# KGiSL INSTITUTE OF TECHNOLOGY

## Coimbatore – 641035

## Institution code: 7117

# Big Data Analysis Using IBM Cloud Databases

**MENTOR:**
MRS.INDU POORNIMA.R
**TEAM MEMBERS:**
YOGANATHAN.R
SABAREESAN.R
VENKATESH.R
SIBIRAJ.M

# Big Data Analysis Using IBM Cloud Databases

**Problem Definition:** The project involves delving into big data analysis using IBM Cloud Databases. The objective is to extract valuable insights from extensive datasets, ranging from climate trends to social patterns. The project includes designing the analysis process, setting up IBM Cloud Databases, performing data analysis, and visualizing the results for business intelligence.

## DEVELOPMENT PART 1

### Step 1: Import Libraries and Load Data :

```
import pandas as pd

data = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/GlobalLandTemperaturesByCity.csv')

from google.colab import drive

drive.mount('/content/drive')
```

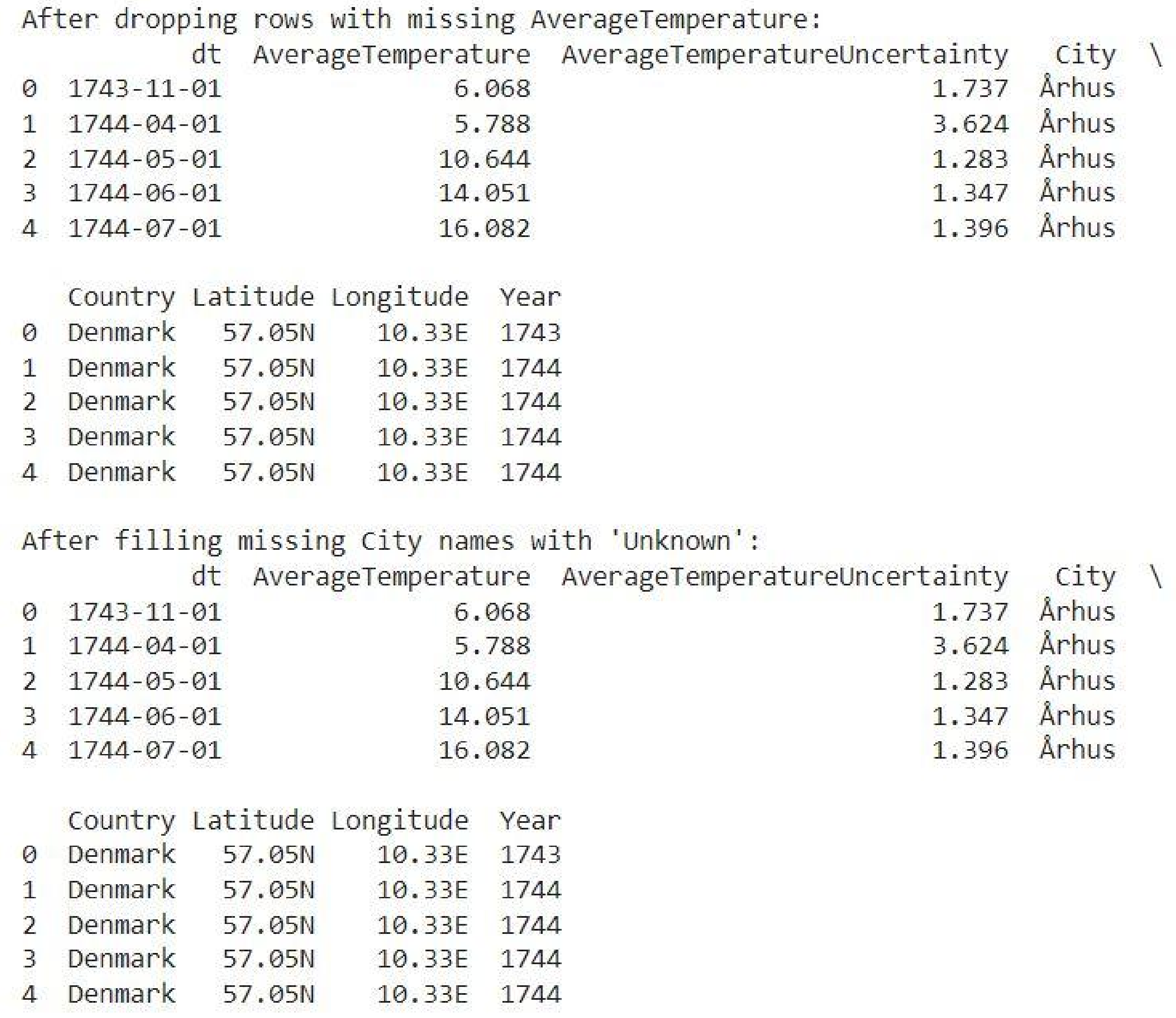
### OUTPUT :

```
Mounted at /content/drive
```

