HW3

Zichun Ye

2014/09/26

In this assignment, we will continue to explore the data set Gapminder excerpt.

Preparation

Before preceeding to the exciting parts, we need some preparation like loading the data and library.

```
# load the data
gdURL <-"http://tiny.cc/gapminder"
gDat <- read.delim(file = gdURL)

# load the library
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggthemes)
library(knitr)
library(reshape2)

# change data.frame to tbl_df
gtbl <- tbl_df(gDat)
glimpse(gtbl)
```

```
## Variables:
## $ country   (fctr) Afghanistan, Afghanistan, Afghanistan, Afghanistan,...
## $ year      (int) 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992...
## $ pop       (dbl) 8425333, 9240934, 10267083, 11537966, 13079460, 1488...
## $ continent (fctr) Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asi...
## $ lifeExp   (dbl) 28.80, 30.33, 32.00, 34.02, 36.09, 38.44, 39.85, 40....
## $ gdpPercap (dbl) 779.4, 820.9, 853.1, 836.2, 740.0, 786.1, 978.0, 852...
```

Our exploration

- TASK NO.1: Get the maximum and minimum of GDP per capita for all continents.

```r
# Get the maximum and minimum of GDP per capita for all continents.
gdp_int <- gtbl %>%
  group_by(continent) %>%
  summarize(min_gdpPercap = min(gdpPercap), max_gdpPercap = max(gdpPercap))

# resharp the data for plot
gdp_int.r = melt(gdp_int)
```

```
## Using continent as id variables
```

continent	min_gdpPercap	max_gdpPercap
Africa	241.2	21951
Americas	1201.6	42952
Asia	331.0	113523
Europe	973.5	49357
Oceania	10039.6	34435

- TASK NO.2: Look at the spread of GDP per capita within the continents.

```r
# first we look at the range of gdp
gdp_spread <- gtbl %>%
  group_by(continent) %>%
  summarize(spread_gdpPercap = max(gdpPercap)-min(gdpPercap))
```

continent	spread_gdpPercap
Africa	21710
Americas	41750
Asia	113192
Europe	48384
Oceania	24396

```r
#  Then look at the sd and iqr of the data of GDP per capita within the continents.
gdp_spread2 <- gtbl %>%
  group_by(continent) %>%
  summarize(sd_gdp = sd(gdpPercap), iqr_gdp = IQR(gdpPercap))
gdp_spread2.r = melt(gdp_spread2)
```

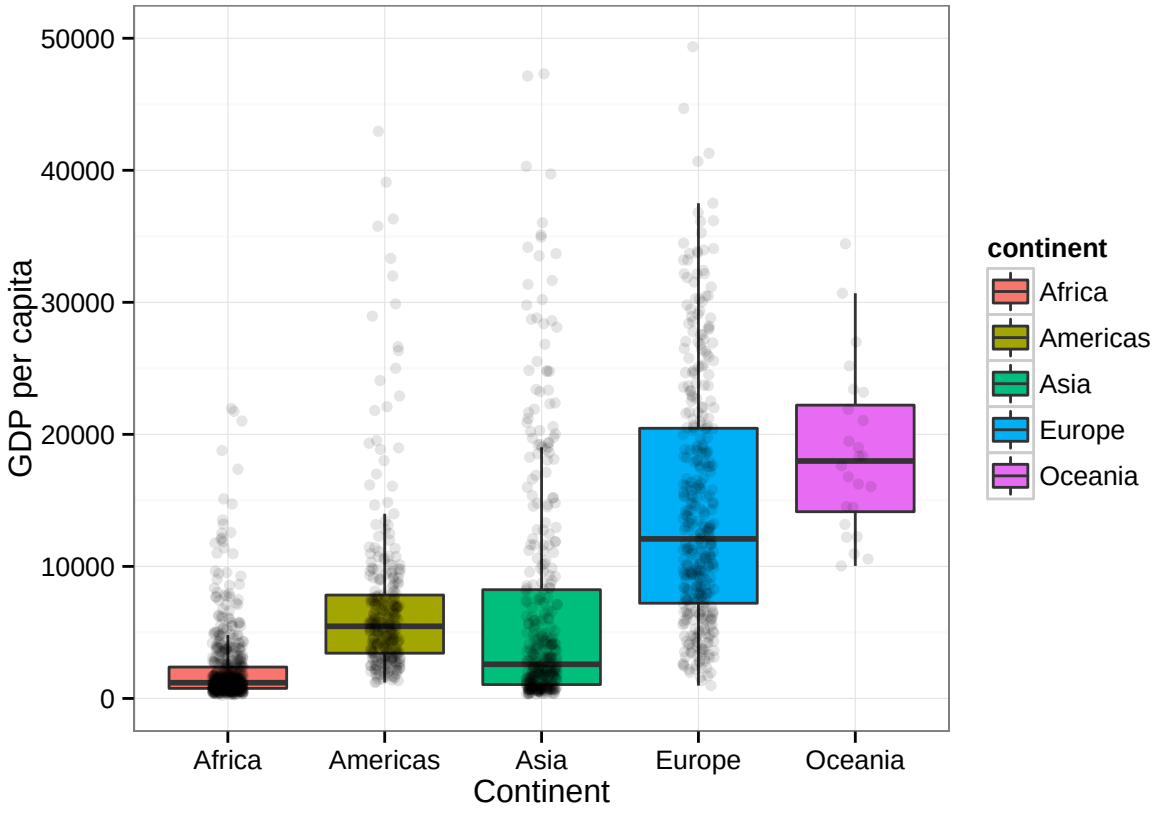
```
## Using continent as id variables
```

