Predictive Analytics MGS 616 Group Project

Team FP_8: Sai Kartheek Mahankali, Nihareeka Suneel Gote, Viraj Vhatkar, Aditya Manjunath Naik

Datasets

1. WHO Mortality Data – All causes
2. WHO Mortality Data – Communicable Diseases
3. WHO Mortality Data – Non-Communicable Diseases
4. Global Health Expenditure Data

Data Cleaning and Pre-processing

Step 1: Load the data into R

We are using the R programming language to perform the exploratory data analysis. We start by importing required libraries into the R environment.

```
library(data.table)
library(dplyr)
library(ggplot2)
library(plotly)
library(tidyr)
```

We then use the 'fread' function to read the csv data into the R environment, for WHO cause of deaths datasets. The all_causes.df contains 300K rows and represent data about deaths due to all causes from 1950-2020. The non_com.df and com.df contain 298K rows each and represent data about deaths due to non-communicable diseases and communicable diseases respectively. Please note that com.df also includes deaths due to maternal, perinatal, and nutritional conditions. The global health expenditure dataset which has 4.2K rows is loaded as global_exp.df

Data	
all_causes.df	300258 obs. of 9 variables
com.df	298326 obs. of 10 variables
global_exp.df	4224 obs. of 19 variables
non_com.df	298851 obs. of 10 variables

Step 2: Merging the datasets

Since the all_causes.df, com.df, and non_com.df have the same column names, we merge them into a single dataframe named merged_df.

```
merged_df <- merge(com.df, non_com.df,
                   by = c("region_name","country_code",
                          "country_name" ,"year","sex",
                          "age_group_code","age_group" ), all = FALSE)
merged_df <- merge(merged_df,all_causes.df,
                   by = c("region_name","country_code",
                          "country_name" ,"year","sex",
                          "age_group_code","age_group"), all = FALSE)
```

This new dataframe has 297K rows and 15 columns. Below are the summary statistics of the merged dataframe. We observe that there are many variables which currently belong to the class "character" that need to be changed to a categorical variable – such as region_name, country_name, sex, age_group_code etc. Also, we notice that there are a lot of NA values, which need to be treated accordingly.

```
> summary(merged_df)
 region_name        country_code       country_name            year          sex            age_group_code      age_group
 Length:297633      Length:297633      Length:297633      Min.   :1950   Length:297633      Length:297633      Length:297633
 Class :character   Class :character   Class :character   1st Qu.:1979   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :1995   Mode  :character   Mode  :character   Mode  :character
                                                          Mean   :1992
                                                          3rd Qu.:2007
                                                          Max.   :2020

   com_deaths      percentage_of_com_deaths com_death_rate_per_100000_pop non_com_deaths     percentage_of_non_com_deaths
 Min.   :     0   Min.   :  0.000          Min.   :    0.00              Min.   :      0   Min.   :  0.00
 1st Qu.:     7   1st Qu.:  3.390          1st Qu.:    7.28              1st Qu.:     46   1st Qu.: 39.91
 Median :    51   Median :  7.012          Median :   34.59              Median :    360   Median : 64.49
 Mean   :  1172   Mean   : 12.973          Mean   :  253.16              Mean   :   8735   Mean   : 61.00
 3rd Qu.:   329   3rd Qu.: 14.679          3rd Qu.:  150.46              3rd Qu.:   2578   3rd Qu.: 84.61
 Max.   :503077   Max.   :100.000          Max.   :38333.33              Max.   :2634041   Max.   :100.00
 NA's   :1593     NA's   :10203            NA's   :25587                 NA's   :1593      NA's   :10203
 non_com_death_rate_per_100000_pop  total_deaths     total_death_rate_per_100000_pop
 Min.   :     0.00                 Min.   :      0   Min.   :     0.0
 1st Qu.:    35.54                 1st Qu.:     99   1st Qu.:   100.5
 Median :   281.11                 Median :    722   Median :   469.1
 Mean   :  1842.10                 Mean   :  11454   Mean   :  2449.4
 3rd Qu.:  1481.47                 3rd Qu.:   4126   3rd Qu.:  2061.5
 Max.   :136842.11                 Max.   :3383729   Max.   :239215.7
 NA's   :25587                     NA's   :1593      NA's   :25587
```

