Covid Data Analysis Project

Sanyasi Sai Sandeep Pothala- 22M0792

Pavan Kumar Marotu - 22M0767

November 2022

Chapter 1

Description

This project aims in finding the top3 most effected countries over the globe based on cumulative cases and deaths gathered from the dataset.

1.1 Preprocessing data

The initial data set contains lots of null values and data across all countries , to remove the rows which contains more null values we have used bash tools AWK and SED which are best suitable for handling csv files. Now using the python scripts we have read the data into list of lists and perfomed manipulations to construct cumulative cases and deaths columns , then sorted this list based on number of deaths and extracted the top 3 countries

1.2 Visualisations and Description

First graph was a cumulative line plot of 3 lines corresponding to 3 countries. Covid Pandemic stayed around 2 years,so we divide it into 4 parts[Jan 2020-Jun2020,July2020-Dec2020,Jan2021-June2021,July2021-Dec2021] and plotted 3 pie plots corresponding to 3 countries to find which time period has most covid cases.
The bar graph shows percentage of male and female deaths of USA and Brazil.
All of this graphs were inserted in a html web page for the high level frontend view.
The other half of this project focuses on covid affect in India where we apply similar techniques used above to find the top 3 most affected states due to covid.
Then we examine the four time periods of these 2 years pandemic period to find which time period has caused the most number of deaths due to covid using the pie plots plotted for 3 states differently and see if same time period has caused most number of deaths across countries .

1.3 Concepts used from the course

AWK and SED from bash,lambda function to sort list of lists,manipulations on list of lists,csv module usage,plotting lineplots,barplots,pieplots using pyplot,latex to generate this report,HTML and CSS for basic frontend to insert and display the graphs.

Chapter 2

What we have learned from this project:-

2.1 Conclusions on spread of covid over globe

The top 3 most affected countries due to covid according to our dataset are
1.USA(United States of America)
2.India
3.Brazil
Eventhough the population of India is higher than USA,USA has considerably more number of deaths than India,this shows that USA has clearly failed in its countermeasures to restrict the covid spread,this is inferred from the cumulative line plot of the 3 cuntries.
From the pie plots of the 3 countries across 4 time periods in the 2 years 2020 and 2021 [2020-Jun2020,July2020-Dec2020,Jan2021-June2021,July2021-Dec2021].All of these 3 coutries have shown some similar trend in these pie plots.
These 3 countries have seen more of its deaths due to covid in the year 2021 with slight variations upto mid 2021 and towards the end of 2021.
A)USA have most of its deaths due to covid in the time period post July 2021
B)India has most of its deaths due to covid in the time period Jan 2021 to June 2021
C)Brazil has most of its deaths due to covid in the time period Jan 2021 to June 2021 ,infact more than 50 percent of its deaths occured in this time period for Brazil.

2.2 Conclusions on ratio of male and female deaths due to covid

The bar graph that we plotted for death percentages of males and females of USA and Brazil shows that death ratios of USA and Brazil were approximately same.But males were around 55percent while females were around 45 percent this may be because population of males was higher than females also men tend to be exposed to outer world for jobs,buy grocesseries etc.

2.3 Conclusions on spread of covid in India

The top 3 most affected states in India due to covid according to our dataset based on the number of deaths in the period 2020 to 2021.
1.Maharashtra
2.Karnataka
3.Kerala
The first graph which has cumulative line curves of 3 states illustrates that Maharashtra has its curve in a bit higher position than Kerala and Karnataka.Maharastra has more deaths compared to Kerala and Karnataka.Maharashtra has second most highest population in India, and this can be a factor which could explain the curve.

In Maharashtra as we can see half of its total deaths due to covid have happend in same period as remaining states that is May 2021 to August 2021 . The other half of the deaths were some what evenly distributed across remaining 3 time periods . This shows Maharashtra has been facing deaths due to covid from the beginning of the pandemic . The huge population and the conjusted streets might be a factor of its wide spread of covid virus.
Karnataka also has similar trend and has most of its death in the time period May 2021 to August 2021. In the rest of the 3 time periods also it has considerable amounts of deaths .
In kerala it is May to August in the year 2021. Also Kerala has less deaths in the period March 2020 to June 2020. It shows Kerala was not that effected in initial Phases of covid but later it has increased.

Chapter 3

Possible future work Pending

we have observed that there are more cases in the year 2021 compared to 2020 in worldwide as well as india wide,we want to analyse why this change has happend,like because of no lockdown,health issues or climate change etc.And we also wish to find what are the precautions we need to take inorder to overcome this kind of pandemics.We can also analyse from least 3 countries which suffered less from covid and educate other countries to defend the pandemic.Countries with huge population needs this to be done.

3.1 High level Documentation

Major part of the code was written in a single python script with name covidanaysis.py.As this file takes 3 csv files as inputs we have hardcoded the paths in code for simplicity.
This code functionalities have broadly 3 parts
1.Preprocessing part
2.Global data analysis part
3.India data analysis part
Preprocessing part reads the file "WHO-COVID-19-global-data.csv" which is a raw data of all countries so we read the data into list of lists,and sorted this data in ascending order based on the number of cases.From this sorted data we took last 3 countries as they are most affected countries and written this data into anothar csv file called top3data.csv.
The global data analysis part reads "top3data.csv" into list of lists and it has code which generates cumulative line graph,then 3 piecharts corresponding to 3 countries.Here we also plotted one bar graph for male females deaths data which was read from "top3malefemale.csv",this file was preprocesed from other csv file called Dataset-historic.csv using sed,awk scripts.
In India data analysis part we read "covid19india.csv" into list of lists and grouped the data by statename then sorted it based on the date.From this we created 3 lists each containing data of 3 states and plotted the cumulative line graph,3 pie plots corresponding to each state as done with the countries.
Directory Structure
——-awkscript.awk
——preprocessing.sh
——covid-analysis.py
——(allinput .csv files)
——-graphsimagesand-html———covid-webpage.html(also graph images)
——-latexfilefor-report———(latex file for report pdf generation)
——-projectreport.pdf

3.2 compilation/running instructions

As all the input files have been hardcoded we can simply run the following command which will generate 9 graph images as output.
"python3 covidanalysis.py"
with above directory structure mentioned. Also in folder graphsimages-html we can open covid-webpage.html which will display all the above generated graphs in a web page abstracting out the inner details with description written for each graph