continent	sd_gdp	iqr_gdp
Africa	2828	1616

continent	sd_gdp	iqr_gdp
Americas	6397	4402
Asia	14045	7492
Europe	9355	13248
Oceania	6359	8072

```
# also box plot is a good way to see the spread of data
ggplot(gtbl, aes(continent, gdpPercap))+
  geom_boxplot(aes(fill = continent), outlier.shape = NA)+
  geom_jitter(alpha = 0.1, position = position_jitter(width = 0.1))+
  xlab("Continent")+
  ylab("GDP per capita")+
  ylim(c(0,5e4))+
  theme_bw()
```

```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
## Warning: Removed 64 rows containing missing values (geom_point).
## Warning: Removed 25 rows containing missing values (geom_point).
## Warning: Removed 45 rows containing missing values (geom_point).
## Warning: Removed 4 rows containing missing values (geom_point).
## Warning: Removed 1 rows containing missing values (geom_point).
## Warning: Removed 6 rows containing missing values (geom_point).
```



- TASK NO.3: Compute a trimmed mean of life expectancy for different years. Or a weighted mean, weighting by population.

```
# Compute 90% trimmed mean of life expectancy for different years.
lifeExp_tmean <- gtbl %>%
  group_by(year) %>%
  summarize(tmean_lifeExp= mean(lifeExp,trim = 0.05))
```

year	tmean_lifeExp
1952	48.85
1957	51.42
1962	53.64
1967	55.80
1972	57.85
1977	59.89
1982	61.85
1987	63.61
1992	64.81
1997	65.56
2002	66.20
2007	67.56

```
# a weighted mean, weighting by population.
lifeExp_wmean <- gtbl %>%
  group_by(year) %>%
  summarize(wmean_lifeExp= weighted.mean(lifeExp,pop))
```

year	wmean_lifeExp
1952	48.94
1957	52.12
1962	52.32
1967	56.98
1972	59.51
1977	61.24
1982	62.88
1987	64.42
1992	65.65
1997	66.85
2002	67.84
2007	68.92

- TASK NO4: How is life expectancy changing over time on different continents?

```
# use weighted average lifeExp here
lifeExp_mean <- gtbl %>%
  group_by(continent, year) %>%
  summarize(wmean_lifeExp= weighted.mean(lifeExp,pop))
```

continent	year	wmean_lifeExp
Africa	1952	38.80
Africa	1957	40.94
Africa	1962	43.10
Africa	1967	45.18
Africa	1972	47.21
Africa	1977	49.21
Africa	1982	51.02
Africa	1987	52.82
Africa	1992	53.37
Africa	1997	53.28
Africa	2002	53.30
Africa	2007	54.56
Americas	1952	60.24
Americas	1957	62.02
Americas	1962	63.44
Americas	1967	64.51
Americas	1972	65.70
Americas	1977	67.61
Americas	1982	69.19
Americas	1987	70.36
Americas	1992	71.72
Americas	1997	73.19
Americas	2002	74.25
Americas	2007	75.36
Asia	1952	42.94
Asia	1957	47.29
Asia	1962	46.57
Asia	1967	53.88
Asia	1972	57.52
Asia	1977	59.56
Asia	1982	61.57
Asia	1987	63.54

continent	year	wmean_lifeExp
Asia	1992	65.15
Asia	1997	66.77
Asia	2002	68.14
Asia	2007	69.44
Europe	1952	64.91
Europe	1957	66.89
Europe	1962	68.46
Europe	1967	69.55
Europe	1972	70.47
Europe	1977	71.54
Europe	1982	72.56
Europe	1987	73.45
Europe	1992	74.44
Europe	1997	75.71
Europe	2002	77.02
Europe	2007	77.89
Oceania	1952	69.17
Oceania	1957	70.32
Oceania	1962	70.99
Oceania	1967	71.18
Oceania	1972	71.92
Oceania	1977	73.26
Oceania	1982	74.58
Oceania	1987	75.98
Oceania	1992	77.36
Oceania	1997	78.62
Oceania	2002	80.16
Oceania	2007	81.06