Step 3: Data Cleaning

- **Drop the rows with NA values**

After careful examination of the dataset, we have decided to drop the rows with NA values. Below is a screenshot where we can observe that there are no NA values in the dataset.

```
> merged_df <- na.omit(merged_df)
> null_count <- colSums(is.na(merged_df))
> print(null_count)
                      region_name                      country_code                      country_name                              year
                                0                                 0                                 0                                 0
                              sex                    age_group_code                         age_group                       com_deaths
                                0                                 0                                 0                                 0
         percentage_of_com_deaths     com_death_rate_per_100000_pop                  non_com_deaths      percentage_of_non_com_deaths
                                0                                 0                                 0                                 0
non_com_death_rate_per_100000_pop                      total_deaths   total_death_rate_per_100000_pop
                                0                                 0                                 0
> print(dim(merged_df))
[1] 271554     15
```

- **Filter out redundant data**

We observed that there is redundant data in the dataset and on further examination, have found that “sex” and “age_group_code” columns contain the an “All” value that is just the sum of values belonging to other categories. Hence, these can be filtered out.

```
> unique(merged_df$sex)
[1] "All"    "Female" "Male"
> unique(merged_df$age_group_code)
 [1] "Age00"      "Age01_04"   "Age_all"    "Age05_09"   "Age10_14"   "Age15_19"   "Age20_24"
 [8] "Age25_29"   "Age30_34"   "Age35_39"   "Age40_44"   "Age45_49"   "Age50_54"   "Age55_59"
[15] "Age60_64"   "Age65_69"   "Age70_74"   "Age75_79"   "Age80_84"   "Age85_over"

> death_counts_data.df <- merged_df %>%
+   filter(sex %in% c("Male","Female") &
+            !age_group_code %in% c( "Age_all"))
> death_counts_data.df <- subset(death_counts_data.df, select = -c(7))
> names(death_counts_data.df)[3] <- "country"
> print(dim(death_counts_data.df))
[1] 171793     14
```

After filtering the required data, we now are left with 171K rows and 14 variables.

- **Convert variables to categorical type**

We convert “country”, “country_code”, “region_name”, “sex”, “age_group_code” into categorical variables.

```
> cols <- c('country','country_code','region_name','sex','age_group_code')
> death_counts_data.df <- death_counts_data.df %>% mutate_at(cols, as.factor)
> str(death_counts_data.df)
Classes 'data.table' and 'data.frame':  171793 obs. of  14 variables:
 $ region_name                    : Factor w/ 6 levels "Africa","Asia",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ country_code                   : Factor w/ 112 levels "ALB","ARG","ARM",..: 22 22 22 22 22 22 22 22 22 22 ...
 $ country                        : Factor w/ 112 levels "Albania","Antigua and Barbuda",..: 18 18 18 18 18 18 18
...
 $ year                           : int  1980 1980 1980 1980 2011 2011 2011 2011 2011 2011 ...
 $ sex                            : Factor w/ 2 levels "Female","Male": 1 1 2 2 1 1 1 1 1 1 ...
 $ age_group_code                 : Factor w/ 19 levels "Age00","Age01_04",..: 1 2 1 2 1 2 3 4 5 6 ...
 $ com_deaths                     : int  215 87 237 77 58 4 1 2 3 2 ...
 $ percentage_of_com_deaths       : num  68.7 65.4 63.2 58.3 67.4 ...
 $ com_death_rate_per_100000_pop  : num  4019 458 4237 401 1120 ...
 $ non_com_deaths                 : int  58 27 85 33 24 6 0 3 1 2 ...
 $ percentage_of_non_com_deaths   : num  18.5 20.3 22.7 25 27.9 ...
 $ non_com_death_rate_per_100000_pop: num  1084 142 1520 172 463 ...
 $ total_deaths                   : int  313 133 375 132 86 14 2 6 6 11 ...
 $ total_death_rate_per_100000_pop  : num  5852 700 6705 687 1661 ...
 - attr(*, ".internal.selfref")=<externalptr>
 - attr(*, "sorted")= chr [1:6] "region_name" "country_code" "country" "year" ...
```

