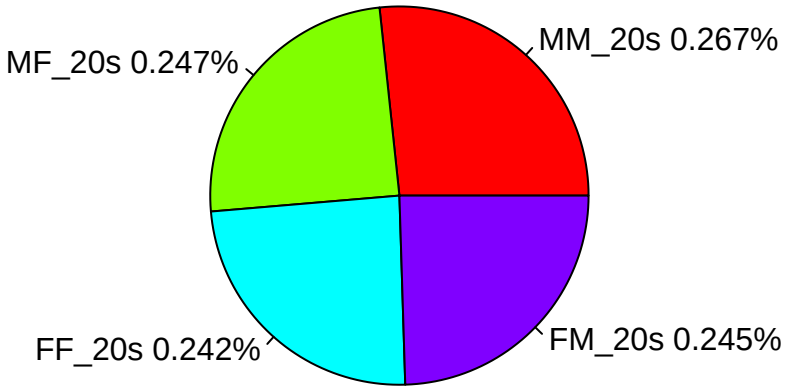
Are the frequencies different for women in their 20s and wemen who are older than 29?

```
Fertility_1 <- Fertility %>% filter(age<30)
Fertility_2 <- Fertility %>% filter(age>29)
MM_20s <- Fertility_1 %>% filter(gender1=="male"& gender2=="male")
MF_20s <- Fertility_1 %>% filter(gender1=="male" & gender2=="female")
FF_20s <- Fertility_1 %>% filter(gender1=="female" & gender2=="female")
FM_20s <- Fertility_1 %>% filter(gender1=="female" & gender2=="male")

MM_30 <- Fertility_2 %>% filter(gender1=="male"& gender2=="male")
MF_30 <- Fertility_2 %>% filter(gender1=="male" & gender2=="female")
FF_30 <- Fertility_2 %>% filter(gender1=="female" & gender2=="female")
FM_30 <- Fertility_2 %>% filter(gender1=="female" & gender2=="male")
```

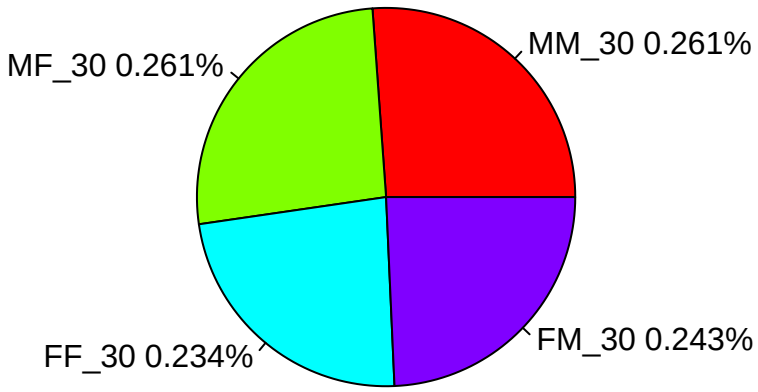
```
slices <- c(24505, 22653,22183,22508)
lbls <- c("MM_20s","MF_20s","FF_20s","FM_20s")
pct <- round(slices/sum(slices),3)
lbls <- paste(lbls,pct)
lbls <- paste(lbls,"%",sep = "")
pie(slices,labels = lbls,col = rainbow(length(lbls)),
    main = "Frequency of four combinations with age under 30")
```

Frequency of four combinations with age under 30




```
slices <- c(43294, 43294,38763,40216)
lbls <- c("MM_30","MF_30","FF_30","FM_30")
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 four combinations with age over 30")
```

