DEATH RATE ANANLYSIS

Anirban Ghosh
Digvijay Sarak-Patil
Pankajh Jhamtani
Gaurish Bansal

Guidance of Dr. Dootika vats
MTH208A - Data Science Lab-I
Indian Institue of Technology, Kanpur

Table of Contents

Approach to the project

Problem Statement:

Assume that you are in 2004, and you have the responsibility to improve the health policies for the country. This guide is intended to help build analytical capacity to assess the quality of mortality statistics currently being collected to improve their value in informing health policies and programs.

Introduction:

Mortality rate or **Death rate** is a measure of number of deaths (in general, or due to a specific cause) in a particular population, scaled to the size of that population ,per unit of time. Mortality rate is typically expressed in units of deaths per $1,000$ individuals per year; thus a mortality rate of 11.5 (out of $1,000$) in a population of $1,000$ would mean 11.5 deaths per year in that entire population, or 1.15 out of total. It is distinct from "**morbidity**",which is either the prevalence or incidence of a disease, and also from the incidence rate.

The project **'Death rate analysis'** focuses on analyzing data sets for various causes of death in different countries. Gender-wise separation of data gave us more insight into it. We have studied the death rate data with respect to the Human Development Index and Gross Domestic Product data. We looked into the properties of data with the help of various visualizations created using R.

Age Standardized Death Rate:

The numbers of deaths per $100,000$ population are influenced by the age distribution of the population. Two populations with the same age-specific mortality rates for a particular cause of death will have different overall death rates if the age distributions of their populations are different. Age-standardized mortality rates adjust for differences in the age distribution of the population by applying the observed age-specific mortality rates for each population to a standard population.

Definition:

The age-standardized mortality rate is a weighted average of the age-specific mortality rates per 100 000 persons, where the weights are the proportions of persons in the corresponding age groups of the standard population.

Calculation of Age-Standardized Death Rate:

To Calculate age standardized Death Rate , We must first calculate the age-specific mortality rate for each age group by dividing the number of deaths by the respective population, and then multiplying the resulting number by $100,000$

Data For the Project

Three datasets were worked upon for this project :

- Gender-wise death rates due to 18 causes for 151 countries in the year 2004
- Human Development Index for 181 countries in the year 2004
- Gross Domestic Product for 172 countries in the year 2004

Final death rate data had the following features :

- 18 data frames each each for different cause
- 5 columns in each data frame representing Country code, Gender, Death rate value, Country, Continent
- 453 rows, each country had 3 rows for three different categories: Male, Female and Other

GDP and HDI data frames have two columns for country codes and for their respective values.

Extraction and Cleaning of the data

1. Death Rates

Major part of the data, i.e. , Death rates of various countries was extracted using an API. The GHO(Global Health Observatory) portal provides a simple query interface to the World Health Organization's data and statistics content, using OData (Open Data Protocol). The WHO website does not allow web scraping but provides access to some data using with a simple API. The GHO API provides access to a hige amount of data which also includes data sets which don't have any relation with the topic of this project. Data sets were filtered such that the name of the data set should contain the word 'Death'. The JSON files of the filtered data sets were converted to R data frames for future use.

The extracted data frames were quite random. Different data sets had data for different years and different spatial dimensions while some didn't had gender wise separation. So we handpicked the data sets such that they should have gender wise segregation and maximum amount of country wise data. This handpicked data, consisting of 18 data frames, was refined such that it only contained the intersection of countries in all the data frames, which was a total of 151 countries.

2. Human Development Index

Human Development Index data was scraped from countryeconomy.com. A total of 172 countries' HDI values for the year 2004 were available.

3. Gross Domestic Product

Gross Domestic Product Data was downloaded as a CSV file from data.worldbank.org. A total of 181 countries' GDP values for the year 2004 were available.

A glimpse of Final Data of Death Rate due to Alcohol use disorders:

```
  index gender Death_Rate country continent
1   AGO   BTSX        1.3  Angola    Africa
2   AGO   FMLE        0.7  Angola    Africa
3   AGO    MLE        1.9  Angola    Africa
4   AND   FMLE        0.1 Andorra    Europe
5   AND    MLE        0.8 Andorra    Europe
6   AND   BTSX        0.4 Andorra    Europe
```

Biases

Possible Bias in the Data :

1. There might be some amount of biases aroused in our data due to methods used in the cleaning of data.Initially We had data on the gender-wise death rate for different countries for 18 different causes, but in that data there were many missing observations on the country ,Death Rate,Gender variable of the 18 different dataframes of 18 different causes. So First we subset all the dataframes which contain death rate of countries for **three gender**. Then we subset the 18 dataframes such that they contain the death rate of only the **common countries** by which we get a list of 18 dataframes and each dataframe contains the death rate of 151 countries. So deleting the missing observations will lead to the biases in the data.
2. There might have been bias due to lack of proper survey in developing countries in the data on GDP of 2004.

