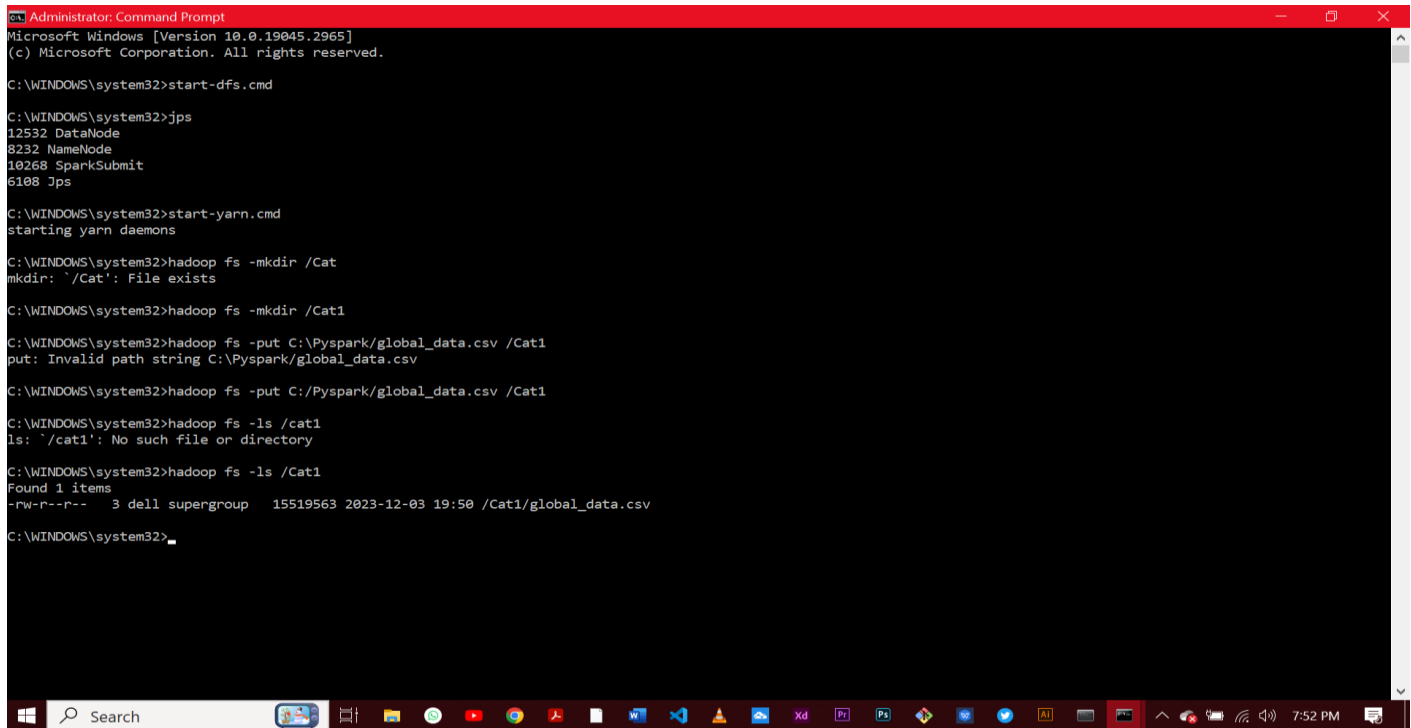


DBMS CAT

Abdirahman Abdiaziz - SCT222-0313/2019
Michael Wainaina - SCT222-0147/2019
Chrispin Mageto - Sct222-0162/2019
Brian Njuguna - SCT222-0125/2020
Ishmael Kimani - Sct222-0248/2020

Starting hadoop and moving the data file to hdfs.



```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.19045.2965]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>start-dfs.cmd

C:\WINDOWS\system32>jps
12532 DataNode
8232 NameNode
10268 SparkSubmit
6108 Jps

C:\WINDOWS\system32>start-yarn.cmd
starting yarn daemons

C:\WINDOWS\system32>hadoop fs -mkdir /Cat
mkdir: '/Cat': File exists

C:\WINDOWS\system32>hadoop fs -mkdir /Cat1

C:\WINDOWS\system32>hadoop fs -put C:\Pyspark/global_data.csv /Cat1
put: Invalid path string C:\Pyspark/global_data.csv

C:\WINDOWS\system32>hadoop fs -put C:/Pyspark/global_data.csv /Cat1

C:\WINDOWS\system32>hadoop fs -ls /cat1
ls: '/cat1': No such file or directory

C:\WINDOWS\system32>hadoop fs -ls /Cat1
Found 1 items
-rw-r--r-- 3 dell supergroup 15519563 2023-12-03 19:50 /Cat1/global_data.csv

C:\WINDOWS\system32>
```