Step 4: Basic statistics of the final dataset

Below are the summary statistics of our dataset –

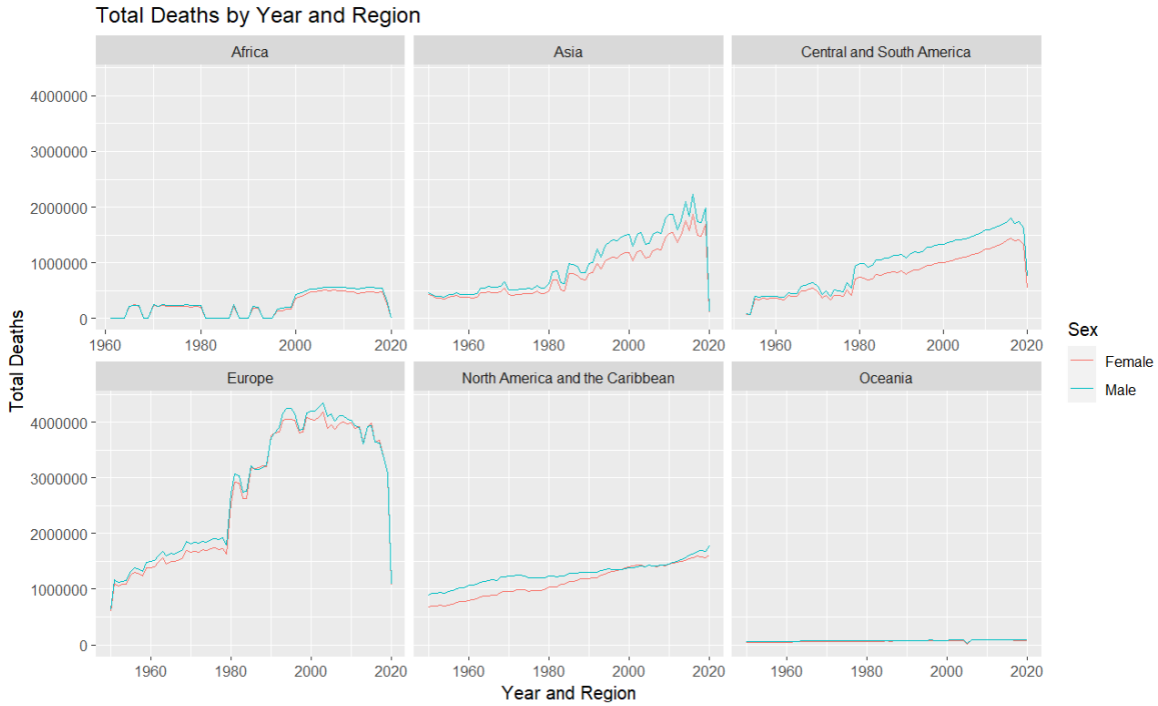
```
> summary(death_counts_data.df)
                         region_name     country_code                       country            year
 Africa                        : 5607   NLD     :  2698   Netherlands             :  2698   Min.   :1950
 Asia                          :30574   USA     :  2698   United States of America:  2698   1st Qu.:1979
 Central and South America     :32383   AUS     :  2660   Australia               :  2660   Median :1995
 Europe                        :73021   CAN     :  2660   Canada                  :  2660   Mean   :1992
 North America and the Caribbean:24546   ESP     :  2660   Japan                   :  2660   3rd Qu.:2007
 Oceania                       : 5662   GBR     :  2660   Spain                   :  2660   Max.   :2020
                                        (Other):155757   (Other)                 :155757
    sex          age_group_code     com_deaths      percentage_of_com_deaths com_death_rate_per_100000_pop
 Female:85827   Age00    :  9150   Min.   :    0.0   Min.   :  0.000          Min.   :    0.00
 Male  :85966   Age01_04:  9093   1st Qu.:    8.0   1st Qu.:  3.276          1st Qu.:    6.47
                Age65_69:  9066   Median :   43.0   Median :  6.789          Median :   30.61
                Age55_59:  9065   Mean   :  502.8   Mean   : 12.979          Mean   :  264.55
                Age45_49:  9064   3rd Qu.:  238.0   3rd Qu.: 14.582          3rd Qu.:  160.91
                Age75_79:  9064   Max.   :88427.0   Max.   :100.000          Max.   :38333.33
                (Other) :117291
 non_com_deaths   percentage_of_non_com_deaths non_com_death_rate_per_100000_pop  total_deaths
 Min.   :     0   Min.   :  0.00               Min.   :     0.00                 Min.   :     0
 1st Qu.:    53   1st Qu.: 40.17               1st Qu.:    33.18                 1st Qu.:   114
 Median :   325   Median : 64.74               Median :   241.50                 Median :   647
 Mean   :  3753   Mean   : 61.44               Mean   :  1933.43                 Mean   :  4916
 3rd Qu.:  1910   3rd Qu.: 85.25               3rd Qu.:  1671.00                 3rd Qu.:  3014
 Max.   :503555   Max.   :100.00               Max.   :136842.11                 Max.   :617885

 total_death_rate_per_100000_pop
 Min.   :     0.0
 1st Qu.:    91.6
 Median :   408.5
 Mean   :  2568.9
 3rd Qu.:  2284.4
 Max.   :239215.7
```

