Project Proposal

Pooja Kabber, Dingkun Yang, Echo Chen, Andrew Kroening

October 21st, 2022

For part 1, you will conduct exploratory data analysis on your selected dataset. You are required to produce a report of your exploratory data analysis findings in R Markdown. The report should be at most five pages. Tables and figures should be well formatted with clear labels and descriptions. You can organize the report as follows

Data Overview

Characteristics of the dataset, sample size, number of variables. Include questions here.

Primary Relationship of Interest

Present descriptive statistics and exploratory plots in whichever format you think is best (tables, figures) for your primary relationship of interest (dependent variable and primary independent variable, if applicable). Describe your findings.

Other Characteristics

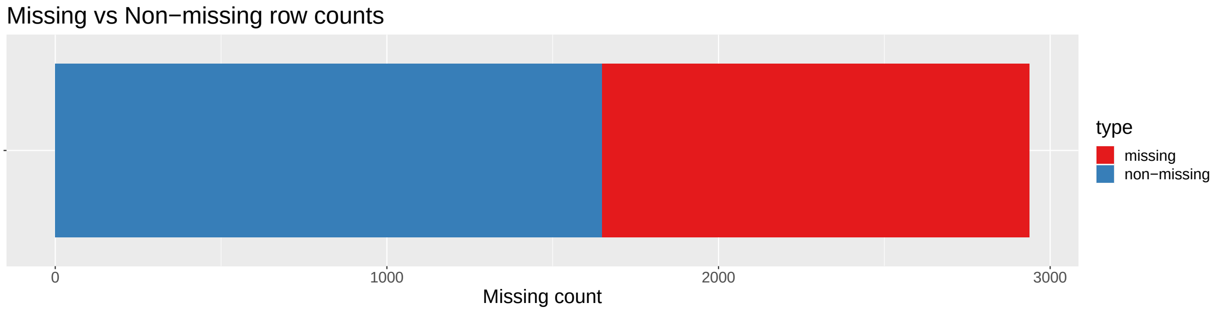
Briefly describe other variables in the data. If there are many, do not list them all. Rather, describe the types of variables that are present (e.g., "demographic information").

