Homework Assignment 7: Possible Answer

Maria Kunevich

2022-11-15

Introduction

In this report I am going to explore the dataset `gapminder` from the `R` package `gapminder`. Additional documentation can be found here:
https://www.rdocumentation.org/packages/gapminder/versions/0.3.0

The `gapminder` data frame includes six variables:
- `country`
- `continent`
- `year`
- `lifeExp` (meaning Life Expectancy at birth)
- `pop` (meaning Population)
- `gdpPercap` (meaning per-capita GDP given in international dollars)

First, I am downloading the package from CRAN (activating the library) to explore the dataset.

```
## # A tibble: 6 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.
```

There are 6 columns with the above mentioned variables and 1,704 entries.
For the `continent` variable there are 5 levels:

```
## [1] "Africa"   "Americas" "Asia"     "Europe"   "Oceania"
```

For the `country` variable there are 142 levels, which means we have 142 different countries in the dataset:

For the `year` variable we have 1704 entries which repeat years 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, 2002, 2007 for each country:

```
## [1] 1704
```

Questions for EDA

Using `dplyr` package, I am building a summary to view the average `lifeExp` for each continent.

```
##   continent lifeExp
## 1    Africa 47.7920
## 2  Americas 67.0480
## 3      Asia 61.7915
## 4    Europe 72.2410
```

```
## 5   Oceania 73.6650
```

In the same line, I am building a summary to view the average `gdpPercap` for each continent.

```
##   continent gdpPercap
## 1    Africa  1192.138
## 2  Americas  5465.510
## 3      Asia  2646.787
## 4    Europe 12081.749
## 5   Oceania 17983.304
```

It looks like there is some relationship between life expectancy and the amount of GDP for each continent (the continents have the same order for both variables) going from lowest to the highest:
Africa, Asia, Americas, Europe, Oceania

However, there are few data points (5 for each variable), so any conclusion based on the correlation between these two variables might be misleading due to high sampling error.

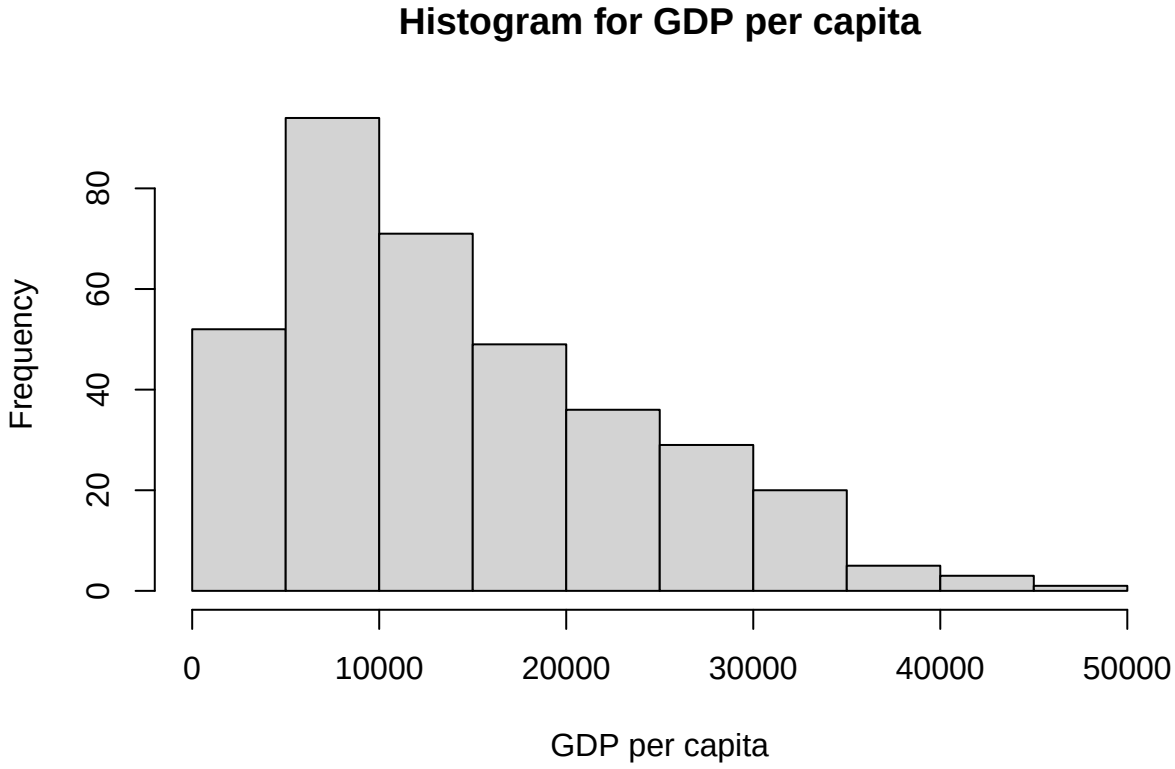
In my further analysis I will concentrate only on one continent: *Europe* and explore whether there is a relationship between `Life expectancy` and `GDP per Capita` for this continent.

