Exercise 6.1 (Revised)

Mnguni Zulu

13.01.2024

Life Expectancy Data

Datasets:

1. **World Bank Life Expectancy Data**

Data Source:

The data is open source and was made available for download from Kaggle.com. It is a dataset external to the project author. The dataset was compiled from separate datasets available from the Worldbank open-source database. It can be considered as moderately reliable because although the dataset was not compiled by myself, it has as its underlying source the Worldbank database. This is something I was able to confirm, by accessing some of the relevant information directly from the Worldbank webpage.

I chose this data source because it has many numeric variables which can be plotted, to reveal insights. The relationship between the variables and their impact on life expectancy is interesting and relevant event today.

Data Characteristics:

The dataset in question from an external data source. It is a dataset containing over 2,469 rows and contains 17 columns. Excluding 'Year', there are 13 numerical columns, and the rest are categorical [Country, Continent, Least Developed]. The dataset is timely as it was released in December 2023.

The data provides variables such as life expectancy, beer consumption per capita, forest area, income and more, for 119 countries.

Data Owner: Ramin Rzayev

Links to data:

https://www.kaggle.com/datasets/raminrzayev/life-expectancy-2000-2020/data

Data Limitations & Ethics:

The dataset is from 2000 to 2020, which excludes the recent COVID pandemic. Although it is good to exclude the period from analysis (because it will skew the analysis), it would e nice to see whether the post-COVID data will have changed significantly since the pandemic.

The dataset being worked with was compiled using various datasets at the Worldbank. The Worldbank itself, had to gather the data from various sources in different countries. Of course, this means we cannot be 100% certain of every aspect of this dataset, but we can trust that the Worldbank will have the best chance of retrieving reliable data.

There is no sensitive personal data in the dataset. Care should be taken to not exert bias in the analysis, which could happen. Especially when looking at income, or similar variables which seem to favour one set of countries over another.

Data Profile

Worldbank Life Expectancy Dataset

A) **Shape**
   - 2,469 rows
   - 17 columns

B) **Missing Values**
   - A breakdown of missing values by column:

Column Name	**Nr.**
Health Expenditure	9
Electric Power Consumption	565
Obesity	452
Beer Consumption	117

Action Taken: The data frame was limited to the years 2000 to 2015. In this way the missing values have been excluded from the dataset, without the need for imputation of calculated averages.

C) **Mixed values present**
   - There are no mixed data types in any of the columns in the dataset.

D) **Duplicates**
   - There were no duplicate rows found in the dataset.

Possible Analysis Questions:

- How does higher or lower beer consumption per capita influence the life expectancy?
- What is the average life expectancy in least developed countries?
- How is obesity linked to beer consumption per capita?
- How is GDP per capita linked to CO2 emission, and does this affect life expectancy?
- Is there a positive correlation between Health Care expenditure and Life Expectancy? Which direction and how strong is the correlation?