IDS 2021. Problem Set 2

Muhammad Zahidul Islam Miaji

Last updated at 2021-04-01 10:00:50

```
#install.packages("gapminder")
library(gapminder)
library(tidyverse)

gapminder07 <- filter(gapminder, year %in% 2007)
```

Q1

How many variables and how many observations are in the original gapminder data? How about for the data subset for 2007?

```
glimpse(gapminder)
```

```
## Rows: 1,704
## Columns: 6
## $ country   <fct> Afghanistan, Afghanistan, Afghanistan, Afghanistan, Afgha...
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asi...
## $ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 199...
## $ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 4...
## $ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372,...
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.113...
```

```
observ_gapminder <- gapminder
observ_gapminder
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Afghanistan Asia       1957    30.3  9240934      821.
##  3 Afghanistan Asia       1962    32.0 10267083      853.
##  4 Afghanistan Asia       1967    34.0 11537966      836.
##  5 Afghanistan Asia       1972    36.1 13079460      740.
##  6 Afghanistan Asia       1977    38.4 14880372      786.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # ... with 1,694 more rows
```

```
gapminder
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Afghanistan Asia       1957    30.3  9240934      821.
##  3 Afghanistan Asia       1962    32.0 10267083      853.
##  4 Afghanistan Asia       1967    34.0 11537966      836.
##  5 Afghanistan Asia       1972    36.1 13079460      740.
##  6 Afghanistan Asia       1977    38.4 14880372      786.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # ... with 1,694 more rows
```

```
gapminder07 <- filter(gapminder, year %in% 2007)
```

Total observation for "gapminder" is 1704 of 6 variables. And year of 2007 (subset 2007) is 142 observation of 6 variables

Q2

Let's create a new variable `real_gdp` by multiplying the `gdpPercap` variable with the `pop` variable (hint: 1. GDP per capita is calculated by dividing the real GDP by population, 2. use the `mutate` function to create the new variable)

```
gapminder_realGdp <- gapminder %>%
  mutate(real_gdp = gdpPercap * pop)
gapminder_realGdp
```

```
## # A tibble: 1,704 x 7
##    country     continent  year lifeExp      pop gdpPercap     real_gdp
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>        <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.  6567086330.
##  2 Afghanistan Asia       1957    30.3  9240934      821.  7585448670.
##  3 Afghanistan Asia       1962    32.0 10267083      853.  8758855797.
##  4 Afghanistan Asia       1967    34.0 11537966      836.  9648014150.
##  5 Afghanistan Asia       1972    36.1 13079460      740.  9678553274.
##  6 Afghanistan Asia       1977    38.4 14880372      786. 11697659231.
##  7 Afghanistan Asia       1982    39.9 12881816      978. 12598563401.
##  8 Afghanistan Asia       1987    40.8 13867957      852. 11820990309.
##  9 Afghanistan Asia       1992    41.7 16317921      649. 10595901589.
## 10 Afghanistan Asia       1997    41.8 22227415      635. 14121995875.
## # ... with 1,694 more rows
```

Q3

Next, let's compute the average life expectancy by continent in the year 2007. (hint: 1. so you'll need to use the `gapminder07` data. 2. you wilcontinl probably need to use functions such as `filter`, `group_by`,

`summarize`, and `mean`, 3. make sure to use `na.rm=TRUE` for your mean function to avoid observations from dropping out of your data inadvertantly).

```
avg_life <- gapminder07 %>%
  filter(!is.na(lifeExp), !is.na(continent)) %>%
  group_by(lifeExp, continent) %>%
  summarise(avg_LE = round(mean(lifeExp, na.rm = TRUE)))
```

```
## `summarise()` has grouped output by 'lifeExp'. You can override using the `.groups` argument.
```

```
avg_life
```

