

ADVANCED DB CAT

MARK NYANGADA – SCT222-0244/2020

JOSEPH OGURI – SCT222-0304/2020

VALENTINE WANJIKU - SCT222-0128/2020

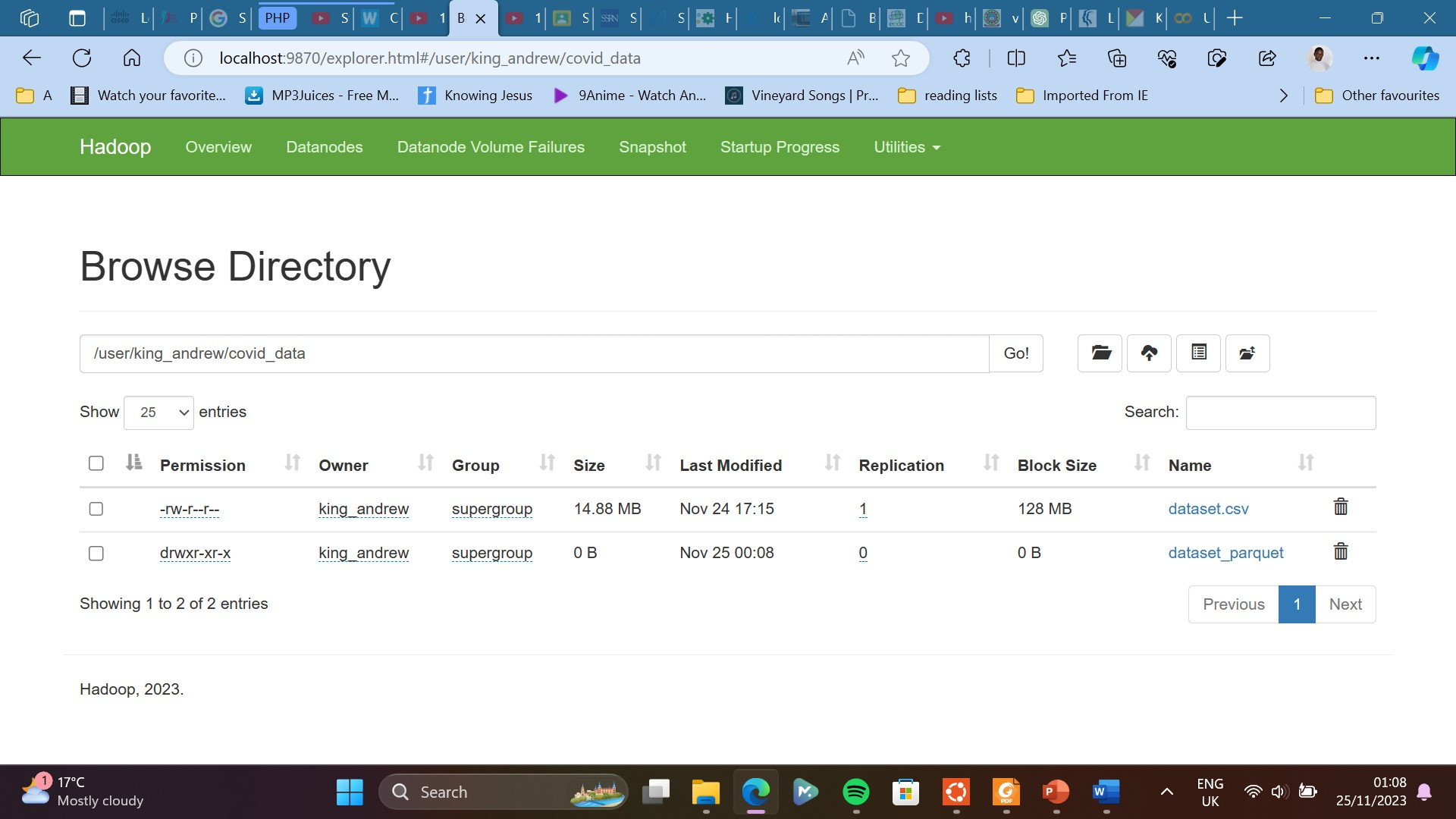
ALPHONCE KOTIENO - SCT222-0181/2020

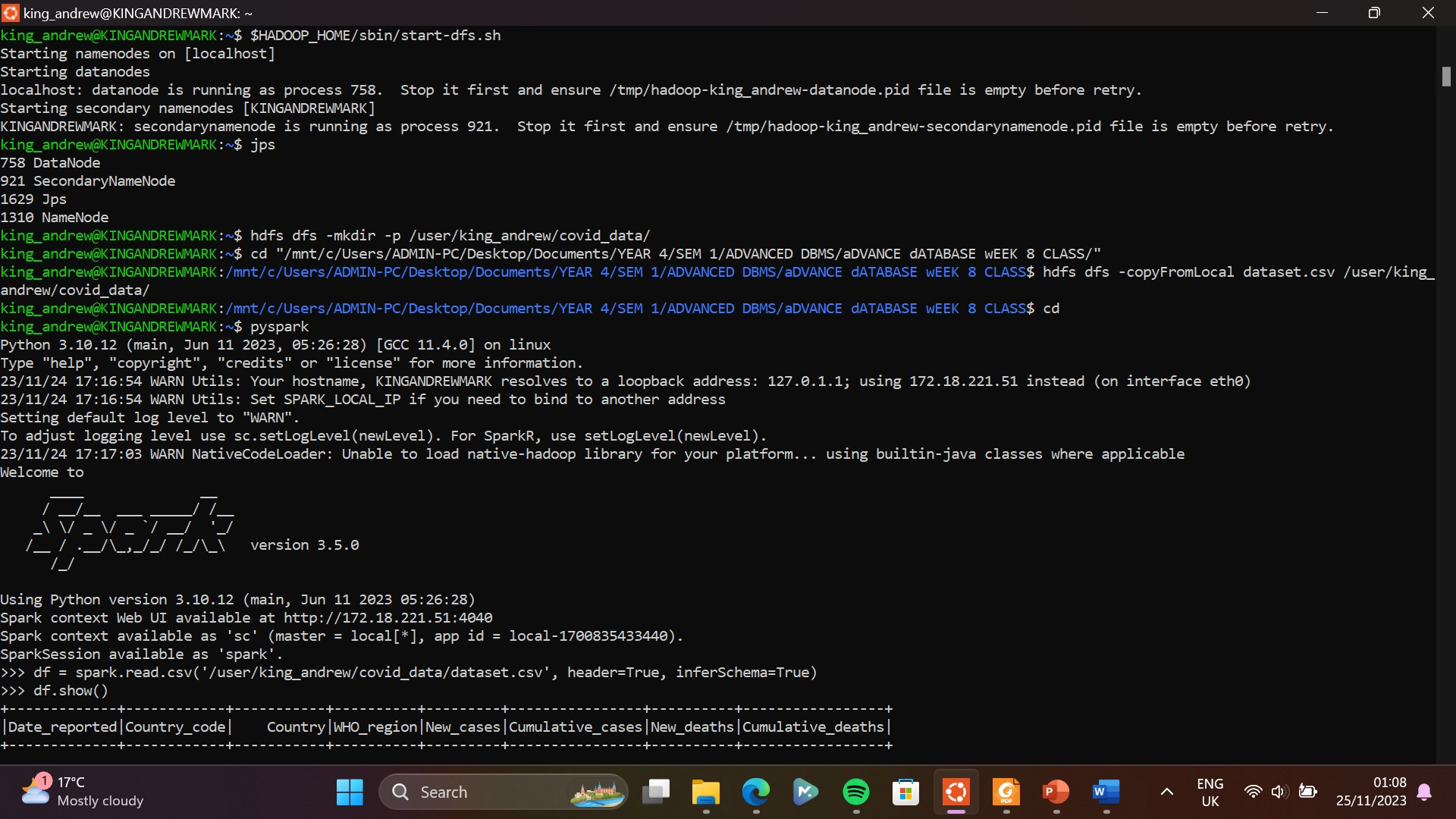
November 24, 2023

1. **Describe how the data was compiled in task 1 and include Screen captures of both code and output**

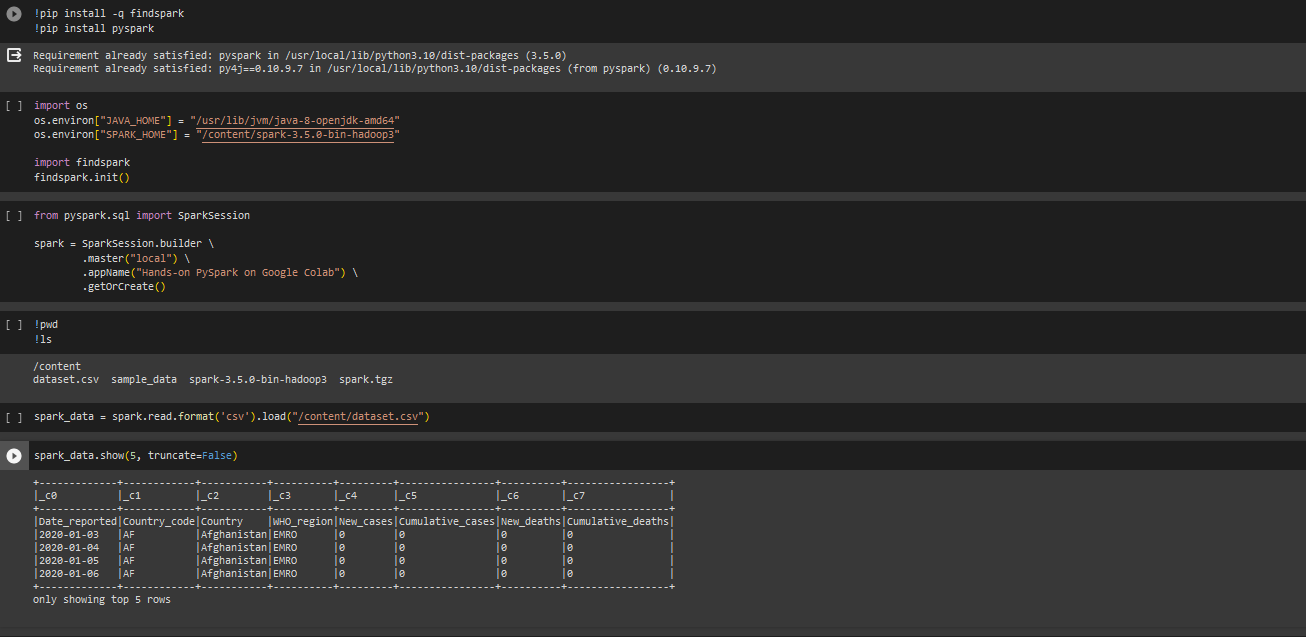
Data was compiled through hadoop's core components: Hadoop Distributed File System (HDFS) for distributed storage and MapReduce for processing. Data was stored in smaller blocks across a cluster using HDFS, allowing parallel storage and retrieval. In the MapReduce programming model, data was processed and compiled through the Map and Reduce phases, enabling parallel and distributed computation

