SUBMITTED TO:

Mr. JEFF GEOFF ODOKONYERO

Recess supervisor

PREPARED BY:

GROUP P

**COLLEGE OF COMPUTING AND INFORMATION SCIENCES**

**SOFTWARE ENGINEERING YEAR TWO (2)**

**YEAR TWO FINAL RECESS PROJECT**

**(PYTHON DATA ANALYSIS)**

|  |  |
| --- | --- |
| RECESS REPORT FOR YEAR 2 DATA ANALYSIS  (COVID\_19 DATASET) | Abstract  This is a final report for the recess project 2022, Software Engineering Year 2 prepared by recess group P  COURSE TITLE  BSE2301 SOFTWARE ENGINEERING MINI PROJECT 2 |

September, 24th 2022

**GROUP MEMBERS**

|  |  |  |  |
| --- | --- | --- | --- |
| S/NO | NAME | STUDENT NUMBER | REGNO |
| 1 | MAMGBI GEOFREY | 2000722748 | 20/U/22748 |
| 2 |  |  |  |
| 3 |  |  |  |
| 4 |  |  |  |
| 5 | ATUHEIRE JOAN | 2000702022 | 20/U/2022/EVE |

**Executive summary**

This report summarizes the statistical and visual modelling and analysis results associated with the (COVID\_19 DATASET) provided including most if not all the countries in the world. The purpose of this report is to document all the results of analysis performed on the dataset, as in the findings of the various COVID\_19 cases in each of the counties and regions of the world that were provided on the dataset. This also includes all corresponding data modelling and inference techniques used during the subsequent statistical and visual analysis performed on the dataset.

1. **INTRODUCTION**

In this report, we include the results of the analysis done on the COVID\_19 dataset provided according to the team of the analysts in group P and serves as a guideline to the client.

The dataset provided for analysis is about the various corona cases in most countries of the world if not all countries and different methods of analysis are used as will be seen in the subsequent sections.

The dataset consists of a total of 15 columns and 187 rows. The columns include New cases, New deaths, New recovered, Deaths / 100 Cases, Recovered / 100 Cases, Deaths / 100 Recovered, Confirmed last week, 1 week change,1 week % increase, WHO Region.

The dataset has countries from six WHO Regions namely Africa, Europe, Americas, Western Pacific, South east Asia, Eastern Mediterranean

1. **OBJECTIVE OF ANALYSIS**

Due to it being hard to get information from a raw data, the data was put into analysis to generate value and information from the raw data provided.

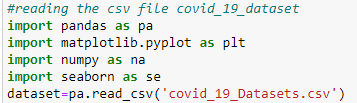
This analysis was to aid the statistical and modelling analysis and come up with final quantitative and visualization results that will aid the clients in effective decision making based on the results of the analysis.

This analysis basically was aimed at generating a conclusion about the corona virus cases of confirmed, active, recovered and death cases, and therefore help the clients to plan, and respond accordingly.

1. **PRE-PROCESSING OF THE DATASET**

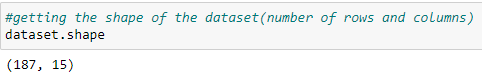
Before the analysis on the given dataset, some preprocessing was applied to the dataset in order to make it clean and ready for analysis. Among the preprocessing techniques used included the following if not all for the reasons described under each of the methods/commands used.

1. The dataset was first read using the relevant python libraries. This is to make available to us the analyst the dataset. The dataset was read and stored in a variable called **dataset** as is shown in the screenshot below.



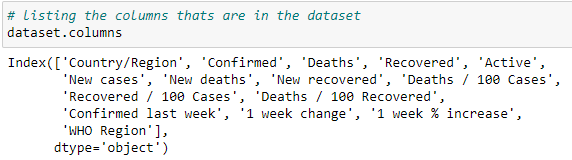
1. Shape of the dataset using the dataset.shape method.

We also printed out the shape of the dataset which give a matrix output of row by columns indicating how many rows and columns the dataset has as shown in the screenshot below.



From the screenshot attached, it indicates that the dataset has 187 rows /records which means the COVID\_19 dataset was collected from 187 countries/regions and 15 columns

1. We viewed the columns that are in the dataset using the dataset.colums method and below is the screen shot of the output

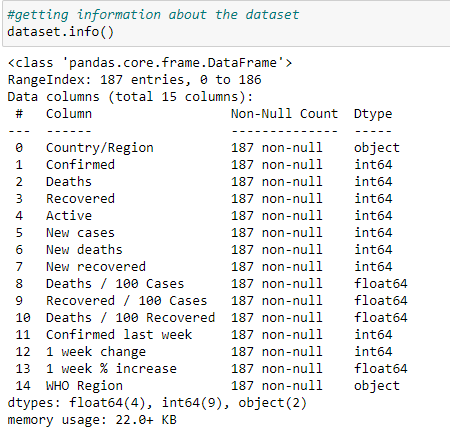


1. Viewing the first five and the last five records in the dataset using the head() and tail() methods respectively on the dataset. This gives us a representation of how the dataset looks like as illustrated in the screenshot below.



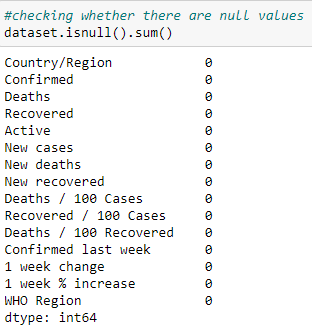
1. Viewing information about the dataset using the dataset.info() method

The output generated from this method includes the number of columns in the dataset, the column labels, column data types, memory usage by the dataset, the range index, and the number of cells in each column which are non-null values as indicated in the screen shot below.



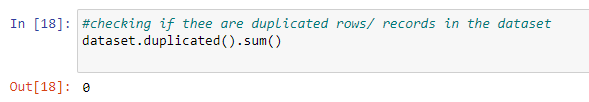
1. Checking null values in the dataset.

We also checked the given dataset against null values to see if there are missing values and get the total of the missing values in each of the columns and the output showed that there were no null/missing values and hence there was no need to drop null values since they don’t exist as shown in the output screenshot below.



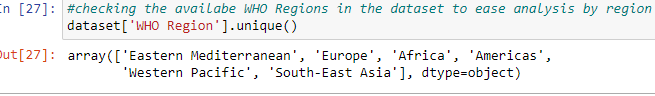
1. Checking duplicate values in the dataset.

We also checked the dataset against duplicate records or rows to see if records of a country are being repeated. However, the output showed that there were no duplicates and hence there was no need to drop the duplicated values since they don’t exist as shown in the screenshot below



1. Checking the number of regions in the dataset.

During analysis, we realized the number of countries to analyze were many that they couldn’t fit in visualization charts making it hard to visualize the dataset better. We therefore decided to analyze the dataset according to each WHO Region available in the dataset and below is a screenshot of the regions available



1. **KEY ANALYSIS (COLUMNS ANALYZED)**

In this analysis, the key columns analyzed