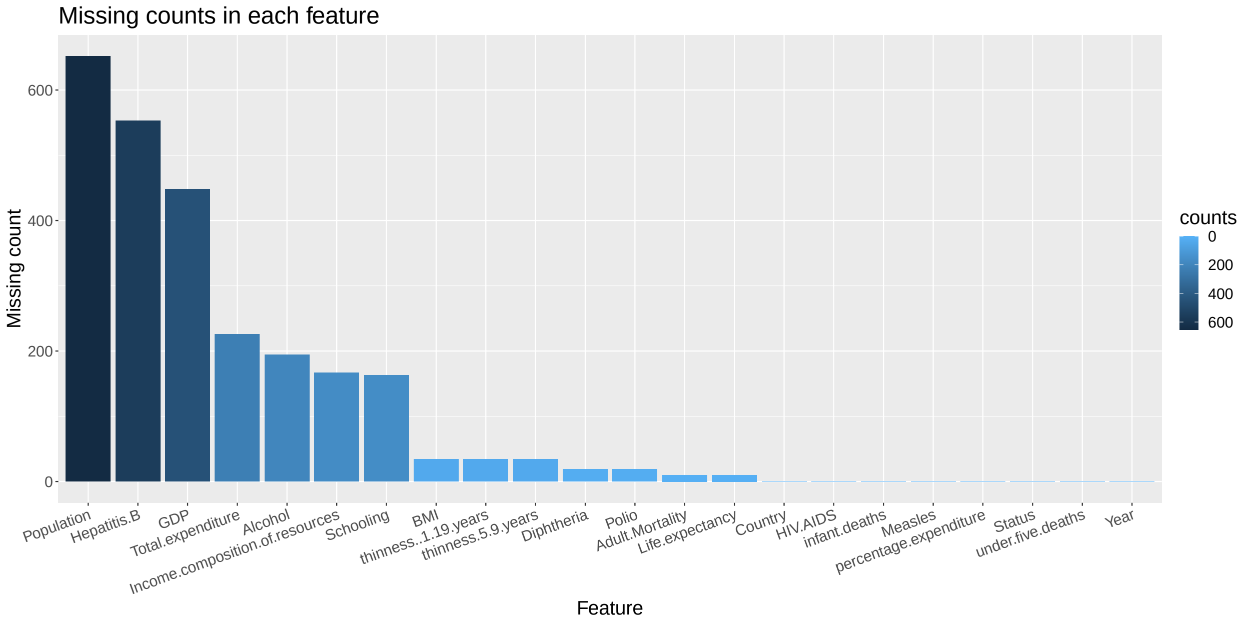
Potential Challenges

Describe aspects of the data that may present challenges in the modeling stage. For example, might certain categorical variables need to be collapsed? Is there a lot of missingness? Could the size of the dataset present model selection challenges?

Challenge 1

First challenge is the huge amount of missing data.As we can see, around 44% of the total data is missing, We want to study the characteristics.





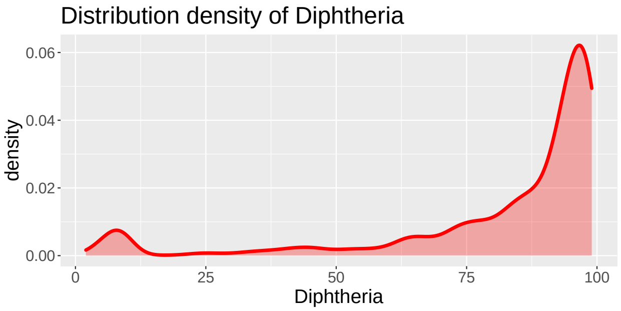
Potential solution The most significant amount of missing values is recorded in population, GDP, Hepatitis B, followed by Total expediture, Achohol, Income.composition.of.resources and Schooling.

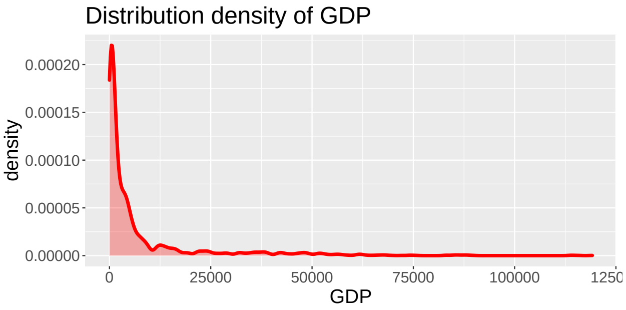
- Cosidering 40% of the misiing data, we can apply data imputation by checking outliers in each variable that contains missings using boxplots:
- for the variables with high outliers will apply imputation with median
- for the variables with low outliers will apply imputation with mean.

Challenge 2

When we check the density of different variables. For example, we find diphtheria is left-skewed,and GDP if right-skewed.

Potential solution We can use log transformation on skewed variables. We can also remove outliers or normalize(min-max) our dataset.





Appendix

Variable	Type	Description
Country	factor	Country name
Year	numeric	Year of the data
Status	factor	Country status of developed or developing
Life_Expectancy	numeric	Life expectancy in age
Adult_Mortality	numeric	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
infant.deaths	numeric	Number of Infant Deaths per 1000 population
Alcohol	numeric	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
percentage.expenditure	numeric	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
Hepatitis.B	numeric	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	numeric	number of reported cases per 1000 population
BMI	numeric	Average Body Mass Index of entire population
under.five.deaths	numeric	Number of under-five deaths per 1000 population
Polio	numeric	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total.expenditure	numeric	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	numeric	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)room)
HIV.AIDS	numeric	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	numeric	Gross Domestic Product per capita (in USD)
Population	numeric	Population of the country
thinness..1.19.years	numeric	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
thinness.5.9.years	numeric	Prevalence of thinness among children for Age 5 to 9(%)
Income.composition.of.resources	numeric	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	numeric	Number of years of Schooling(years)