Ann Kalondu Mutua -sct221-0177/2021

BIT2203

# ADVANCED DATABASE MANAGEMENT SYSTEM

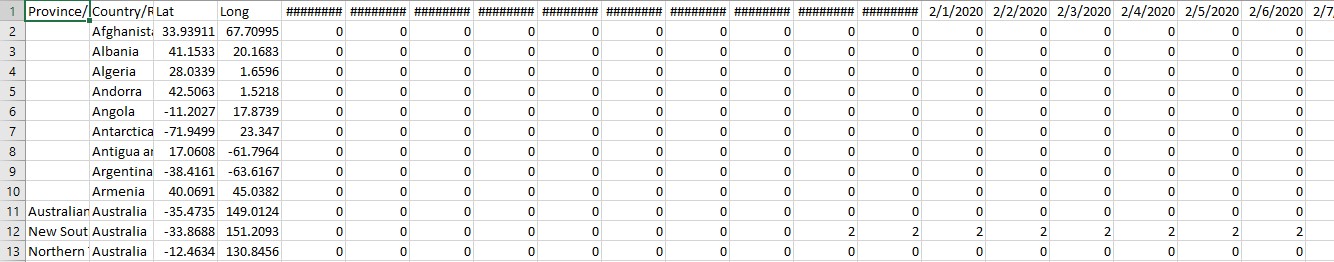
GROUP 10

**1. How data was compiled:**

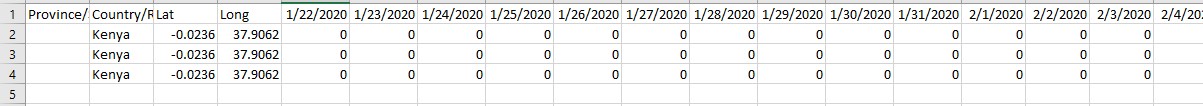
We got our data source from this link; <https://coronavirus.jhu.edu/>

The data came in three excel sheets that included; time\_series\_covid19\_recovered\_global.csv time\_series\_covid19\_deaths\_global.csv time\_series\_covid19\_confirmed\_global.csv

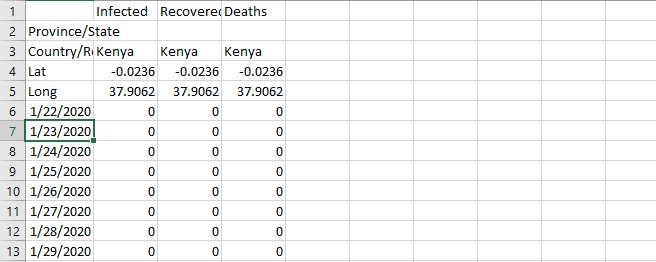
The data sources included data from all the countries in the world.



We created a separate excel sheets and copied Kenya covid 19 data to it from the three csv files.



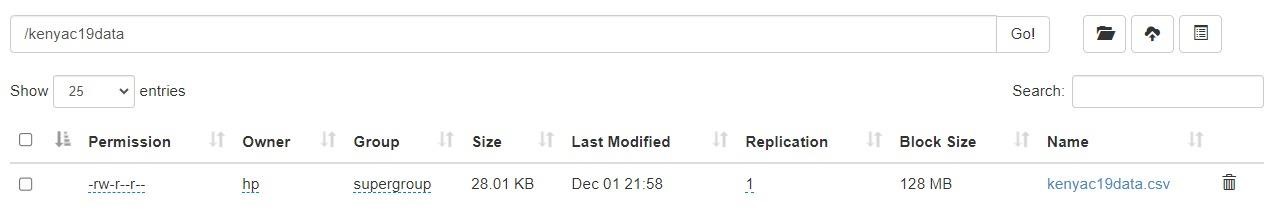
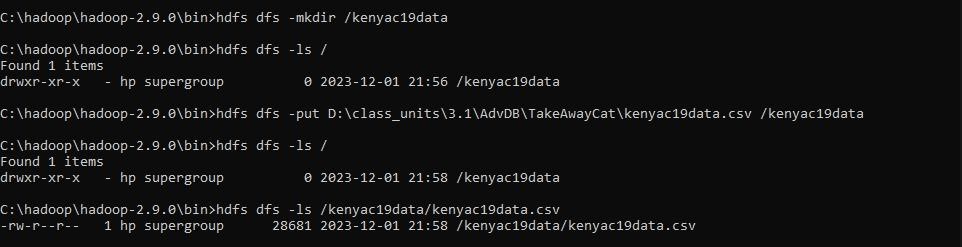
We then transposed the data cells and added new column heads the transformed the excel sheet to a csv file.