Visualizations

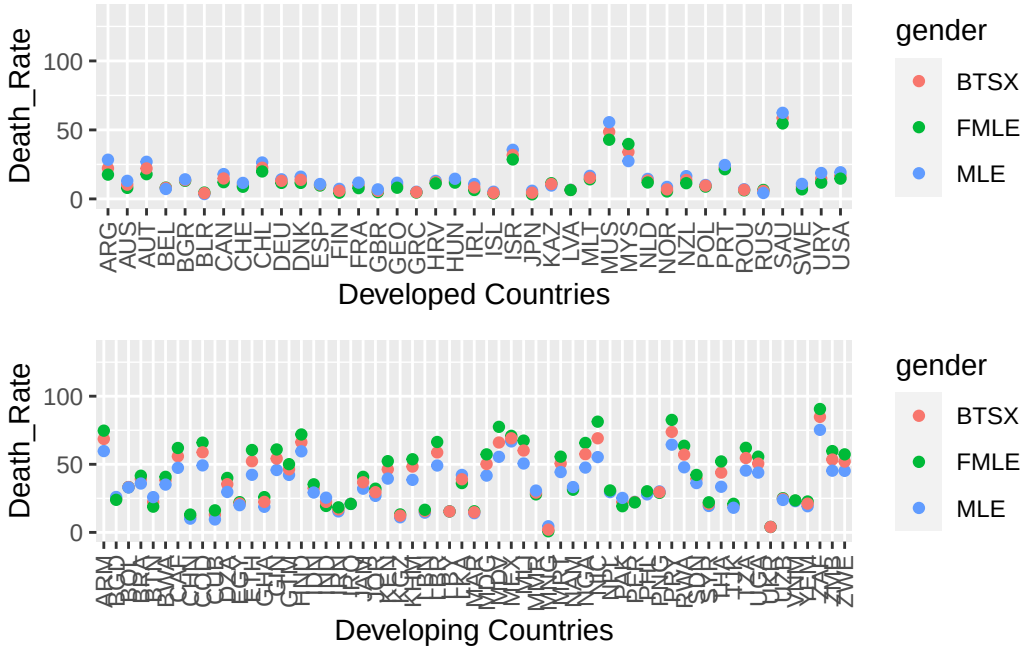
To understand various trends, outliers, patterns in the data we start with data visualizations. Data becomes more understandable with visualizations. Keeping this in mind, we made some visualizations for the analysis.

We begin with trying to find a relation between HDI and death rates.

Countries with HDI more than 0.8 are considered to be developed and rest are developing countries.

Separate gender wise scatter plot for developed and developing countries shows us what role different genders might have in different types of countries

Death rates of diabetes mellitus vs Countries :