### Step 2: Data Cleaning :

```
# Drop rows with missing temperature data

data = data.dropna(subset=['AverageTemperature'])

print("\nAfter dropping rows with missing AverageTemperature:")

print(data.head(5))  # Display the first 5 rows of the DataFrame


# Fill missing city names with 'Unknown'

data['City'] = data['City'].fillna('Unknown')

print("\nAfter filling missing City names with 'Unknown':")

print(data.head(5))
```

**OUTPUT :**



```
After dropping rows with missing AverageTemperature:
           dt  AverageTemperature  AverageTemperatureUncertainty   City  \
0  1743-11-01               6.068                          1.737  Århus
1  1744-04-01               5.788                          3.624  Århus
2  1744-05-01              10.644                          1.283  Århus
3  1744-06-01              14.051                          1.347  Århus
4  1744-07-01              16.082                          1.396  Århus

   Country Latitude Longitude  Year
0  Denmark   57.05N    10.33E  1743
1  Denmark   57.05N    10.33E  1744
2  Denmark   57.05N    10.33E  1744
3  Denmark   57.05N    10.33E  1744
4  Denmark   57.05N    10.33E  1744

After filling missing City names with 'Unknown':
           dt  AverageTemperature  AverageTemperatureUncertainty   City  \
0  1743-11-01               6.068                          1.737  Århus
1  1744-04-01               5.788                          3.624  Århus
2  1744-05-01              10.644                          1.283  Århus
3  1744-06-01              14.051                          1.347  Århus
4  1744-07-01              16.082                          1.396  Århus

   Country Latitude Longitude  Year
0  Denmark   57.05N    10.33E  1743
1  Denmark   57.05N    10.33E  1744
2  Denmark   57.05N    10.33E  1744
3  Denmark   57.05N    10.33E  1744
4  Denmark   57.05N    10.33E  1744
```

```
# Remove duplicate records based on all columns
data = data.drop_duplicates()
print("\nAfter removing duplicates based on all columns:")
```

