[Devastating Impact of Covid Pandemic]

**Data Science Capstone Project   
Data Acquisition and Pre-Processing Report**

Date

[May/12/2023]

Team Members:

Name: Shanti Mutyala

**Identifying Data**

**Data Sources:**

I have used following data sources:

1.Following data will be considered for studying the effect of COVID pandemic and number of deaths and number of cases in different countries :

<https://www.kaggle.com/datasets/imdevskp/corona-virus-report>

2. Also used the following data in conjugation with the previous one, so I can get population for each country which can help us in further analysis

<https://www.kaggle.com/datasets/josephassaker/covid19-global-dataset>

3. For the tweet related data, I will be using the following data, which will be containing information about major tweets around that time

<https://www.kaggle.com/datasets/gpreda/covid19-tweets/code>

4. Next I will be using the following data which will be useful for studying the number of hours lost during covid period related to job.

<https://www.kaggle.com/datasets/vineethakkinapalli/impact-of-covid19-on-employment-ilostat>

**Acquisition Process:**

All the data were in already downloadable form, I downloaded them manually and extracted the zip folder and used pandas csv reader for reading those files in the form of dataframes.

All the data sources were public and I have used multiple data sources as mentioned in the previous section. I have used each data source separately and didn’t involve any data merging.

**Issues:**

1. For the tweets related data source I first tried for a MegaCov dataset (<https://www.kaggle.com/datasets/vineethakkinapalli/impact-of-covid19-on-employment-ilostat> ) but its size was heavy(greater than 10 GB) so I cannot load it directly in our pc (due to computational requirements) and also the file was in json format so selective reading was also not possible so I have turned to another smaller public dataset. So finally this issue was solved with a small dataset.
2. I have not been able to find any public data source related to financial revenue for the companies, for which I will be experimenting more and try to integrate them to make a proper data for this.

**Data-Processing**

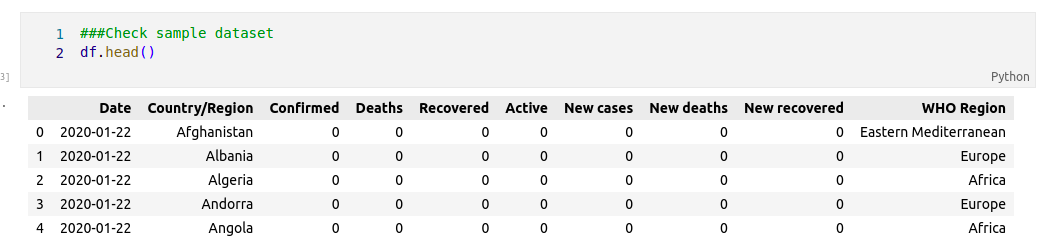
[Examine the data you have acquired and understand the data properties. Is there any pre-processing you need to do before you can start analyzing the data? For example, missing data, sparsity, noise, veracity, ambiguity, interoperability, etc. Describe each data issue in a sub-section and explain how you clean up the data.]

1. For corona-virus-report dataset there were multiple tables so I have to make through analysis to understand the data structure
2. I also did grouping and sorting operations to fetch top infected countries information from the dataset.
3. I also created some new feature by dividing ‘TotalDeaths’ with ‘Population’ for computing ‘Mortality Rate’
4. For text data related to tweets I also did some preprocessing like splitting texts and manually segregating months July and August Data for better study.
5. I also generated some new features by checking with ‘Job’ or ‘Health’ related terms in the tweets to understand the conditions prevalent at that time.
6. Also as I did study there were no major missing data found in our datasets so cleaning was not much needed for this datasets.

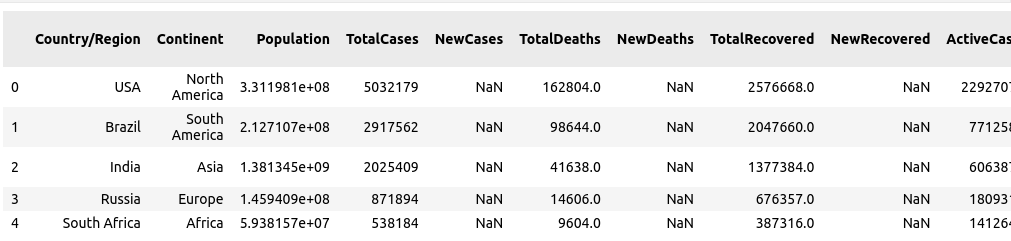
**Appendix**

Sample datasets has been provided below:

1. Related to Covid-report-data, I can see that date and number of Confirmed Cases, Death, Recovered and Active Cases are given as below for various countries.



1. In the worldometer data for covid, I have population detail as well along with the covid related cases information as shown below:



1. Related to Employment Related Dataset, where I have information related to hours lost due to covid for various countries. Also the dataset has been divided for percentage basis as well for better study.

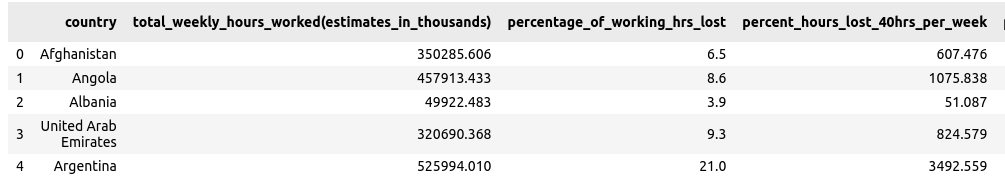


Table of Contributions

The table below identifies contributors to various sections of this document.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Data Sources** | **Shanti** | **Shanti** |
| **2** | **Data Pre-Processing** | **Shanti** | **Shanti** |
| **3** | **Appendix** | **Shanti** | **Shanti** |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.