Pre-class work (week 3)

Kari Kusler

9/26/2018

Getting Started

We will work with the dataset called gapminder, this is a cleaned up version from Gapminder Data. Gapminder contains a lot of great data on all of the nations of the world. We first need to install the gapminder package in R.

1. How many unique countries are represented per continent?

continent	countries
Africa	52
Americas	25
Asia	33
Europe	30
Oceania	2

2. Which European nation had the lowest GDP per capita in 1997?

country
Albania

3. According to the data available, what was the average life expectancy across each continent in the 1980s?

Continent	Mean Life Expectancy
Africa	52.5
Americas	67.2
Asia	63.7
Europe	73.2
Oceania	74.8

4. What 5 countries have the highest total GDP over all years combined?

Country	Total GDP per Capita
Kuwait	783994.9
Switzerland	324892.0
Norway	320967.7
United States	315133.8
Canada	268929.0

5. What countries and years had life expectancies of at least 80 years? N.b. only output the columns of interest: country, life expectancy and year (in that order).

Country	Life expectancy (years)	Year
Australia	80.370	2002
Australia	81.235	2007
Canada	80.653	2007
France	80.657	2007
Hong Kong, China	80.000	1997
Hong Kong, China	81.495	2002
Hong Kong, China	82.208	2007
Iceland	80.500	2002
Iceland	81.757	2007
Israel	80.745	2007
Italy	80.240	2002
Italy	80.546	2007
Japan	80.690	1997
Japan	82.000	2002
Japan	82.603	2007
New Zealand	80.204	2007
Norway	80.196	2007
Spain	80.941	2007
Sweden	80.040	2002
Sweden	80.884	2007
Switzerland	80.620	2002
Switzerland	81.701	2007

6. What 10 countries have the strongest correlation (in either direction) between life expectancy and per capita GDP?

country	correlation
France	0.9962239
Austria	0.9929642
Belgium	0.9927496
Norway	0.9921416
Oman	0.9907526
United Kingdom	0.9898930
Italy	0.9897600
Israel	0.9884894
Denmark	0.9870896
Australia	0.9864457

7. Which combinations of continent (besides Asia) and year have the highest average population across all countries? N.b. your output should include all results sorted by highest average population. With what you already know, this one may stump you. See this Q&A for how to ungroup before arrangeing. This also behaves differently in more recent versions of dplyr.

continent	year	avgpop
Americas	2007	35954847
Americas	2002	33990910
Americas	1997	31876016
Americas	1992	29570964
Americas	1987	27310159
Americas	1982	25211637
Americas	1977	23122708

continent	year	avgpop
Americas	1972	21175368
Europe	2007	19536618
Europe	2002	19274129
Americas	1967	19229865
Europe	1997	18964805
Europe	1992	18604760
Europe	1987	18103139
Africa	2007	17875763
Europe	1982	17708897
Americas	1962	17330810
Europe	1977	17238818
Europe	1972	16687835
Europe	1967	16039299
Africa	2002	16033152
Americas	1957	15478157
Europe	1962	15345172
Europe	1957	14596345
Africa	1997	14304480
Europe	1952	13937362
Americas	1952	13806098
Africa	1992	12674645
Oceania	2007	12274974
Oceania	2002	11727414
Oceania	1997	11120715
Africa	1987	11054502
Oceania	1992	10459826
Oceania	1987	9787208
Africa	1982	9602857
Oceania	1982	9197425
Oceania	1977	8619500
Africa	1977	8328097
Oceania	1972	8053050
Africa	1972	7305376
Oceania	1967	7300207
Oceania	1962	6641759
Africa	1967	6447875
Oceania	1957	5970988
Africa	1962	5702247
Oceania	1952	5343003
Africa	1957	5093033
Africa	1952	4570010

The Americas in 2007 had the highest average population.

8. Which three countries have had the most consistent population estimates (i.e. lowest standard deviation) across the years of available data?

country	sd
China	264394873
India	251724253
Indonesia	49157536

9. Subset gm to only include observations from 1992 and store the results as gm1992. What kind of object is this?

```
## [1] "list"
```

The gm1992 object is a list

10. Which observations indicate that the population of a country has decreased from the previous year and the life expectancy has increased from the previous year? See the vignette on window functions.