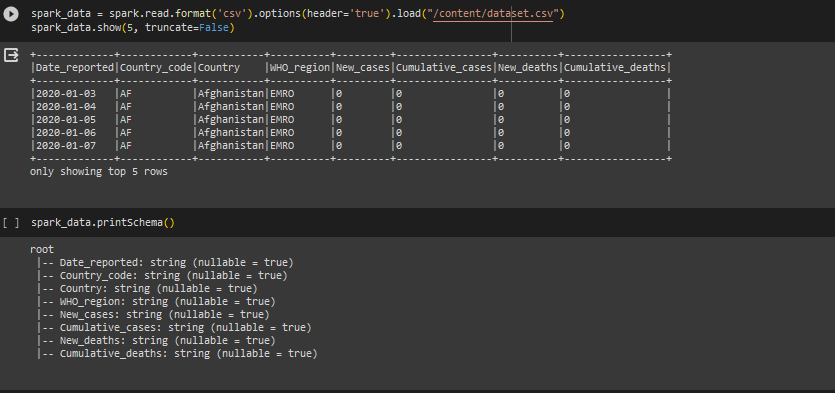
1. **Describe how the data was ingested into Hadoop data lake and include screenshots**

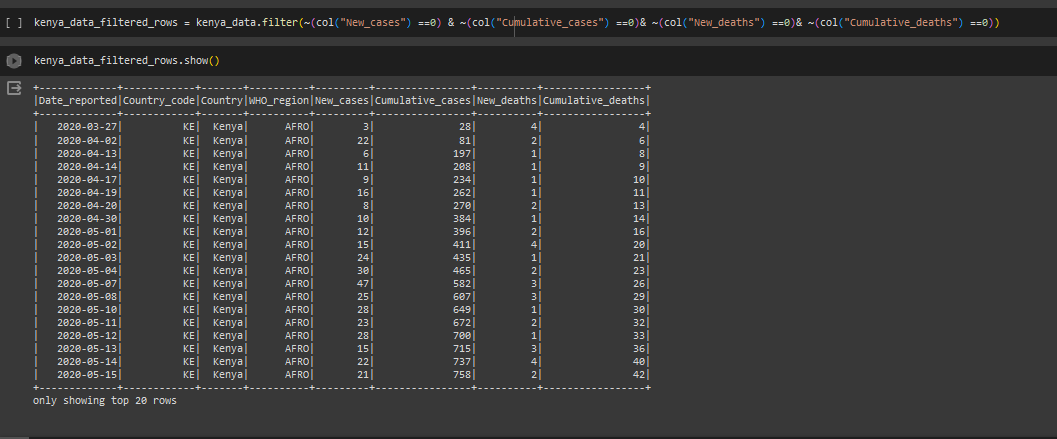


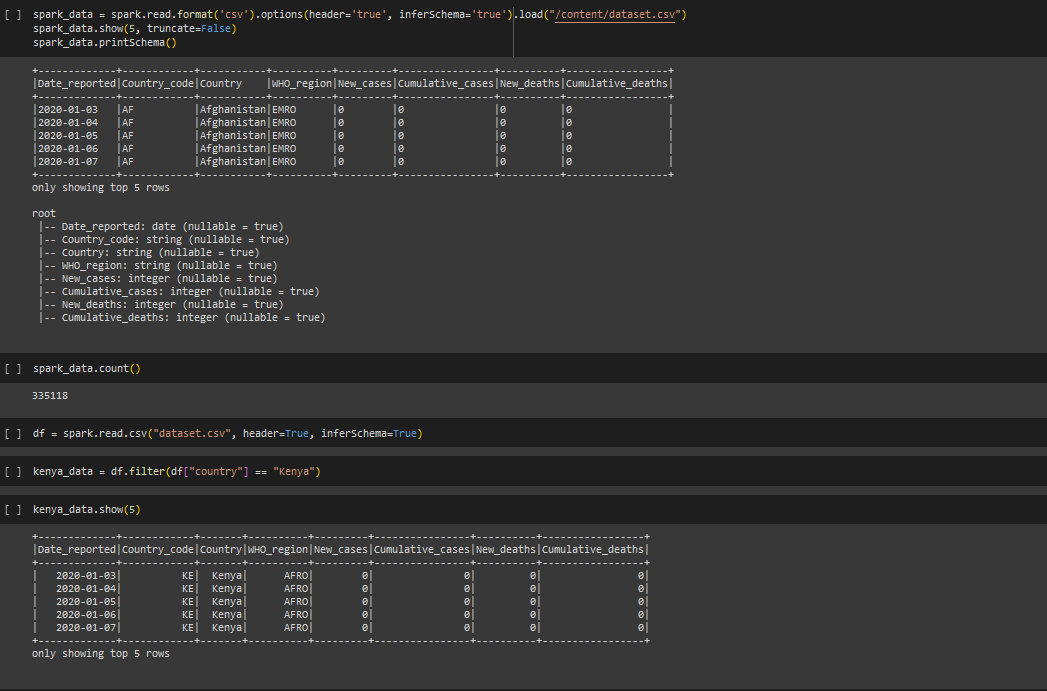


1. **Describe how data was extracted using pyspark and include associated screen shots**

