Analysis of Total wealth of countries across the world

Team Members: Shucheng Zhang, Weihua Pan, Yufei Sun, Rishabh Chawla

Overview Abstract

Based on the rapid development of technology, it's been an amazing start to 2023 with the growth of artificial intelligence, neural science, medical and pharmaceutical engineering, nuclear and sustainable clean energy. However, there has always been a significant problem for the human beings during the recent decades because of the continued fossil energy exploitation, the problem about our future, our destiny should be considered priorly. Hence, in this project, the Wealth Accounts dataset will be analyzed and a comprehensive view of different country's wealth composition including natural, produced, human capital will be provided and researched. We are dedicated to first visualize the dataset by performing EDA, data cleaning and preprocessing, generating scatterplots, bar charts, histograms, etc. Then, find any correlations and patterns between data by applying multiple machine learning models, providing insights towards the results. Finally, predict the future trend of a country's wealth status among their natural, human, produced resources to give suggestions and help the policymakers to make decisions on sustainable development and resource allocation.

Data Preparation

1. Collect and import data: This dataset originates and has been collected from THE WORLD BANK, DATA BANK. Variables including 146 countries, 52 attributes, and 24 available years of data from 1995 to 2018 will be retrieved as a csv file. Based on the enormous number of the various attributes among different capital, we will only analyze the typical and significant features of each country to generate a more concise and clearer conclusion. (https://github.com/rdc2697/WB)
2. Exploratory Data Analysis:
    a. This data set has 52 attributes and 146 countries will be assessed.
    b. Data has 4 categorical columns and rest are time-related columns. The gather function will be applied initially to understand the data better.
    c. "Country Name", "Country Code", "Series Name", and "Series Code" columns are categorical columns.
    d. Economic indicators include measures of human capital, such as human capital per capita and human capital per employed female, as well as measures of total wealth, such as total wealth in constant 2018 US dollars.
3. Data Cleaning and Transformation:
    a. Removal of unnecessary columns: We can adjust the dataset with lesser columns while preserving the shape of the data using the subset() function.
    b. Reshape the data: We will need to adjust the data from the wide to long format using functions like gather() and spread().
    c. Scale or normalize the data: Ensuring the variables are on the same scale using functions like scale() or normalize().
    d. Handle Outliers: Outlier detection is a common technique that will be employed to ensure outliers are handled appropriately while avoiding affecting the data set.

Modeling

As specified already, the goal is to predict the Total Wealth of a country given the other variables in the dataset. In the first step, we are going to perform the EDA to give the big picture of the data such as the description of the columns and columns type, checking the missing value and doing the outlier detection, and showing the distribution of the features. In the data cleaning step, we plan to fill out the missing value with the mean of corresponding columns. Next, doing correlation analysis is necessary for the preprocessing. We will pick up the ideal variable by using the filter under tidyverse to generate an appropriate plot by using the function ggplot2.

In the fitting model process, we will consider using a time series model to conclude the previous performance of each country in the 24 years. Besides, we can fit the dataset to the lasso regression model to select the most accurate predictors for the prediction on the future trend of a country's wealth status. Some possible predictors could be:

a. Human capital per capita
b. GDP per capita
c. Gross national expenditure per capita
d. Labor force participation rate
e. Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population)
f. CO2 emissions (metric tons per capita)

We could employ Multiple linear regression, where Total Wealth is the response variable and the other variables (e.g. Human capital, GDP, etc.) are the predictor variables as well. Ny linear and non-linear relationships between the predictor variables will be handled appropriately. Following which an evaluation of the model using performance metrics like R squared, mean squared error, etc. The last thing is to establish them into interactive mode through R Shiny.

References

https://databank.worldbank.org/home: Original data set and reports collected
https://chat.openai.com: Definitions of some terms, R functions info