Data Engineering Report

Jordan Marshall - 18256716

Purpose

The purpose of this project is to visualise each country's confirmed and death cases daily. The user will be able to select a day they will like to visualise and display data in relevant graphs and charts. Other statistics such as total confirmed cases, total death cases, mean confirmed cases, mean death cases and ratio between total confirmed and total death cases are also shown to the user.

Dataset

John Hopkins University COVID-19 Daily Reports

https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports

Data Exploration

Data exploration is the process of exploring a large dataset to identify patterns, characteristics, or points of interest. Data exploration can be split be done manually or automatically. The main aim is to identify trends in the data. The data can be visualised into charts, plots, reports, or other forms of visualisations. This high level approach can help build a bigger picture of the data and help increase chances of identifying trends. Data exploration can also help cut down the dataset by identifying data that isn't required for the intended output. This can help make the dataset a manageable size.

In the case of the COVID-19 John Hopkins University data, I chose to use the daily reports. Each daily report consists of the following attributes:

- FIPS
- Admin2
- Province_State
- Country_Region
- Last_Update
- Lat,Long_
- Confirmed
- Deaths
- Recovered

- Active
- Combined_Key,Incident_Rate
- Case_Fatality_Ratio

In some cases, data values for FIPS orAdmin2 were not available. Missing data makes it hard to graph certain trends. I chose Confirmed cases and Death cases due to its consistent availability. Daily case reports only go back as far as March 2020.

Data Pre-processing

Sometimes raw data isn't understandable in its current state and must be processed first to a readable state. Data needs to be checked and validated before any further steps. It is important that the data is accurate as it could skew the result of a machine learning model. Data must be complete and must not be missing any information. It must be consistent and kept to the intended structure. The data also must be believable and timely. Most importantly the data must be interpretable so to get an accurate result. There are different steps to achieve data pre-processing: data cleaning, data integration, data reduction, data transformation. Data cleaning removes incorrect, incomplete, and irregular data. It can also input the missing values if it can be inputted. Data integration is combining multiple data sources into one dataset. Data reduction helps with making the dataset a manageable size by reducing the volume of data. This can be achieved through data compression to help reduce storage size. Data transformation entails formatting or restructuring the data to fit the requirements.

After I explore my data, I will remove unused attributes. This includes FIP, Admin2, Last_Update, Province_State, Combined_Key. This will help with loading times. I will rename some of the attributes so it will be easier to understand and comprehend. I will order the dataset by country to get a better understanding of the data. One additional feature I will add is a check to see if the file the user is looking for is there. If there is a 404 error, meaning the file is not available from the repo, I will inform the user so they can view a different date. The date the user selects needs to be formatted to either be printed or used to retrieve the correct file the user is looking for. Due to the large amount of countries and data, I will find it difficult to fit everything into a single graph. An example of this is a pie chart. I won't be able to fit all the countries so I will find the top 10 or 20 counties with the data I am looking for and display that instead.

Standardisation

Data standardisation is converting data into a standard format for processing and analysis. The main aim is to make data clear, concise, and consistent defined attributes throughout a collection of data. It can help make understanding the data a lot easier.

Labels provide an identity to each element in the dataset. Through this method, data can be easily and quickly accessed to get the most up to date data. Data standardisation allows for scaling. One scaling technique used during standardisation is to establish the mean and the standard deviation at 0 and 1, respectively. This can mean all data is consistent and organised.