1. **How data was ingested into Hadoop data lake:**

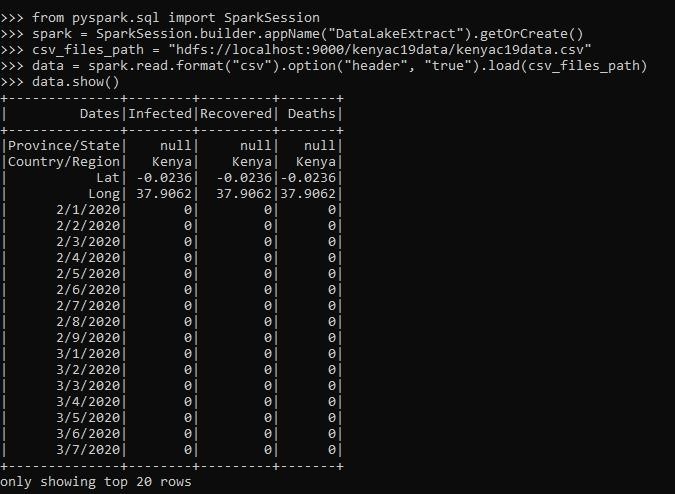
We created a directory using the command **hdfs dfs -mkdir**

**/kenyac19data** and inserted the prepared csv file into it using the command **hdfs dfs -put D:\filepath /kenyac19data**



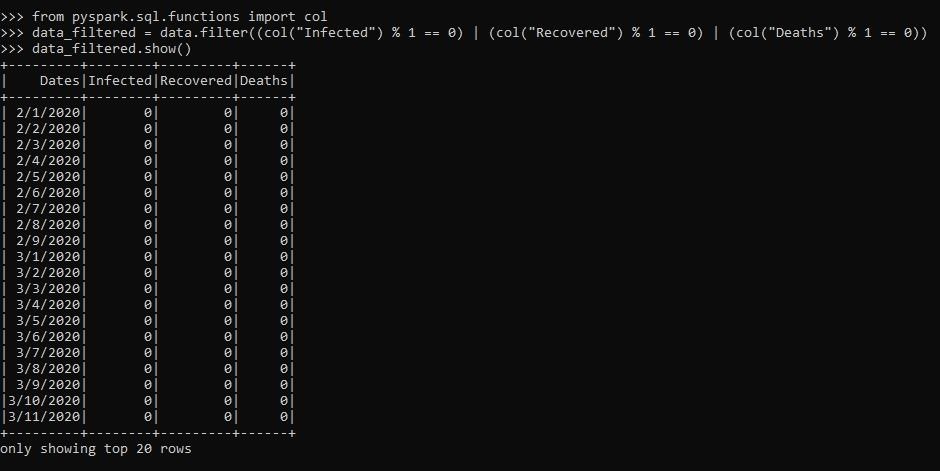
1. **How data was extracted using pyspark:**

We imported the SparkSession library, created a session then loaded our file from our Hadoop data lake into pyspark by specifying the files path. We then displayed the csv file.

