TidyR Problem Set

Kathryn Haglich

10/4/2019

Problem 1 - Gapminder

1) How many continents are included in the data set?

Five continents are included in the data set.

2) How many countries are included in the data set?

One hundred forty two countires are included in the data set.

3) How many countries per continent?

Continent	Number of Countries
Africa	52
Americas	25
Asia	33
Europe	30
Oceania	2

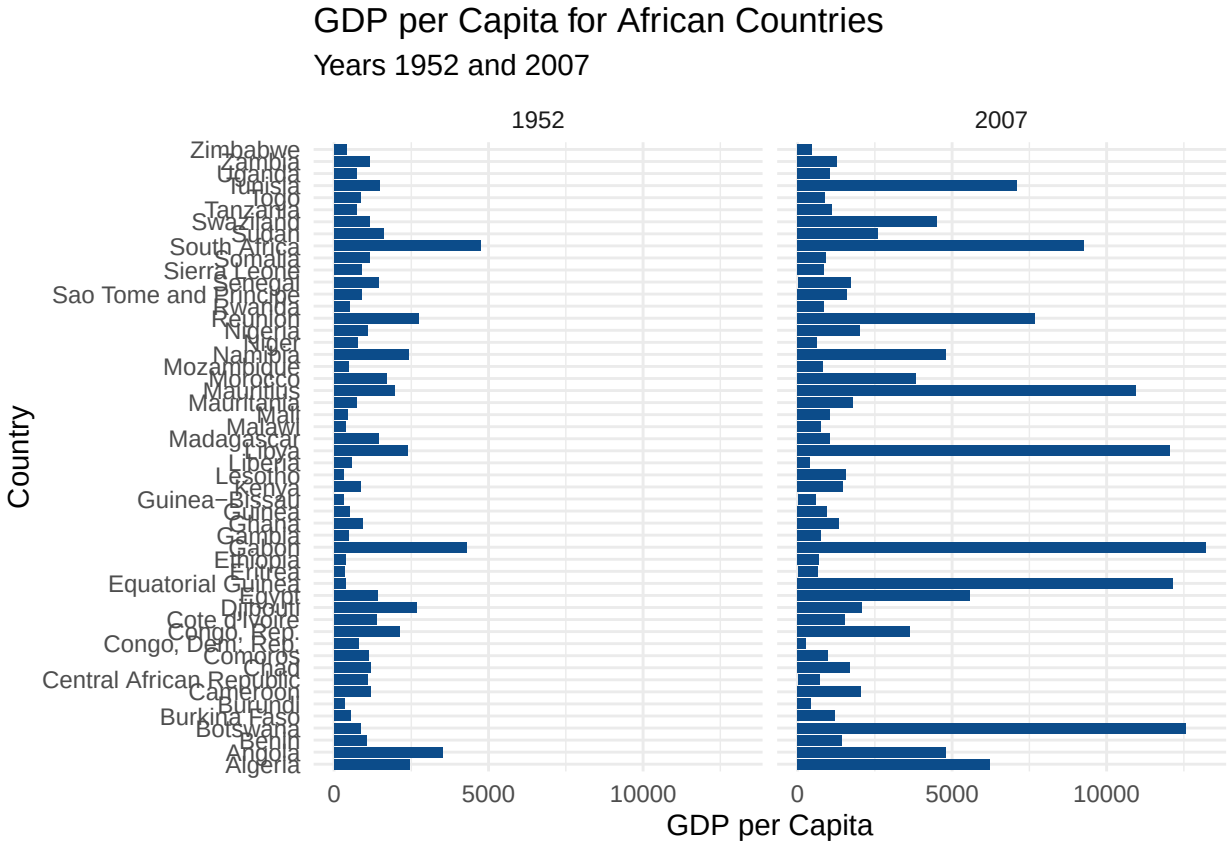
4) Produce a report showing the continents in the dataset, total population per continent, and GDP per capita. Be sure that the table is properly labeled and suitable for inclusion in a printed report.

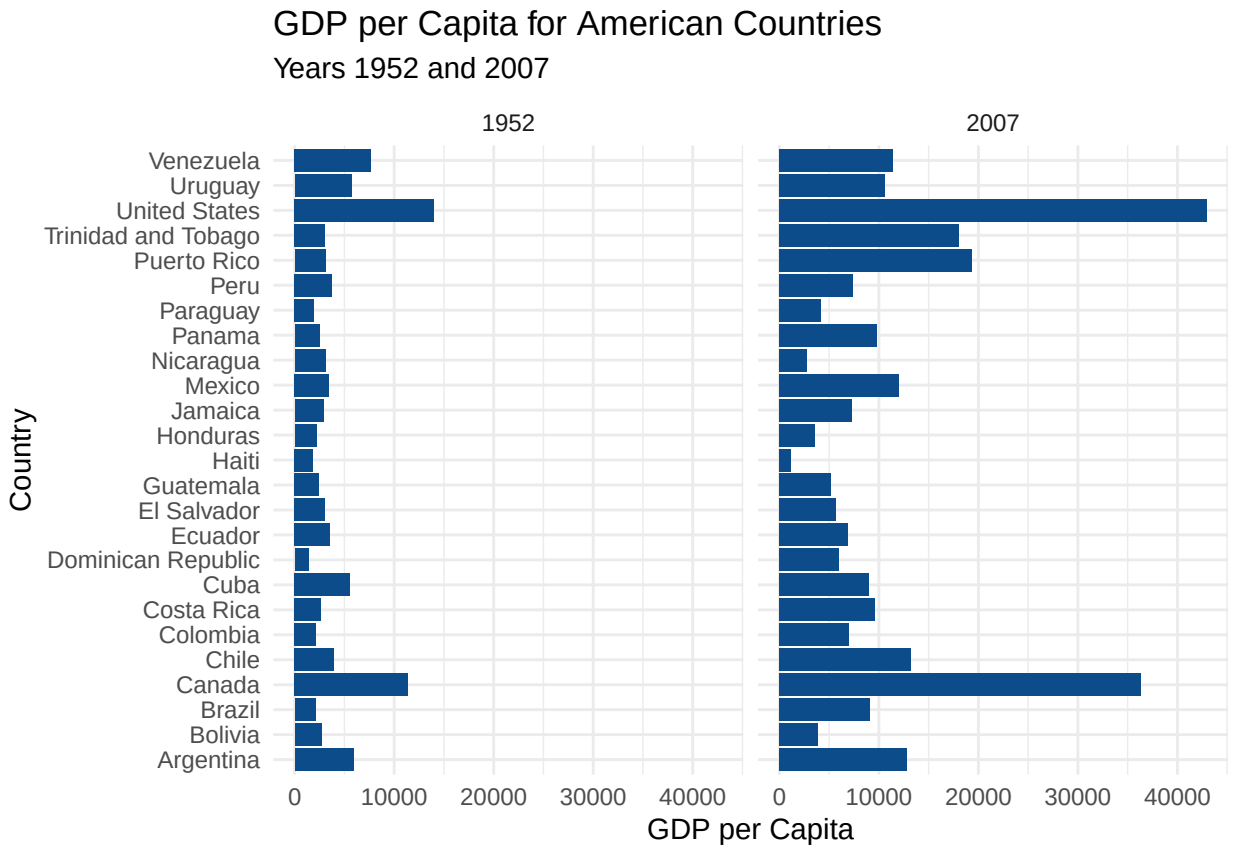
Continent	Total Population	Total GDP per Capita
1952		
Africa	237640501	65133.77
Americas	345152446	101976.56
Asia	1395357351	171450.97
Europe	418120846	169831.72
Oceania	10686006	20596.17
1957		
Africa	264837738	72032.28
Americas	386953916	115401.09
Asia	1562780599	190995.19
Europe	437890351	208890.38
Oceania	11941976	23197.04
1962		
Africa	296516865	83100.10
Americas	433270254	122538.55
Asia	1696357182	189069.20
Europe	460355155	250964.60
Oceania	13283518	25392.90
1967		
Africa	335289489	106618.92
Americas	480746623	141706.34
Asia	1905662900	197048.72
Europe	481178958	304314.71
Oceania	14600414	28990.04
1972		
Africa	379879541	121660.02
Americas	529384210	162283.35
Asia	2150972248	270186.47
Europe	500635059	374387.26
Oceania	16106100	32834.67
1977		
Africa	433061021	134468.80
Americas	578067699	183800.18
Asia	2384513556	257113.36
Europe	517164531	428519.37
Oceania	17239000	34567.92
1982		
Africa	499348587	129042.83
Americas	630290920	187668.43
Asia	2610135582	245326.46
Europe	531266901	468536.90
Oceania	18394850	37109.42
1987		
Africa	574834110	118698.79
Americas	682753971	194835.01
Asia	2871220762	251071.47
Europe	543094160	516429.32
Oceania	19574415	40896.08
1992		
Africa	659081517	118654.14
Americas	739274104	201123.36
Asia	3133292191	285109.78
Europe	558142797	511847.04
Oceania	20919651	41788.09
1997		
Africa	743832984	123695.50
Americas	796900410	222232.52

