STAT 545 hw-01 (gapminder)

Hannah J McSorley

September 11, 2019

Homework 1, Exercise 2

This is the second exercise in assignment one for STAT 545. The goal of this exercise is to explore a dataset and create a GitHub-readable output using R Markdown.

'gapminder' dataset

This exercise uses the dataset 'gapminder', which includes information about life expectancy, GDP per capita and population by country and continent. First, load the 'gapminder' dataset into the workspace by calling it into the library.

```
# load the packages required for this analysis
library(gapminder)
```

Data structure

Explore the data structure and type by executing some simple commands.

```
str(gapminder)  # structure
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1704 obs. of  6 variables:
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num  28.8 30.3 32 34 36.1 ...
##  $ pop      : int  8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16317921 22
##  $ gdpPercap: num  779 821 853 836 740 ...
```

```
class(gapminder)  # data type
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

Note that calling 'class()' was redundant because the output from calling 'str()' included the data class as well as the dataset's structure.

Dataset Exploration

Countries

Identify how many countries are included in the gapminder dataset, and what they are.

```r
#  how many countries are included?
length(unique(gapminder$country))
```

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

```r
# see if Canada is a country in the dataset
any(gapminder$country == "Canada")
```

```
## [1] TRUE
```

Span of Dataset (Years)

This dataset includes life expectancy ("lifeExp") in each country. What is the range of this data and how does it vary? Using the 'range()' function yeilds the lowest and highest end members in the data (min & max).

```r
# what is the range of years included?
range(gapminder$year)
```

```
## [1] 1952 2007
```

```r
# what is the range of life expectancies across the world?
range(gapminder$lifeExp)
```

```
## [1] 23.599 82.603
```

Life Expectancy

Try to locate which country has the lowest and highest life expectancy...

```r
# which country had the lowest life expectancy?
which(gapminder$lifeExp == min(gapminder$lifeExp))
```

```
## [1] 1293
```

Hmmm...that command returned the index location for the lowest life expectancy, but can I identify the information associate with that index (e.g. country, year, GDP) by subsetting the dataframe with that index?

```r
# subset using the index command from above
gapminder[which(gapminder$lifeExp == min(gapminder$lifeExp)), ]
```

```
## # A tibble: 1 x 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Rwanda  Africa     1992    23.6 7290203      737.
```

```
# what about the maximum life expectancy?
gapminder[which(gapminder$lifeExp == max(gapminder$lifeExp)), ]
```

```
## # A tibble: 1 x 6
##   country continent  year lifeExp       pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>     <int>     <dbl>
## 1 Japan   Asia       2007    82.6 127467972    31656.
```

The lowest life expectancy (23.6 years) was in 1992 in Rwanda, Africa. The highest life expectacy (82.6 years) was in Japan, Asia in 2007.

GDP match with Life Expectancy?

Let's see if there is a match between min/max GDP and life expectancy, or population and life expectancy.

```
# is there a pattern between GDP and life expectancy?
min(gapminder$gdpPercap) == min(gapminder$lifeExp)
```

```
## [1] FALSE
```

```
max(gapminder$gdpPercap) == max(gapminder$lifeExp)
```

```
## [1] FALSE
```

```
# is there a pattern between population size and life expectancy?
min(gapminder$pop) == min(gapminder$lifeExp)
```

```
## [1] FALSE
```

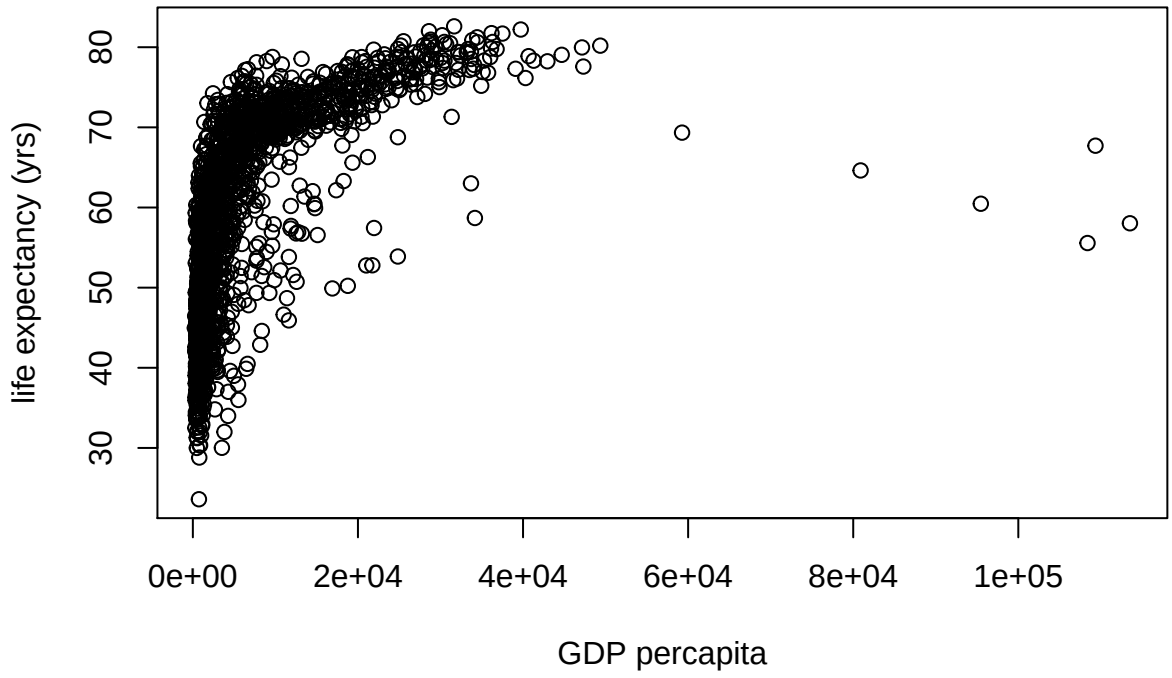
```
max(gapminder$pop) == max(gapminder$lifeExp)
```

```
## [1] FALSE
```

No, there was not a direct match between min/max GDP nor population with life expectancy.

I imagine that life expectancy would have some relationship to GDP though... what if we plot those data together to visually explore that relationship?

```
# use base R to plot global life expectancy with GDP
plot(y = gapminder$lifeExp, x = gapminder$gdpPercap,
     ylab = "life expectancy (yrs)", xlab = "GDP percapita")
```



Canada, eh

I am Canadian, what was the most recent life expectancy in Canada?

```
# subset the gapminder data for Canada
# save as an object
Canada_df <- gapminder[gapminder$country == "Canada", ]

# find minimum and maximum life expectancy
Canada_df[which(Canada_df$lifeExp == min(Canada_df$lifeExp)), ]
```

```
## # A tibble: 1 x 6
##   country continent  year lifeExp      pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Canada  Americas   1952    68.8 14785584    11367.
```

```
Canada_df[which(Canada_df$lifeExp == max(Canada_df$lifeExp)), ]
```

```
## # A tibble: 1 x 6
##   country continent  year lifeExp      pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Canada  Americas   2007    80.7 33390141    36319.
```

The life expectancy in Canada was lowest in the first recorded year of this dataset (1952, 68.8 years), and highest in the most recent recorded year (2007, 80.7 years). So, life expectancy has increased in Canada over the span of this dataset.

Let's plot that also!

```
# use base R to plot Canadian life expectancy over time
plot(y = Canada_df$lifeExp, x = Canada_df$year,
     ylab = "life expectancy (yrs)", xlab = "Year",
     type = "b")
```