Frequency of four combinations with age over 30



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

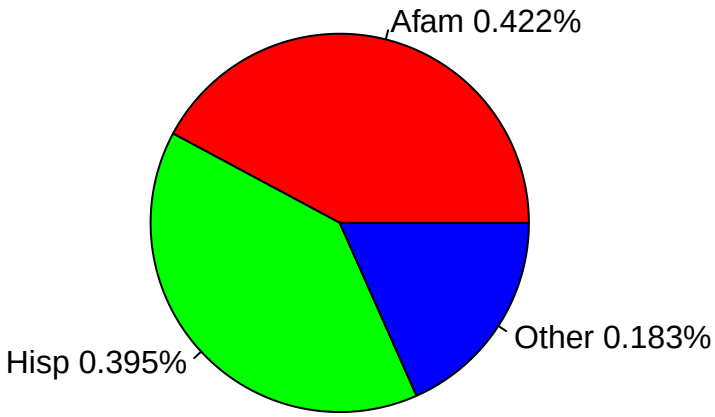
```
New_Fertility <- Fertility %>% filter(morekids == "yes")
```

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

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

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

slices <- c(5933, 5555,2581)
lbls <- c("Afam","Hisp","Other")
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")
```



Problem 3

Use the mtcars and mpg datasets.
How many times does the letter "e" occur in mtcars rownames?

```
count_e = sum(str_count(rownames((mtcars),"e")))
count_e
```

```
## [1] 381
```

There are 381 cars whose names contain "e".

How many cars in mtcars have the brand Merc?

```
Merc = sum(str_detect(rownames(mtcars),"Merc"))
Merc
```

```
## [1] 7
```

There are 7 cars in mtcars have the brand Merc.

How many cars in mpg have the brand("manufacturer" in mpg) Merc?

```
manu = sum(str_detect(mpg$manufacturer,"merc"))
manu
```

```
## [1] 4
```

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

Table 5: Mileage data for Merc cars in mpg

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

*There are 4 cars in mpg have the brand Merc.

Contrast the mileage data for Merc cars as reported in mtcars and mpg. Use tables, plots, and a short explaination.

```
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", booktabs=TRUE, ,caption = "Mileage data for Merc cars in mtc
```

```
kable(tbl_mpg, digits = 2, align = "c",  booktabs=TRUE, ,caption = "Mileage data for Merc cars in mpg",
```

Problem 4

Install the babynames package.

Draw a sample of 500,000 rows from the babynames data

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

```
## # A tibble: 500,000 x 5
##     year sex   name          n       prop
##    <dbl> <chr> <chr>     <int>      <dbl>
##  1  1912 F     Villa         9 0.0000153
##  2  1998 M     Prince      172 0.0000848
##  3  1995 M     Alphonzo     11 0.00000547
##  4  2005 M     Tamarrion     5 0.00000235
##  5  2005 F     Yarelin       6 0.00000296
##  6  1942 F     Donnarae      5 0.0000036
```

```
##  7  1958 M     Andre      1120 0.000520
##  8  2017 M     Kanai        15 0.00000764
##  9  1914 F     Cola          8 0.0000100
## 10  1935 F     Osia          5 0.0000046
## # ... with 499,990 more rows
```

Produce a table that displays the five most popular boy names and girl names in the years 1880,1920, 1960, 2000.

What names overlap boys and girls?

```
names = subset %>% group_by(name) %>% summarise(lap = length(sex)) %>% filter(lap>1)
names
```

