**DBMSCAT.**

**JEROME OBENJO-SCT222-0154/2020**

1. Compile COVID-19 related Data(preferably data from Kenya) from relevant online sources such as <https://coronavirus.jhu.edu/>

[https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distributioncovid-19-cases-worldwide](https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide)

<https://www.worldometers.info/coronavirus/#countries>

[https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9 ecf6](https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6)

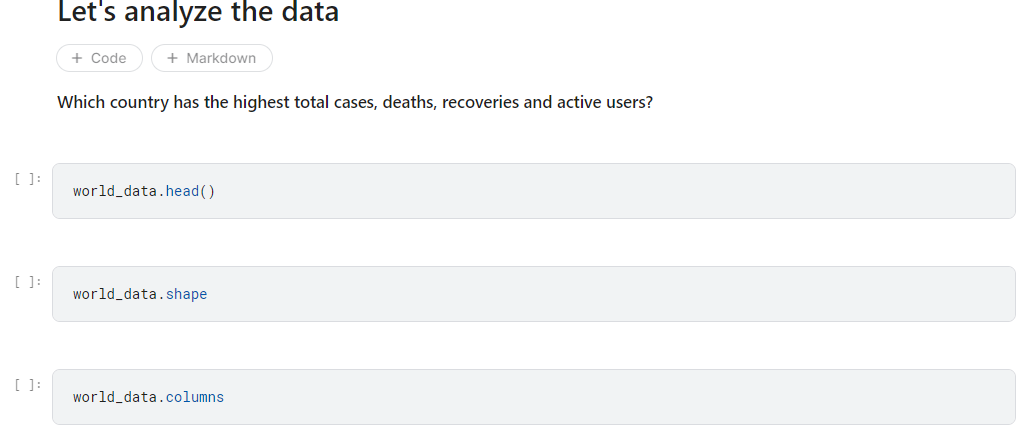
<https://vincentarelbundock.github.io/Rdatasets/datasets.html><https://africaopendata.org/group/kenya> <http://www.opendata.go.ke/>

1. Ingest the data into Hadoop DFS Data Lake //jupyter notebook



1. Use pyspark package to extract the data from the data lake

1. Choose appropriate techniques to Pre- process the extracted data
2. Apply one predictive analytics technique to generate a model for predicting any of the following cases: a) Number of Death cases or Mortality rate
   1. Number of confirmed cases
   2. Number of recovery cases or Recovery rate



1. Visualize the model

1. Test the model
2. ***Validate the Model***
3. Compile pdf processed document that has the following content:
4. Describe how the data was compiled in task 1 and include Screen captures of both code &and output) (3 Marks)
5. Describe how data was extracted using pyspark and include associated screen shots
   1. Marks)
6. Describe pre-processing tasks/techniques used to prepare the data (include screen shots) and

|  |  |  |  |
| --- | --- | --- | --- |
| give reason (s) to justify your choices  **Data Cleaning:**  Removing duplicates: If there are any duplicate entries in the dataset, they need to be identified and removed. Duplicates can skew the analysis results.  **Data Transformation:**  Feature scaling: Scaling numerical features to a common range (e.g., normalization or standardization) can prevent certain features from dominating the analysis due to their larger magnitude.  **Data Integration:**  Combining multiple datasets: If there are multiple sources of COVID-19 data, integrating them can provide a comprehensive view for analysis. This can involve merging datasets based on common identifiers (e.g., country or date).  **Data Aggregation:**  Summarizing data at a higher level: Aggregating COVID-19 data by country, region, or time period can provide a broader perspective and facilitate analysis. | | (3 Marks). | |
|  |
| (vi) Test results and interpretations | |  | (3 Marks) |
| (vii) Validation Results and interpretations | |  | (3 Marks) |
| (viii) Potential applications of the interpreted results | |  |  |
| 10. Present your work in class on 23rd NOV. 2023 | |  | (5 Marks) |

11. Host your **PDF**-processed **document** a **text file** of list of commands used to GitHub and submit your details and link by filling in the form here:  by 1st DEC 2023.





