Homework Assignment 7

Maria Kunevich

2022-11-10

Assignment Description

Download the `gapminder` data set into `R Markdown` session.

1. Perform quick EDA and pick up two variables you want to explore in more depth (for example, life expectancy and gdp) and a subset of data set (for instance, only certain continents, or countries, etc).

```
## # 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.
```

2. Prepare the data set that includes only variables of your interest in a suitable format for analysis (use `dlyr` package and `tidyr` when necessary).

3. Explore two variables and how they are associated with each other (correlation analysis).

3.1 Visualize your data using a scatter plot and include the description of assumptions for correlation analysis:
- Is the co-variation linear?
- Are the data from each of the 2 variables (x, y) follow a normal distribution (visual inspection of the data normality using histograms)?

3.2 Calculate correlation coefficient and provide your interpretation.

4. Hypothesis testing.
   4.1 State the null hypothesis and the alternative hypothesis.
   4.2 Report on collected data and sample size.
   4.3 Perform Pearson correlation test between two variables.
   4.4 Decide whether to reject or fail to reject your null hypothesis, report selected significance level.
   4.5 Interpret and report the results.

5. Create a report in `R Markdown` with the following sections:

- Introduction (brief description of the data set and variables)
- Description of data transformation
- Correlation analysis (steps for visualisation, checking assumption for correlation analysis, interpretation of correlation coefficient)
- Hypothesis testing: using the Pearson r statistic to conduct hypothesis tests about population correlation values
- Reporting the results