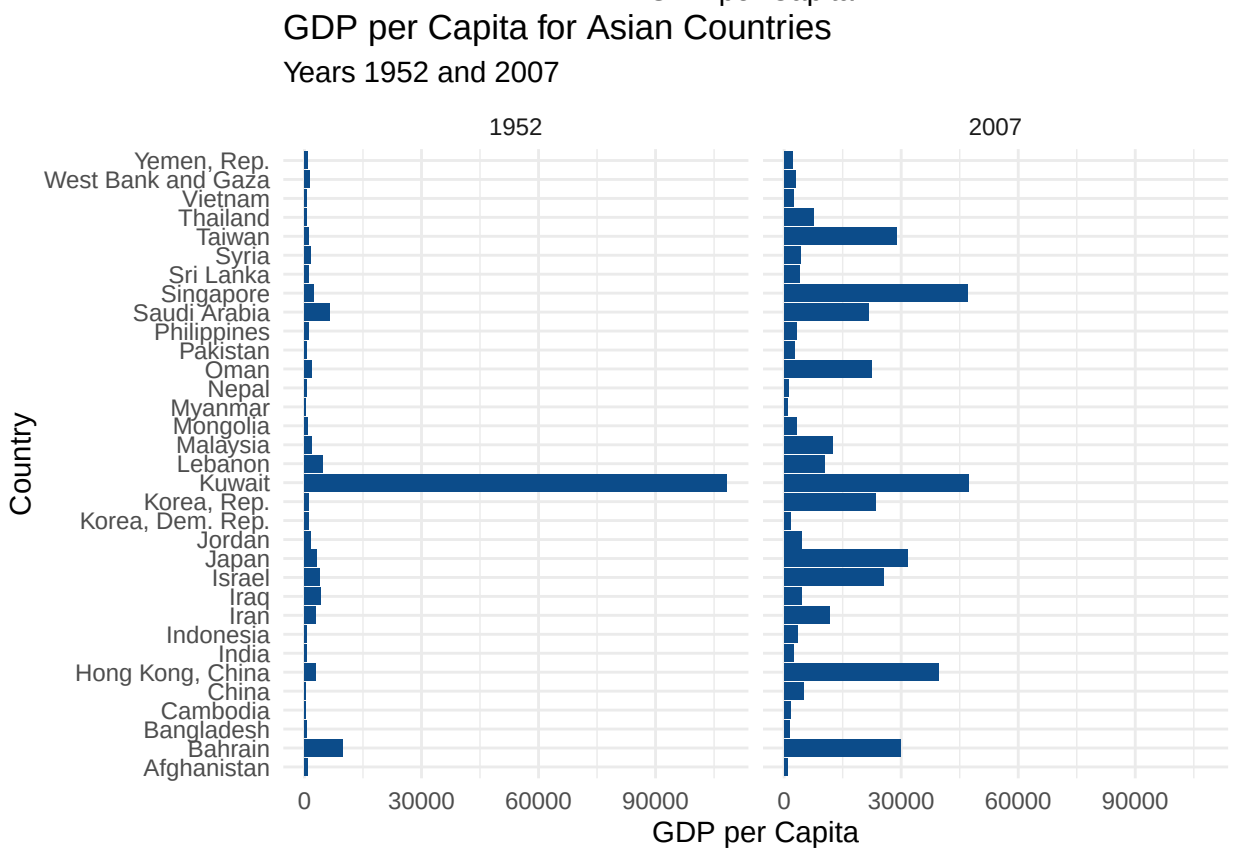
5) Produce a well-labeled table that summarizes GDP per capita for the countries in each continent, contrasting the years 1952 and 2007.

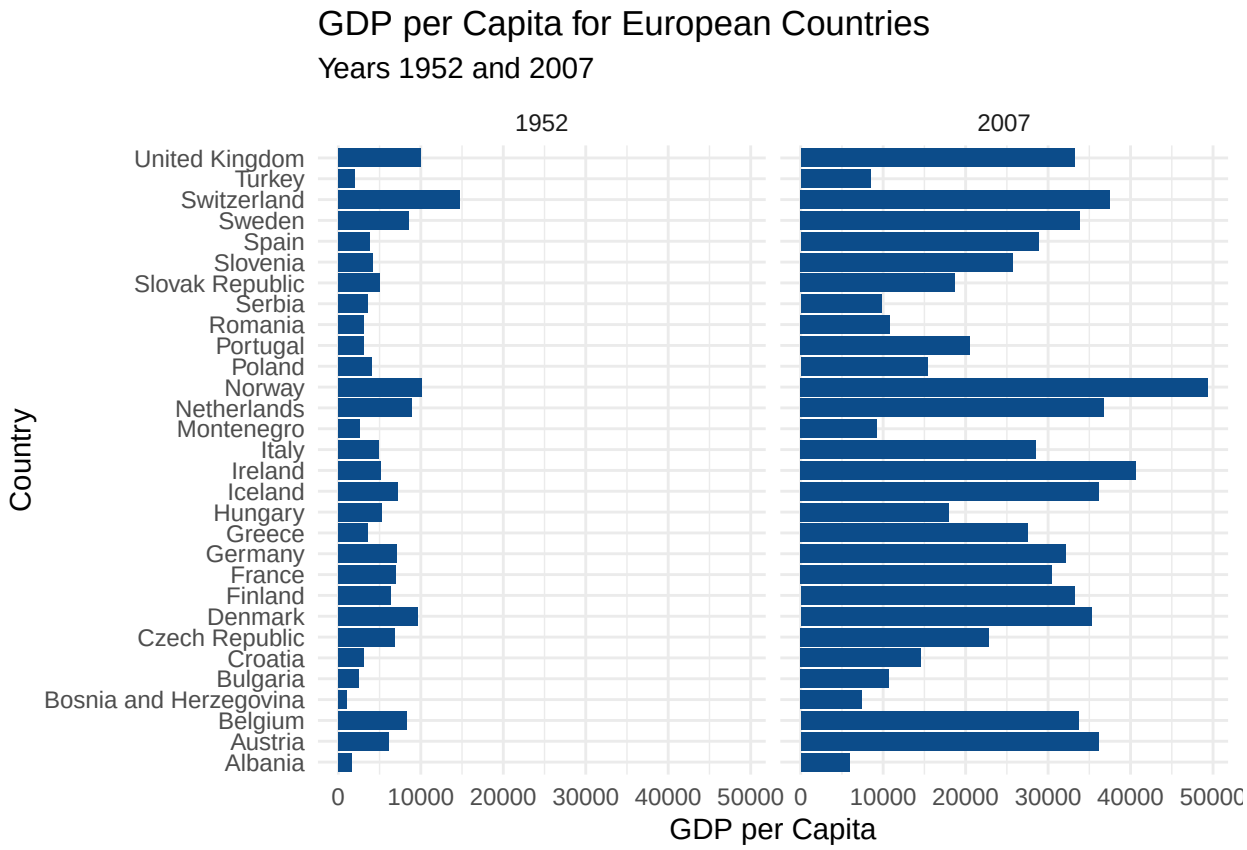
country	year	gdpPercap
Africa		
Algeria	1952	2449.0082
Algeria	2007	6223.3675
Angola	1952	3520.6103
Angola	2007	4797.2313
Benin	1952	1062.7522
Benin	2007	1441.2849
Botswana	1952	851.2411
Botswana	2007	12569.8518
Burkina Faso	1952	543.2552
Burkina Faso	2007	1217.0330
Burundi	1952	339.2965
Burundi	2007	430.0707
Cameroon	1952	1172.6677
Cameroon	2007	2042.0952
Central African Republic	1952	1071.3107
Central African Republic	2007	706.0165
Chad	1952	1178.6659
Chad	2007	1704.0637
Comoros	1952	1102.9909
Comoros	2007	986.1479
Congo, Dem. Rep.	1952	780.5423
Congo, Dem. Rep.	2007	277.5519
Congo, Rep.	1952	2125.6214
Congo, Rep.	2007	3632.5578
Cote d'Ivoire	1952	1388.5947
Cote d'Ivoire	2007	1544.7501
Djibouti	1952	2669.5295
Djibouti	2007	2082.4816
Egypt	1952	1418.8224
Egypt	2007	5581.1810
Equatorial Guinea	1952	375.6431
Equatorial Guinea	2007	12154.0897
Eritrea	1952	328.9406
Eritrea	2007	641.3695
Ethiopia	1952	362.1463
Ethiopia	2007	690.8056
Gabon	1952	4293.4765
Gabon	2007	13206.4845
Gambia	1952	485.2307
Gambia	2007	752.7497
Ghana	1952	911.2989
Ghana	2007	1327.6089
Guinea	1952	510.1965
Guinea	2007	942.6542
Guinea-Bissau	1952	299.8503
Guinea-Bissau	2007	579.2317
Kenya	1952	853.5409
Kenya	2007	1463.2493
Lesotho	1952	298.8462
Lesotho	2007	1569.3314
Liberia	1952	575.5730
Liberia	2007	414.5073
Libya	1952	2387.5481
Libya	2007	12057.4993
Madagascar	1952	1443.0117
Madagascar	2007	1044.7701