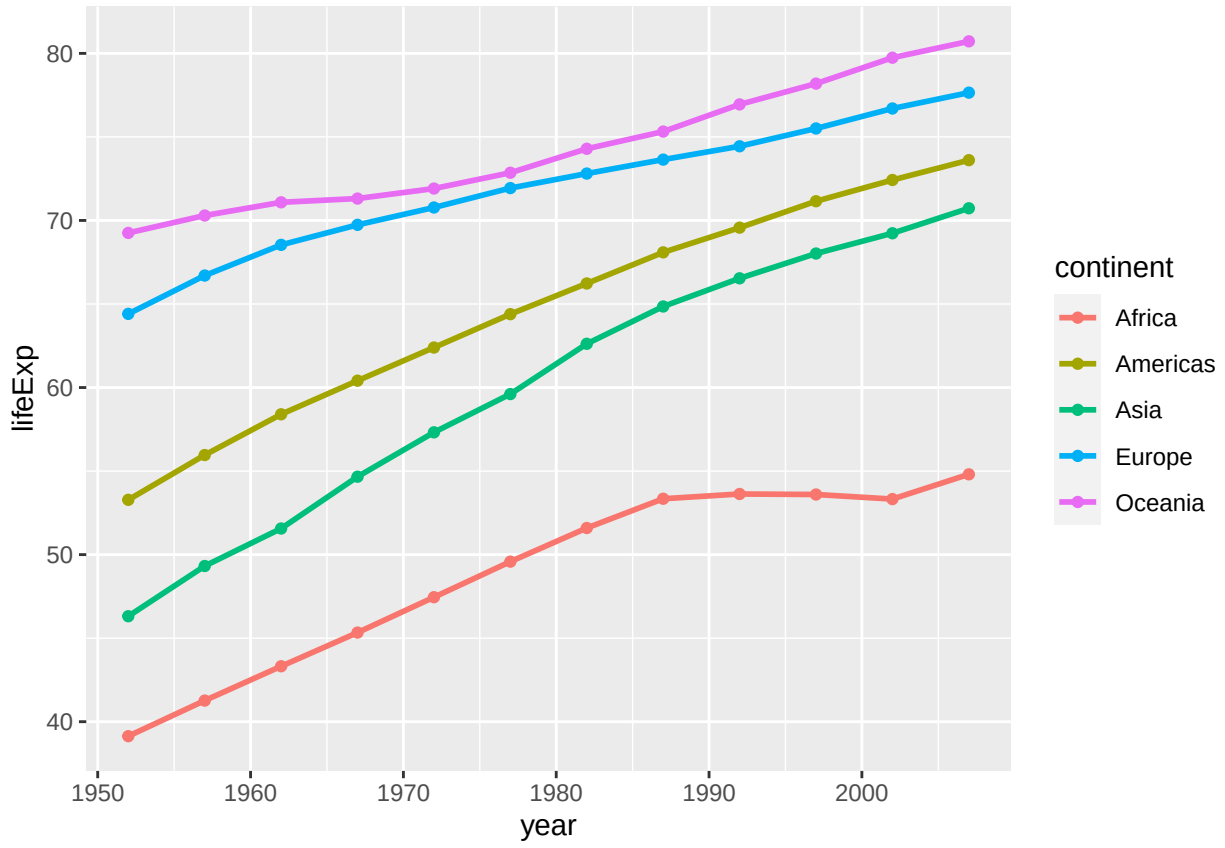
```
## # A tibble: 142 x 3
## # Groups:   lifeExp [142]
##    lifeExp continent avg_LE
##      <dbl> <fct>      <dbl>
##  1    39.6 Africa        40
##  2    42.1 Africa        42
##  3    42.4 Africa        42
##  4    42.6 Africa        43
##  5    42.6 Africa        43
##  6    42.7 Africa        43
##  7    43.5 Africa        43
##  8    43.8 Asia          44
##  9    44.7 Africa        45
## 10    45.7 Africa        46
## # ... with 132 more rows
```

Q4

Next, let's compute the average life expectancy by continent and year from the full dataset (`gapminder`). Draw a line plot over time to examine the trend.