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1982	39.854	12881816	978.0114
Bosnia and Herzegovina	Europe	1992	72.178	4256013	2546.7814
Bosnia and Herzegovina	Europe	1997	73.244	3607000	4766.3559
Bulgaria	Europe	2002	72.140	7661799	7696.7777
Bulgaria	Europe	2007	73.005	7322858	10680.7928
Croatia	Europe	1997	73.680	4444595	9875.6045
Czech Republic	Europe	1997	74.010	10300707	16048.5142
Czech Republic	Europe	2002	75.510	10256295	17596.2102
Czech Republic	Europe	2007	76.486	10228744	22833.3085
Equatorial Guinea	Africa	1977	42.024	192675	958.5668
Germany	Europe	1977	72.500	78160773	20512.9212
Germany	Europe	1987	74.847	77718298	24639.1857
Guinea-Bissau	Africa	1967	35.492	601287	715.5806
Hungary	Europe	1987	69.580	10612740	12986.4800
Hungary	Europe	1997	71.040	10244684	11712.7768
Hungary	Europe	2002	72.590	10083313	14843.9356
Hungary	Europe	2007	73.338	9956108	18008.9444
Ireland	Europe	1957	68.900	2878220	5599.0779
Ireland	Europe	1962	70.290	2830000	6631.5973
Kuwait	Asia	1992	75.190	1418095	34932.9196
Lebanon	Asia	1982	66.983	3086876	7640.5195
Montenegro	Europe	2007	74.543	684736	9253.8961
Poland	Europe	2002	74.670	38625976	12002.2391
Poland	Europe	2007	75.563	38518241	15389.9247
Portugal	Europe	1972	69.260	8970450	9022.2474
Romania	Europe	1997	69.720	22562458	7346.5476
Romania	Europe	2002	71.322	22404337	7885.3601
Romania	Europe	2007	72.476	22276056	10808.4756
Rwanda	Africa	1997	36.087	7212583	589.9445
Serbia	Europe	2002	73.213	10111559	7236.0753
Slovenia	Europe	2002	76.660	2011497	20660.0194
Slovenia	Europe	2007	77.926	2009245	25768.2576
Switzerland	Europe	1977	75.390	6316424	26982.2905
Trinidad and Tobago	Americas	1992	69.862	1183669	7370.9909
Trinidad and Tobago	Americas	2007	69.819	1056608	18008.5092
West Bank and Gaza	Asia	1972	56.532	1089572	3133.4093

R code:

```
#install.packages("gapminder")
library(dplyr)
library(gapminder)
gapminder = gapminder # dataset is called gapminder
```

```r
#to use kable to make tables look nice
library(knitr)
```

```r
#Question 1
#removing extraneous information
cntry.cont <- gapminder[,1:2]
distinct <- distinct(cntry.cont)
#grouping by continent and counting number of countries
cntry.per.cont <- distinct %>%
  group_by(continent) %>%
  summarise(countries=n())
#printing output
kable(cntry.per.cont,format="markdown")
```

```r
#Question 2
#filtering to Europe and 1997
lowgdp <- gapminder %>%
  filter(continent == 'Europe' & year == '1997') %>%
arrange(gdpPercap) %>%
  head(1)
kable(lowgdp[1,1],format="markdown")
```

```r
#Question 3
#filter to include only years of interest, group by continent, and summarize mean life expectancy
lifeexp <- gapminder %>%
  filter(year >= 1980 & year < 1990) %>%
  group_by(continent) %>%
  summarize("avg" = mean(lifeExp))

#round to one decimal point
lifeexp[,2] <- round(lifeexp[,2],1)
colnames(lifeexp)<-c("Continent","Mean Life Expectancy")
kable(lifeexp,format="markdown")
```

```r
#Question 4
#group by country, summarize gdp for all years, and sort by descending gdp
high.gdp <- gapminder %>%
  group_by(country) %>%
  summarize(total.gdp = sum(gdpPercap)) %>%
  arrange(desc(total.gdp))
colnames(high.gdp) <- c("Country","Total GDP per Capita")
kable(high.gdp[1:5,],format="markdown")
```

```r
#Question 5
#filter by life expenctancy at least 80 and only keep columns of interest
oldlife <- gapminder %>%
  filter(lifeExp >= 80) %>%
  select(country, lifeExp, year)

colnames(oldlife) <- c("Country","Life expectancy (years)","Year")
kable(oldlife,format="markdown")
```

```r
#Question 6
life.gdp.cor <- gapminder %>%
  group_by(country)  %>%
  summarize(correlation=cor(lifeExp,gdpPercap)) %>%
  arrange(desc(abs(correlation)))

kable(life.gdp.cor[1:10,],format="markdown")
```

```r
#Question 7
#remove Asia, group by continent and year, find average population, and arrange results
highpop <- gapminder %>%
  filter(continent != "Asia") %>%
  group_by(continent,year) %>%
  summarize(avgpop = mean(pop)) %>%
  arrange(desc(avgpop))

kable(highpop,format="markdown")
```

```r
#Question 8
# grouped by country, calculated SD, arrange by descending SD, printed top 3
lowsd <- gapminder %>%
  group_by(country) %>%
  summarize(sd = sd(pop)) %>%
  arrange(desc(sd))

kable(lowsd[1:3,],format="markdown")
```

```r
#Question 9
gm1992 <- gapminder %>%
  filter(year==1992)
  typeof(gm1992)
```

```r
#Question 10
#arrange by country and year, group by country, and use lag functions to compare observations to previo
gap <- gapminder %>%
  arrange(country,year) %>%
group_by(country) %>%
filter(pop < lag(pop) & lifeExp > lag(lifeExp))
kable(gap,format="markdown")
```