**Introduction**

|  |  |
| --- | --- |
| A group of different colors of the same weather  Description automatically generated with medium confidence  WEATHER DATA  FINAL PROJECT | Abstract  Understanding and analysing historical weather data to provide valuable insights into regional climate trends, aiding in decision-making processes across different sectors.  Elsie Dumashie  BDAT 1002 |

**Exploring Weather Patterns in the Southeast Region**

Weather patterns play a crucial role in various aspects of our lives and overall environmental conditions. The purpose of this project was to help in the understanding and analysing of weather data from the southeast region which can provide valuable insights into regional climate trends, aiding in decision-making processes across different sectors.

The dataset spans multiple years and encompasses a variety of meteorological parameters, including precipitation, atmospheric pressure, temperature, wind direction, and more. These observations are collected from a network of weather stations strategically positioned throughout the region. The data was found on Kaggle and was 2.3 GB. It includes hourly weather observations from a network of meteorological stations across the Southeast region. The projects goals included to be able to conduct a thorough exploration of the dataset to identify patterns, trends, and anomalies in weather-related variables. This step is crucial for gaining a holistic understanding of the region's climate and in enhancing the understanding of the Southeast region's climate and its implications on various sectors. The analysis will involve a combination of statistical techniques, time-series analysis, and visualization to derive meaningful insights from the dataset. Exploratory data analysis will serve as the foundation for more in-depth investigations The findings may help in better understanding of the weather data to be able to answer research questions and the project aims to produce a comprehensive report detailing the climatic characteristics of the Southeast region, highlighting noteworthy trends.

**Data Acquisition and Cleaning**

As required, I ensured I found a dataset that was at least 1 GB. The dataset under consideration in this analysis pertains to the southeastern region, offering significant information specific to the environmental and meteorological conditions in this geographical area. the focus is on factors such as precipitation, temperature, wind speed, and atmospheric pressure, providing insights into the climate dynamics of the southeastern part of the globe.

Prior to analysis, the dataset underwent a cleaning and preprocessing process using Hive on the Google Cloud Platform (GCP). A bucket called bdat1002\_bucket was created, and the data was stored in there. Then using VM instances in the hive environment the data was cleaned with a focus on addressing inaccurate records, handling missing values, and standardizing data formats to ensure the reliability and accuracy of the dataset for subsequent analysis.The cleaned data was then saved and uploaded into a Spark environment, and by leveraging Apache Spark's capabilities for efficient and scalable data processing, I continued to analyze the data and focused on data transformation.

The main objectives of the analysis and visualization stages are to extract meaningful insights from the southeastern dataset. Emphasis was placed on exploring trends related to precipitation, temperature fluctuations, and variations in wind speed. These insights are crucial for developing a comprehensive understanding of the region's climate patterns.

**Blockers**

The transfer of large datasets across networks, Hive and Spark, GCP was extremely slow speed with lots of errors encountered. Downloading the dataset was also strenuous to the system causing the process to be longer than expected. The solution was to avoid using multiple tabs at once. Analyzing a substantial 2GB dataset initially strained the system, leading to sluggish performance and resource-intensive operations due to limitations in memory and processing capabilities. Another blocker was time constraint since the project involved several parts making it challenging to perform an exhaustive exploration of the Southeast weather dataset. I prioritized key aspects of the analysis and focused on extracting meaningful insights within the given time constraints.

|  |  |
| --- | --- |
| **PROJECT**-*project was done in project called “My first project* |  |
| **BUCKET-** *bucket was created and data stored* | A screenshot of a computer  Description automatically generated |
| **CLUSTER-** *cluster was created and VM instances* |  |
| **DIRECTORY CREATED** – *hive\_data* | A screenshot of a computer  Description automatically generated |
| *Uploading DATA* |  |
| *Confirming that data is in directory* | A black screen with white text  Description automatically generated |
|  | A screen shot of a computer  Description automatically generated |
| *Using hive to create data base* | A screen shot of a computer  Description automatically generated |
| To create table in hive |  |
| Query to Select top 10 | A screenshot of a computer program  Description automatically generated |
| Queries to clean table, Remove duplicate, Remove nulls, Removing unwanted characters and negatives | A screen shot of a computer  Description automatically generated |
| Select 10 from cleaned. | A screenshot of a computer screen  Description automatically generated |
| moving data to local directory | A screenshot of a computer  Description automatically generated |
|  | A screenshot of a computer  Description automatically generated |
|  | A screen shot of a computer program  Description automatically generated |
| Data Set uploaded as *Southeast2.csv* | A screenshot of a computer  Description automatically generated |