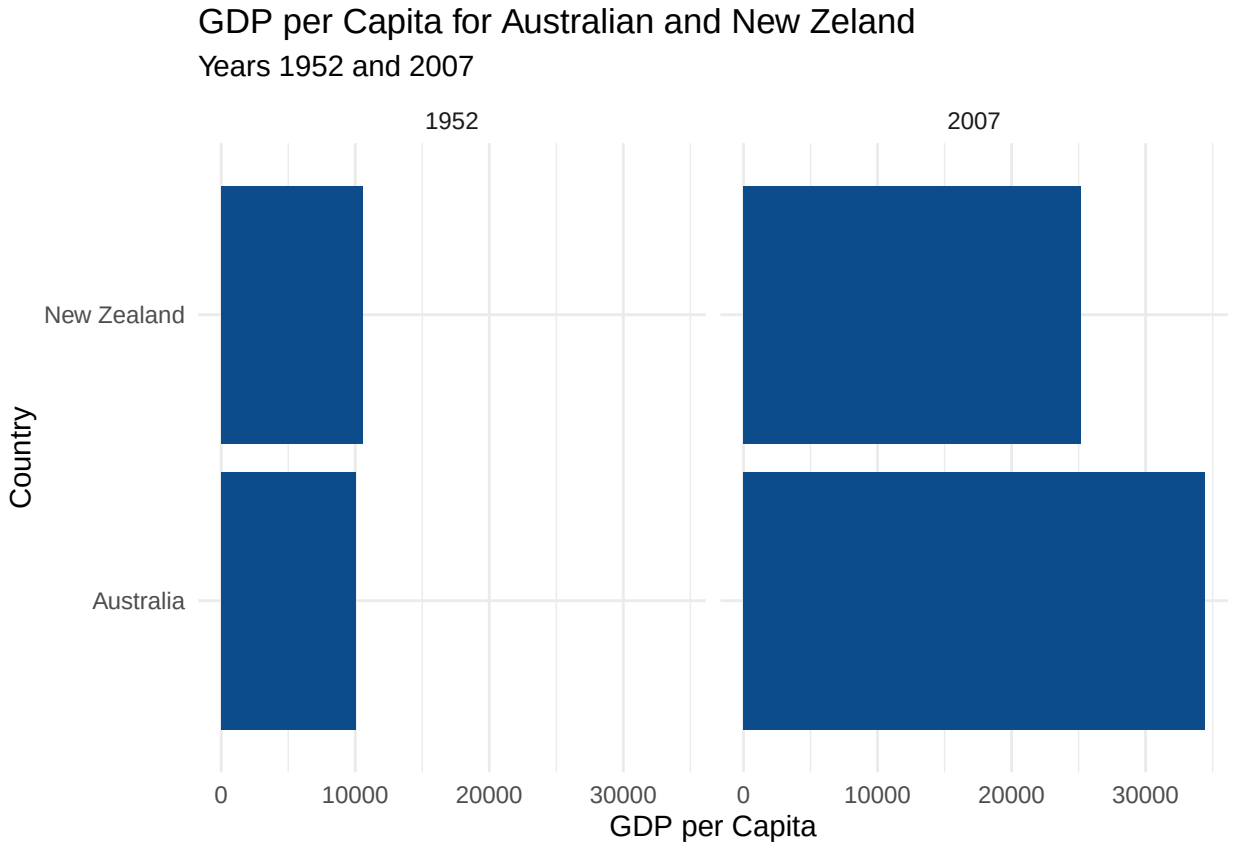
6) Product a plot that summarizes the same data as the table. There should be two plots per continent.




GDP per Capita for American Countries
Years 1952 and 2007
1952
2007
Venezuela
Uruguay
United States
Trinidad and Tobago
Puerto Rico
Peru
Paraguay
Panama
Nicaragua
Mexico
Jamaica
Honduras
Haiti
Guatemala
El Salvador
Ecuador
Dominican Republic
Cuba
Costa Rica
Colombia
Chile
Canada
Brazil
Bolivia
Argentina
Country
0
10000
20000
30000
40000
GDP per Capita


GDP per Capita for Asian Countries
Years 1952 and 2007
1952
2007
Yemen, Rep.
West Bank and Gaza
Vietnam
Thailand
Taiwan
Syria
Sri Lanka
Singapore
Saudi Arabia
Philippines
Pakistan
Oman
Nepal
Myanmar
Mongolia
Malaysia
Lebanon
Kuwait
Korea, Rep.
Korea, Dem. Rep.
Jordan
Japan
Israel
Iraq
Iran
Indonesia
India
Hong Kong, China
China
Cambodia
Bangladesh
Bahrain
Afghanistan
Country
0
30000
60000
90000
GDP per Capita


GDP per Capita for European Countries
Years 1952 and 2007
1952
2007
United Kingdom
Turkey
Switzerland
Sweden
Spain
Slovenia
Slovak Republic
Serbia
Romania
Portugal
Poland
Norway
Netherlands
Montenegro
Italy
Ireland
Iceland
Hungary
Greece
Germany
France
Finland
Denmark
Czech Republic
Croatia
Bulgaria
Bosnia and Herzegovina
Belgium
Austria
Albania
Country
0
10000
20000
30000
40000
50000
GDP per Capita


GDP per Capita for Australian and New Zeland
Years 1952 and 2007
1952
2007
New Zealand
Australia
Country
0
10000
20000
30000
GDP per Capita

7) Which countries in the datset have had periods of negative population growth? Illustrate your answer with a table or plot.