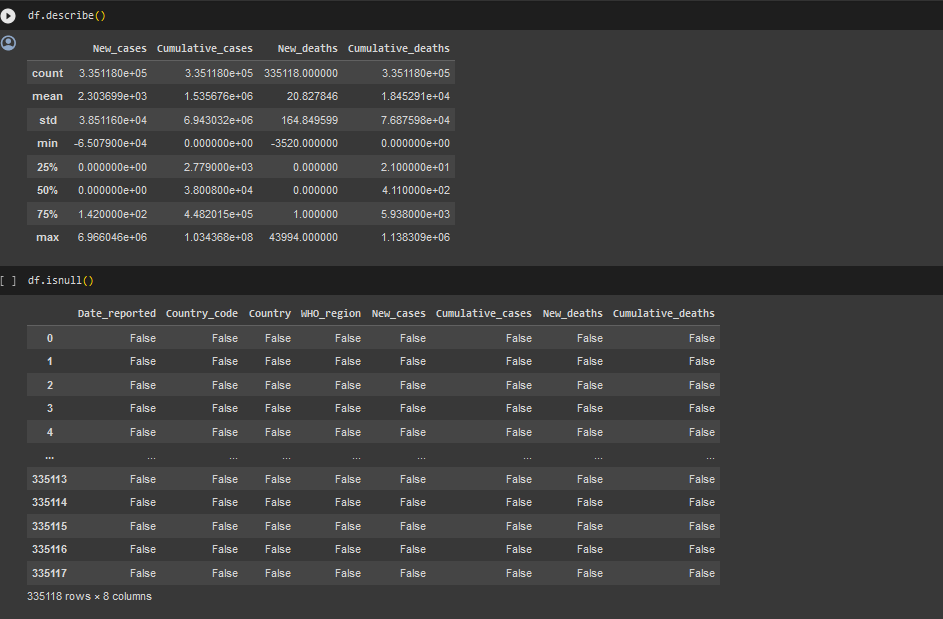
****

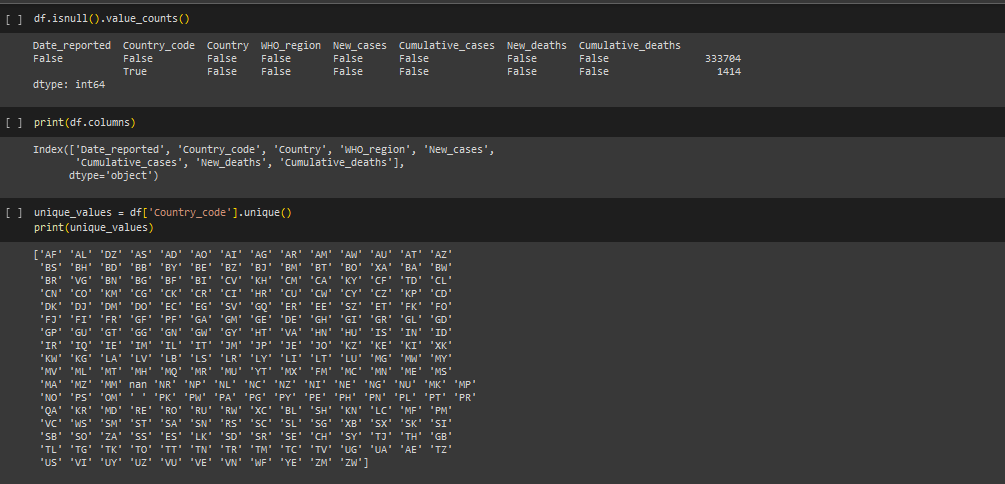
****

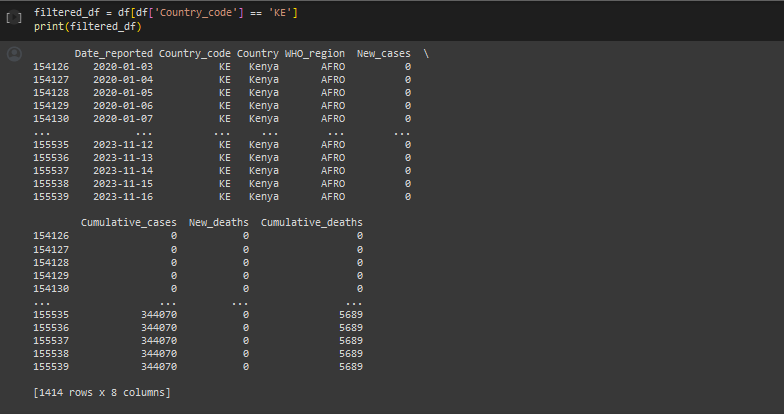
****

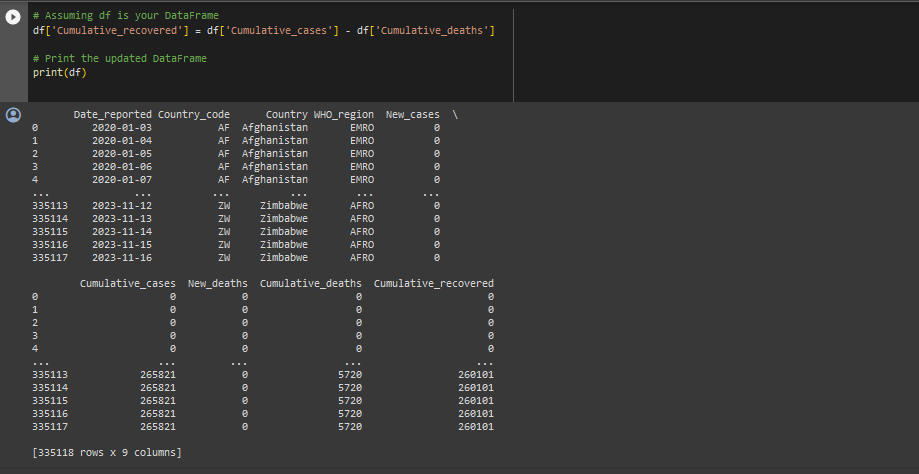
****

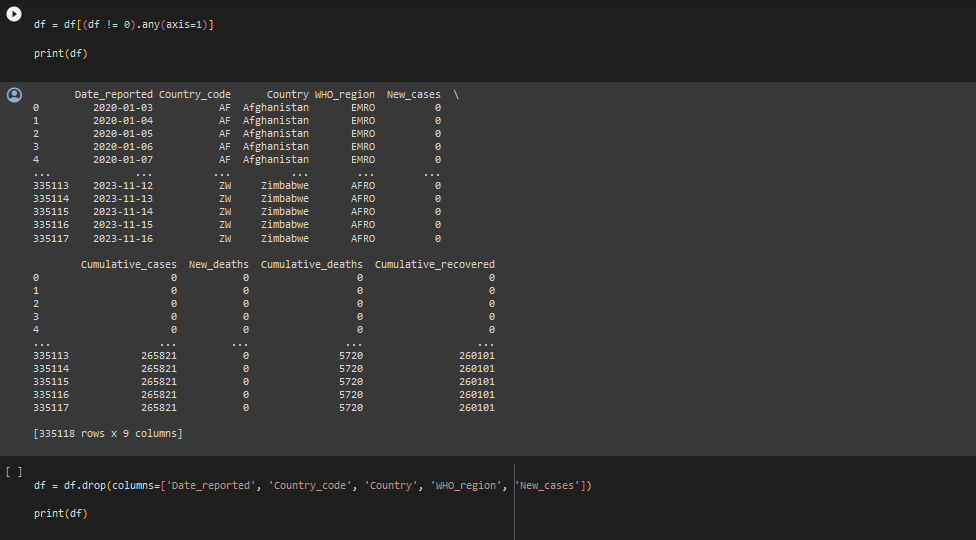
1. **Describe pre-processing tasks/techniques used to prepare the data (include screenshots) and give reasons to justify your choices**
2. Collection and loading data into python: this is the first step in the data analysis process. Python with libraries like pandas, provide powerful tools for data manipulation and analysis. Using pandas to load data allows for efficient handling of datasets making it easy to perform various operations.
3. Isnull and unique values: isnull identifies and handles missing values (NaN values). This is essential because many machine learning algorithms cannot handle missing values. Unique values checks unique values in categorical columns helping in understanding the diversity of data. Useful for identifying cardinality of categorical features and deciding on encoding strategies. It assists in identifying potential issues like typos or inconsistencies in categorical data.
4. Filtering only values from Kenya: is essential when you are interested in analysing or modelling data for that specific entity. It helps to focus the analysis, gain insights into country-specific trends and is crucial for tasks like building country specific predictive models or generating reports tailored to a particular region.
5. Getting recovered by subtracting deaths from cases: assumes that individuals who aren’t active cases are recovered. It provides a quick estimate and is useful for certain types of analyses, especially when detailed recovery data is not available.