- TASK NO5: Report the absolute and/or relative abundance of countries with low life expectancy over time by continent: Compute some measure of worldwide life expectancy – you decide – a mean or median or some other quantile or perhaps your current age. The determine how many countries on each continent have a life expectancy less than this benchmark, for each year.

```
# use median as benchmark
benchmark<-median(gtbl$lifeExp)
lifeExp_abu<-gtbl %>%
  group_by(continent, year) %>%
  filter(lifeExp < benchmark) %>%
```

```
  summarize(n_countries = n_distinct(country))
```

continent	year	n_countries
Africa	1952	52
Africa	1957	52
Africa	1962	52
Africa	1967	51
Africa	1972	50
Africa	1977	50
Africa	1982	46
Africa	1987	41
Africa	1992	40
Africa	1997	44
Africa	2002	41
Africa	2007	41
Americas	1952	19
Americas	1957	16
Americas	1962	13
Americas	1967	13
Americas	1972	10
Americas	1977	7
Americas	1982	5
Americas	1987	2
Americas	1992	2
Americas	1997	1
Americas	2002	1
Asia	1952	30
Asia	1957	27
Asia	1962	26
Asia	1967	25
Asia	1972	20
Asia	1977	16
Asia	1982	12
Asia	1987	10
Asia	1992	8
Asia	1997	7
Asia	2002	5
Asia	2007	3

continent	year	n_countries
Europe	1952	7
Europe	1957	3
Europe	1962	1
Europe	1967	1
Europe	1972	1
Europe	1977	1

- TASK NO6: Find countries with interesting stories.

```
gtbl %>%
  filter(continent == "Asia") %>%
  select(year, country, lifeExp) %>%
  arrange(year) %>%
  group_by(year) %>%
  filter(min_rank(desc(lifeExp)) < 2 | min_rank(lifeExp) < 2)
```

```
## Source: local data frame [24 x 3]
## Groups: year
##
##    year     country lifeExp
## 1  1952 Afghanistan   28.80
## 2  1952      Israel   65.39
## 3  1957 Afghanistan   30.33
## 4  1957      Israel   67.84
## 5  1962 Afghanistan   32.00
## 6  1962      Israel   69.39
## 7  1967 Afghanistan   34.02
## 8  1967       Japan   71.43
## 9  1972 Afghanistan   36.09
## 10 1972       Japan   73.42
## 11 1977    Cambodia   31.22
## 12 1977       Japan   75.38
## 13 1982 Afghanistan   39.85
## 14 1982       Japan   77.11
## 15 1987 Afghanistan   40.82
## 16 1987       Japan   78.67
## 17 1992 Afghanistan   41.67
## 18 1992       Japan   79.36
## 19 1997 Afghanistan   41.76
## 20 1997       Japan   80.69
## 21 2002 Afghanistan   42.13
## 22 2002       Japan   82.00
## 23 2007 Afghanistan   43.83
## 24 2007       Japan   82.60
```

We see that (min = Afghanistan, max = Japan) is the most frequent result.

```
#Compare Afghanistan, Japan and world average
data_avg<- gtbl %>%
  group_by(year) %>%
  summarize(lifeExp= mean(lifeExp))
data_life<-tbl_df(data.frame(year = data_avg$year,
                             Jap = filter(gtbl, country == "Japan")$lifeExp,
                             Afg = filter(gtbl, country == "Afghanistan")$lifeExp,
                             Avg = data_avg$lifeExp))
```

year	Jap	Afg	Avg
1952	63.03	28.80	49.06
1957	65.50	30.33	51.51
1962	68.73	32.00	53.61
1967	71.43	34.02	55.68
1972	73.42	36.09	57.65
1977	75.38	38.44	59.57
1982	77.11	39.85	61.53
1987	78.67	40.82	63.21
1992	79.36	41.67	64.16
1997	80.69	41.76	65.01
2002	82.00	42.13	65.69
2007	82.60	43.83	67.01