Country	YearsDecreaseOccured	AmountPopulationDecreasedBy
Equatorial Guinea	1972-1977	84928
Guinea-Bissau	1962-1967	26533
Lesotho	2002-2007	34123
Liberia	1987-1992	356440
Rwanda	1992-1997	77620
Somalia	1987-1992	822059
South Africa	2002-2007	435794
Trinidad and Tobago	1987-1992	7667
Trinidad and Tobago	1992-1997	45568
Trinidad and Tobago	1997-2002	36269
Trinidad and Tobago	2002-2007	45224
Afghanistan	1977-1982	1998556
Cambodia	1972-1977	471999
Kuwait	1987-1992	473392
Lebanon	1977-1982	28911
West Bank and Gaza	1967-1972	53064
Bosnia and Herzegovina	1987-1992	82964
Bosnia and Herzegovina	1992-1997	649013
Bulgaria	1987-1992	313452
Bulgaria	1992-1997	592449
Bulgaria	1997-2002	404258
Bulgaria	2002-2007	338941
Croatia	1992-1997	49418
Czech Republic	1992-1997	14995
Czech Republic	1997-2002	44412
Czech Republic	2002-2007	27551
Germany	1952-1957	616968
Germany	1972-1977	556315
Germany	1982-1987	616968
Hungary	1952-1957	92795
Hungary	1982-1987	92795
Hungary	1987-1992	264056
Hungary	1992-1997	104000
Hungary	1997-2002	161371
Hungary	2002-2007	127205
Ireland	1957-1962	48220
Montenegro	2002-2007	35494
Poland	1997-2002	28981
Poland	2002-2007	107735
Portugal	1967-1972	132550
Romania	1992-1997	234569
Romania	1997-2002	158121
Romania	2002-2007	128281
Serbia	1997-2002	225035
Slovenia	1997-2002	115
Slovenia	2002-2007	2252
Switzerland	1972-1977	84976

8) Which countries in the dataset have had the highest rate of growth in per capita GDP? Illustrate your

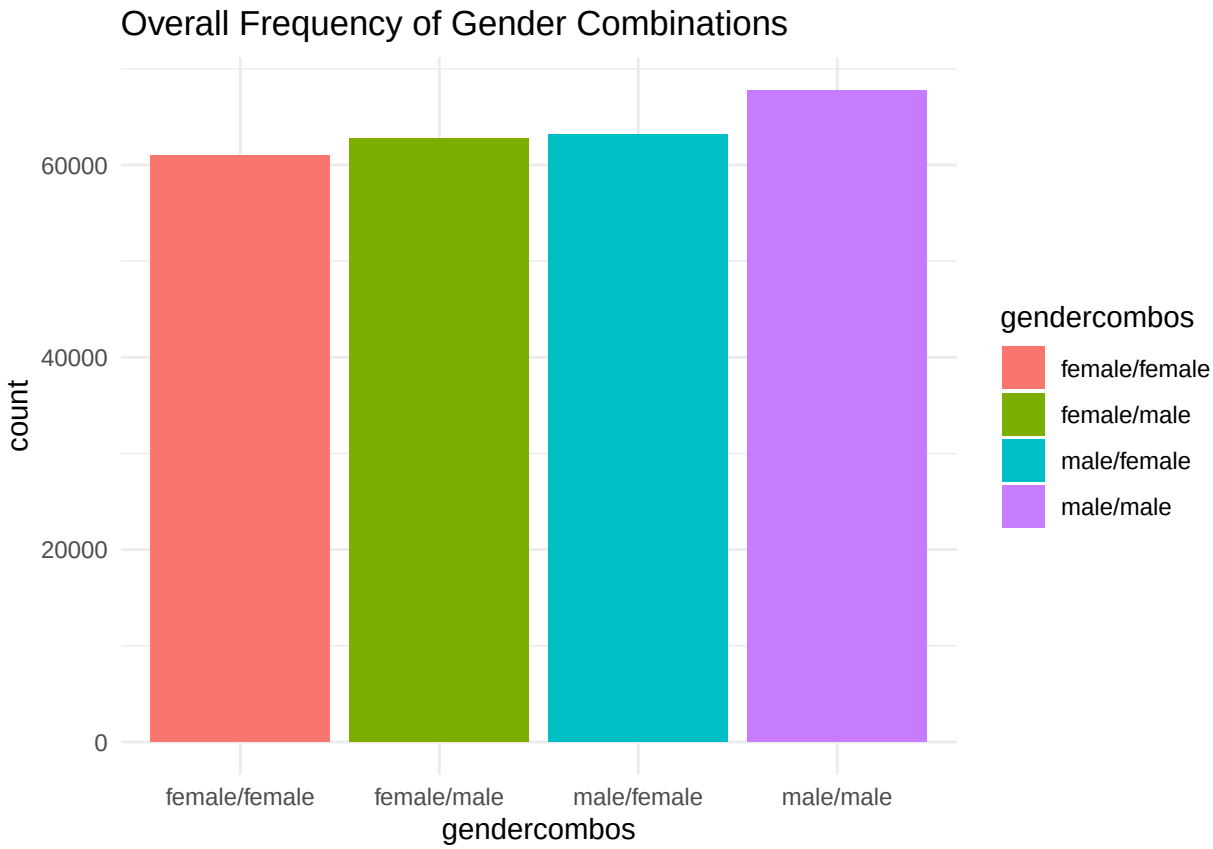
answer with a table or plot. (Just going to focus on big picture from 1952 to 2007.)

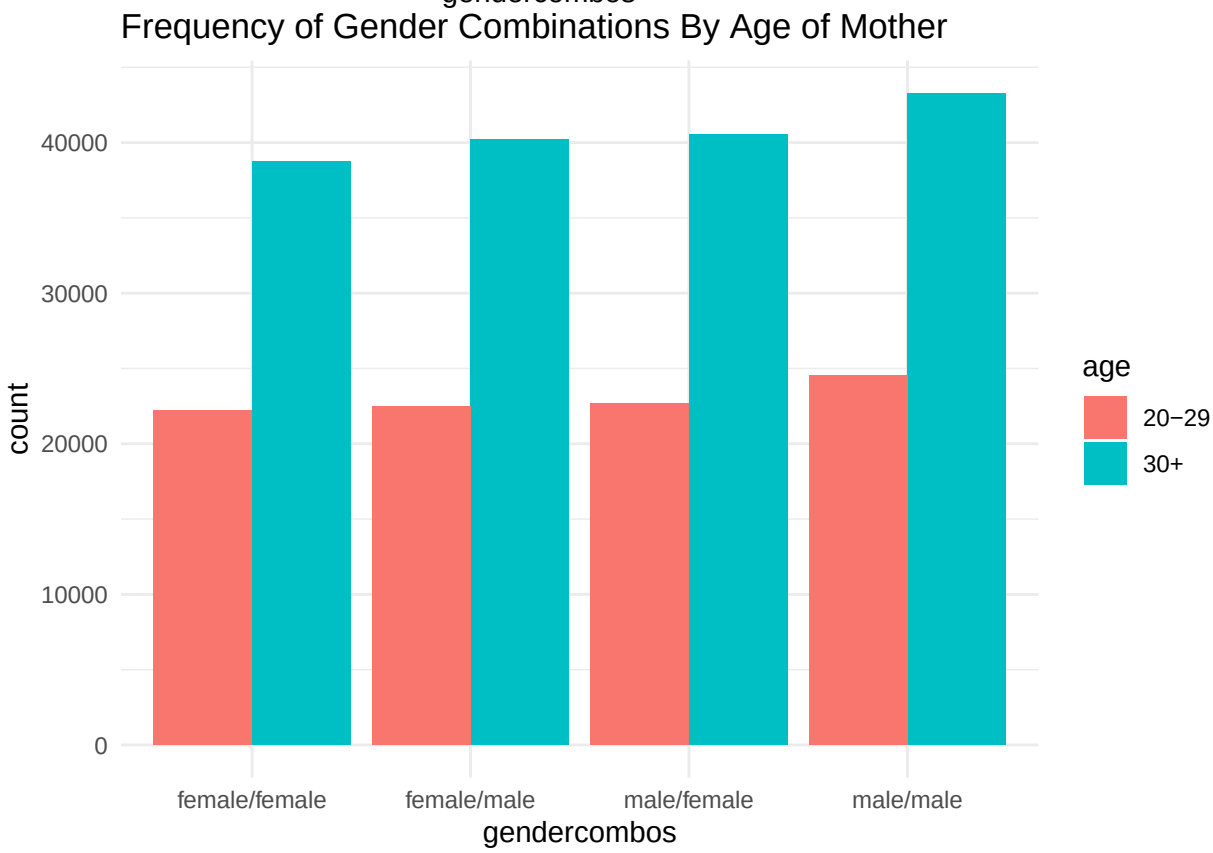
Country	Continent	GDP per Capita - 1952	GDP per Capita - 2007	Rate of GDP Growth
Equatorial Guinea	Africa	375.6431	12154.090	31.355417
Taiwan	Asia	1206.9479	28718.277	22.794131
Korea, Rep.	Asia	1030.5922	23348.140	21.655071
Singapore	Asia	2315.1382	47143.180	19.363009
Botswana	Africa	851.2411	12569.852	13.766499
Hong Kong, China	Asia	3054.4212	39724.979	12.005730
China	Asia	400.4486	4959.115	11.383898
Oman	Asia	1828.2303	22316.193	11.206445
Thailand	Asia	757.7974	7458.396	8.842203
Japan	Asia	3216.9563	31656.068	8.840378
Ireland	Europe	5210.2803	40675.996	6.806873
Greece	Europe	3530.6901	27538.412	6.799725
Bosnia and Herzegovina	Europe	973.5332	7446.299	6.648736
Spain	Europe	3834.0347	28821.064	6.517163
Malaysia	Asia	1831.1329	12451.656	5.799974
Portugal	Europe	3068.3199	20509.648	5.684325
Puerto Rico	Americas	3081.9598	19328.709	5.271564
Israel	Asia	4086.5221	25523.277	5.245721
Slovenia	Europe	4215.0417	25768.258	5.113405
Trinidad and Tobago	Americas	3023.2719	18008.509	4.956629
Austria	Europe	6137.0765	36126.493	4.886596
Italy	Europe	4931.4042	28569.720	4.793425
Mauritius	Africa	1967.9557	10956.991	4.567702
Lesotho	Africa	298.8462	1569.331	4.251301
Finland	Europe	6424.5191	33207.084	4.168805