print(data.head(5))  # Display the first 5 rows of the DataFrame

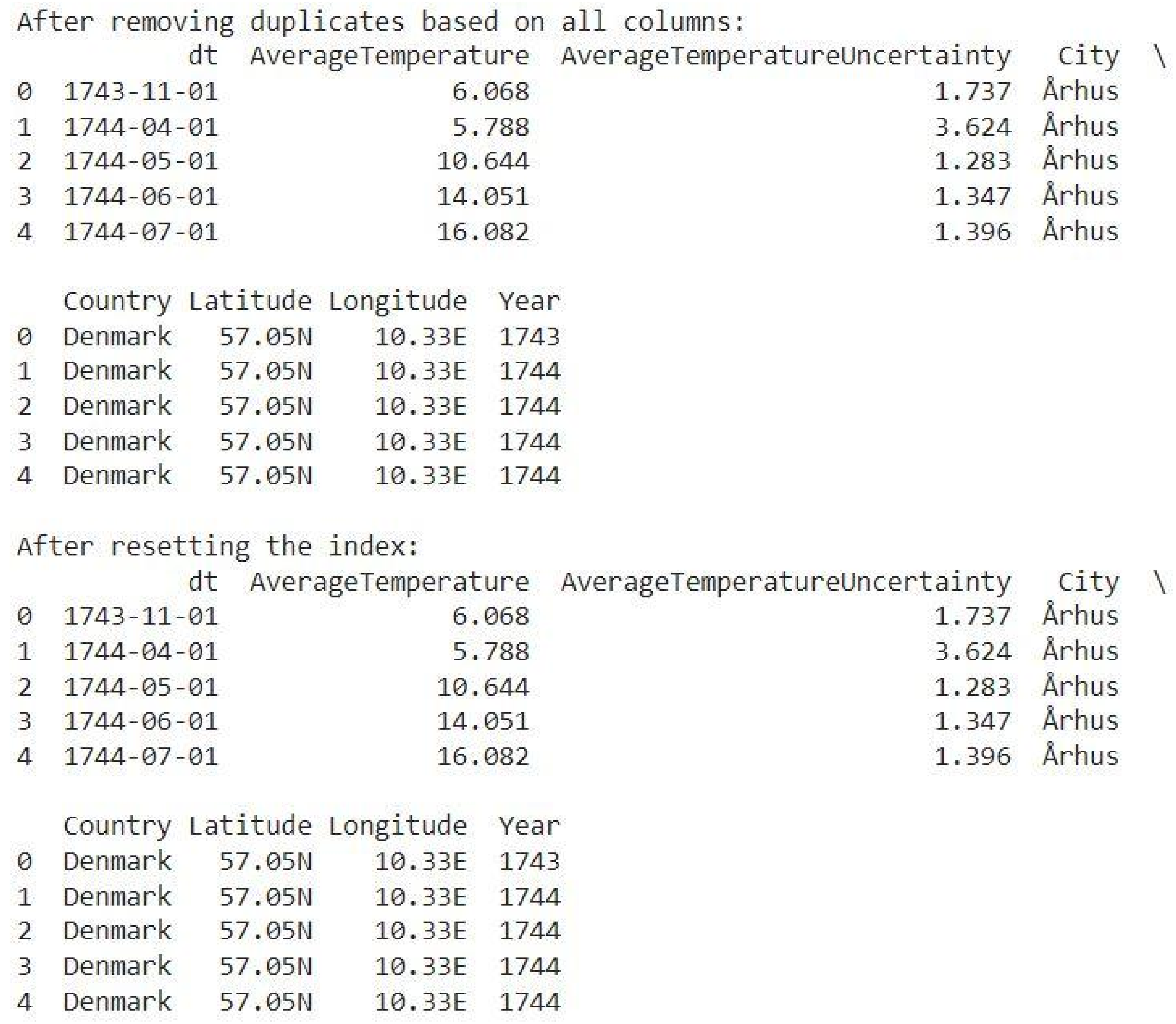
# Optionally, reset the index

data = data.reset_index(drop=True)

print("\nAfter resetting the index:")

print(data.head(5))

