Week 5 - Visualizations Activity

Srilekha Gurrapu

2023-06-22

Description of the data

The dataset I am using is the COVID-19 dataset, which provides comprehensive information about the COVID-19 pandemic across multiple countries. The data measures various aspects of the pandemic, including the number of cases, deaths, and recoveries. The dataset was collected from reliable sources such as national health agencies and international organizations tracking the pandemic.

The data is saved in a CSV (Comma-Separated Values) format. It is a flat file that uses commas as the delimiter to separate the values in each row. The dataset is organized in a tabular format, where each row represents a specific observation (e.g., a specific country and date), and each column represents a variable or attribute related to that observation (e.g., cases, deaths, recoveries). The CSV format is a widely used and easily accessible format for storing structured data. It can be opened and processed using various tools and programming languages, including R, Python, and spreadsheet software.

To read this data into R, I will use the **read_csv** function, which is a base R function specifically designed for reading CSV files. This function automatically handles the parsing of the CSV format, including the detection of the delimiter and the conversion of the data into appropriate data types. The resulting data will be stored in a dataframe, which is a common data structure in R for handling tabular data.

Reading the data into R

In this code, I used the **read_csv()** function to read the data from the CSV file and assign it to the **data** dataframe object. The **read_csv()** function from the readr package is used to read the data from the zip file. The **read_csv()** function is designed to read CSV files and is a part of the readr package.

The **unzip()** function is used to extract the file from the zip archive, and the extracted data is directly passed to the **read_csv()** function to read it as a CSV file.

Cleaning the data

In the code chunk, I performed some basic data cleaning operations on the dataset.

First, I renamed the columns of the dataframe using the **colnames** function. I provided a vector of new column names that correspond to the desired names for each column.

Next, I converted the "Date" column to the "Date" format using the **as.Date** function. This ensures that the date values are treated as dates in R, allowing for easier manipulation and analysis.

Finally, I subsetted the dataset to keep only the "Country", "Date", "Total_Cases", and "Total_Deaths" columns using indexing with square brackets ([]). This creates a new dataframe called **data_subset** that contains the selected columns.

To verify the changes and check the cleaned dataset, I used the **head** function to display the first few rows of the **data_subset** dataframe.

```
## Warning: package 'dplyr' was built under R version 4.2.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## # A tibble: 6 × 4
##   Date       Country             Total_Cases Total_Deaths
##   <date>     <chr>                     <dbl>        <dbl>
## 1 2020-01-22 Afghanistan                   0            0
## 2 2020-01-22 Albania                       0            0
## 3 2020-01-22 Algeria                       0            0
## 4 2020-01-22 Andorra                       0            0
## 5 2020-01-22 Angola                        0            0
## 6 2020-01-22 Antigua and Barbuda           0            0
```

Characteristics of Data

```
## Warning: package 'knitr' was built under R version 4.2.3
```

Column Name	Description
Date	The date of the recorded data for COVID-19 cases.
Country	The name of the country where the COVID-19 cases were reported.
Total_Cases	The total number of confirmed COVID-19 cases in a specific country or region.
Total_Deaths	The total number of deaths caused by COVID-19 in a specific country or region.
Total_Recoveries	The total number of individuals who have recovered from COVID-19 in a specific country or region.
Active_Cases	The number of active COVID-19 cases in a specific country or region.
New cases	The number of new COVID-19 cases reported on a specific date.

Column Name	Description
New deaths	The number of new deaths due to COVID-19 reported on a specific date.
New recovered	The number of new recoveries from COVID-19 reported on a specific date.
WHO Region	The World Health Organization (WHO) region to which the country or region belongs.