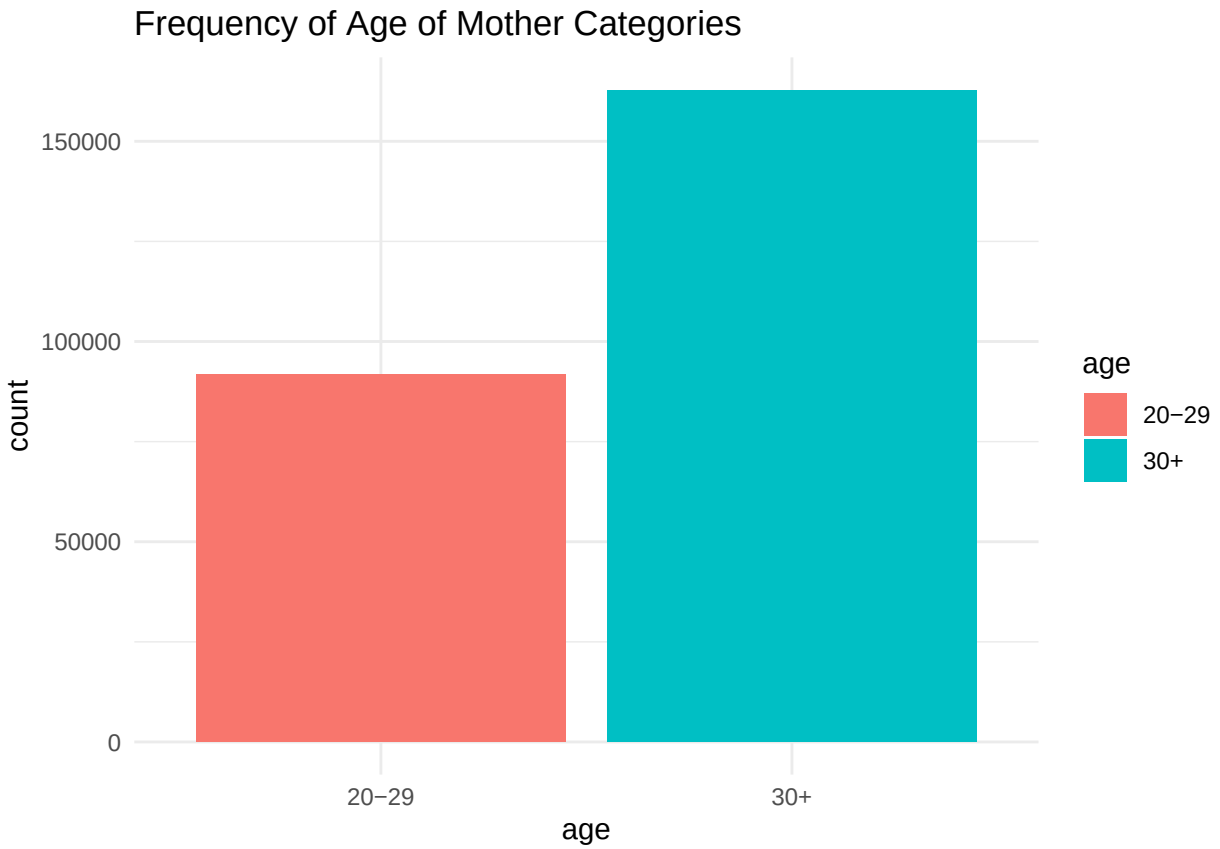
Problem 2 - Fertility Data

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.

1) There are four possible gender combinations for the first two Children. Product a plot the contracts the frequency of these four combinations. Are the frequencies different for women in their 20s and women who are older than 29?


Overall Frequency of Gender Combinations
60000
40000
20000
0
count
female/female
female/male
male/female
male/male
gendercombos
gendercombos
female/female
female/male
male/female
male/male