Initializing Pyspark and using Jupyter notebook from conda environment for data manipulation.

```
Anaconda Prompt (Anaconda3) - "C:\Users\dell\anaconda3\condabin\conda.bat" activate pyspark - jupyter notebook

(base) C:\Users\dell>conda activate pyspark

(pyspark) C:\Users\dell>jupyter notebook
[I 2023-12-03 14:29:15.780 ServerApp] Package notebook took 0.0000s to import
[I 2023-12-03 14:29:15.967 ServerApp] Package jupyter_lsp took 0.1912s to import
[W 2023-12-03 14:29:15.967 ServerApp] A `_jupyter_server_extension_points` function was not found in jupyter_lsp. Instead, a `_jupyter_server_extension_paths` function was found and will be used for now. This function name will be deprecated in future releases of Jupyter Server.
[I 2023-12-03 14:29:16.311 ServerApp] Package jupyter_server_terminals took 0.3361s to import
[I 2023-12-03 14:29:16.311 ServerApp] Package jupyterlab took 0.0000s to import
[I 2023-12-03 14:29:21.748 ServerApp] Package notebook_shim took 0.0000s to import
[W 2023-12-03 14:29:21.748 ServerApp] A `_jupyter_server_extension_points` function was not found in notebook_shim. Instead, a `_jupyter_server_extension_paths` function was found and will be used for now. This function name will be deprecated in future releases of Jupyter Server.
[I 2023-12-03 14:29:21.780 ServerApp] jupyter_lsp | extension was successfully linked.
[I 2023-12-03 14:29:21.795 ServerApp] jupyter_server_terminals | extension was successfully linked.
[I 2023-12-03 14:29:21.811 ServerApp] jupyterlab | extension was successfully linked.
[I 2023-12-03 14:29:21.826 ServerApp] notebook | extension was successfully linked.
[I 2023-12-03 14:29:30.905 ServerApp] notebook_shim | extension was successfully linked.
[I 2023-12-03 14:29:31.389 ServerApp] notebook_shim | extension was successfully loaded.
[I 2023-12-03 14:29:31.436 ServerApp] jupyter_lsp | extension was successfully loaded.
[I 2023-12-03 14:29:31.436 ServerApp] jupyter_server_terminals | extension was successfully loaded.
[I 2023-12-03 14:29:31.873 LabApp] JupyterLab extension loaded from C:\Users\dell\anaconda3\envs\pyspark\Lib\site-packages\jupyterlab
[I 2023-12-03 14:29:31.873 LabApp] JupyterLab application directory is C:\Users\dell\anaconda3\envs\pyspark\share\jupyterlab
[I 2023-12-03 14:29:31.905 LabApp] Extension Manager is 'pypi'.
[I 2023-12-03 14:29:31.905 ServerApp] jupyterlab | extension was successfully loaded.
[I 2023-12-03 14:29:31.920 ServerApp] notebook | extension was successfully loaded.
[I 2023-12-03 14:29:31.967 ServerApp] Serving notebooks from local directory: C:\Users\dell
[I 2023-12-03 14:29:31.967 ServerApp] Jupyter Server 2.11.1 is running at:
[I 2023-12-03 14:29:31.967 ServerApp] http://localhost:8888/tree?token=2df3977fde0cf983d575a0c5ac325e8a1de0e61fc3129e1
[I 2023-12-03 14:29:31.967 ServerApp] http://127.0.0.1:8888/tree?token=2df3977fde0cf983d575a0c5ac325e8a1de0e61fc3129e1
[I 2023-12-03 14:29:31.967 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 2023-12-03 14:29:33.608 ServerApp]

To access the server, open this file in a browser:
file:///C:/Users/dell/AppData/Roaming/jupyter/runtime/jpserver-3048-open.html
Or copy and paste one of these URLs:
http://localhost:8888/tree?token=2df3977fde0cf983d575a0c5ac325e8a1de0e61fc3129e1
http://127.0.0.1:8888/tree?token=2df3977fde0cf983d575a0c5ac325e8a1de0e61fc3129e1
[W 2023-12-03 14:29:33.952 ServerApp] Could not determine npm prefix: [WinError 193] %1 is not a valid Win32 application
[I 2023-12-03 14:29:34.311 ServerApp] Skipped non-installed server(s): bash-language-server, dockerfile-language-server-nodejs, javascript-typescript-langserver, jedi-language-server, julia-language-server, pyright, python-language-server, python-lsp-server, r-languageserver, sql-language-server, texlab, typescript-language-server, unified-languageserver
```

