[Devastating Impact of Covid Pandemic]

[Shanti Mutyala]

**Data Science Capstone Project   
Exploratory Data Analytics Report**

Date:

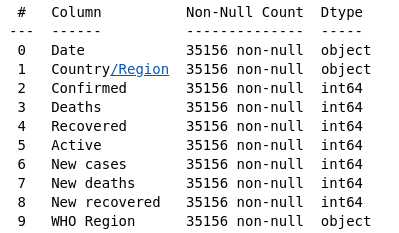
[May /18/2023]

Team Members: Shanti Mutyala

**Analysis the basic metrics of variables**

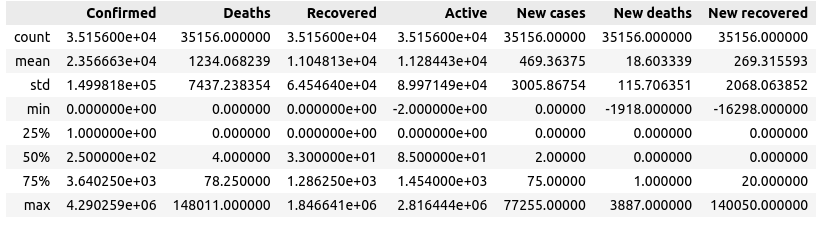
[In this section, we identify all the variables in the dataset and conduct the basic metrics of the variables. What are the data types (numerical/categorical, discrete or continuous, ordinal or nominal) and size? Provide the descriptive statistics of the variables such as mean, standard deviation, min, max, percentiles, etc.]

**COVID-19 Dataset:** For this dataset we can see that Date, Country and WHO Region are categorical values, rest all are continuous values as we can see in the below table.

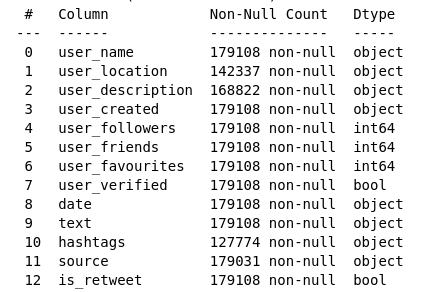


We have also observed that the 187 unique categories are present for country features and 6 for WHO Region.

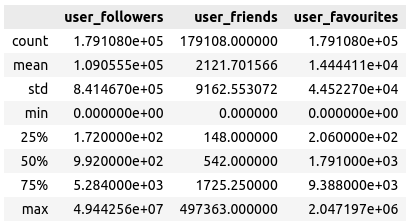
We have also checked the descriptive statistics of continuous variables as below:



**COVID Tweets Dataset:** For the covid dataset we can see following data information, where most of the features are object type as they are carrying english words/tweets. Detailed list can be seen below:



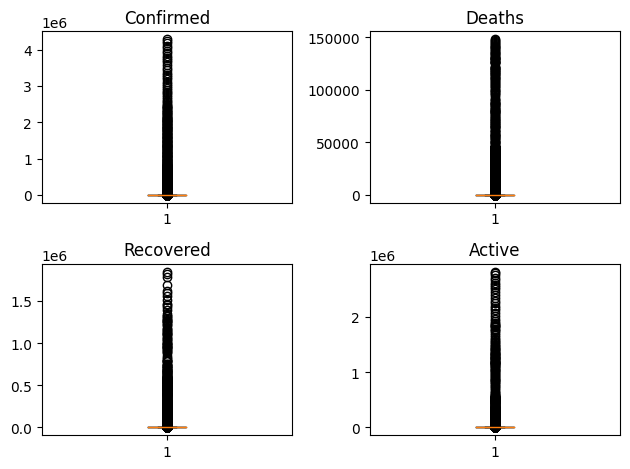
Also the descriptive statistics for continuous variables are as below:

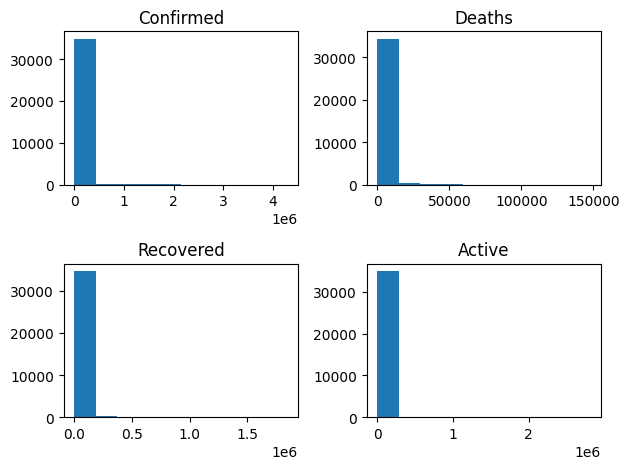


**Non-graphical and graphical univariate analysis**

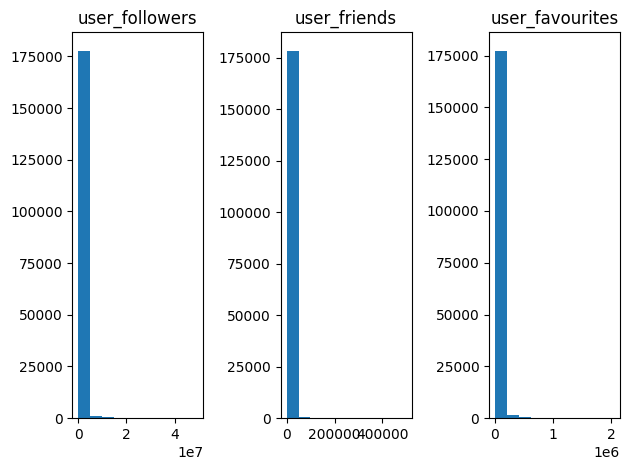
[In this section, we identify the list and number of unique values for each variable and provide the histogram and box plots to understand the distribution of the data.]

**COVID-19 Dataset:** We can see from below plots that mostly all the features are skewed.

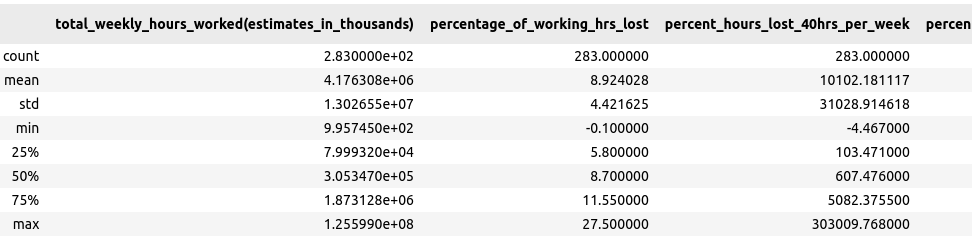
****

****

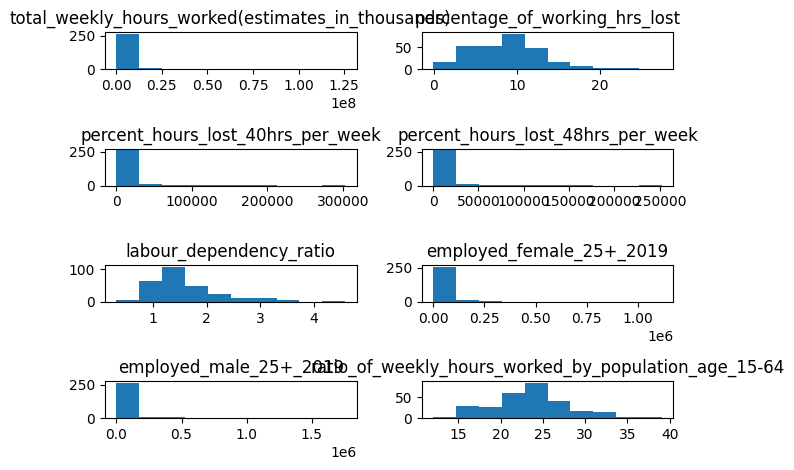
**Covid-19 Tweets Dataset:** We can see the histogram plots for the user\_friends, user\_followers and user\_favorites dataset as below.



**Job Market Dataset:** We can see descriptive statistics data here.



And the boxplots can be seen as below:

****

**Missing value analysis and outlier analysis**

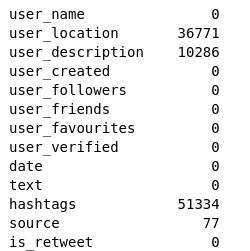
[In this section, we identify the missing values and outliers and determine how we handle these values before analysis.]

**COVID-19 Dataset:** We have checked no null values are present in this dataset.

We have also found the outliers using IQR Method where and found the below result:

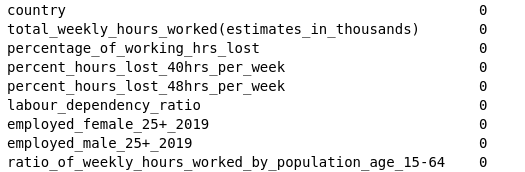
* Percentage of Outliers in Confirmed Data is 17.96 %
* Percentage of Outliers in Deaths Data is 18.43 %
* Percentage of Outliers in Recovered Data is 17.68 %
* Percentage of Outliers in Active Data is 17.93 %
* Percentage of Outliers in New cases Data is 18.06 %
* Percentage of Outliers in New deaths Data is 20.96 %
* Percentage of Outliers in New recovered Data is 18.98 %

**Covid-19 Tweets Dataset:** We can see that the following number of data points are missing. Mostly user\_location information and hashtags are the missing data.



**Job Market Analysis Dataset:**

We checked there are no null data found in this dataset.



**Feature engineering and analysis**

[In this section, we identify the variables that are useful for predictive modeling and machine learning through correlation analysis. You may also reduce the dimension or derive new variables so that the predictive modeling can be more efficient and effective.]