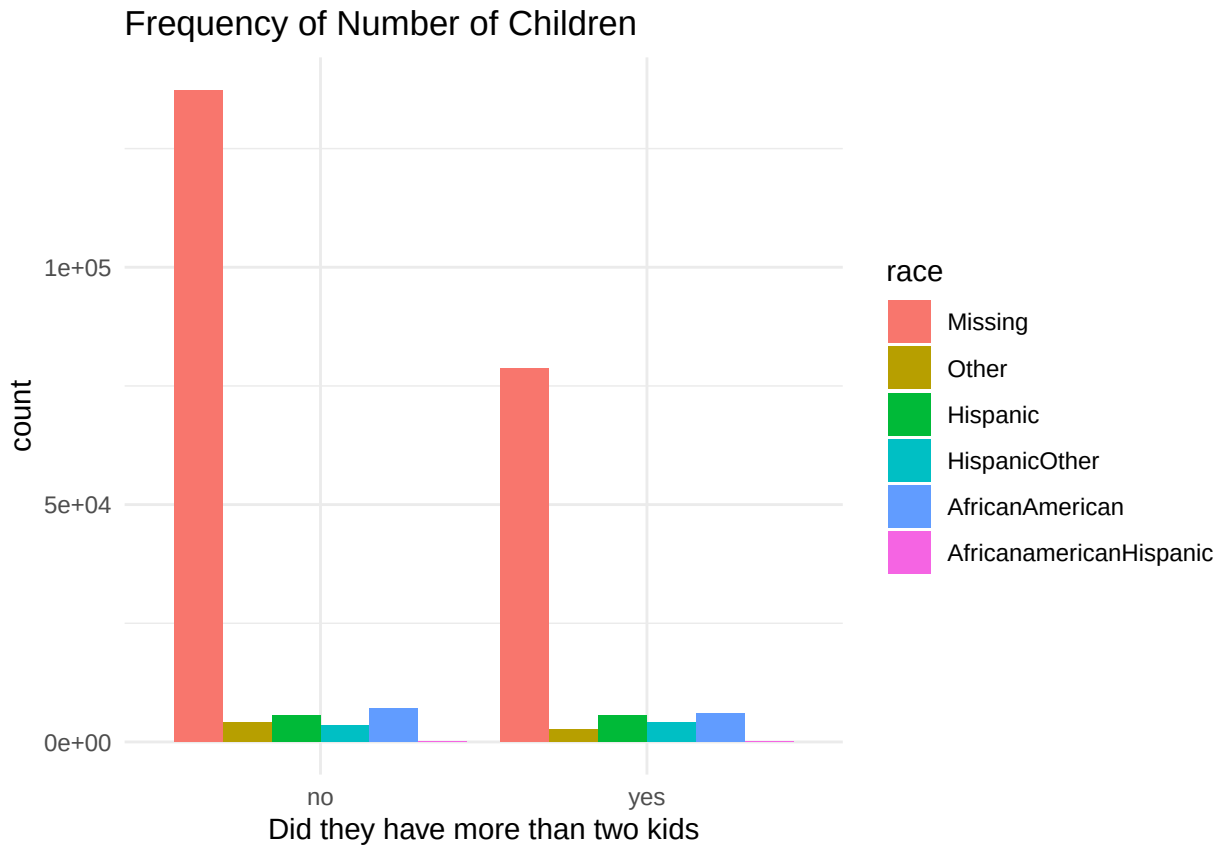

Frequency of Gender Combinations By Age of Mother
40000
30000
20000
10000
0
count
female/female
female/male
male/female
male/male
gendercombos
age
20−29
30+

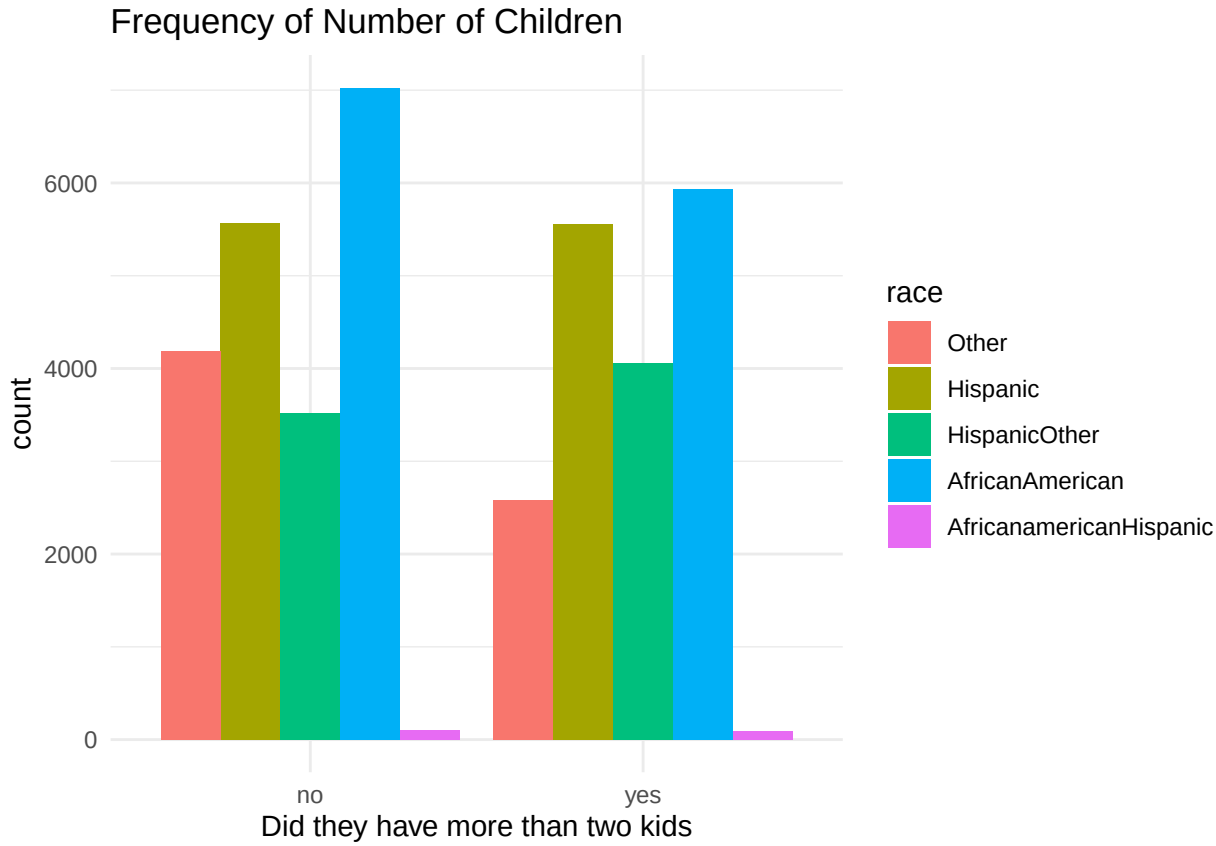


According to the plot, the frequencies do differ for women in their 20s and women who are older than 29. While the general distribution of combinations is the same, there are more higher counts for all combinations for women 30 and over. This is probably because, as shown in the last plot, more women 30 and over were surveyed that women inclusively between 20 and 29.

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

As noted in the documentation, there was some confusion with the participants on how to answer the race related questions on the survey. Therefore, some individuals answered "no" for all options, which has been noted as "missing" in this data. The first plot includes the missing variable, but obscures any patterns among the other known races. Thus, it was removed for the final graph.


Frequency of Number of Children
count
1e+05
5e+04
0e+00
no
yes
Did they have more than two kids
race
Missing
Other
Hispanic
HispanicOther
AfricanAmerican
AfricanamericanHispanic


Frequency of Number of Children
count
6000
4000
2000
0
no
yes
Did they have more than two kids
race
Other
Hispanic
HispanicOther
AfricanAmerican
AfricanamericanHispanic

Problem 3 - Mtcars and Mpg

Use the mtcars and mpg datasets.

1)How many times does the letter "e" occur in mtcars rownames?

The letter e/E occcurs 28 times in mtcars rownames.

2) How many cars in mtcars have the brand Merc?

Seven cars in the mtcars have the brand Merc.

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

In the literal sense, 0 cars have the brand "Merc", but when human logic is applied, 4 cars have the brand "merc" (mercury).