Search

Home CAT Jupyter Server CAT

localhost:8888/notebooks/CAT.ipynb

Gmail YouTube Maps News Translate Oto255 Flat UI Palette v... Icons: The premi... transform - CSS: Ca... Unicode Character...

jupyter CAT Last Checkpoint: 41 minutes ago

File Edit View Run Kernel Settings Help

Trusted

JupyterLab Python 3 (ipykernel)

```
[6]: import findspark
    findspark.init()
    findspark.find()
    import pyspark

[7]: from pyspark.sql import SparkSession

[12]: print (spark.version)
3.5.0

[10]: spark = SparkSession.builder.appName("Cat").getOrCreate()

[13]: local_file_path = "pyspark/global_data.csv"

[14]: df = spark.read.csv(local_file_path, headers=True, inferSchema=True)

[15]: df.printSchema()
df.show(5)

root
 |-- Date_reported: date (nullable = true)
 |-- Country_code: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- WHO_region: string (nullable = true)
 |-- New_cases: integer (nullable = true)
 |-- Cumulative_cases: integer (nullable = true)
 |-- New_deaths: integer (nullable = true)
 |-- Cumulative_deaths: integer (nullable = true)
```

Search

7:53 PM

Home CAT Jupyter Server CAT

localhost:8888/notebooks/CAT.ipynb

Gmail YouTube Maps News Translate Oto255 Flat UI Palette v... Icons: The premi... transform - CSS: Ca... Unicode Character ... All Bookmarks

Jupyter CAT Last Checkpoint: 41 minutes ago

File Edit View Run Kernel Settings Help Trusted

JupyterLab Python 3 (ipykernel)

```
-- Cumulative_deaths: integer (nullable = true)

|Date_reported|Country_code|Country|WHO_region|New_cases|Cumulative_cases|New_deaths|Cumulative_deaths|
|-----|-----|-----|-----|-----|-----|-----|-----|
|2020-01-03|AF|Afghanistan|EMRO|0|0|0|0|
|2020-01-04|AF|Afghanistan|EMRO|0|0|0|0|
|2020-01-05|AF|Afghanistan|EMRO|0|0|0|0|
|2020-01-06|AF|Afghanistan|EMRO|0|0|0|0|
|2020-01-07|AF|Afghanistan|EMRO|0|0|0|0|

only showing top 5 rows

[16]: # Q2. Pre-processing data
      # Drop rows with missing values
      df = df.dropna()

[27]: #Data Cleaning
      # Remove duplicate rows
      df = df.dropDuplicates()

[25]: !pip install numpy
```

0.0/61.2 kB ? eta --:--
0.0/61.2 kB ? eta --:--
0.0/61.2 kB ? eta --:--
0.0/61.2 kB ? eta --:--
0.0/61.2 kB ? eta --:--
0.0/61.2 kB ? eta --:--
0.0/61.2 kB ? eta --:--
0.0/61.2 kB ? eta --:--
0.0/61.2 kB ? eta --:--
0.0/61.2 kB ? eta --:--

Home CAT Jupyter Server CAT

localhost:8888/notebooks/CAT.ipynb

Gmail YouTube Maps News Translate Oto255 Flat UI Palette v... Icons: The premi... transform - CSS: Ca... Unicode Character ... All Bookmarks