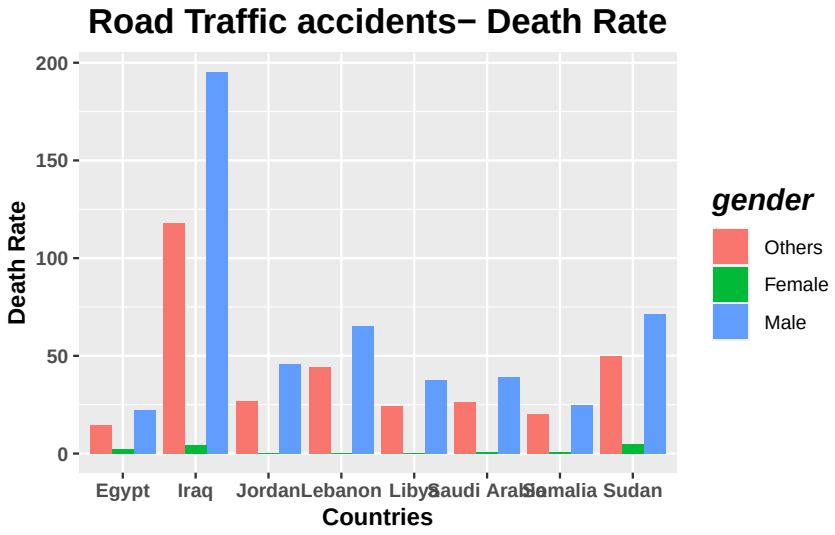


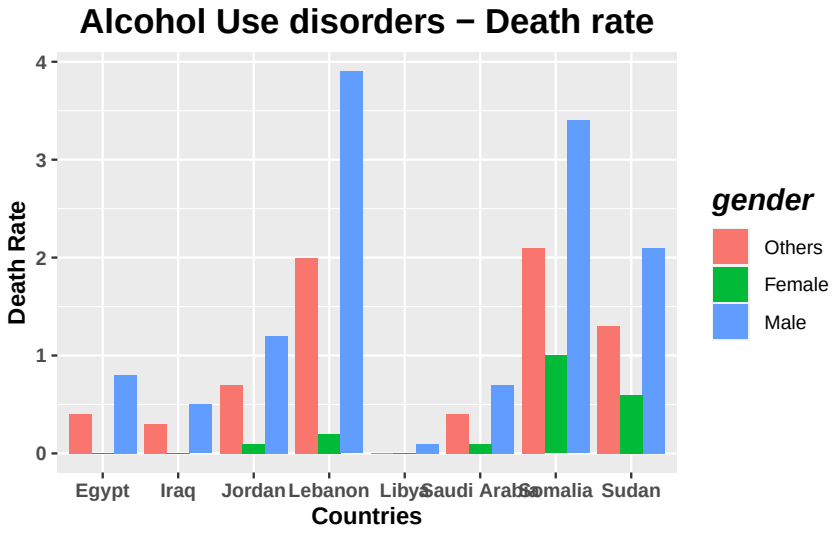
We observe a clear female dominance in death rates of Developing countries unlike in Developed countries. A possible explanation for this would be that, as obesity among women increases, diabetes in pregnancy is becoming increasingly common in LMIC(Low and Low-middle income countries). Because of lack of resources and trained personnel, and other priorities related to reducing maternal, feotal, and neonatal mortality, diagnosing and providing care to women with diabetes in pregnancy is not high on the priority lists in many LMIC.

Looking at the plot we can ask the question : *How does socio-economic conditions effect death rates of a particular disease?*