Description of data transformation

I am using `dlyr` package to transform the dataset and create a new dataset with the data for only one continent - *Europe*.

The new tibble includes 360 observations for 4 variables: `continent` (Europe), `year`, `lifeExp`, `gdpPercap`. By briefly exploring all variables, no abnormalities are detected, so I move to the analysis of correlation. To make analysis easier, I also convert both variables to integers and plot histograms for `Life expectancy` and `GDP per Capita`.

```
## # A tibble: 360 x 4
##    continent  year lifeExp gdpPercap
##    <fct>     <int>   <int>     <int>
##  1 Europe     1952      55      1601
##  2 Europe     1957      59      1942
##  3 Europe     1962      64      2312
##  4 Europe     1967      66      2760
##  5 Europe     1972      67      3313
##  6 Europe     1977      68      3533
##  7 Europe     1982      70      3630
##  8 Europe     1987      72      3738
##  9 Europe     1992      71      2497
## 10 Europe     1997      72      3193
## # ... with 350 more rows
```



It looks like the histogram for `lifeExp` is negatively skewed and the histogram for `gpdPercap` is positively skewed.

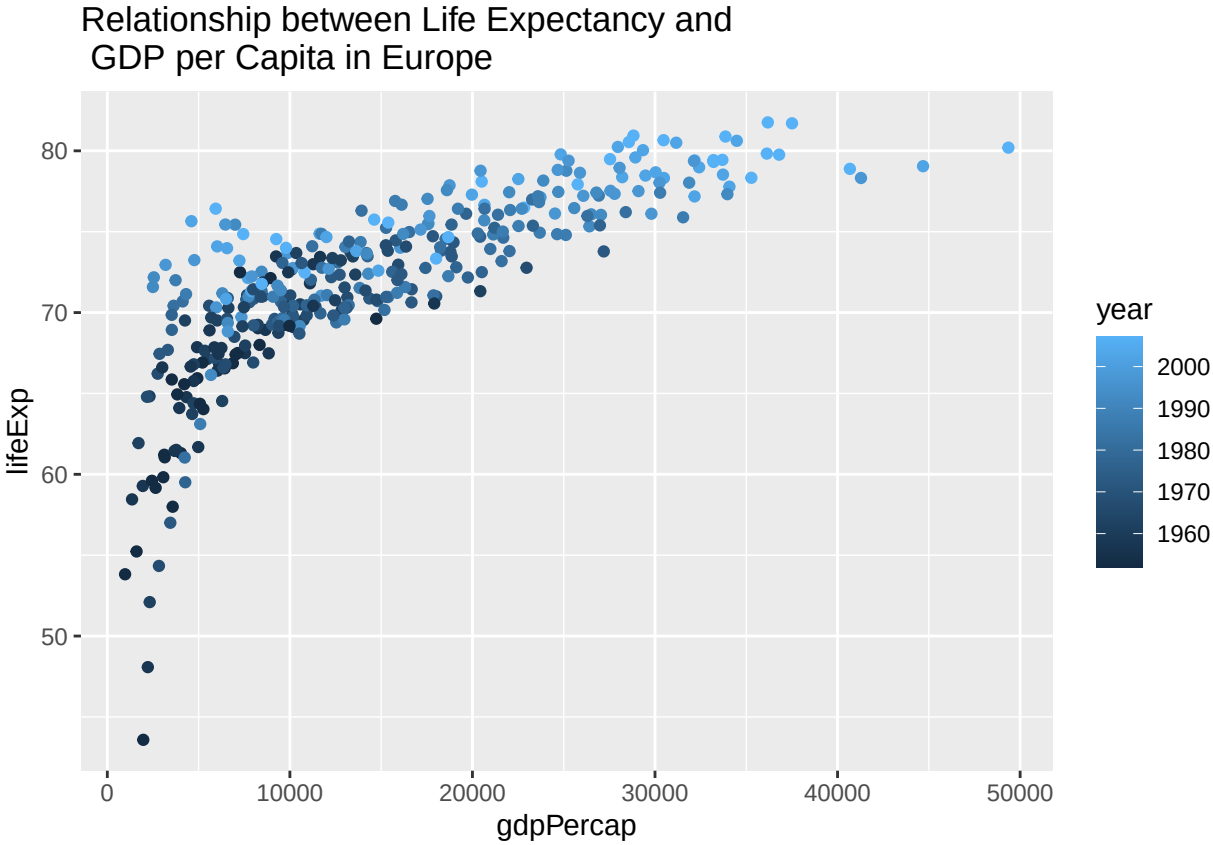
Pearson correlation analysis

Correlation test is used to evaluate the association between two or more variables. In my new dataset there are two interval continuous variables: *lifeExp* and *gdpPercap*, so I am interested in the correlation between these two variables. My *predictor* variable is `gdpPercap` and my dependent variable is `lifeExp`. So I expect that

Research Question:

Is there a relationship between life expectancy in different countries in Europe and the amount of gross domestic product in international dollars for these countries?