Jupyter CAT Last Checkpoint: 42 minutes ago

File Edit View Run Kernel Settings Help Trusted

JupyterLab Python 3 (ipykernel)

```
[28]: from pyspark.sql import SparkSession
      from pyspark.sql.functions import mean

[ ]: # Calculate mean values for all numeric columns
      mean_values = df.agg(*[mean(c).alias(c) for c in df.columns if df[c].dtypes[0] in ['int', 'double']]).collect()[0]

[ ]: # Fill missing values with mean
      for col in mean_values.asDict():
          df = df.na.fill(mean_values[col], [col])

[43]: #Data transformation
      from pyspark.ml.feature import StandardScaler, OneHotEncoder
      from pyspark.ml.feature import VectorAssembler
      from pyspark.ml import Pipeline

[58]: # Define the features you want to include in the vector
      features_col_names = ["New_cases", "Cumulative_cases", "New_deaths", "Cumulative_deaths"]

[59]: # Use VectorAssembler to create a vector of features
      assembler = VectorAssembler(inputCols=features_col_names, outputCol="features")

[60]: # Standardize the features using StandardScaler
      scaler = StandardScaler(inputCol="features", outputCol="scaled_features", withStd=True, withMean=True)

[65]: # Filter data for Kenya (country code is "KE")
      kenya_data = df.filter(df["Country_code"] == "KE")

[66]: # Show the resulting DataFrame
      kenya_data.show()
```