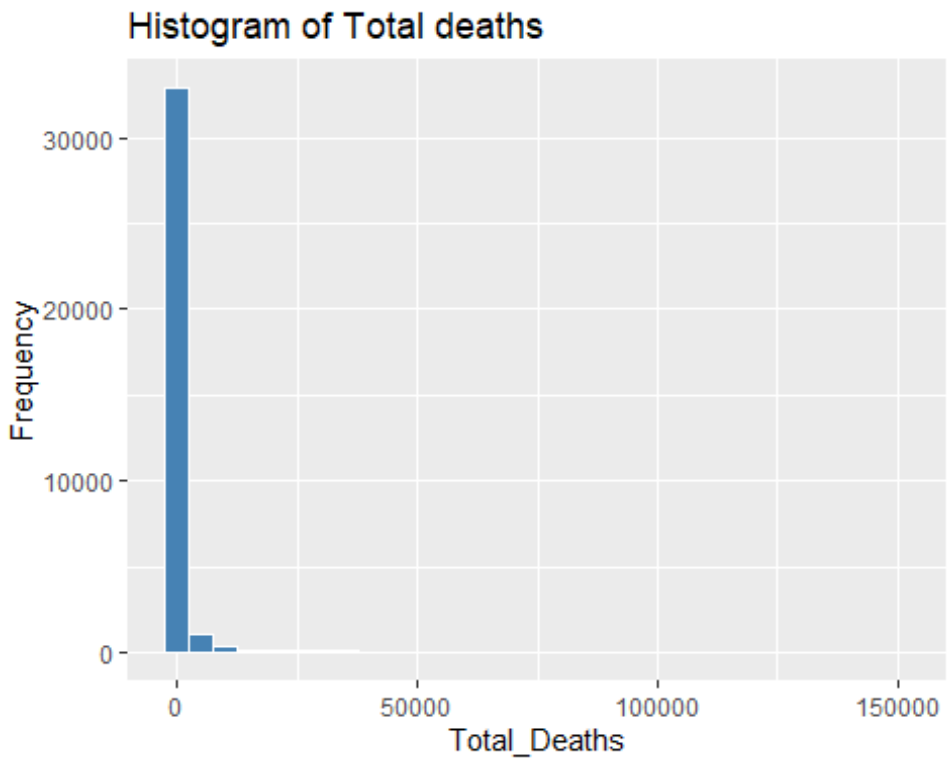
#Inline code “This data set has 35156 rows and 10 columns. The names of the columns and a brief description of each are in the table above.”

Summary Statistics:

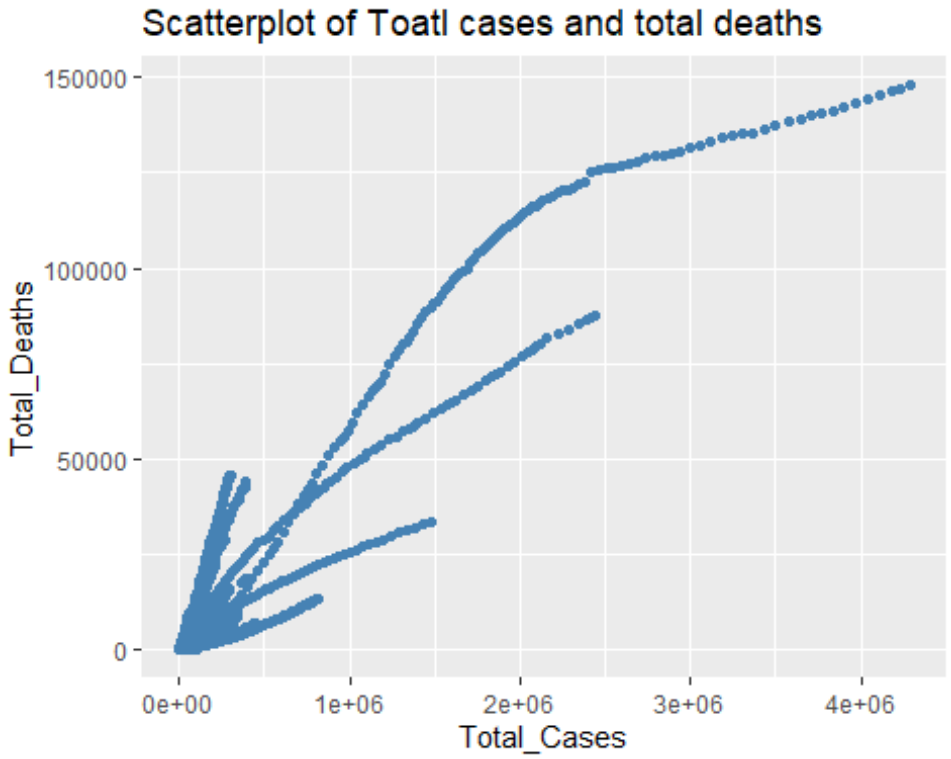
```
##                              Column Minimum Maximum      Mean Missing_Values
## Total_Cases             Total_Cases       0 4290259 23566.631              0
## Total_Deaths           Total_Deaths       0  148011  1234.068              0
## Total_Recoveries Total_Recoveries       0 1846641 11048.135              0
```

Histogram of Total Deaths

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```



Scatterplot of Total Cases and Total Deaths



Bar graph