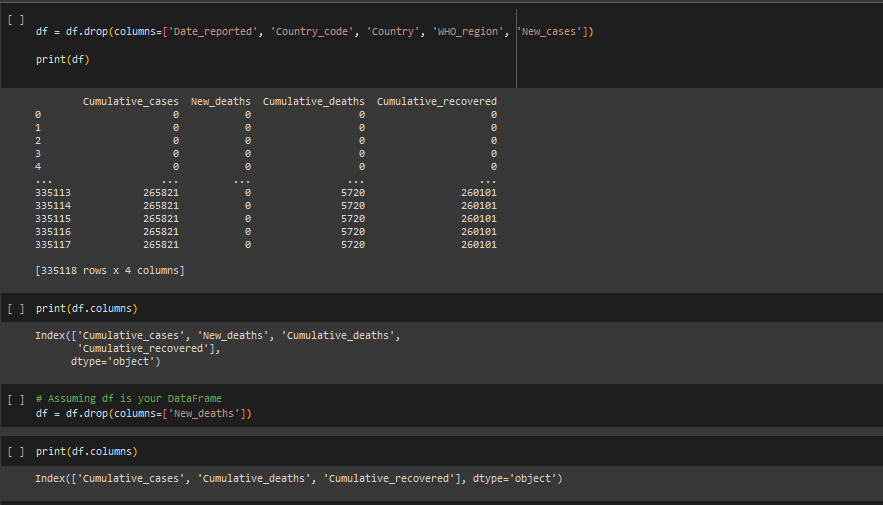
****

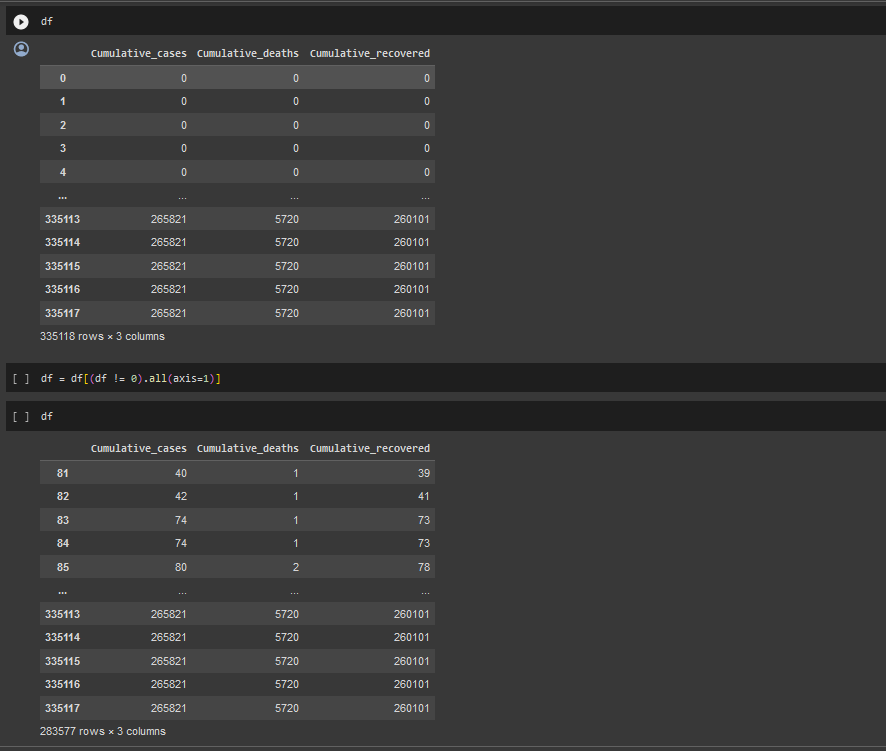
****

****

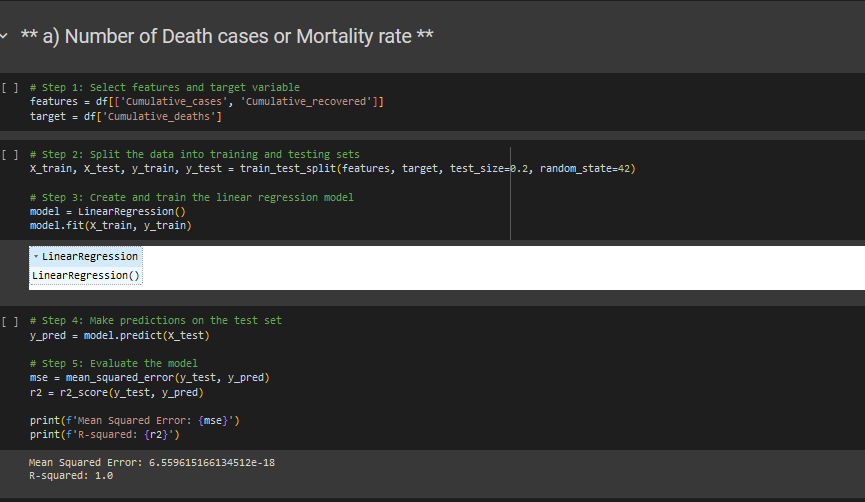
****

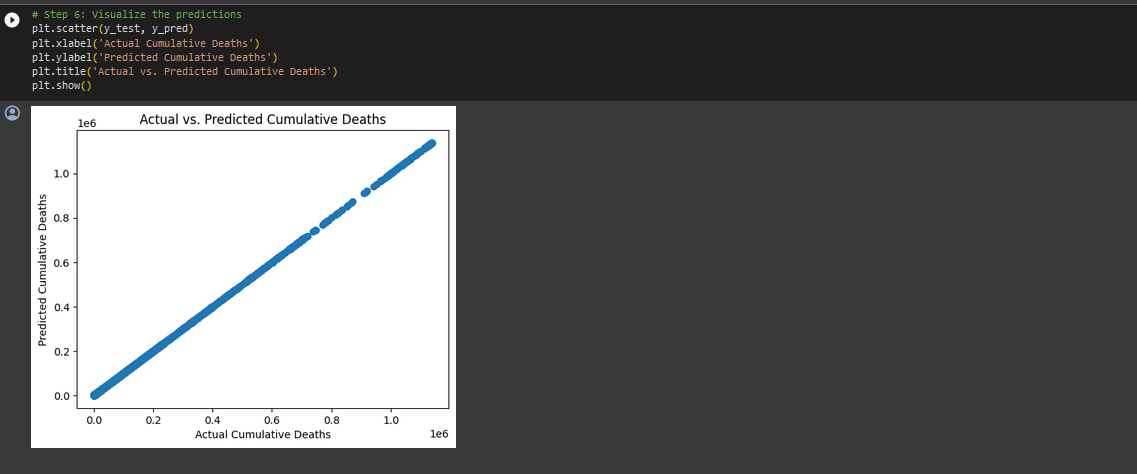