The screenshot displays a JupyterLab environment with a notebook titled 'CAT'. The interface includes a top navigation bar with various application icons, a browser address bar showing 'localhost:8888/notebooks/CAT.ipynb', and a JupyterLab header with a 'Trusted' status indicator. The notebook content shows two code cells. The first cell (line 67) contains a comment and a Spark SQL query to collect new deaths. Its output is a list of 15 'Number of New Deaths: 0' entries. The second cell (line 69) contains a comment and a Spark SQL query to calculate the total number of new deaths in Kenya. The third cell (line 70) contains a comment and a print statement to display the total number of new deaths in Kenya.

```
[67]: # Extract the number of new deaths
new_deaths = df.select("New_Deaths").collect()

[68]: # Display the extracted values
for row in new_deaths:
    print(f"Number of New Deaths: {row['New_Deaths']}")

Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0
Number of New Deaths: 0

[69]: from pyspark.sql import functions as F

# Calculate the total number of new deaths in Kenya
total_new_deaths_kenya = kenya_data.select(F.sum("New_Deaths")).collect()[0][0]

[70]: # Display the total number of new deaths in Kenya
print(f"Total New Deaths in Kenya: {total_new_deaths_kenya}")
```

```
Home CAT Jupyter Server CAT
localhost:8888/notebooks/CAT.ipynb
jupyter CAT Last Checkpoint: 42 minutes ago
File Edit View Run Kernel Settings Help
JupyterLab Python 3 (ipykernel)

[70]: # Display the total number of new deaths in Kenya
print(f"Total New Deaths in Kenya: {total_new_deaths_kenya}")

Total New Deaths in Kenya: 5689

[73]: # Calculate the total number of confirmed cases in Kenya
total_confirmed_cases_kenya = kenya_data.select(F.max("Cumulative_cases")).collect()[0][0]

[74]: # Display the total number of confirmed cases in Kenya
print(f"Total Confirmed Cases in Kenya: {total_confirmed_cases_kenya}")

Total Confirmed Cases in Kenya: 344190

[80]: # Calculate the total number of cumulative deaths in Kenya