Gender wise Multiple bar plot of countries in East Mediterranean Countries.

```
Attaching package: 'dplyr'

The following object is masked from 'package:gridExtra':

    combine

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```



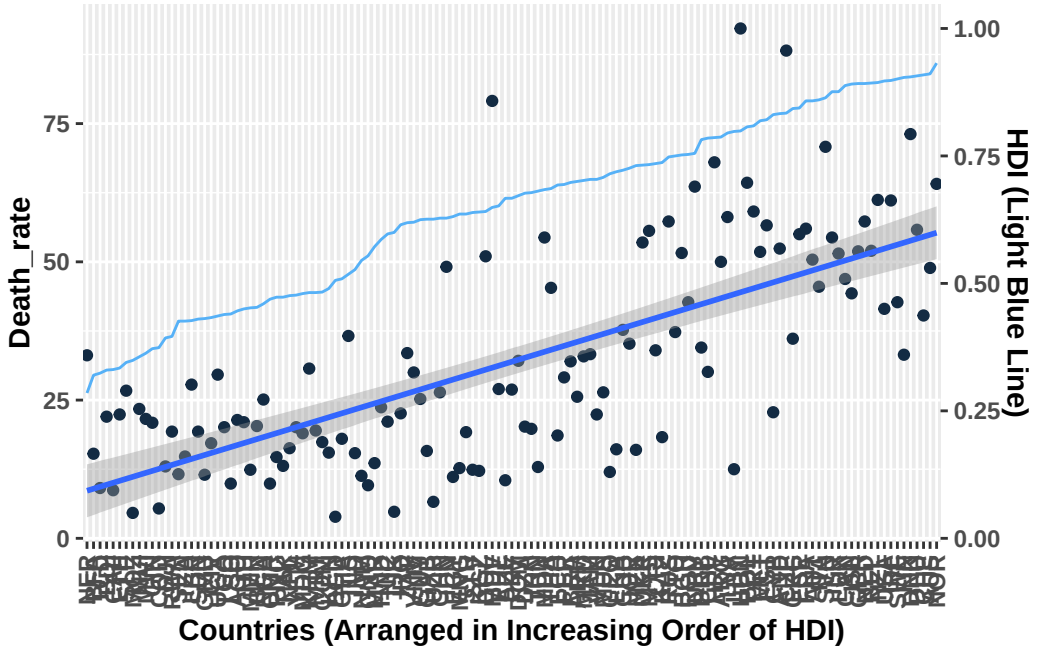


We observe that the females have unusually low values as compared to Males. Possible explanations for this would be that the death rates are affected by laws and less independence in of Women in East Mediterranean Countries.

We can ask the following question from this :

How does socio-economic conditions of countries effect the death rates of a class of people?

Scatterplot of Death rates of countries along with Linear model plotted with line plot of HDI of various countries

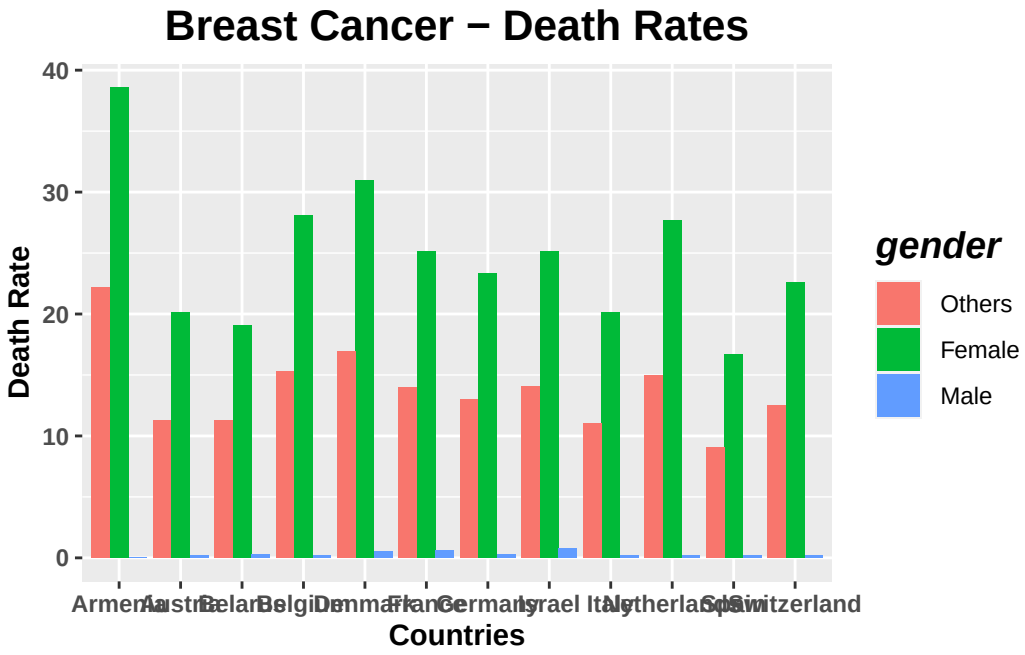


Colon and rectum cancer death rates showed quite unusual behavior. All the death rates showed negative or no correlation with HDI and GDP except this one. It has positive correlation. A possible reason for the same would be the mortality to incidence ratio for this is negatively correlated with HDP and GDP. HDI increases as diagnosed cases increases , so are the deaths. In low HDI countries diagnosed cases are low as treatment and diagnosis needs costly resources.

Gender wise Multiple bar plot of Breast cancer in some countries

```
#| echo: false
cont <- "Europe" ###
datasetNUM <- 2 ###

i<-array(c(49,50,51,52 ,53,54,55,57,58,59,60,61,62,63,64,65,66,74))
load("/Users/pj/Library/CloudStorage/OneDrive-IITKanpur/IITK sem 3/MTH208A/group-project-20/FinalDat

  df<- Final_data_corrected_2[[datasetNUM]]
  dfsub <- subset(df,df$continent==cont)
  ###
  counts<-c("Armenia","Austria","Belgium","Belarus","Switzerland","Germany","Denmark","Spain","Franc