****

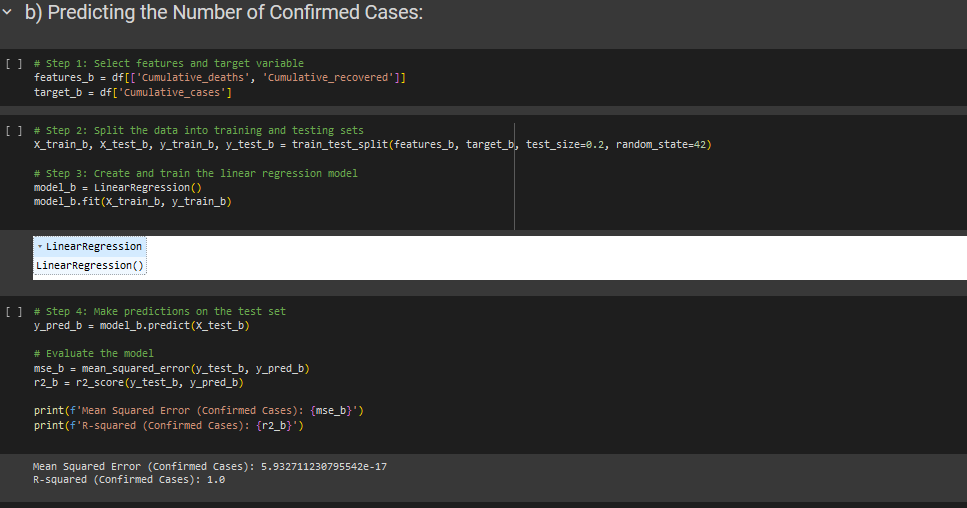
****

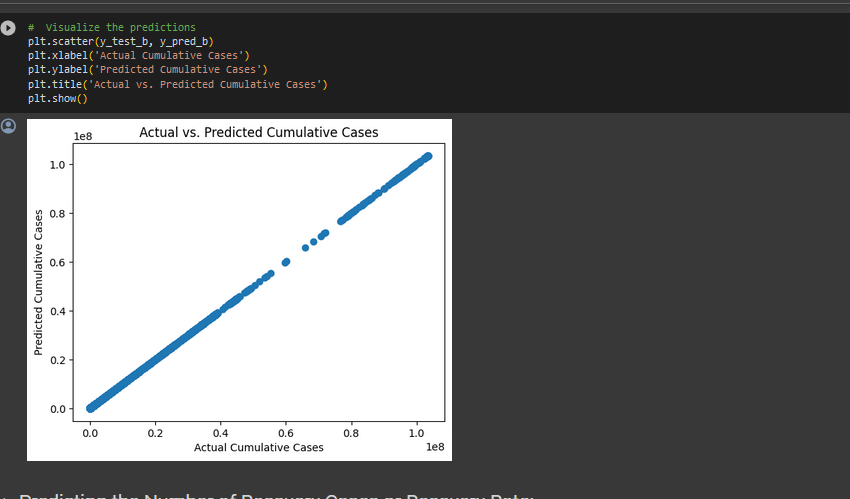
****

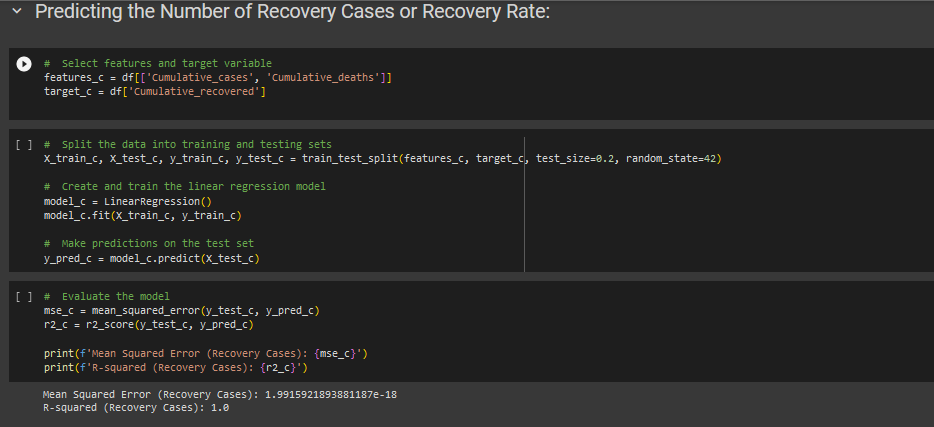
1. **Test results and interpretations and Validation results and interpretations**