total_cumulative_deaths_kenya = kenya_data.select(F.max("Cumulative_deaths")).collect()[0][0]

[81]: # Display the total number of cumulative deaths in Kenya
print(f"Total Cumulative Deaths in Kenya: {total_cumulative_deaths_kenya}")

Total Cumulative Deaths in Kenya: 5689

[82]: pip install matplotlib

Note: you may need to restart the kernel to use updated packages.
Collecting matplotlib
  Downloading matplotlib-3.8.2-cp312-cp312-win_amd64.whl.metadata (5.9 kB)
Collecting contourpy>=1.0.1 (from matplotlib)
  Downloading contourpy-1.2.0-cp312-cp312-win_amd64.whl.metadata (5.8 kB)
Collecting cycler>=0.10 (from matplotlib)
  Downloading cycler-0.12.1-py3-none-any.whl.metadata (3.8 kB)
Collecting fonttools>=4.22.0 (from matplotlib)
  Downloading fonttools-4.46.0-cp312-cp312-win_amd64.whl.metadata (159 kB)
----- 0.0/159.4 kB ? eta -:--
----- 0.0/159.4 kB ? eta -:--
```

```
Home CAT Jupyter Server CAT
localhost:8888/notebooks/CAT.ipynb
jupyter CAT Last Checkpoint: 43 minutes ago
File Edit View Run Kernel Settings Help
JupyterLab Python 3 (ipykernel)

[84]: pip install pandas

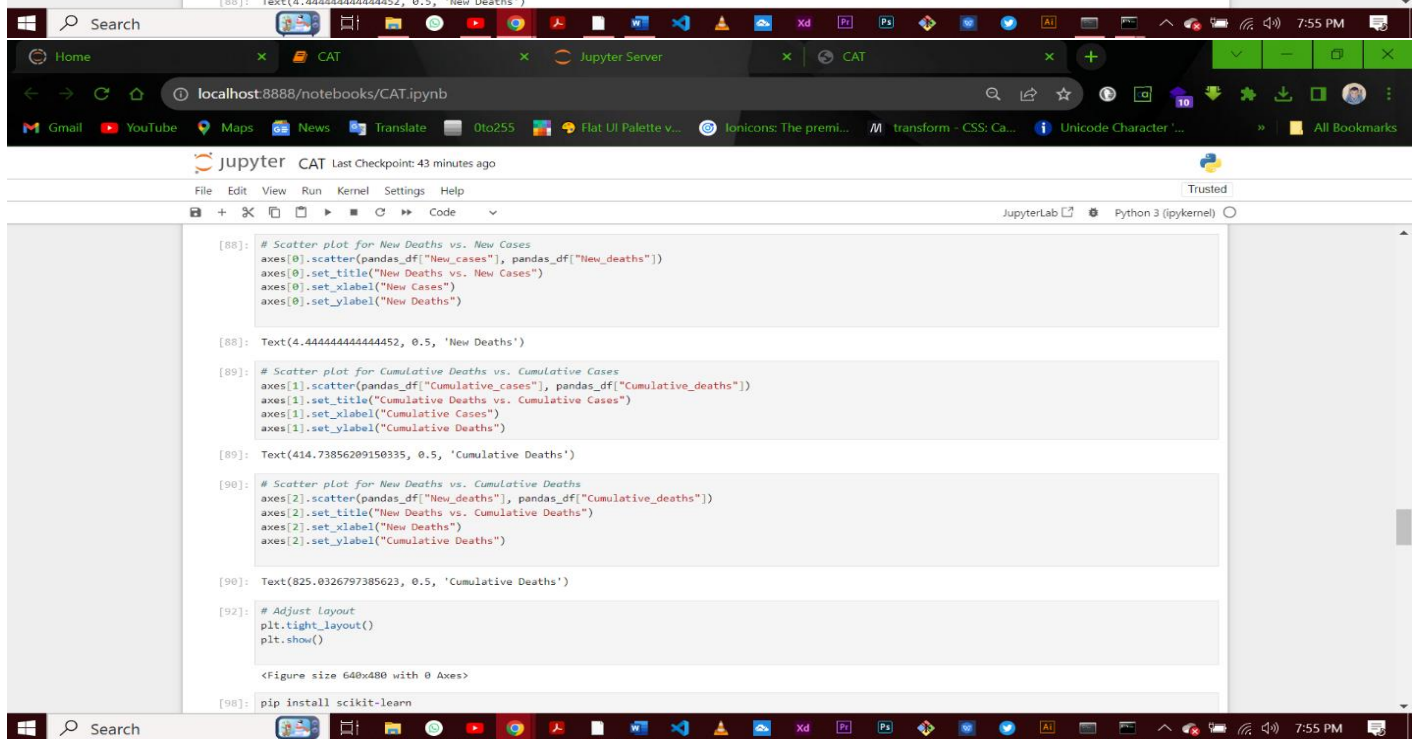
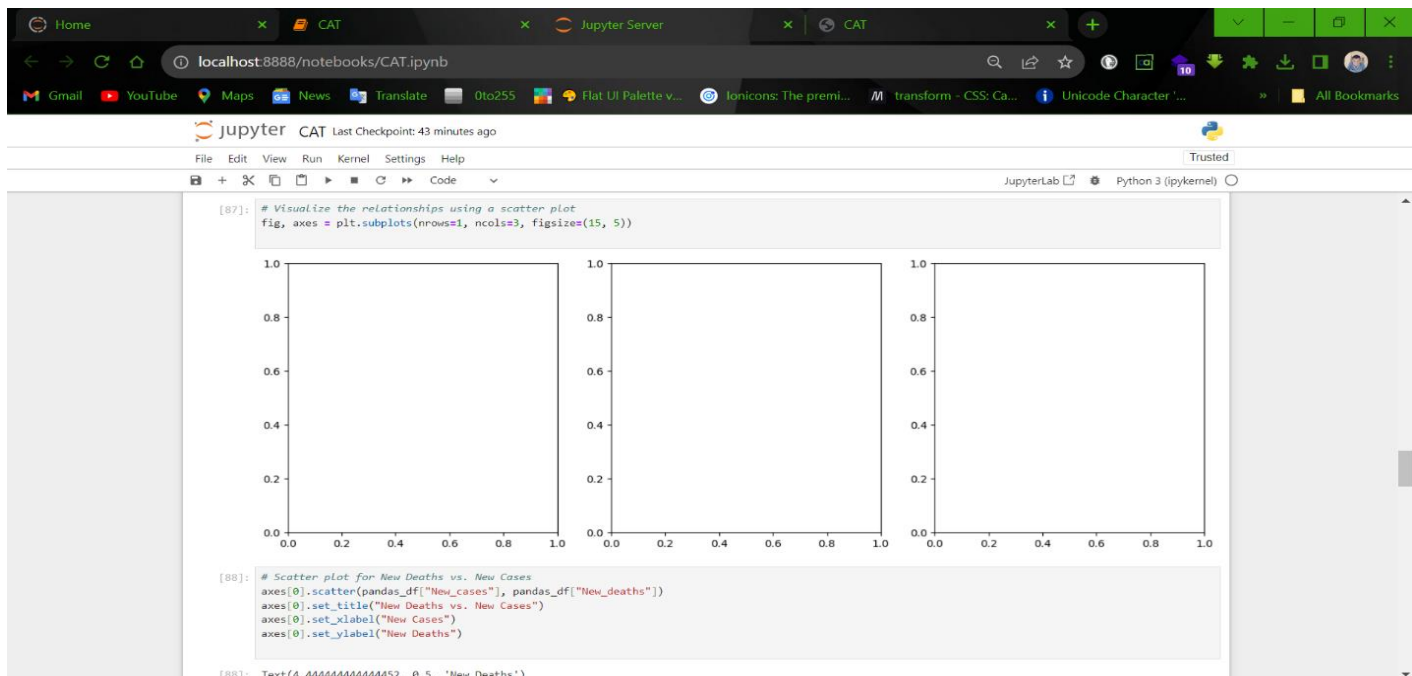
Collecting pandas: Note: you may need to restart the kernel to use updated packages.
  Downloading pandas-2.1.3-cp312-cp312-win_amd64.whl.metadata (18 kB)
Requirement already satisfied: numpy<2, >=1.26.0 in c:\users\dell\anaconda3\envs\pyspark\lib\site-packages (from pandas) (1.26.2)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\dell\anaconda3\envs\pyspark\lib\site-packages (from pandas) (2.8.2)
Collecting pytz>=2020.1 (from pandas)
  Downloading pytz-2023.3.post1-py2.py3-none-any.whl.metadata (22 kB)
Collecting tzdata>=2022.1 (from pandas)
  Downloading tzdata-2023.3-py2.py3-none-any.whl (341 kB)
----- 0.0/341.8 kB ? eta -:--
----- 10.2/341.8 kB ? eta -:--
----- 41.0/341.8 kB 653.6 kB/s eta 0:00:01
----- 81.9/341.8 kB 919.0 kB/s eta 0:00:01
----- 81.9/341.8 kB 919.0 kB/s eta 0:00:01
----- 81.9/341.8 kB 919.0 kB/s eta 0:00:01
----- 92.2/341.8 kB 374.1 kB/s eta 0:00:01
----- 122.9/341.8 kB 450.6 kB/s eta 0:00:01
