Internship Report

Team Members
Anupam Kumar Goswami
342/cse/2k21
Arham Eqbal
21/CSE/381
Ritesh Kumar
- Roll: 2k21/CSE/330
Couse: B.tech. In Computer Science & Technology
Sem : VI

Mentor: Deshbhandhu Mishra

College : Cambridge Institute of technology,Ranchi

University: Jharkhand University of Technology

Internship Firm: DINGIR(Online Mode)

Topic: Data Analysis of Temperature Variation from 1901-2001

Temperature Analysis Report(1901-2001)

Objective

The objective of this internship was to perform a comprehensive data analysis on the temperature records in India from 1901 to 2021. The analysis aimed to identify trends, patterns, and anomalies in temperature changes over this period. This project involved the use of Databricks, PySpark, Spark SQL, and various visualization tools to process, analyze, and visualize the data.

Tools Used

- **Databricks:** For creating a collaborative environment and executing data analysis workflows.
- **Databricks File System (DBFS):** For storing and managing the dataset.
- **PySpark:** For large-scale data processing and transformation.
- **Spark SQL:** For performing SQL-based queries on the dataset.
- **Python:** For scripting and automation tasks.
- **Visualization Libraries:** Matplotlib, Seaborn, Plotly for creating insightful visualizations.

Data Analysis Process

1. Data Ingestion

- Set up Databricks Environment: Created a Databricks workspace and set up a cluster with appropriate configurations.

- Upload the Dataset to DBFS: Uploaded the temperature dataset to Databricks File System (DBFS) and verified the upload by listing the contents of the DBFS directory.

2. Data Processing and Transformation

- Load the Dataset into a Spark DataFrame: Used PySpark to read the dataset from DBFS into a Spark DataFrame.

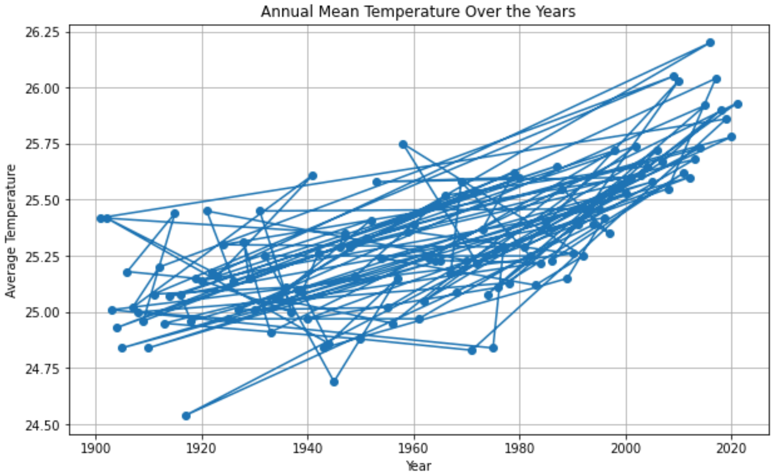
- Data Cleaning and Transformation: Cleaned the dataset by handling missing values and renaming columns for easier access.

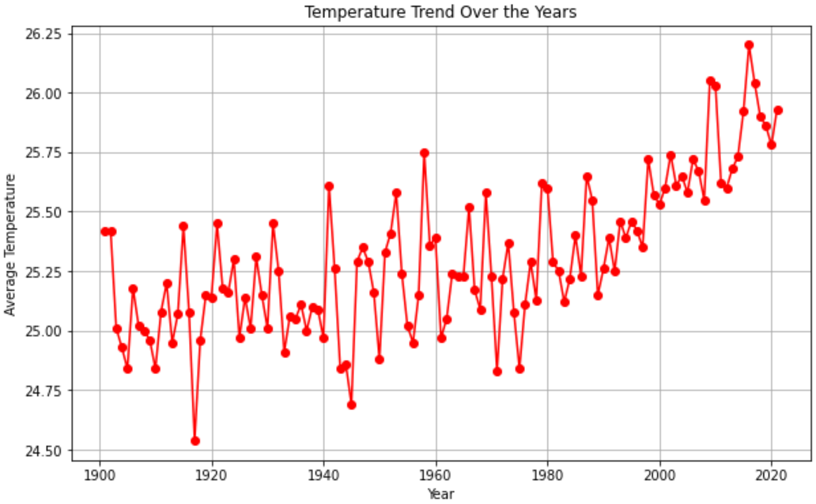
3. Data Analysis

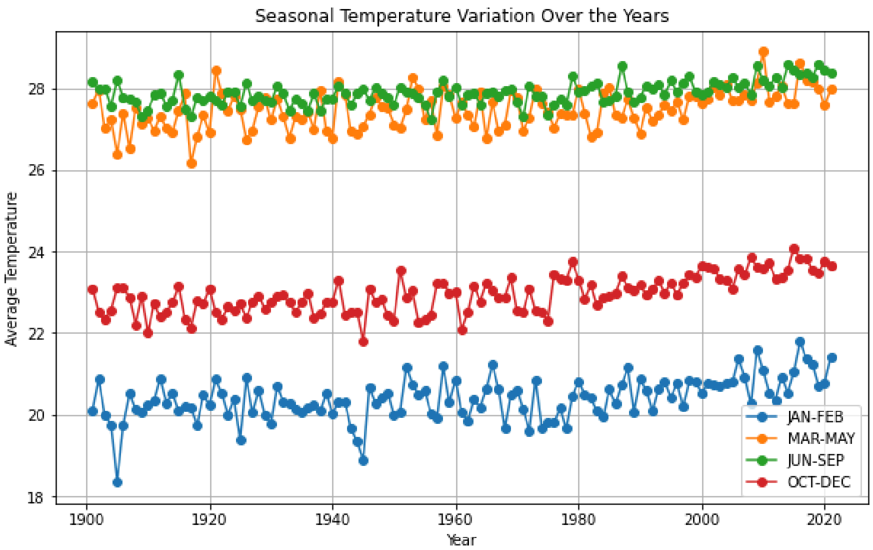
- Analyze Temperature Trends using Spark SQL: Used Spark SQL to perform various analyses, such as calculating the average temperature per decade and identifying any significant trends or anomalies.

4. Data Visualization

- Visualize the Data using Databricks Notebooks: Converted Spark DataFrames to Pandas DataFrames for visualization and used Matplotlib and Seaborn to create charts.







Conclusion🎯

This internship provided a thorough understanding of data analysis processes using Databricks and PySpark. By analyzing the temperature data from 1901 to 2021, we identified key trends and patterns, gaining insights into how temperatures have changed over time in India. The hands-on experience with data processing, analysis, and visualization has enhanced our skills and prepared us for future projects.

Team members

Anupam Kumar Goswami
-Roll:342/CSE/2k21
Arham Eqbal
-Roll:381/CSE/2k21
Ritesh Kumar
- Roll: 330/CSE/2k21

Responsibilities

-Coding:Ritesh Kumar
- GitHub: Arham Eqbal
- Report: Anupam Kumar Goswami

Future Aspects

- Enhanced Functionality: Incorporate additional datasets, such as precipitation and humidity, for a more comprehensive climate analysis.
- Advanced Analytics: Apply machine learning algorithms to predict future temperature trends based on historical data.
- Interactive Dashboards: Develop interactive dashboards using tools like Plotly Dash or Power BI for real-time data exploration and visualization.
- Automation: Automate the data ingestion, processing, and analysis workflows to improve efficiency and scalability.

Important Links

Data Source: https://data.gov.in/

Github: My github

Databricks Community Edition: https://community.cloud.databricks.com/