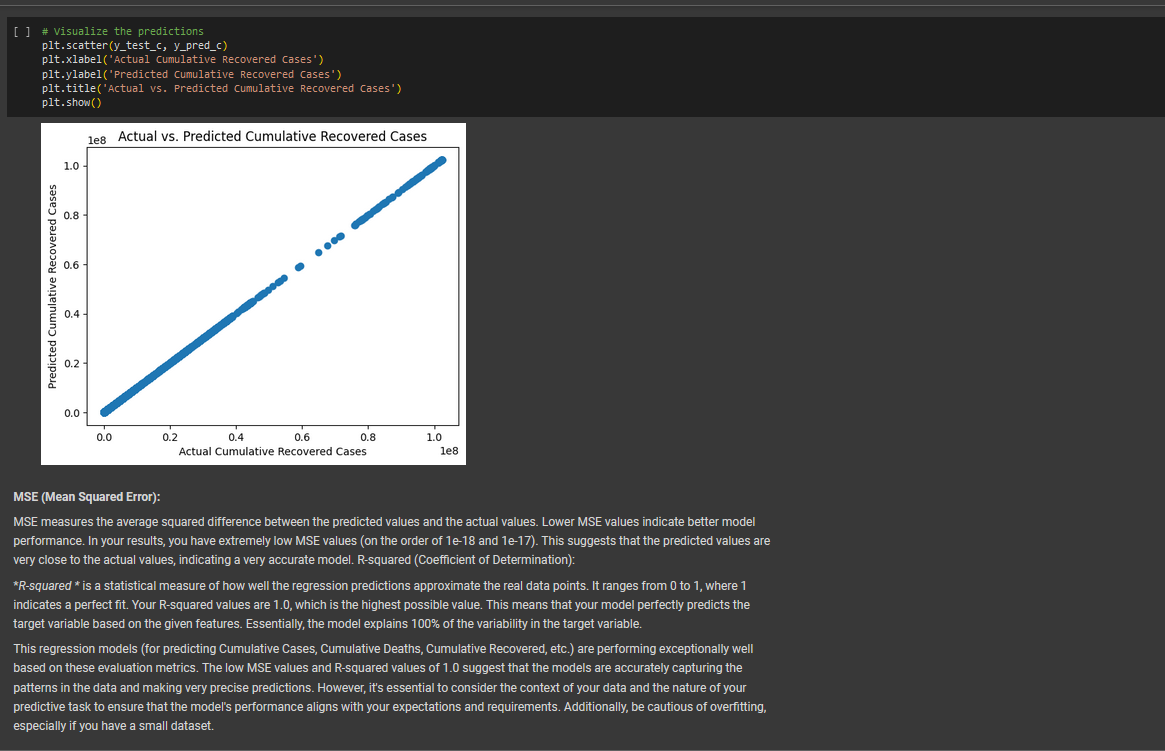
****

****

****

****

****

****

1. **Potential applications of the interpreted results**
2. Public Health Policy Planning:

* Mortality Rate Analysis: Understanding the mortality rate provides critical insights into the severity of the disease. Governments and health organizations can use this information to plan and implement public health policies, allocate resources effectively, and prepare healthcare systems for potential surges in cases.
* Confirmed Cases Analysis: Analyzing the confirmed cases helps in assessing the overall disease burden. This information is valuable for resource allocation, determining testing and healthcare infrastructure needs, and understanding the scale of the outbreak.
* Recovery Rate Analysis: Monitoring the recovery rate is crucial for evaluating the effectiveness of healthcare interventions. High recovery rates can indicate the success of treatment protocols and guide decisions on patient care.

1. Resource Allocation and Preparedness:

* Mortality Rate Forecasting: Predictive analysis can be used to forecast mortality rates based on current trends. This information assists in proactively allocating medical resources, including hospital beds, ventilators, and medical personnel, to regions at higher risk.
* Scenario Planning: Understanding potential future scenarios based on predictive analysis helps authorities plan for different outcomes. This includes estimating the number of hospitalizations, ICU admissions, and overall healthcare needs.

1. Community Awareness and Education:

* Communication Strategies: Interpreted results can inform public communication strategies. Health authorities can use mortality rates, confirmed cases, and recovery rates to communicate the seriousness of the situation, promote preventive measures, and provide realistic expectations to the public.
* Targeted Interventions: Knowing the mortality rate and distribution of confirmed cases allows for targeted interventions in specific communities or demographics. This can include focused testing campaigns, vaccination drives, and educational programs.

1. Research and Development:

* Vaccine and Treatment Research: Predictive analysis can guide research efforts, helping prioritize vaccine development and treatment strategies. It can identify areas with high mortality rates where interventions are urgently needed.
* Epidemiological Studies: Understanding the patterns of confirmed cases and recovery rates contributes to epidemiological studies. Researchers can explore factors influencing the spread of the disease and identify characteristics associated with higher recovery rates.

1. International Collaboration:

* Global Response Coordination: Sharing interpreted results internationally facilitates collaboration in response efforts. Countries with successful strategies can share insights and support regions facing challenges.
* Policy Harmonization: Consistent data interpretation enables better coordination of policies and strategies across borders. This is especially important in the context of a global pandemic where international cooperation is essential.