----- 163.8/341.8 kB 517.2 kB/s eta 0:00:01

[85]: import matplotlib.pyplot as plt
import pandas as pd
from pyspark.sql import SparkSession

[86]: # Convert PySpark DataFrame to Pandas DataFrame for visualization
pandas_df = df.select("New_deaths", "Cumulative_deaths", "New_cases", "Cumulative_cases").toPandas()

[87]: # Visualize the relationships using a scatter plot
fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(15, 5))

1.0 1.0 1.0
```

```
Home CAT Jupyter Server CAT
localhost:8888/notebooks/CAT.ipynb
jupyter CAT Last Checkpoint: 43 minutes ago
File Edit View Run Kernel Settings Help
JupyterLab Python 3 (ipykernel)

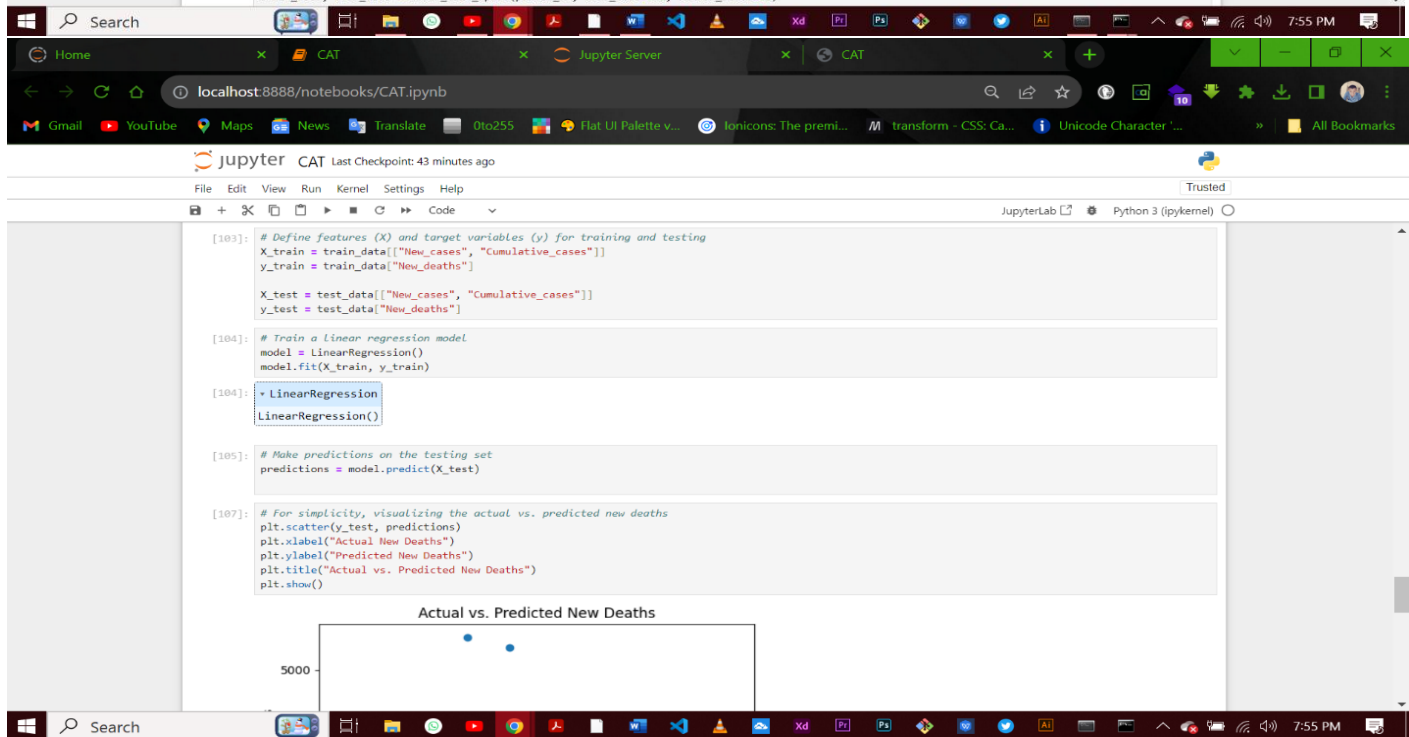
[98]: pip install scikit-learn

Collecting scikit-learn
  Downloading scikit_learn-1.3.2-cp312-cp312-win_amd64.whl.metadata (11 kB)
Requirement already satisfied: numpy<2.0,=>1.17.3 in c:\users\de\l\anaconda3\envs\pyspark\lib\site-packages (from scikit-learn) (1.26.2)
Collecting scipy>=1.5.0 (from scikit-learn)
  Downloading scipy-1.11.4-cp312-cp312-win_amd64.whl.metadata (60 kB)
----- 0.0/60.4 kB ? eta -:--:--
----- 0.0/60.4 kB ? eta -:--:--
----- 0.0/60.4 kB ? eta -:--:--
----- 10.2/60.4 kB ? eta -:--:--
----- 30.7/60.4 kB 220.2 kB/s eta 0:00:01
----- 30.7/60.4 kB 220.2 kB/s eta 0:00:01
----- 51.2/60.4 kB 238.1 kB/s eta 0:00:01
----- 51.2/60.4 kB 238.1 kB/s eta 0:00:01
----- 51.2/60.4 kB 238.1 kB/s eta 0:00:01
----- 60.4/60.4 kB 178.5 kB/s eta 0:00:00
Collecting joblib>=1.1.1 (from scikit-learn)
  Downloading joblib-1.3.2-py3-none-any.whl.metadata (5.4 kB)
Collecting threadpoolctl>=2.0.0 (from scikit-learn)

[99]: # ***** TEST *****
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import pandas as pd
from pyspark.sql import SparkSession

[100]: # Convert PySpark DataFrame to Pandas DataFrame for testing
pandas_df = df.select("New_deaths", "Cumulative_deaths", "New_cases", "Cumulative_cases").toPandas()

[102]: # Split the data into training and testing sets
train_data, test_data = train_test_split(pandas_df, test_size=0.2, random_state=42)
```





