STAT 545 - Assignment 2

Sean La

23/09/2019

Exercise 1

Exercise 1.1

Let's filter the dataset for the countries Canada, Vietnam, and China in the 1970s.

```
dataset <- gapminder %>% filter(country %in% c('Canada','Vietnam','China'),
                                year >= 1970,
                                year < 1980)
```

Exercise 1.2

Now we select the variables `country` and `gdpPercap`.

```
dataset <- dataset %>% select('country','gdpPercap')
```

Exercise 1.3

Let's add a new column containing the difference in life expectancy from the previous entry.

```
dataset <- gapminder %>% mutate(lifeExpDiff = c(0,diff(lifeExp)))
```

Now, let's filter for rows where the life expectancy difference is negative and the previous row concerns the same country.

```
negative_life_exp_dataset <- dataset %>% filter(lifeExpDiff < 0,
                                                country == lag(country,1))
```

Exercise 1.4

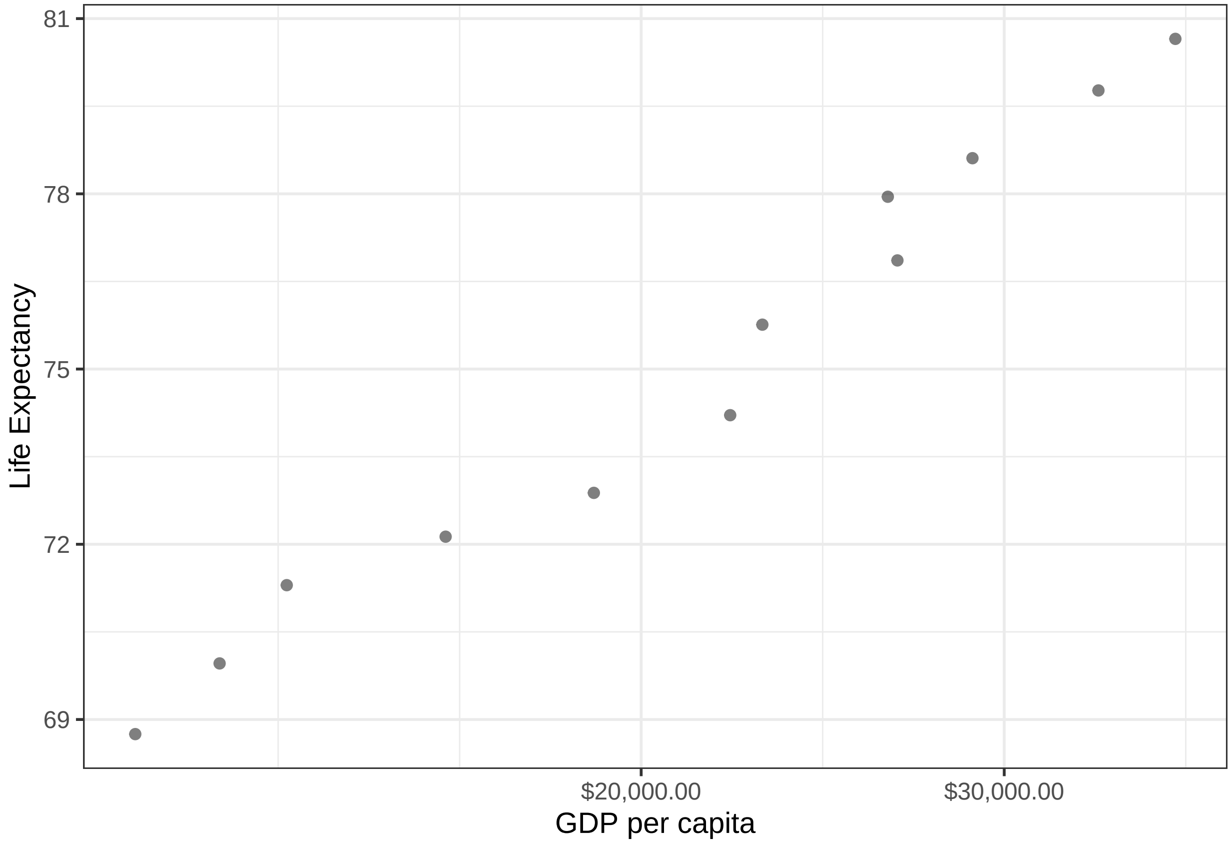
We will find the rows corresponding to each countries' maximum GDP per capita.

```
max_gdpPercap <- gapminder %>% group_by(country) %>% filter(gdpPercap == max(gdpPercap))
```