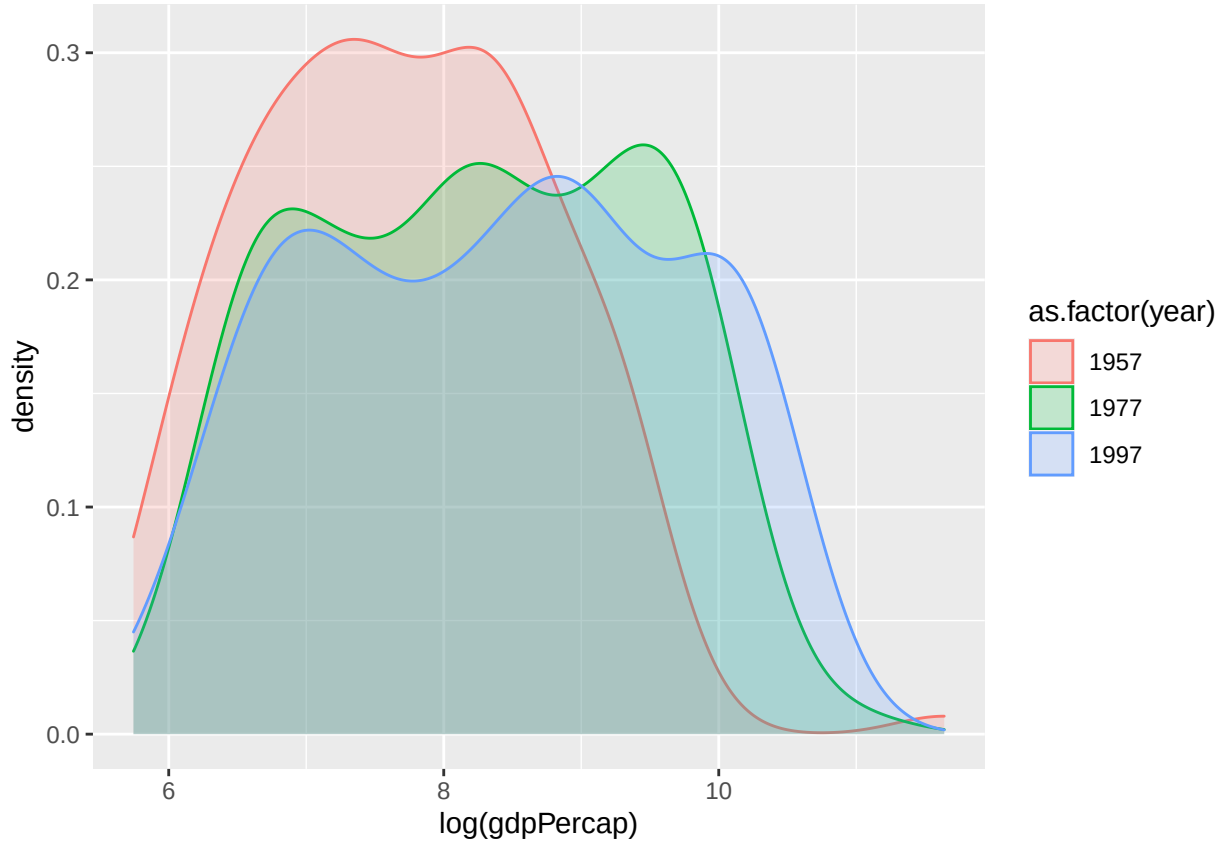
```
gapminder %>%
    group_by(continent, year) %>%
    summarise(lifeExp=mean(lifeExp)) %>%
    ggplot(aes(x=year, y=lifeExp, color=continent)) +
     geom_line(size=1) +
     geom_point(size=1.5)
```

```
## `summarise()` has grouped output by 'continent'. You can override using the `.groups` argument.
```



Q5

```
gapminder%>%
  filter (year %in% c(1957, 1977, 1997))%>%
  ggplot(aes(x = log(gdpPercap), color= as.factor(year), fill= as.factor(year))) +
  geom_density(alpha= 0.2)
```



Q6.

Create a table that shows the Life Expectancy in 2007 for the countries in the Americas. Report only the country and the life expectancy variables in the table. (hint: you will probably need to use functions such as `filter`, `select`, `tableGrob`, `grid.arrange`, or `kable`).

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
gapminder %>%
  filter(year== 2007, continent== "Americas")%>%
  select(country, lifeExp)%>%
  tableGrob(cols = c("country", "lifeExp")) %>%
  grid.arrange()
```

	country	lifeExp
1	Argentina	75.320
2	Bolivia	65.554
3	Brazil	72.390
4	Canada	80.653
5	Chile	78.553
6	Colombia	72.889
7	Costa Rica	78.782
8	Cuba	78.273
9	Dominican Republic	72.235
10	Ecuador	74.994
11	El Salvador	71.878
12	Guatemala	70.259
13	Haiti	60.916
14	Honduras	70.198
15	Jamaica	72.567
16	Mexico	76.195
17	Nicaragua	72.899
18	Panama	75.537
19	Paraguay	71.752
20	Peru	71.421
21	Puerto Rico	78.746
22	Trinidad and Tobago	69.819
23	United States	78.242
24	Uruguay	76.384
25	Venezuela	73.747