**SPARK**

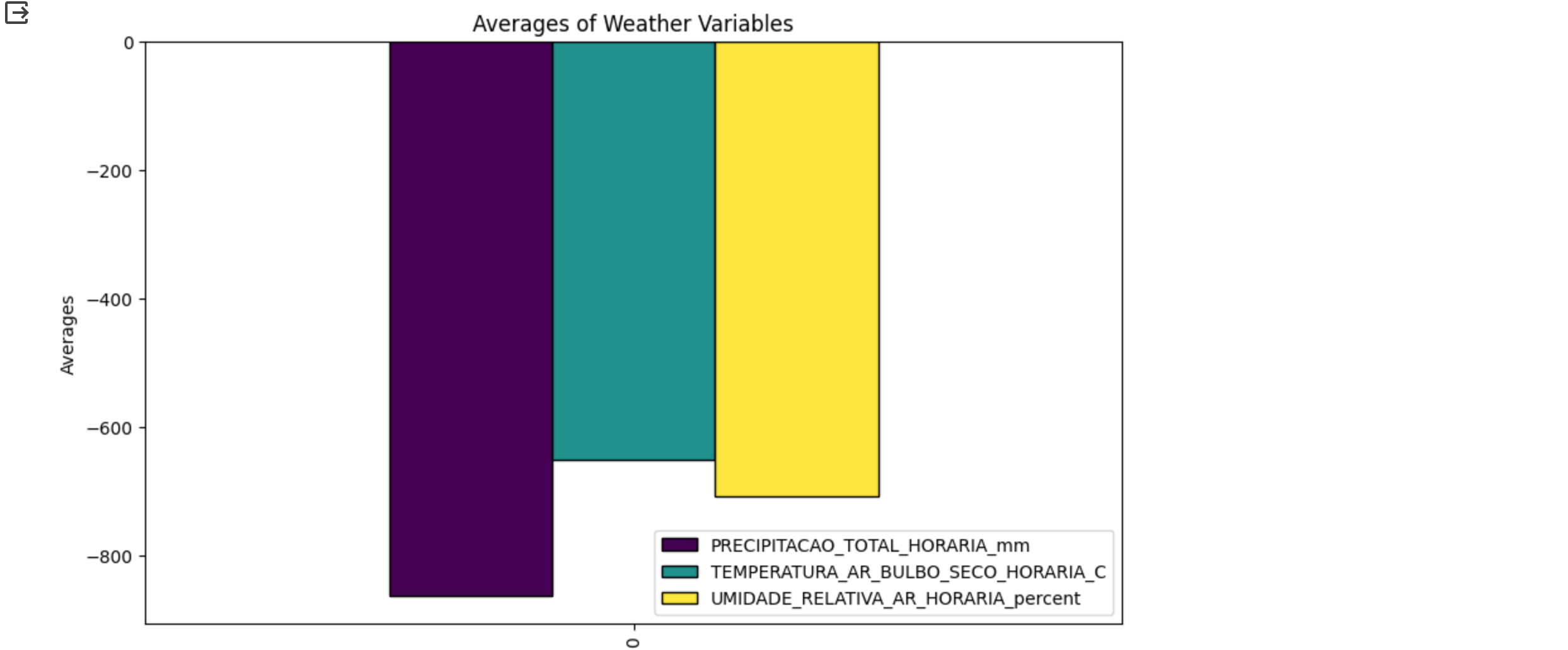
|  |  |
| --- | --- |
| **Importing necessary libraries** | A white background with black and white clouds  Description automatically generated with medium confidence |
| **Create Schema** | A screenshot of a computer program  Description automatically generated |
| **Importing data using path** | A screenshot of a computer  Description automatically generated |
| **Top 5 of data set** | A white rectangular object with black text  Description automatically generated |
| **To print Schema** | A screenshot of a computer program  Description automatically generated |
| **Aggregation by region, state, and station** | A screenshot of a computer  Description automatically generated |
|  | A white square with black lines  Description automatically generated |
|  | A screenshot of a computer  Description automatically generated |
| **aggregate date by year using panda.** | A screenshot of a computer  Description automatically generated |
|  | A white rectangular object with black text  Description automatically generated with medium confidence |
|  | A white sheet with numbers and lines  Description automatically generated with medium confidence A screenshot of a data  Description automatically generated |
|  | A screenshot of a document  Description automatically generated |

