Introduction:

You are a data analyst tasked with analyzing a temperature dataset. The dataset contains information about average temperatures, uncertainties, cities, countries, and geographical coordinates. Your goal is to derive meaningful insights from the data using PySpark.

Dataset :

```Plain Text
https://github.com/nag9s/awesome-datasets/blob/main/Datasets/GlobalLandTemperatures/GlobalLandTemperatures_GlobalLandTemperaturesByMajorCity.csv
```

Notebook :

```Apache
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8321264234333142/3603809094474529/5395886
```

Task:

1. Load the provided dataset into a PySpark DataFrame.
2. Analyze the data using various PySpark operations to create insightful reports.

Step 1: Load the Dataset

```Python
from pyspark.sql.functions import year, month

# Load the dataset
# Assume 'df' is the DataFrame
df.show()
```

Step 2: Analyze the Data

Report 1: Average Temperature by City

```
Plain Text


report1 = df.groupBy("City").agg({"AverageTemperature": "avg"})
report1.show()
```

Questions:

- What information does Report 1 provide?
- Can you identify the city with the highest average temperature?

Report 2: Average Temperature by Country

```
Plain Text


report2 = df.groupBy("Country").agg({"AverageTemperature": "avg"})
report2.show()
```

Questions:

- How does the average temperature vary between countries?
- Identify a country with a significant temperature difference from others.

Report 3: Maximum Temperature by City and Year

```
Vbscript


report3 = df.withColumn("Year", year("Date")).groupBy("City", "Year").agg({"AverageTemperature": "max"})
report3.show()
```

Questions:

- What insights can you draw from the maximum temperatures reported for each city?
- Identify a city with a noticeable temperature increase over the years.

Report 4: Minimum Temperature by Country and Month

```
Vbscript


report4 = df.withColumn("Month", month("Date")).groupBy("Country", "Month").agg({"AverageTemperature":
"min"})
report4.show()
```

Questions:

- How does the minimum temperature vary by month within a country?
- Identify a country with a significant temperature fluctuation across months.

Report 5: Average Temperature Uncertainty by City

```
report5 = df.groupBy("City").agg({"AverageTemperatureUncertainty": "avg"})
report5.show()
```

Questions:

- What information can we gather from the average temperature uncertainty in different cities?
- How does uncertainty vary between cities?

Report 6: Average Temperature by Latitude and Longitude

```
report6 = df.groupBy("Latitude", "Longitude").agg({"AverageTemperature": "avg"})
report6.show()
```

Questions:

- What geographical patterns can you observe in average temperatures?
- Identify regions with consistently high or low temperatures.

Report 7: Monthly Average Temperature in a Specific City

```
report7 = df.filter(df.City == "Abidjan").groupBy(month("Date").alias("Month")).agg({"AverageTemperature": "avg"})
report7.show()
```

Questions:

- What insights can you gather from the monthly average temperature in a specific city?
- Identify a month with the highest and lowest average temperature in Abidjan.

Report 8: Average Temperature by Year and Country

```
report8 = df.withColumn("Year", year("Date")).groupBy("Year", "Country").agg({"AverageTemperature": "avg"})
report8.show()
```

Questions:

- How does the average temperature vary by year and country?
- Identify a year with a noticeable temperature trend across multiple countries.

Report 9: Cities with the Highest Average Temperature

```
Plain Text


report9 = df.groupBy("City").agg({"AverageTemperature": "avg"}).filter("avg(AverageTemperature) > 25")
report9.show()
```

Questions:

- What cities have consistently high average temperatures?
- Identify factors that might contribute to these high temperatures.

Report 10: Average Temperature Uncertainty by Year

```
Vbscript


report10 = df.withColumn("Year", year("Date")).groupBy("Year").agg({"AverageTemperatureUncertainty":
"avg"})
report10.show()
```

Questions:

- How does the uncertainty in average temperature change over the years?
- Identify a year with unusually high or low uncertainty.