Insights

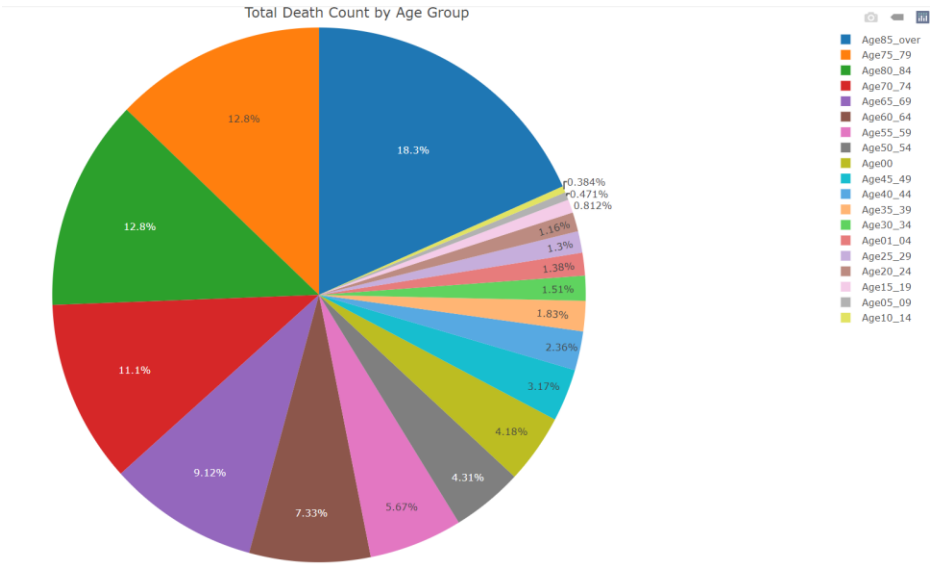
Insight 1: Time Series of death counts per region

Let us visualize total deaths by year per region. We notice that in the span of 70 years from 1950-2020, there seems to be an **increasing trend in the total deaths** in Asia, Central and South America, and North America and the Caribbean regions. The Oceania region appears to have the least deaths among other regions. Also, we observe that all the regions have witnessed more male deaths compared to female deaths.



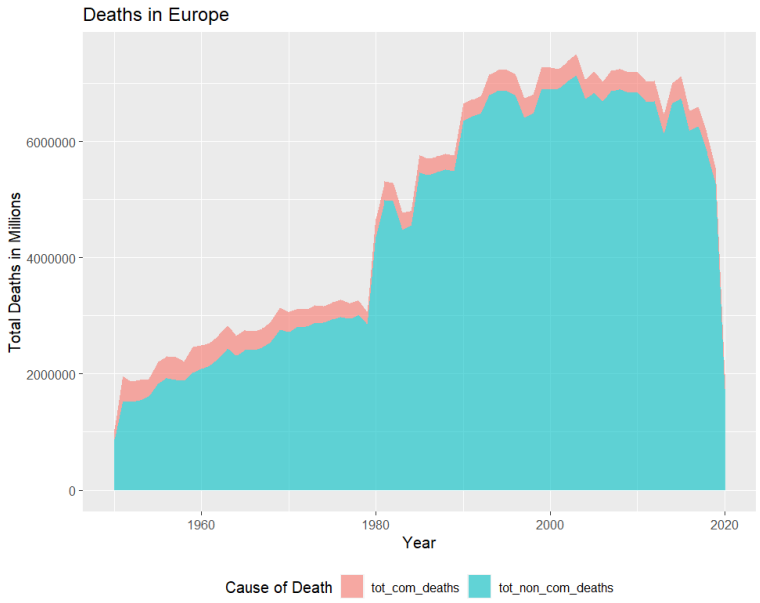
Insight 2: Spread of deaths by age group

Let us now visualize the spread of deaths across different age groups – starting from new born children to people aged over 85 years. From the below pie chart, we observe that **people aged 70 years and above** contribute to almost **55% of total deaths** from 1950 to 2019. This indicates that there hasn't been any catastrophic event like a war or global pandemic that was responsible for disproportionate human deaths in the time period we consider.



Insight 3: Deaths by Non-Communicable Diseases vs Communicable Diseases in Europe

From Insight 1, we can infer that most deaths occurred in Europe region. Let us now visualize the contribution of communicable diseases in Europe's death counts. We observe from the below stacked area chart that most of the deaths in Europe from 1950 – 2020 are caused due to non-Communicable diseases.



Insight 4: Average Health Expenditure as a percentage of GDP for European countries

As we are currently concentrating on the Europe region, let us now check on the average health budget as the percentage of GDP for European countries from 2000 – 2022. According to the Global Economy, the global average of health expenditure as a percentage of GDP is **6.5%.** So, we draw a red line on our plot to check how many of the European countries are spending more than the global average towards the health.