Exercise 1.5

Let's make this plot!

```
gapminder %>%
  filter(country == 'Canada') %>%
  ggplot(aes(gdpPercap,lifeExp)) +
    geom_point(alpha=0.5) +
    scale_x_log10("GDP per capita", labels = scales::dollar_format()) +
    theme_bw() +
    ylab("Life Expectancy")
```



Exercise 2

Let's explore the relationship between continent and change in GDP per capita across years.

First, let's make a new column `gdpPercapChange` in the `gapminder` dataset, which will be the change in GDP per capita from the previous entry/year, for only those rows where the previous row is of the same country. Otherwise, if the previous row describes a different country, we'll just set `gdpPercapChange` to be 0.

```
dataset <- gapminder %>%
  group_by(country) %>%
  mutate(gdpPercapChange = c(0,diff(gdpPercap))) %>% ungroup()
```

Let's make a quick summary of change in GDP per capita vs continent using `dplyr`, where we'll look at the minimum, maximum, and range of change of GDP per capita within each continent.

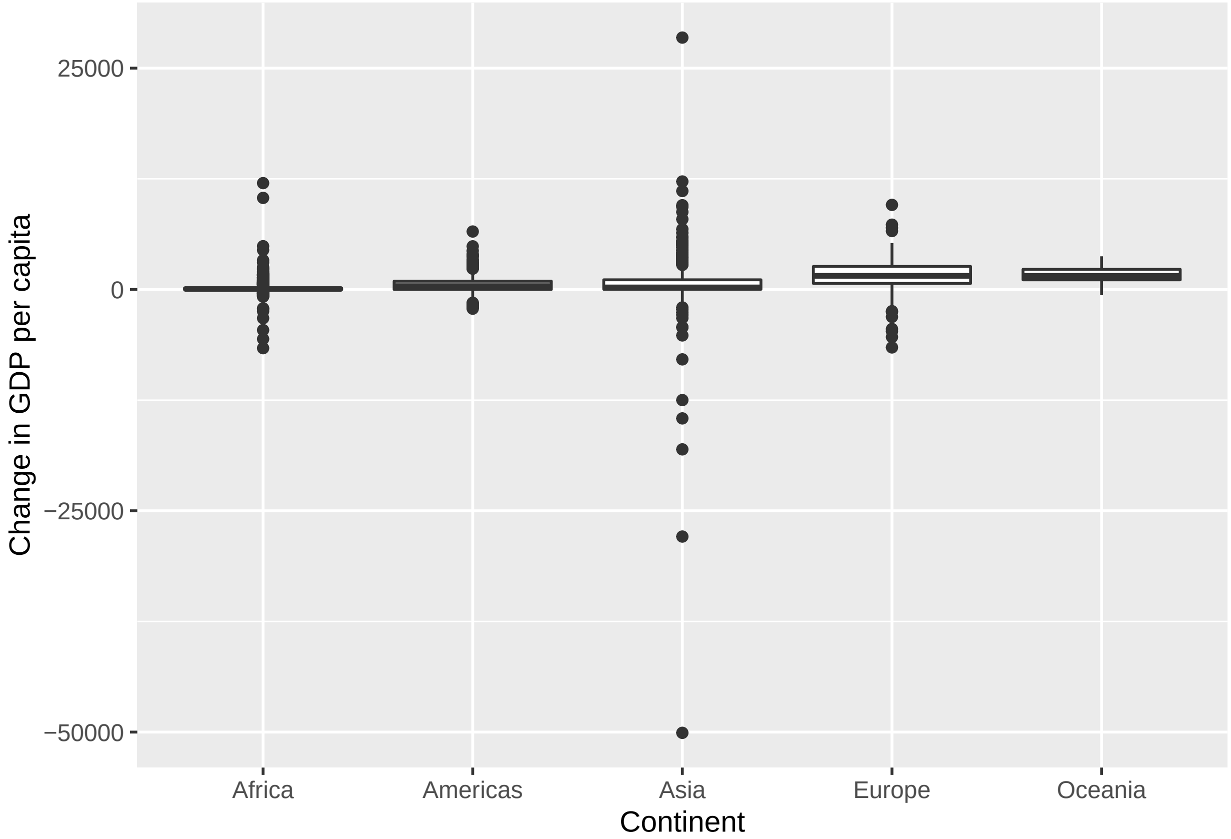
```
dataset %>% group_by(continent) %>% summarize(min_gdpPercapChange = min(gdpPercapChange),
                                              max_gdpPercapChange = max(gdpPercapChange),
                                              range_gdpPercapChange = max(gdpPercapChange) -
                                                                      min(gdpPercapChange),
                                              mean_gdpPercapChange = mean(gdpPercapChange),
                                              sd_gdpPercapChange = sd(gdpPercapChange))
```

```
## # A tibble: 5 x 6
##   continent min_gdpPercapCh~ max_gdpPercapCh~ range_gdpPercap~
##   <fct>                <dbl>            <dbl>            <dbl>
## 1 Africa              -6632.           12016.           18648.
## 2 Americas              -2170.              6548.              8718.
## 3 Asia                 -50082.             28453.             78535.
## 4 Europe                -6546.              9555.             16101.
## 5 Oceania                -644.              3748.              4391.
## # ... with 2 more variables: mean_gdpPercapChange <dbl>,
## #   sd_gdpPercapChange <dbl>
```