```
## # A tibble: 51,452 x 2
##    name        lap
##    <chr>     <int>
##  1 Aaban         2
##  2 Aabriella     4
##  3 Aadam         8
##  4 Aadarsh       5
##  5 Aaden         6
##  6 Aadhav        4
##  7 Aadhavan      3
##  8 Aadhira       3
##  9 Aadhya        3
## 10 Aadi          3
## # ... with 51,442 more rows
```

There are 51367 names that are overlapped.

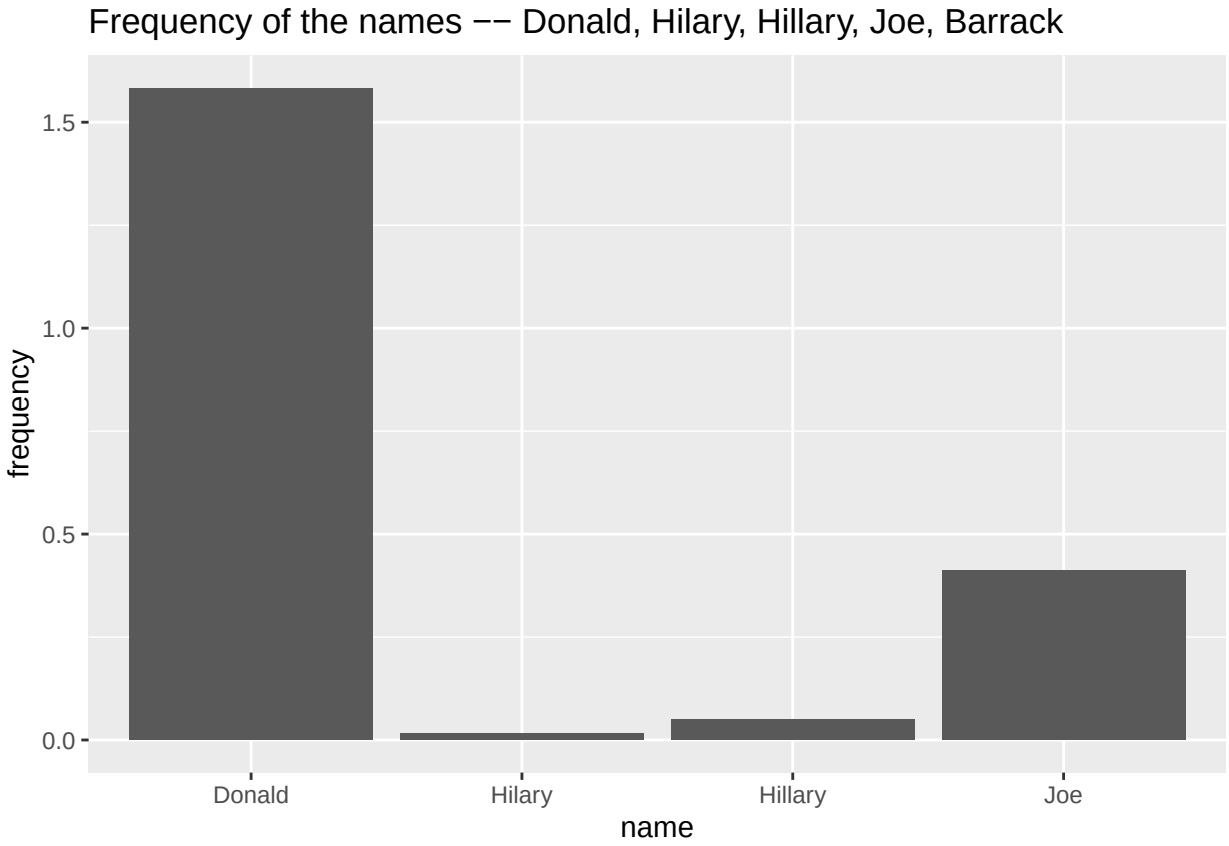
What names were used in the 19th century but have not been used in the 21sth century?

```
names_19 <- subset %>% filter(year<1900)
names_21 <- subset %>% filter(year>1999)
names_19 <- names_19 %>% count(name)
names_21 <- names_21 %>% count(name)
name_dif <- subset(names_19, !(name %in% names_21))
```

Produce a chart that shows the relative frequency of the names “Donald”, “Hilary”, “Hillary”, “Joe”, “Barrack”, over the years 1880 through 2017.

```
Frm_1880_2017 = subset %>% 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"))
```


Frequency of the names −− Donald, Hilary, Hillary, Joe, Barrack
frequency
1.5
1.0
0.5
0.0
Donald
Hilary
Hillary
Joe
name