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

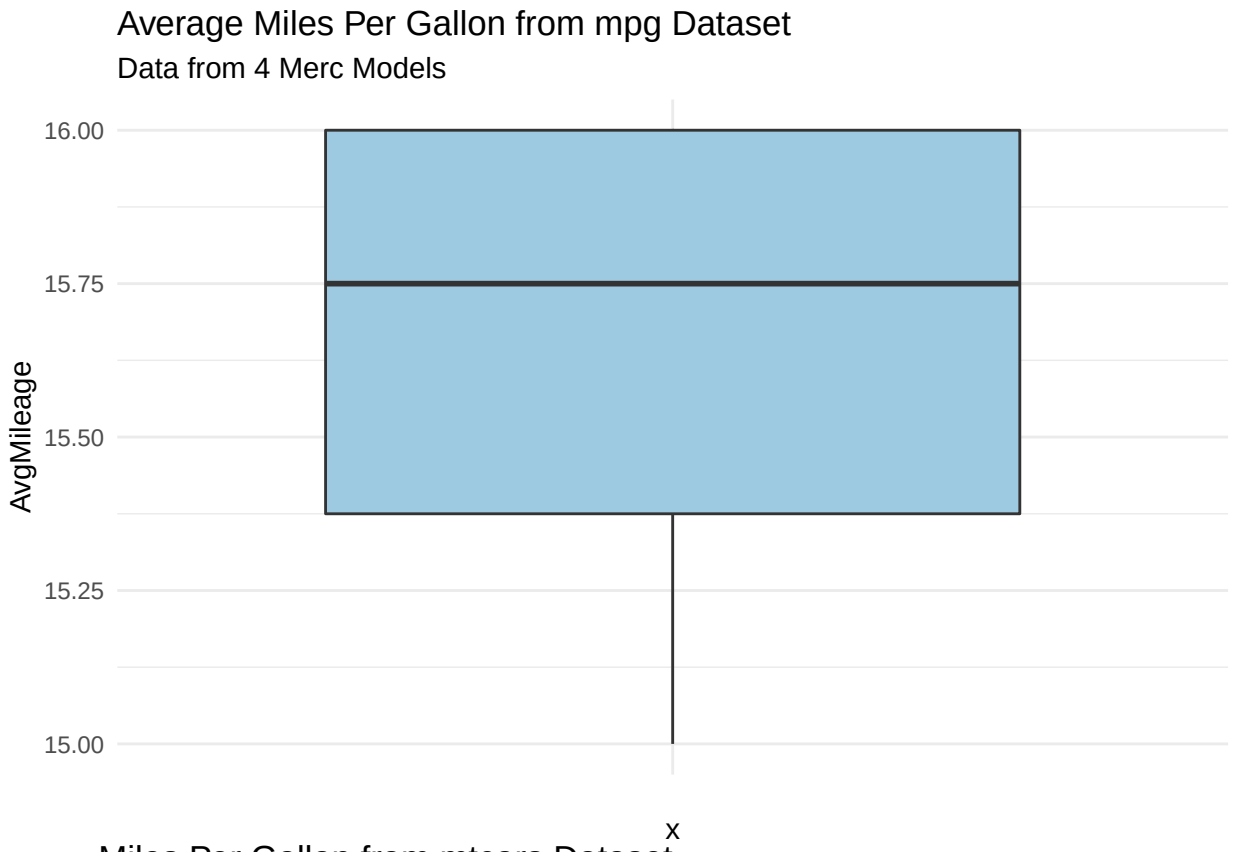
From the data available, the two data sets have similar mileage data as seen in the following summary statistics and box plots with all values falling in the range between 15 and 24.4. The mpg data reports smaller numbers overall, primarily due to the small sample size (4 observations for mpg data set and 7 observations for mtcars data set). However, no solidified conclusions can be drawn from this analysis. Additionally, the mpg data set does not include specific car names, just labels the four observations "mercury". We do not know if these are the same models as the ones being analyzed in the mtcars data set. Therefore, we cannot say that these two data sets are worthy of true comparison.

```
##
## Attaching package: 'data.table'

## The following object is masked from 'package:purrr':
##
##     transpose

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

SummaryStatistic	Original Data Set	Value
Minimum	MPG	15
Minimum	MTCARS	15.2
1st Qu	MPG	15.38
1st Qu	MTCARS	16.85
Median	MPG	15.75
Median	MTCARS	17.8
Mean	MPG	15.62
Mean	MTCARS	19.01
3rd Qu	MPG	16
3rd Qu	MTCARS	21
Max	MPG	16
Max	MTCARS	24.4


Average Miles Per Gallon from mpg Dataset
Data from 4 Merc Models
16.00
15.75
15.50
15.25
15.00
AvgMileage
x

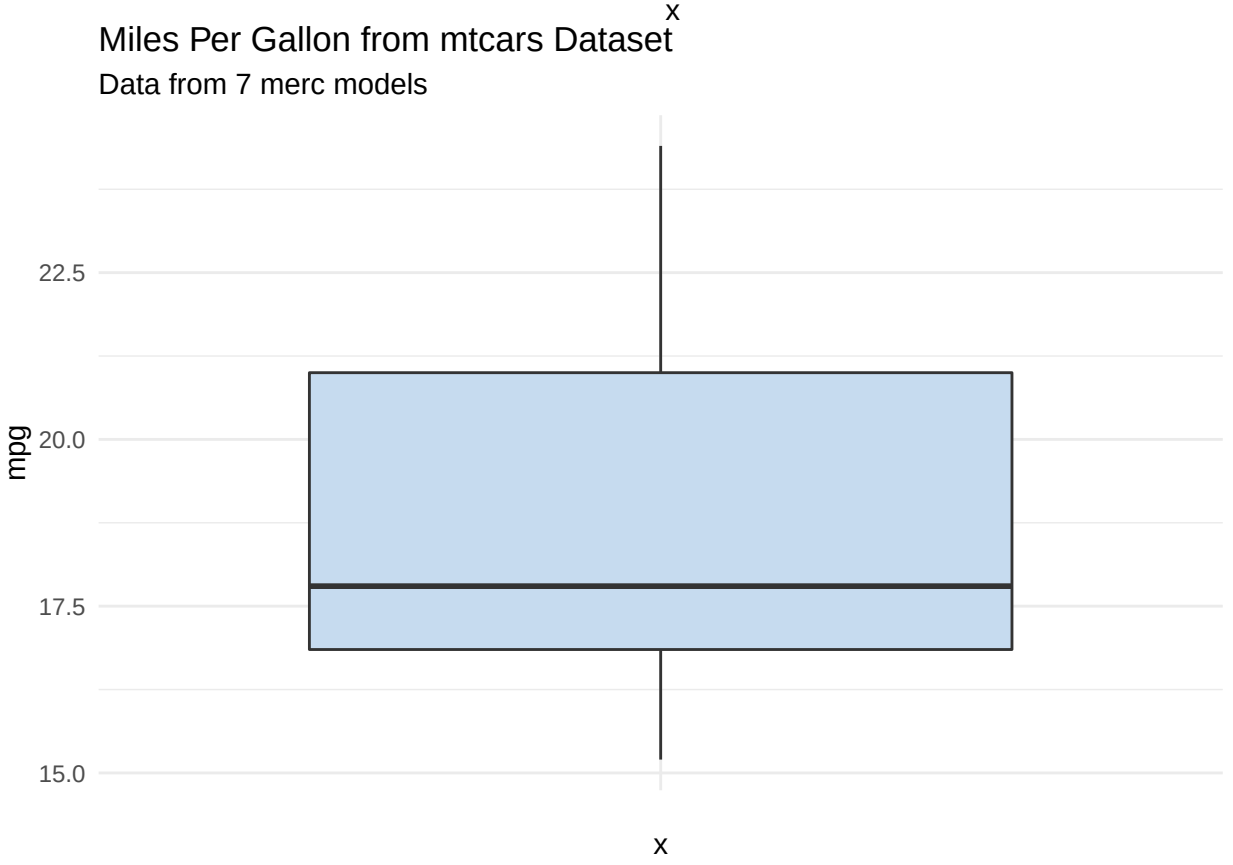

Miles Per Gallon from mtcars Dataset
Data from 7 merc models
22.5
20.0
17.5
15.0
mpg
x

Table 1: First 10 Rows of Random Names

year	sex	name	n	prop
1991	M	Navarre	5	2.40e-06
1986	F	Ricarda	7	3.80e-06
1963	F	Clemmie	9	4.50e-06
1977	M	Garth	105	6.14e-05
1951	M	Windsor	5	2.60e-06
1988	F	Venita	10	5.20e-06
1976	F	Brena	6	3.80e-06
2013	F	Avalee	86	4.47e-05
1979	F	Nikeia	6	3.50e-06
1972	M	Gian	15	9.00e-06

Problem 4 - Babynames

Install the babynames package.

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

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

3) What names overlap boys and girls?

There are 10,663 baby names that overlap boys and girls. Ten of them include John, William, James, Charles, George, Frank, Joseph, Thomas, Henry, and Robert. To keep this document short, the remaining 10,653 can be viewed by uncommenting the last line of code in the R chunk for this question.