First, I visualise the data and run preliminary test to check the test assumptions:



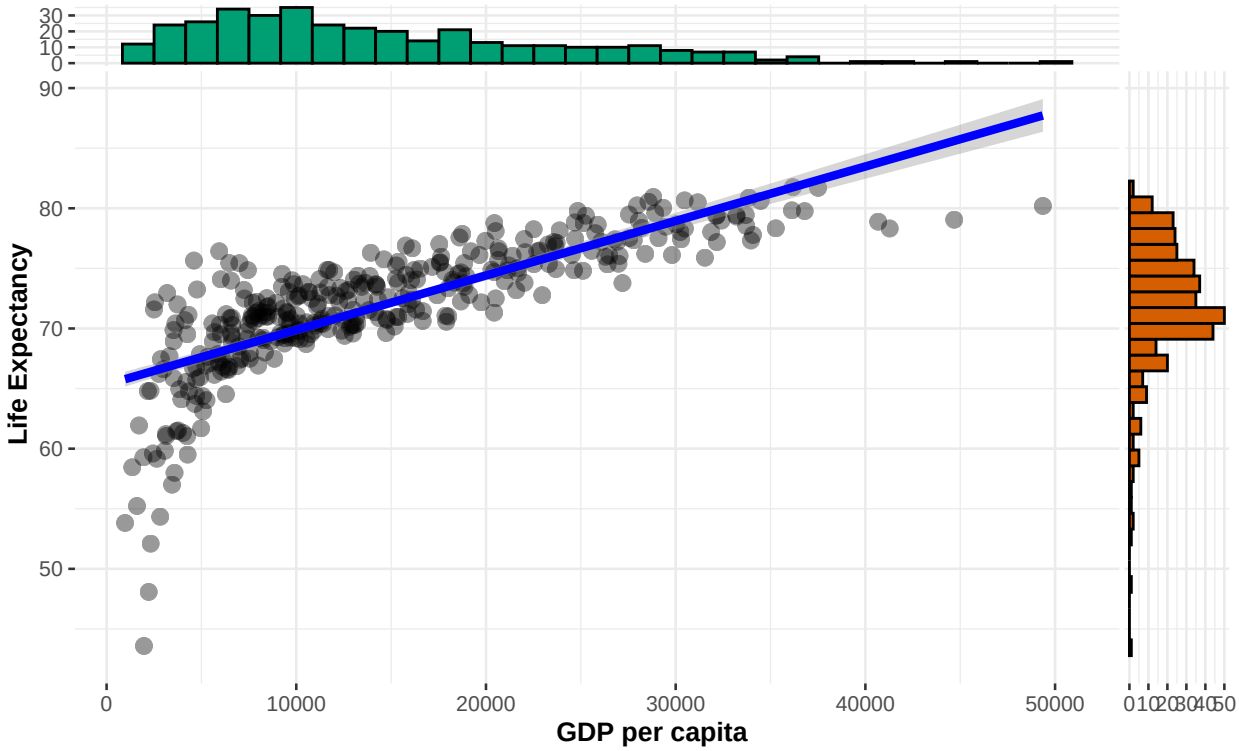
1. Is the covariation linear? Yes, from the plot above, the relationship is linear, it can be characterised as strong positive relationship.

2. Are the data from each of the two variables (*lifeExp* and *gdpPercap*) follow a normal distribution? Since it's a large sample we can assume normality: if data sample is above 30 data points, according to the Central Limit theorem the data can be assumed to be normal.

```
## You can cite this package as:
##      Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach.
##      Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167
```

```
## Registered S3 method overwritten by 'ggside':
##   method from
##   +.gg   ggplot2
```

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



I can conclude that both populations may come from normal distributions even if they are skewed as the sample is large enough. Assuming normality we can run Pearson correlation test.

Null hypothesis: the correlation coefficient is not significantly different from 0. There is no significant linear relationship between `gdpPercap` and `lifeExp` in the population.

Alternative hypothesis: the population correlation coefficient is significantly different from 0. There is a significant linear relationship between `gdpPercap` and `lifeExp` in the population.

```
## [1] 0.7807831
```

```
##
##  Pearson's product-moment correlation
##
## data:  data.europe$lifeExp and data.europe$gdpPercap
## t = 23.644, df = 358, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7368907 0.8181201
## sample estimates:
##       cor
## 0.7807831
```

Correlation test between `lifeExp` and `gdpPercap` for European countries reveals the following:
t is the t-test statistic value (t = 23.64),
df is the degrees of freedom (df= 358),
p-value is the significance level of the t-test (p-value = < 2.2e-16),
conf.int is the confidence interval of the correlation coefficient at 95% (conf.int = [0.7368907, 0.8181201]),
sample estimates is the correlation coefficient (cor = 0.78).

Conclusion:

Pearson correlation test revealed that the amount of gross domestic product in international dollars and life expectancy in European countries are significantly positively correlated, $r(358)=.78$, $p < 0$. This correlation is strong and positive, meaning that with increased amount of GDP the life expectancy also increases.