Interesting! It seems that countries in Africa and Asia, which mostly consist of developing countries, have the greatest rates of change in GDP per capita, as compared to the Americas and Europe. Africa and Asia also have the greatest decrease in GDP per capita, and range of the change in GDP per capita as well. While Europe and Oceania's change in GDP fall within a relatively modest range as compared to Africa and Asia, their mean change in GDP per capita is greater than Africa and Asia's.

Now, let's make some boxplots to visualize the distributions of change in GDP per capita for each continent.

```
dataset %>%
  ggplot(aes(continent,gdpPercapChange)) +
    geom_boxplot() +
    xlab('Continent') +
    ylab('Change in GDP per capita')
```



As we anticipated, the distributions for Africa and Asia are more spread out compred to the Americas and Europe. In particular, Asia seems to have an outlier, with a change in GDP per capita of -50000. Let's find out which entry this corresponds to.

```
dataset %>% filter(gdpPercapChange == min(gdpPercapChange)) %>% print()
```

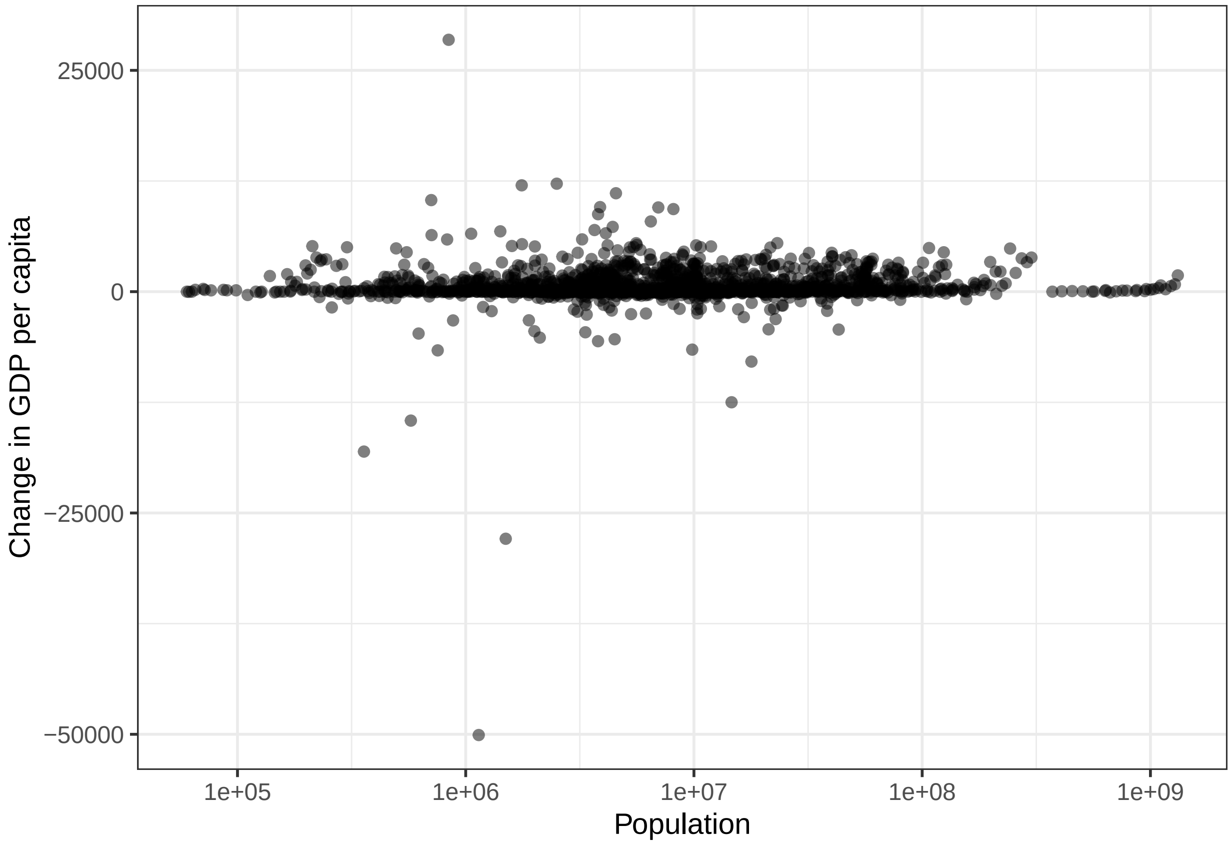
```
## # A tibble: 1 x 7
##   country continent  year lifeExp     pop gdpPercap gdpPercapChange
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>           <dbl>
## 1 Kuwait  Asia       1977    69.3 1140357    59265.         -50082.
```

It's Kuwait! Fascinating.

Exercise 3

Let's continue our analysis from the previous exercise with change in GDP per capita. Let's compare how population size (on a logartithm scale) relates to change in GDP per capita among Asian countries.

```
dataset %>%
  ggplot(aes(pop,gdpPercapChange)) +
  geom_point(alpha=0.5) +
  scale_x_log10() +
  theme_bw() +
  xlab("Population") +
  ylab("Change in GDP per capita")
```

