Multivariate Analysis – Individual Final Project

Kireeti Mantrala – sm2594

About the Data:

This dataset is sourced from the Global Health Observatory (GHO) data repository, which is managed by the World Health Organization (WHO). It covers a wide range of health-related indicators and factors for every country in the world. The dataset spans the years 2000 to 2015 and includes data for a total of 193 countries. The dataset talks about the Life expectancy and the various factors affecting life expectancy like demographic variables, income composition, mortality rates, immunization, human development index, social and economic factors.

Data Dictionary:

Variable Name	Description	Datatype	Accepts Null Values
Country	Country Name	Object	N
Year	Year	Object	N
Status	Developed or Developing	Object	N
Life Expectancy	Life expectancy in age	Object	N
Adult Mortality	Probability of dying between 15 and 60 years per 1000 population	Object	N
infant deaths	Number of infant deaths per 1000 population	Object	N
Alcohol	recorded per capita consumption(in litres)	Object	N
percentage expenditure	Expendidture on health as per GDP(%)	Object	N
Hepatitis B	Immunization coverage among 1 year old(%)	Object	N
Measles	Number of reported cases per 1000 population	Object	N
BMI	Average BMI of entire population	Object	N
under-five deaths	Number of under five deaths per 1000 population	Object	N
Polio	Immunization coverage amoung one year olds(%)	Object	N
Total Expenditure	Government expenditure of health as a percentage of total govt. expenditure(%)	Object	N
Diphtheria	Immunization coverage amoung one year old(%)	Object	N
HIV/AIDS	Deaths per 1000 population	Object	N
GDP	per capita(USD)	Object	N
Population	population of the country	Object	N
thinness 10-19 years	Thinness amomg children from age 10-19(%)	Object	N
thinness 5-9 years	Thinness amomg children from age 5-9(%)	Object	N
Income composition of resources	Index ranging from 0-1	Object	N
Schooling	Number of years of schooling	Object	N

Questions:

1. What insights can we draw from the given factors in the dataset for Life Expectancy?
2. Can we predict life expectancy using these given factors from the dataset?

Answers:

Answer 1:

We cannot classify accurate insights for the given factors in the dataset for Life Expectancy because the clusters after Cluster Analysis are overlapping. But, we understand by analyzing the clusters manually, that there is correlation between these and can be further analyzed to create a prediction model for Life Expectancy.

Answer 2:

We can predict the Life Expectancy of a country by the data of Adult Mortality, infant deaths, Alcohol, percentage expenditure, HepatitisB, Measles, BMI under five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources.

From our Analysis here, the Life Expectancy of a country is vastly dependent on multiple factors and the Insights we have driven have helped us proceed with further analysis to perform Multiple Regression to predict Life Expectancy from these factors.

Hypothetically there are furthermore factors like Immunization, Health Care, Genetics and Environmental Factors too. This Analysis will help us draw a path towards predicting expectancy with these factors and can be used to make equations to consider more factors if needed for other analysis of data.

Process:

We have loaded the dataset and got the Scree Plot to understand if we can perform PCA on the given dataset. Then got the Correlation between these factors. Then, since the percentage is not above 75, we have not proceeded further with PCA and performed the Factor Analysis to reduce and explore the underlying structure of these set of variables to provide insights into the relationships between different aspects of the Life Expectancy dataset. Here after performing Factor Analysis, we have identified the underlying factors that are driving the correlation between the set of variables determining the optimal number of factors to extract from the data. We got the criteria as below:

- RC1 has Schooling, Alcohol and Income Composition of Resources
- RC2 factorized Population, Infant Deaths and Under five deaths
- RC3 has BMI, Diptheria, HepatitisB and Polio
- RC4 has HIV/AIDS and Adult Mortality
- RC5 has GDP and Percentage expenditure
- RC6 only has Total Expenditure
- RC7 only has Measles
- RC8 has thinness 5-9 years and thinness 1-19 years along with BMI being inversely related.

Next we performed Cluster Analysis on the data. We determined that we cannot classify insights for the given factors in the dataset for Life Expectancy because these clusters are overlapping. And performed Multiple Regression to understand if we can predict Life Expectancy from our factors from the dataset, answering our second question. In conclusion, from our Analysis here, the Life Expectancy of a country is vastly dependent on multiple factors and the Insights we have driven have helped us proceed with further analysis to perform Multiple Regression to predict Life Expectancy from these factors. Hypothetically there are furthermore factors like Immunization, Health Care, Genetics and Environmental Factors too. This Analysis will help us draw a path towards predicting expectancy with these factors and can be used to make equations to consider more factors if needed for other analysis of data.