**OUTPUT :**

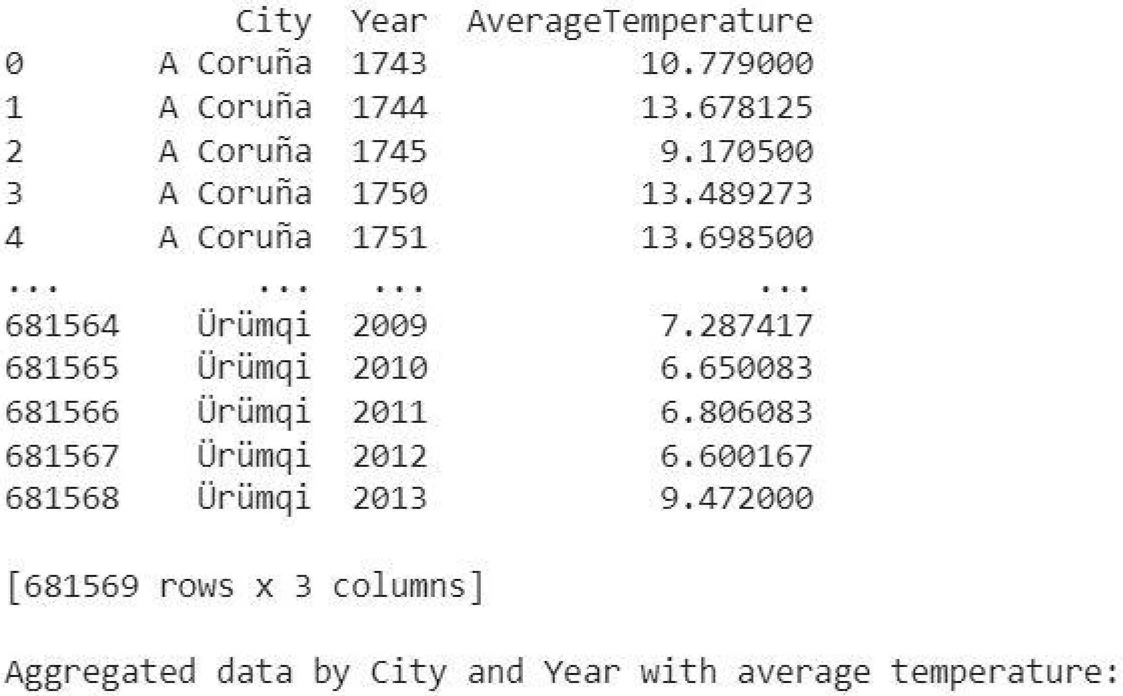


```
After removing duplicates based on all columns:
           dt  AverageTemperature  AverageTemperatureUncertainty   City  \
0  1743-11-01               6.068                          1.737  Århus
1  1744-04-01               5.788                          3.624  Århus
2  1744-05-01              10.644                          1.283  Århus
3  1744-06-01              14.051                          1.347  Århus
4  1744-07-01              16.082                          1.396  Århus

   Country Latitude Longitude  Year
0  Denmark   57.05N    10.33E  1743
1  Denmark   57.05N    10.33E  1744
2  Denmark   57.05N    10.33E  1744
3  Denmark   57.05N    10.33E  1744
4  Denmark   57.05N    10.33E  1744

After resetting the index:
           dt  AverageTemperature  AverageTemperatureUncertainty   City  \
0  1743-11-01               6.068                          1.737  Århus
1  1744-04-01               5.788                          3.624  Århus
2  1744-05-01              10.644                          1.283  Århus
3  1744-06-01              14.051                          1.347  Århus
4  1744-07-01              16.082                          1.396  Århus

   Country Latitude Longitude  Year
0  Denmark   57.05N    10.33E  1743
1  Denmark   57.05N    10.33E  1744
2  Denmark   57.05N    10.33E  1744
3  Denmark   57.05N    10.33E  1744
4  Denmark   57.05N    10.33E  1744
```

**Step 3: Data Transformation :**

```
data['Year'] = data['dt'].str[:4].astype(int)

# Calculate the average temperature for each city and year

agg_data = data.groupby(['City',
'Year'])['AverageTemperature'].mean().reset_index()

print(agg_data)

print("\nAggregated data by City and Year with average temperature:")
```