4) What names were used in the 19th century but have not been used in the 21st century? (Names unique to 19th century... $A - (A \cup B)$)

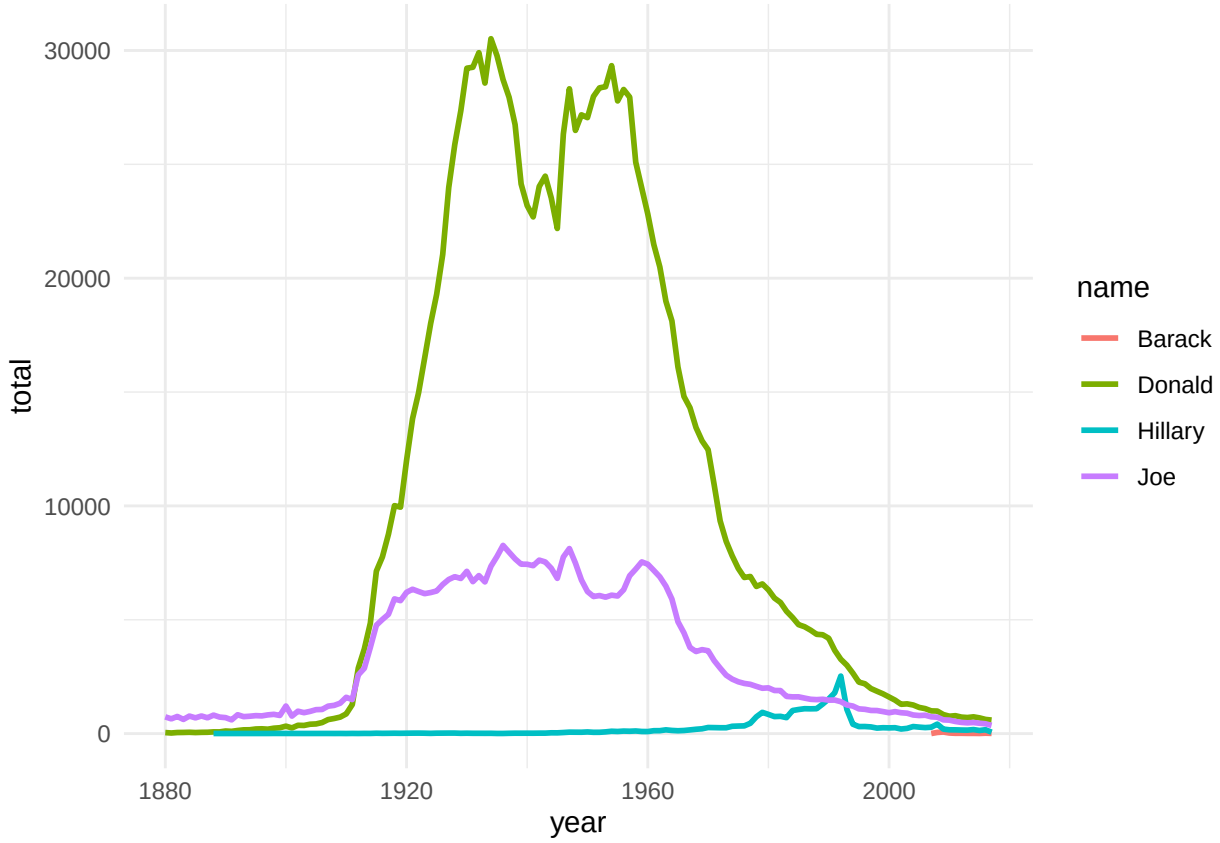
There are 1,362 that were used in the 19th century but have not been used in the 21st century. Again, in order to keep the document short and organized, the list can be viewed by uncommenting the last line of code in the R chunk for this question. However, I will note that my favorite ones on that list include Math, Lemma, Alto, Cathern, Gaylord, Euclid, and Wealthy. An interesting one to notice is that Sister is on that list. I suspect that this was given to babies who died at birth or within a few days after. Instead of giving her a proper name or something of the like, the family decided simply to call her "Sister".

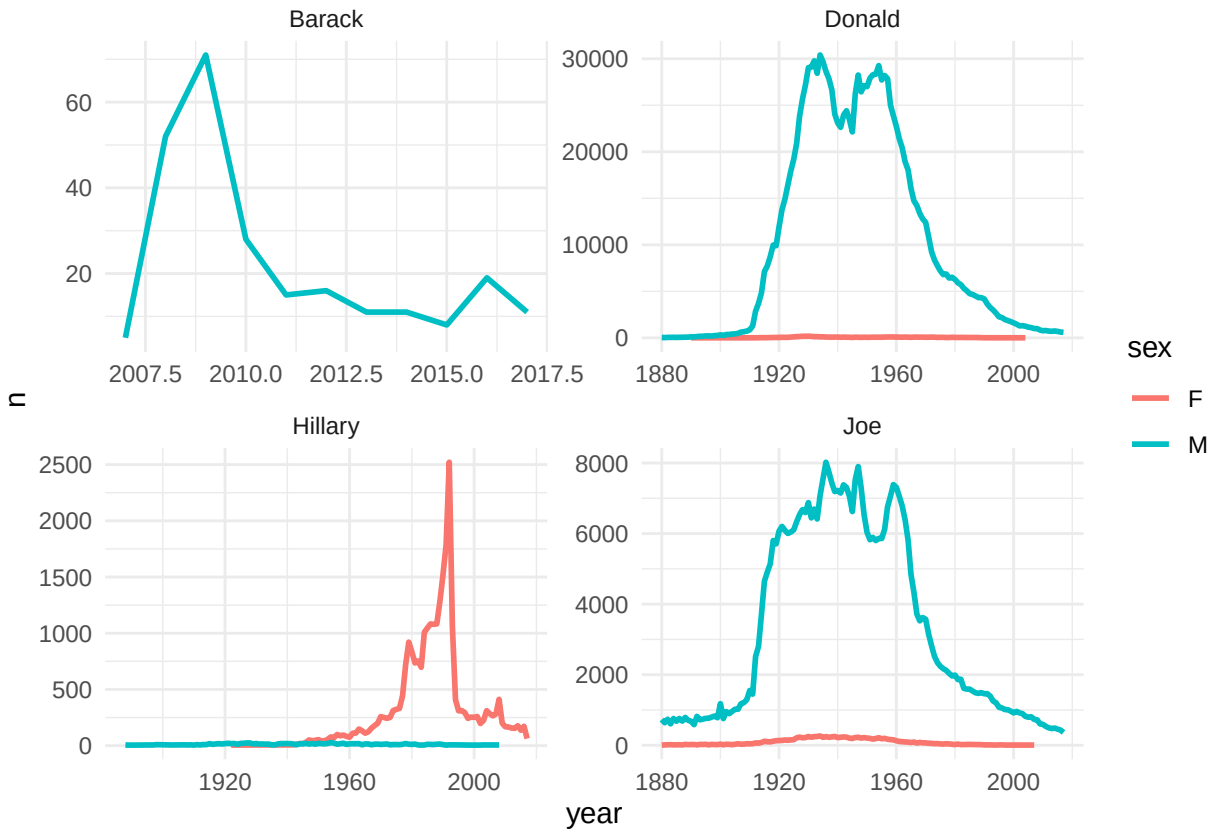
5) Produce a chart that shows the relative frequency of the names "Donald", "Hillary", "Joe", "Barack", over the years 1880 through 2017.

The first plot shows frequency of each name over the years 1880 through 2017 regardless of gender. However, it is difficult to see some of the data (ie data for Barack" when shown this way. To provide some clarity for the specific names, the following plots look at each name individually with gender reconsidered. One note about these plots is that the scales on the x and y axes are different for each name's plot.

Table 2: Top Five Names Per Year

year	sex	name	n	prop
1880	F	Mary	7065	0.0723836
1880	M	John	9655	0.0815456
1880	M	William	9532	0.0805068
1880	M	James	5927	0.0500591
1880	M	Charles	5348	0.0451689
1920	F	Mary	70980	0.0570561
1920	M	John	56913	0.0517007
1920	M	William	50147	0.0455544
1920	M	Robert	48678	0.0442199
1920	M	James	47909	0.0435213
1960	M	David	85928	0.0396768
1960	M	Michael	84183	0.0388711
1960	M	James	76842	0.0354814
1960	M	John	76096	0.0351370
1960	M	Robert	72369	0.0334160
2000	F	Emily	25953	0.0130098
2000	M	Jacob	34471	0.0165139
2000	M	Michael	32035	0.0153469
2000	M	Matthew	28572	0.0136879
2000	M	Joshua	27538	0.0131926




Barack
Donald
Hillary
Joe
n
year
sex
F
M
60
40
20
30000
20000
10000
0
2007.5
2010.0
2012.5
2015.0
2017.5
1880
1920
1960
2000
2500
2000
1500
1000
500
8000
6000
4000