Search

Home CAT Jupyter Server CAT

localhost:8888/notebooks/CAT.ipynb

jupyter CAT Last Checkpoint: 1 hour ago

File Edit View Run Kernel Settings Help Trusted

JupyterLab Python 3 (ipykernel)

```
[111]: # Filter data for Kenya (country code is "KE")  
kenya_data = df.filter(df["Country_code"] == "KE")  
  
[112]: # Select relevant columns  
selected_columns = ["New_deaths", "Cumulative_deaths", "New_cases", "Cumulative_cases"]  
  
[149]: # Drop existing 'features_modified' column if it exists  
if 'features_modified' in kenya_data.columns:  
    kenya_data = kenya_data.drop("features_modified")  
  
[157]: # Assemble features into a single column with a modified output column name  
assembler = VectorAssembler(inputCols=selected_columns[2:], outputCol="features_modified2")  
kenya_data = assembler.transform(kenya_data)  
  
[151]: # Split the data into training and testing sets  
train_data, test_data = kenya_data.randomSplit([0.8, 0.2], seeds=42)  
  
[152]: # Define the linear regression model  
lr = LinearRegression(featuresCol="features_modified", labelCol="New_deaths")  
  
[159]: # Create a pipeline  
pipeline = Pipeline(stages=[  
    VectorAssembler(inputCols=selected_columns[2:], outputCol="features5"),  
    lr  
)  
  
[160]: # Train the model on the training set  
model = pipeline.fit(train_data)  
  
[161]: # Make predictions on the testing set  
predictions = model.transform(test_data)
```

Search

Home CAT Jupyter Server CAT

localhost:8888/notebooks/CAT.ipynb

Gmail YouTube Maps News Translate 0to25 Flat UI Palette v... Icons: The premi... transform - CSS: Ca... Unicode Character '... All Bookmarks

Jupyter CAT Last Checkpoint: 1 hour ago

File Edit View Run Kernel Settings Help Trusted

JupyterLab Python 3 (ipykernel)

```
[157]: # Assemble features into a single column with a modified output column name
assembler = VectorAssembler(inputCols=selected_columns[2:], outputCol="features_modified2")
kenya_data = assembler.transform(kenya_data)

[151]: # Split the data into training and testing sets
train_data, test_data = kenya_data.randomSplit([0.8, 0.2], seeds=42)

[152]: # Define the Linear regression model
lr = LinearRegression(featuresCol="features_modified", labelCol="New_deaths")

[150]: # Create a pipeline
pipeline = Pipeline(stages=[
    VectorAssembler(inputCols=selected_columns[2:], outputCol="features5"),
    lr
])

[160]: # Train the model on the training set
model = pipeline.fit(train_data)

[161]: # Make predictions on the testing set
predictions = model.transform(test_data)

[162]: # Evaluate the model using RegressionEvaluator
evaluator = RegressionEvaluator(labelCol="New_deaths", predictionCol="prediction", metricName="rmse")
rmse = evaluator.evaluate(predictions)

[163]: # Display the Root Mean Squared Error (RMSE)
print(f"Root Mean Squared Error (RMSE): {rmse}")
```

Root Mean Squared Error (RMSE): 5.48301492481005

Search

8:16 PM