  dfsubsub <- dfsub %>% filter(country %in% counts)
  g <- ggplot(dfsubsub,aes(x = country,y = Death_Rate,fill = gender))
  g <- g + geom_bar(position = "dodge",stat="identity") + labs(x = "Countries", y = "Death Rate", ti
  g<-g + theme(plot.title= element_text(size=16,
                                        hjust=0.5,
```

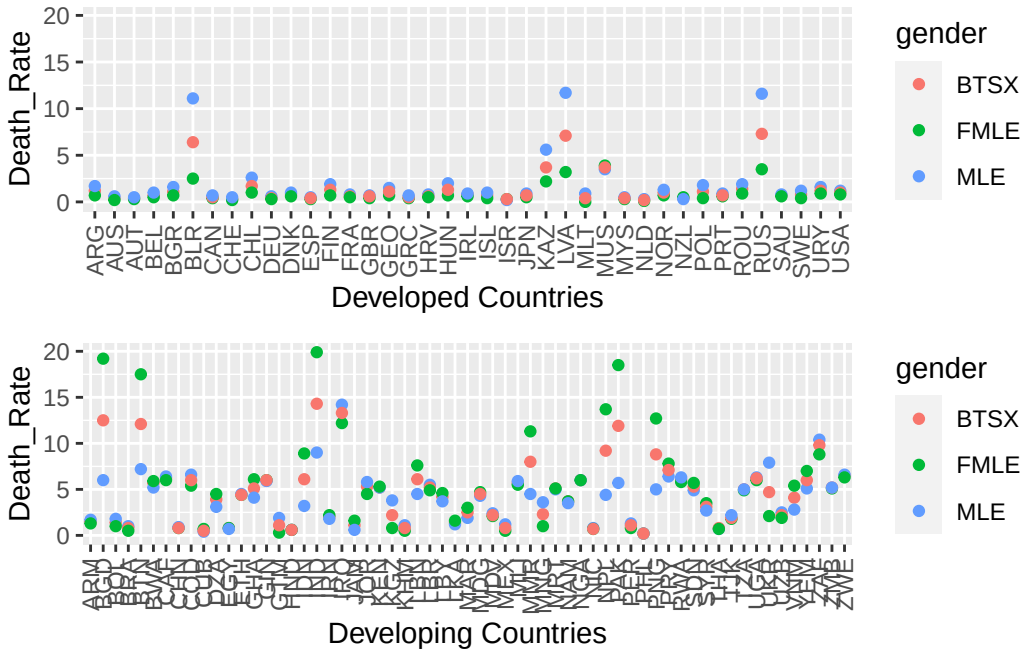
```
                                        face= "bold"),
              legend.title = element_text(size=14,
                                          face="bold.italic"),
              axis.title = element_text(face="bold"),
              axis.text = element_text(face= "bold"))
g+scale_fill_discrete(labels = c("Others", "Female","Male"))
```




```
##We can see that Armenia has comparitively high Breast Cancer Death Rate, this is because Armenia d
```

We can see that Armenia has comparitively high Breast Cancer Death Rate, this is because Armenia does not have a systematic breast cancer screening program.

Gender-wise scatter plot of death rates due to Poisoning



Outlier values in death rate due to Poisoning are seen in India ,Kyrgyzstan, Ukraine, Kazakhstan and Russia These are the outlier countries, also the HDI doesn't have much effect on death due to poisoning. Possible reasonwould be that the Agriculture based economies, easy availability of pesticides, poverty related socioeconomic problems, lack of adequate protective clothing, and limited treatment facilities are some of the factors contributing to the high morbidity and mortality due to poisoning. The highest number of cases among both children and adults occur in Africa and Southeast Asia, where food is more often prepared with unclean water or cannot be properly produced and stored.

Conclusion

Improving the quality of vital statistics will be of inestimable value to public health decision-makers. It will significantly increase confidence in the data.

This guide and the accompanying electronic tool provide guidance on simple actions that can and should be taken to assess the quality of mortality data, particularly vital statistics on deaths and causes of death. Conducting such a data quality review aims to diagnose problems and identify potential solutions. Solutions may include:

- Breast Cancer-related awareness in Armenia.
- Introducing incentives to encourage accurate reporting of all births and deaths
- improving the training of medical doctors in death certification
- improving the laws for women in East Mediterranean countries.

- improving the quality and completeness of medical records so that doctors have all the information they need to certify causes of death correctly.

The guide places emphasis on three particular aspects of data quality:

- The completeness of the data. (Are all deaths registered?)
- The gender pattern of reported deaths. (Is there serious gender-specific misreporting or underreporting?)

Improving the quality of vital statistics will be of inestimable value to public health decision-makers. It will greatly increase confidence in the data, and thereby facilitate and promote the use of mortality and causeof-death statistics to ensure that resource allocation is evidence informed, and focuses on interventions most needed to improve overall population health levels.

Acknowledgement:

We, group 20, would like to express our profound gratitude towards Dr. Dootika Vats, our academic and project instructor for MTH208A(Data Science Lab), for her guidance and constant supervision throughout the process and providing creative ideas and necessary information regarding the project, which led to the completion of this project. We could only have undertaken this journey with your generously offered knowledge and expertise in R programming and Data Science. It has been a great learning experience and provided us with a practical insight into the theoretical knowledge gathered during the course lecture. We are also grateful to our teaching assistants and classmates for their constant feedback and help in clearing doubts. Thanks should also go to the Lab Assistants in the New Core Labs, who helped in lab sessions.

References :

As we have collected data for death rate, HDI, GDP here are the references
https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG

https://countryeconomy.com/hdi https://www.who.int/data/gho/data/indicators