MA317 Final Assignment

Jasper Yang

12/6/2022

## Question 1

1. Exploratory Data Analysis: describe the data and deal with missing values
2. Analyse using descriptive statistics (both graphical and numerical representations) and R the Life Expectancy data1.csv dataset. [12 marks]

## [1] 217 29

## vec  
## character   
## 29

## Table by Continents

table1\_table

| Continent | Number of Countries | Life Expectancy at Birth (years) | Primary Education Attainment Age 25+ (%) | Health Expenditure (% of GDP) | Population Density (People per sq. km) | GDP per Capita (current international $) |
| --- | --- | --- | --- | --- | --- | --- |
| Africa1 | 54 | 64.1 (5.9) | 86.3 (NA) | 86.3 (NA) | 104.8 (133.2) | 3.3 (3.3) |
| Asia | 50 | 74.6 (5.1) | 82.8 (16.2) | 82.8 (16.2) | 949.5 (3069.1) | 3.7 (4.1) |
| Australia/Oceania | 19 | 73.5 (6.0) | NaN (NA) | NaN (NA) | 145.2 (152.8) | 1.7 (4.3) |
| Europe | 48 | 79.3 (3.6) | 98.7 (2.0) | 98.7 (2.0) | 650.7 (2819.2) | 2.9 (1.5) |
| North America | 34 | 76.2 (3.8) | 73.5 (19.9) | 73.5 (19.9) | 281.6 (295.5) | 1.7 (2.5) |
| South America | 12 | 75.1 (3.2) | 86.3 (9.9) | 86.3 (9.9) | 24.7 (18.4) | 1.3 (2.0) |
| Total | 217 | 72.9 (7.5) | 87.7 (15.1) | 87.7 (15.1) | 446.0 (1996.6) | 2.8 (3.2) |
| 1Missing data: | | | | | | |

1. Many predictors in the dataset contain missing values. Is deleting predictor variables with many missing values an appropriate method to deal with missing values? Choose a method to deal with the missing values and then employ this method to the life expectancy data. Justify your choice. Additionally, there are some countries (cases) where the value of Life expectancy is missing. Explain how you will handle this problem. [13 marks]
2. Collinearity increases the variance of the estimators and hence, reduces the adequacy of the model. When collinearity is present, how do you solve this problem? Investigate collinearity between the predictor variables in the LifeExpectancyData1.csv dataset. [12 marks]

library(faraway)  
#vif(LE\_data)

1. To understand better life expectancy and the factors that affect it, suggest the best linear model which predicts life expectancy in 2020. Interpret and evaluate the suggested model. [25 marks]
2. Using the same dataset and using the additional feature Continent, employ an appropriate experimental design to study differences of average life expectancies across the continents: Asia, Europe, North America, South America, Africa, Australia/Oceania. Justify your choice of experimental design and methods. [13 marks]