**COVID-19 Dataset:**  We have created new variables like number\_of\_infected, number\_of\_mortality, mortality\_rates as new variables by grouping the dataset based on Country/Region and then used it for our further study purpose.

**Covid-19 Tweets Dataset:**  For analysis related to this dataset, we created top\_july\_tweets, top\_august\_tweets and top\_all\_tweets which are tweets of July and August month respectively. We used them to study the active users and what their country is for each of the months.

We also created the word cloud which has frequency of words as well from the overall aggregated tweets. This helps to check in one sight what is being discussed during these covid periods.



We also created some new features like count\_job\_tweets and count\_health\_tweets to check the number of tweets which have the word ‘job’ and ‘health’ in it to study which is a more discussed topic in this.

**Job Market Analysis Dataset:** We have created correlation analysis in this dataset to study how features are interrelated to each other in this dataset.

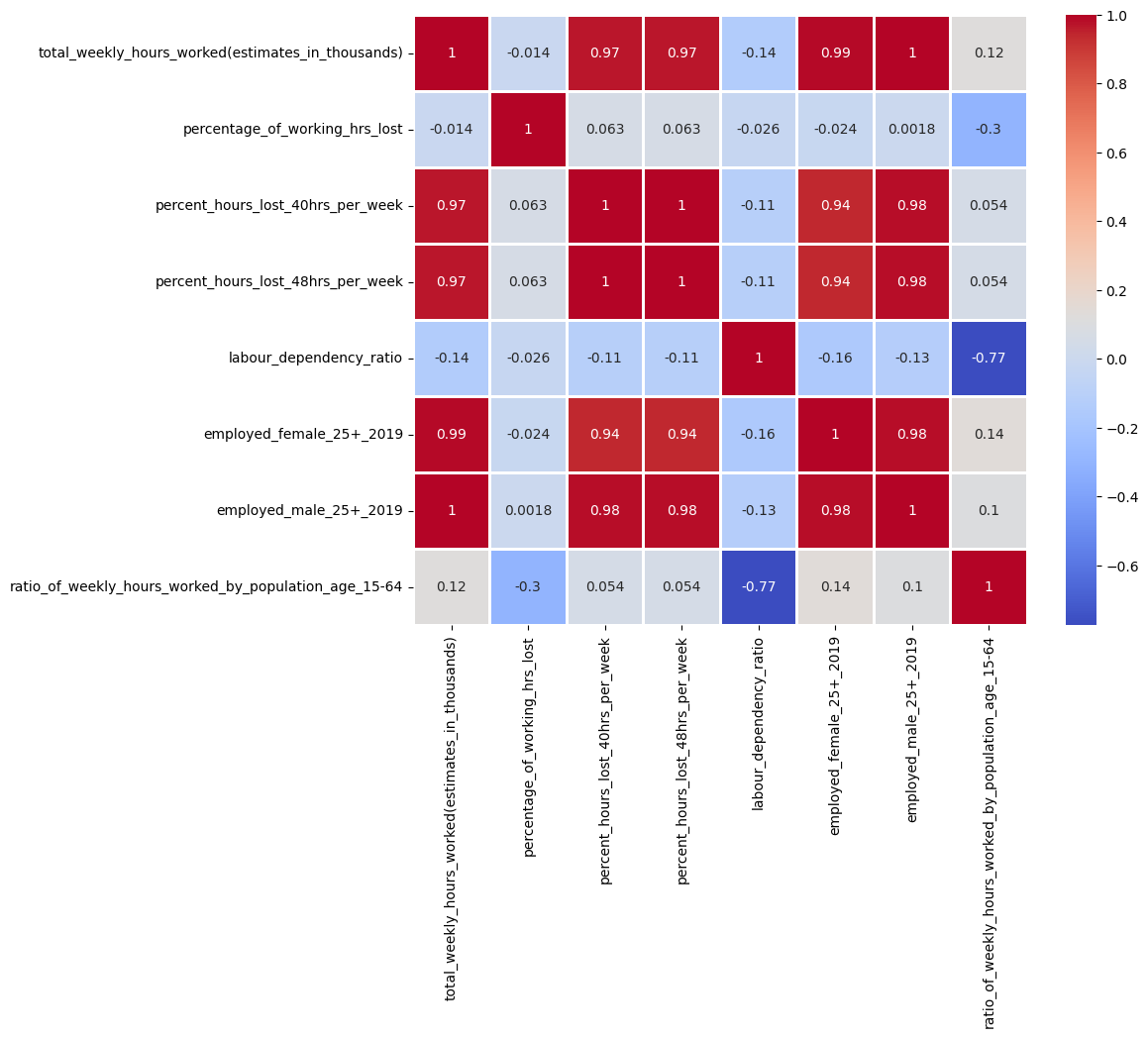


Table of Contributions

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Analysis the basic metrics of variables** |  |  |
| **2** | **Non-graphical and graphical univariate analysis** |  |  |
| **3** | **Feature engineering and analysis** |  |  |
| **4** | **Appendix** |  |  |

Appendix

The Jupyter Notebook Code file as attached with the document has all the code in it along with relevant comments.