**OUTPUT :**

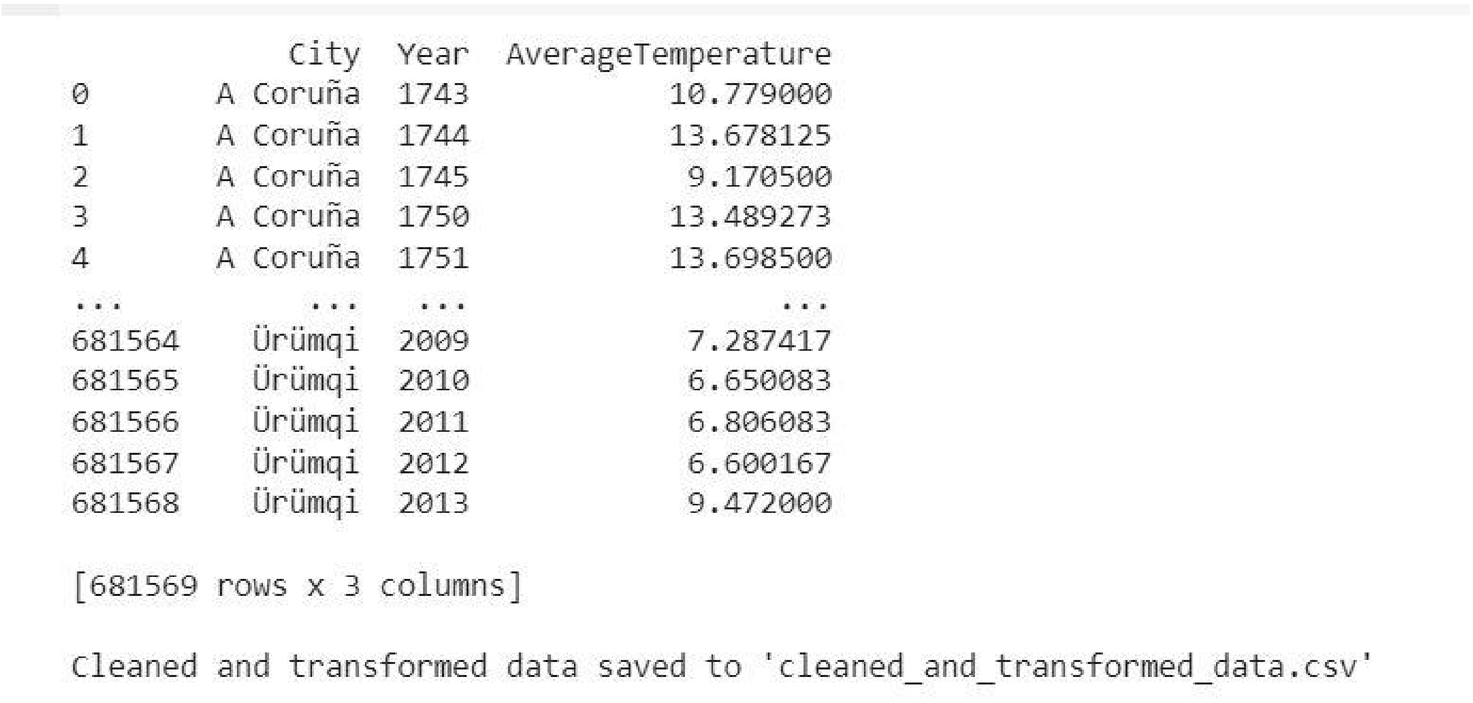


```
              City  Year  AverageTemperature
0        A Coruña  1743           10.779000
1        A Coruña  1744           13.678125
2        A Coruña  1745            9.170500
3        A Coruña  1750           13.489273
4        A Coruña  1751           13.698500
...           ...   ...                 ...
681564     Ürümqi  2009            7.287417
681565     Ürümqi  2010            6.650083
681566     Ürümqi  2011            6.806083
681567     Ürümqi  2012            6.600167
681568     Ürümqi  2013            9.472000

[681569 rows x 3 columns]

Aggregated data by City and Year with average temperature:
```

**Step 4: Save the Cleaned and Transformed Data :**

```
# Save the cleaned and transformed data to a new CSV file
agg_data.to_csv('cleaned_and_transformed_data.csv', index=False)
print(agg_data)
print("\nCleaned and transformed data saved to 'cleaned_and_transformed_data.csv'")
```

**OUTPUT :**



```
              City  Year  AverageTemperature
0        A Coruña  1743           10.779000
1        A Coruña  1744           13.678125
2        A Coruña  1745            9.170500
3        A Coruña  1750           13.489273
4        A Coruña  1751           13.698500
...           ...   ...                 ...
681564     Ürümqi  2009            7.287417
681565     Ürümqi  2010            6.650083
681566     Ürümqi  2011            6.806083
681567     Ürümqi  2012            6.600167
681568     Ürümqi  2013            9.472000

[681569 rows x 3 columns]

Cleaned and transformed data saved to 'cleaned_and_transformed_data.csv'
```