**Visualizations**

To ensure that the data set was providing essential details, I created 3 research questions to help me better understand my data set. This questioned helped with the analyses through additional visualizations.

1. **What is the overall average of temperature and precipitation in the Southeast region, and how can visual representations of these total averages contribute to a comprehensive understanding of the regional climate?**A screenshot of a computer

   Description automatically generated A screen shot of a computer code

   Description automatically generated 

A graph showing different colors

Description automatically generated

**Some values are negative, so we exclude because they appear to be outliers or erroneous. Average temperature values are typically within a certain range, and precipitation values are non-negative.**

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A graph of a number of meters

Description automatically generated with medium confidence

1. **How does the average temperature, wind and precipitation vary across different states in the Southeast region? Are there noticeable differences in the average temperature levels, and if so, what factors might contribute to these variations?**

A screenshot of a computer program

Description automatically generated A graph of different colored bars

Description automatically generated A graph of different colored rectangular shapes

Description automatically generated

A graph of different colored bars

Description automatically generated

c. **How does the distribution of total precipitation vary among selected states in the Southeast region, and what insights can be gained from analysing this variation in precipitation patterns**?A screen shot of a computer

Description automatically generated

A pie chart with numbers and a number of percentages

Description automatically generated

**Insight**

1. I observed that the average temperature is around 22°C, indicating a moderate climate. The average precipitation is relatively low at 0.15 mm, suggesting a drier climate. Additionally, the average relative humidity is around 72.23%, providing insights into the moisture content in the air. This is as expected since Brazil is a relatively humid country with temperatures ranging between 20-26 degrees.
2. The analysis did not reveal significant differences in average temperature levels, wind levels and precipitation among states. A possible explanation for limited difference is that the Southeast region may exhibit geographical similarities, leading to consistent weather patterns across states.
3. After visualization, it is shown through the pie chart that MG (Minas Gerais): 969,369.6 mm has the highest precipitation and ES (Esprito Santo): 166,653.6 mm has the lowest precipitation. This provides a comprehensive understanding of the precipitation dynamics in the Southeast region.

**Conclusion**

In conclusion, the analysis of weather data across different states in the Southeast region has provided valuable insights into various meteorological variables. The exploration focused on key factors such as temperature, precipitation, and wind speed, aiming to understand regional patterns and variations. The bar charts illustrated the average values of temperature, precipitation, and wind speed across different states. These visualizations helped identify states with distinct weather patterns and highlighted the variability in climatic conditions within the Southeast region. The pie chart helps analyst understand which states have the highest and lowest precipitation levels in total and analyze how this affects climates.

The research questions addressed in this analysis were answered through a combination of statistical summaries and visual representations. The utilization of Spark facilitated the efficient processing of large-scale weather data, enabling comprehensive insights into regional weather patterns.

In summary, my exploration of Southeast region weather data contributes to a better understanding of the climatic dynamics in this area.

**Reference**

PROPPG/PPG em Inform&aacute;tica - Doutorado e Mestrado. (2023, January 29). *Climate weather surface of Brazil - hourly*. Kaggle. https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region/

*World Bank Climate Change Knowledge Portal*. Climatology | Climate Change Knowledge Portal. (n.d.). https://climateknowledgeportal.worldbank.org/country/brazil/climate-data-historical