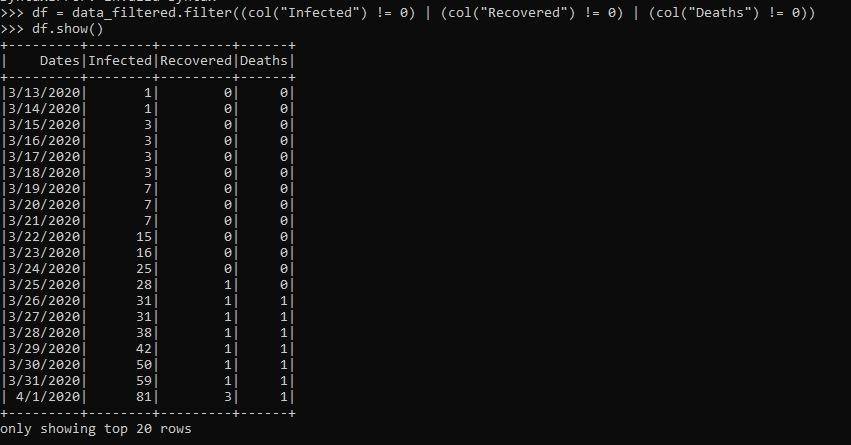


1. **Pre-processing tasks/techniques used to prepare data.**

We removed all the rows with null values, string values and floating point values. We remained with only integer values in the Infected, Recovered and Deaths columns which would be easier to work with.

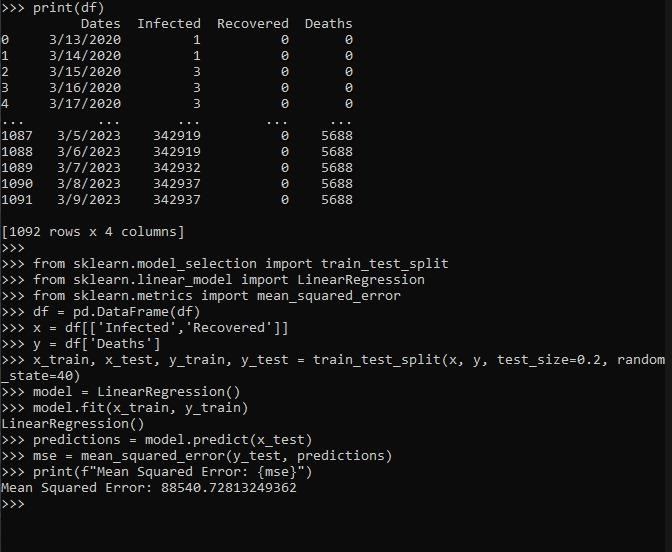


Next we removed all the rows that had zero (0) values in all the columns. The zero values in all columns makes the rows irrelevant as we need real numbers to work with.



1. **Test Results and Interpretation:**

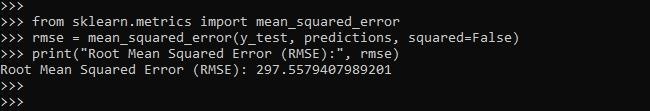
We used the Mean Squared Error predictive analytics method to predict the number of deaths. The MSE we found was as 88540, that for all the infections recorded and the rate of recovery, this is the number of people that are expected to perish from the virus.



1. **Validations results and interpretations:**

We validated the value we got from the Mean Squared Error technique using the Root Mean Squared Error technique. The value we got was the exact root of the value we found in the MSE.

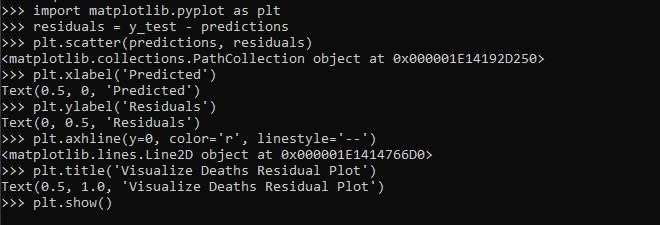
√88540.73 = 297.56



1. **Potential applications of the interpreted results:** 
   1. ***Public Awareness:*** Communicating anticipated death rates based on infection numbers can raise public awareness about the seriousness of the situation. This information can encourage adherence to preventive measures and vaccination, potentially reducing the spread of the virus.
   2. ***Mitigation Strategies:*** Knowing the potential death toll can prompt the implementation of targeted interventions in high-risk areas or among vulnerable populations, such as the elderly or those with preexisting health conditions.
   3. ***Policy Making:*** Predictions can inform policymakers about the potential impact of the virus, guiding decisions on lockdowns, social distancing measures, travel restrictions, and vaccination drives. It helps in creating a more targeted and effective response.

**Data Visualization:**

We imported the python matplotlib library to visualize the model.



The result was;