Next, we want to find the country experiencing the sharpest 5-year drop in life expectancy.

```
gtbl %>%
  group_by(continent, country) %>%
  select(country, year, continent, lifeExp) %>%
  mutate(le_delta = lifeExp - lag(lifeExp)) %>%
  summarize(worst_le_delta = min(le_delta, na.rm = TRUE)) %>%
  filter(min_rank(worst_le_delta) < 2) %>%
  arrange(worst_le_delta)
```

```
## Source: local data frame [5 x 3]
## Groups: continent
##
##   continent     country worst_le_delta
## 1    Africa      Rwanda        -20.421
## 2      Asia    Cambodia         -9.097
## 3  Americas El Salvador         -1.511
## 4    Europe  Montenegro         -1.464
## 5   Oceania   Australia          0.170
```

For above five countries, we have:

```
data_drop <- tbl_df(data.frame(year = data_avg$year,
                               Rwa = filter(gtbl, country == "Rwanda")$lifeExp,
                               Cam = filter(gtbl, country == "Cambodia")$lifeExp,
                               ES  = filter(gtbl, country == "El Salvador")$lifeExp,
                               Mon = filter(gtbl, country == "Montenegro")$lifeExp,
                               Aus = filter(gtbl, country == "Australia")$lifeExp,
                               Avg = data_avg$lifeExp))
```

year	Rwa	Cam	ES	Mon	Aus	Avg
1952	40.00	39.42	45.26	59.16	69.12	49.06
1957	41.50	41.37	48.57	61.45	70.33	51.51
1962	43.00	43.41	52.31	63.73	70.93	53.61
1967	44.10	45.41	55.85	67.18	71.10	55.68
1972	44.60	40.32	58.21	70.64	71.93	57.65
1977	45.00	31.22	56.70	73.07	73.49	59.57
1982	46.22	50.96	56.60	74.10	74.74	61.53
1987	44.02	53.91	63.15	74.86	76.32	63.21
1992	23.60	55.80	66.80	75.44	77.56	64.16
1997	36.09	56.53	69.53	75.44	78.83	65.01
2002	43.41	56.75	70.73	73.98	80.37	65.69
2007	46.24	59.72	71.88	74.54	81.23	67.01

We also want a special analysis of Rwanda.

```
# analysis for Rwanda
data_Rwa = gtbl %>% filter(country == "Rwanda")
```

country	year	pop	continent	lifeExp	gdpPercap
Rwanda	1952	2534927	Africa	40.00	493.3
Rwanda	1957	2822082	Africa	41.50	540.3
Rwanda	1962	3051242	Africa	43.00	597.5
Rwanda	1967	3451079	Africa	44.10	511.0
Rwanda	1972	3992121	Africa	44.60	590.6
Rwanda	1977	4657072	Africa	45.00	670.1
Rwanda	1982	5507565	Africa	46.22	881.6
Rwanda	1987	6349365	Africa	44.02	848.0
Rwanda	1992	7290203	Africa	23.60	737.1
Rwanda	1997	7212583	Africa	36.09	589.9
Rwanda	2002	7852401	Africa	43.41	785.7
Rwanda	2007	8860588	Africa	46.24	863.1

We notice that population and Gdp also experienced a big decrease in 1990s. After googling, we think the reason must be *Rwandan Genocide*.

My experience and workflow

1. `dplyr` is indeed a power tool for the analysis. Some of its grammer are similar that of `sql`. With some experience of using sql, I think I am really quick in understanding the functions in `dplyr`.
2. This time we continue our application of `ggplot`. One preblem I meet is to draw a bar graph with the data in two columns side-by-side. Although I have already learned about `position="dodge"`, it still took me some time as I finally found I need to resharp the data. See this on stackoverflow for more detail.
3. To acheive the task of put a figure and relevant table right next to each other. It need some code in html. The following code may help if anyone need

```
#```{r, results='asis', echo=FALSE, out.extra=''}
#cat("<table class='container'><tr>")
#cat("<td>")
#kable(data_Rwa)
#cat("</td>")
#cat("<td>")
#ggplot(data_Rwa, aes(year, pop)) +
#  ggtitle("Rwanda Population")+
#  geom_point(color="red")+
#  geom_line(color="blue")+
#  theme_bw()+
#  theme(plot.title = element_text(lineheight=.8, face="bold"))
#ggplot(data_Rwa, aes(year, gdpPercap)) +
#  ggtitle("Rwanda gdpPercap")+
#  geom_point(color="red")+
#  geom_line(color="blue")+
#  theme_bw()+
#  theme(plot.title = element_text(lineheight=.8, face="bold"))
#cat("</td>")
#cat("</tr></table>")
#```
```