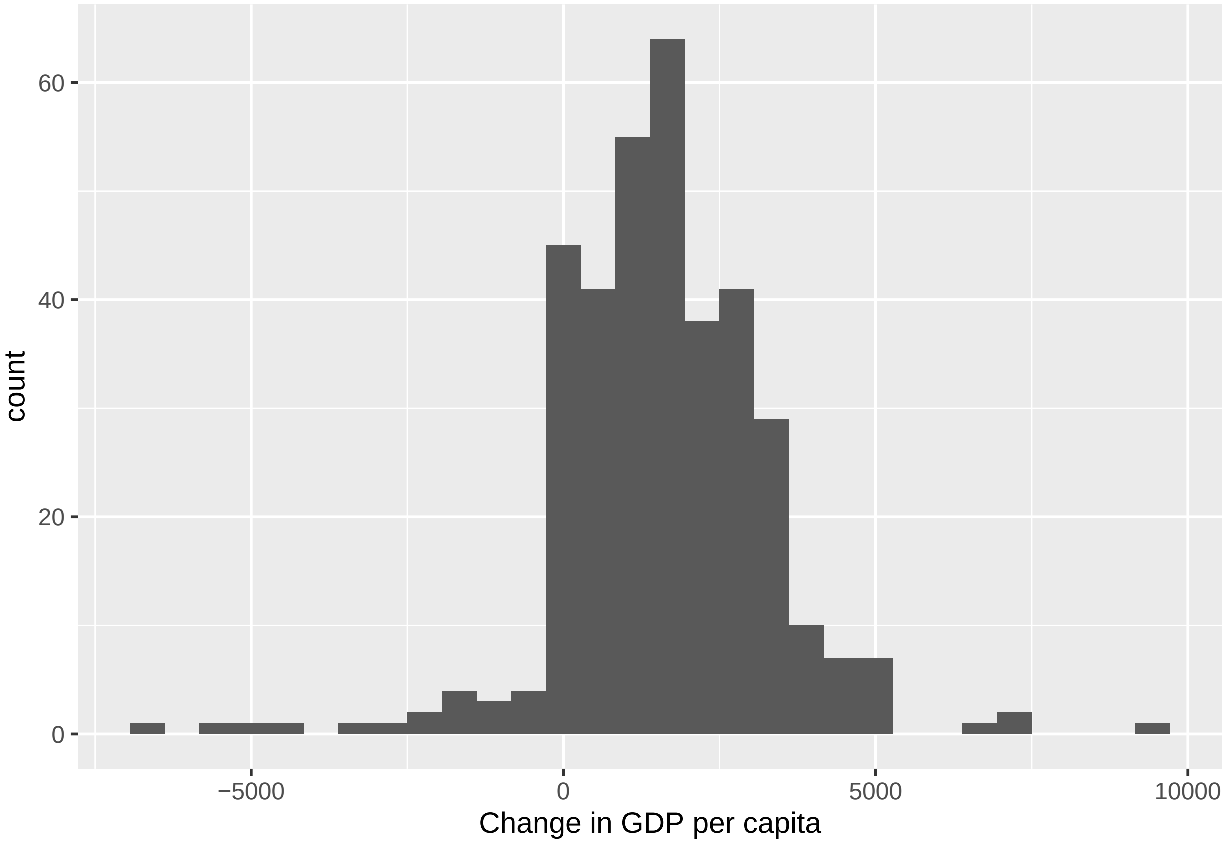


There doesn't seem to be a relationship. Oh well!

I'm still interested in Europe's change in GDP per capita. I wonder if it follows a normal distribution? Let's make a histogram!

```
dataset %>% filter(continent == 'Europe') %>%
  ggplot(aes(gdpPercapChange)) +
  geom_histogram() +
  xlab('Change in GDP per capita')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Ah, I'd say it's roughly normal.

Recycling (Optional)

The line of code is given as

```
filter(gapminder, country == c("Rwanda", "Afghanistan"))
```

It seems that the analyst was hoping to filter the `gapminder` dataset for entries for Rwanda and Afghanistan. They didn't succeed - in fact, the result would be an empty tibble because the filter command is looking for entries whose country is the *vector* `c("Rwanda","Afghanistan")`, of which there is none. The correct thing to do would be to use the command

```
filter(gapminder, country %in% c("Rwanda", "Afghanistan"))
```