**Capstone Project- Weather Analysis**

Name-Arpit Verma

**Project Summary**

The Weather Data Analysis project focuses on systematically collecting, organizing, and examining meteorological data to extract meaningful insights. The project employs a structured relational database to store and manage vital weather attributes such as temperature, humidity, wind speed, and atmospheric pressure.

At the core of this system is the final\_fact table, which holds the primary weather measurements. Supporting this table are several dimension or lookup tables that enrich the dataset with contextual information:

* Time and Date Lookup Tables: time\_lookup and date\_lookup allow for time-series analysis and trend identification across different time scales—hours, days, seasons, and years.
* Geographic Dimension Tables: city\_lookup, country, and city\_attributes provide regional context for the data, enabling geographic analysis and comparisons across locations.

This project aims to be a versatile tool for studying historical weather patterns, forecasting future trends, and guiding decision-making in domains such as environmental science, urban planning, disaster resilience, agriculture, and infrastructure development.

**Workflow Breakdown**

1. **Dataset Acquisition via GitHub**The first step involves downloading a well-prepared dataset from a specified GitHub repository. Although initially designed for university rankings, the process outlines general data sourcing techniques that were adapted to the weather dataset for this project.
2. **Data Transformation and Enrichment**Data is cleaned and transformed to ensure consistency, remove anomalies, and normalize formats. Additional enhancements include the incorporation of new problem statements and context-specific dimensions to make the data more insightful and analysis-ready.
3. **Tool Integration for Analysis**The refined dataset is integrated with powerful data tools including:
   * MySQL Workbench for relational querying and storage
   * Microsoft Excel for spreadsheet-based exploration
   * Power BI for visualization, dashboarding, and interactive reports
4. **Problem Solving via Power BI**Specific analytical questions are addressed using Power BI. It allows users to create dynamic visualizations, filter data by variables like time and location, and derive insights using DAX measures.
5. **Exploratory Data Analysis (EDA)  
   EDA is conducted using either Excel or SQL** queries in MySQL Workbench. This step helps to identify trends, detect anomalies, and uncover relationships in the data before further modeling or reporting.
6. **Presentation and Communication**A PowerPoint deck is developed to communicate the project's purpose, methods, and insights. Each problem statement is addressed with relevant visualizations, conclusions, and implications.
7. **Comprehensive Documentation**A detailed report is created that records every stage of the project—from initial data gathering to final visualizations. This includes methodologies, queries used, transformation logic, tool workflows, and decision-making outcomes.

**Project Objectives**

**The key goals of the project are:**

**1. Efficient Data Storage and Schema Design**

* Organize meteorological data in a normalized relational schema.
* Maintain accurate relationships between cities, timestamps, and weather metrics.
* Ensure high performance for analytical queries and data retrieval.

**2. Historical Weather Trend Analysis**

* Examine patterns such as rising or falling temperatures, shifts in humidity, and changes in wind behavior.
* Identify seasonal fluctuations and long-term climate trends.
* Provide data that supports scientific research into environmental changes.

**3. Geographic-Based Weather Comparison**

* Assess how weather varies between cities and countries.
* Analyze local anomalies and regional climate characteristics.
* Provide the basis for localized weather prediction and hazard identification.

**4. Time-Series Forecasting**

* Use temporal data to understand and predict weather behavior over hours, days, and seasons.
* Enable machine learning applications for forecasting extreme events like heatwaves, floods, or cold snaps.
* Assist in planning and resource management through predictive analytics.

**5. Decision-Making Support Across Sectors**

* Climate Science: Offer reliable data for modeling global warming effects.
* Disaster Preparedness: Inform emergency systems by highlighting high-risk weather patterns.
* Agriculture: Help farmers align planting and irrigation with weather trends.
* Urban Infrastructure: Support the design of weather-resilient cities and smart systems.
* Energy Management: Forecast demand for heating or cooling based on weather predictions.
* Transportation and Logistics: Optimize operations by anticipating disruptions due to adverse conditions.

**Significance of the Dataset**

**This weather dataset holds immense value for:**

**1. Climate Change Research**

* Tracks persistent changes in temperature and humidity.
* Identifies patterns linked to extreme events like droughts or storms.
* Aids international policy development with empirical evidence.

**2. Disaster Risk Reduction**

* Highlights historically vulnerable areas.
* Supports the creation of early warning systems for extreme weather.
* Informs infrastructure resilience planning.

**3. Urban and Environmental Planning**

* Guides smart city development by analyzing environmental stressors.
* Contributes to pollution monitoring and heat island studies.
* Facilitates sustainable city growth strategies.

**4. Precision Agriculture**

* Helps monitor weather effects on crop cycles.
* Guides farmers in selecting appropriate cultivation techniques.
* Supports irrigation management and yield optimization.

**5. Industry and Logistics**

* Informs flight route planning for airlines.
* Supports energy providers in managing grid loads.
* Enhances supply chain stability by mitigating weather-related risks.

**Data Dictionary**

**1. final\_fact (Core Weather Measurements)**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| city\_id | bigint | References city in city\_lookup |
| date\_id | bigint | References date in date\_lookup |
| time\_id | bigint | References time in time\_lookup |
| humidity | double | Humidity percentage |
| pressure | double | Atmospheric pressure (in hPa) |
| temperature | double | Temperature in Celsius |
| weather\_description | text | Textual description of weather (e.g., "clear") |
| wind\_direction | double | Direction of wind in degrees |
| wind\_speed | double | Speed of wind in meters per second |

**2. time\_lookup**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| time\_id | bigint | Unique identifier |
| time | text | Time (e.g., "14:00") |

**3. city\_attributes**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| country\_id | bigint | References country table |
| city\_id | bigint | City identifier |
| latitude | double | Geographical latitude |
| longitude | double | Geographical longitude |

**4. city\_lookup**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| city\_id | bigint | Unique city ID |
| City | text | Name of the city |

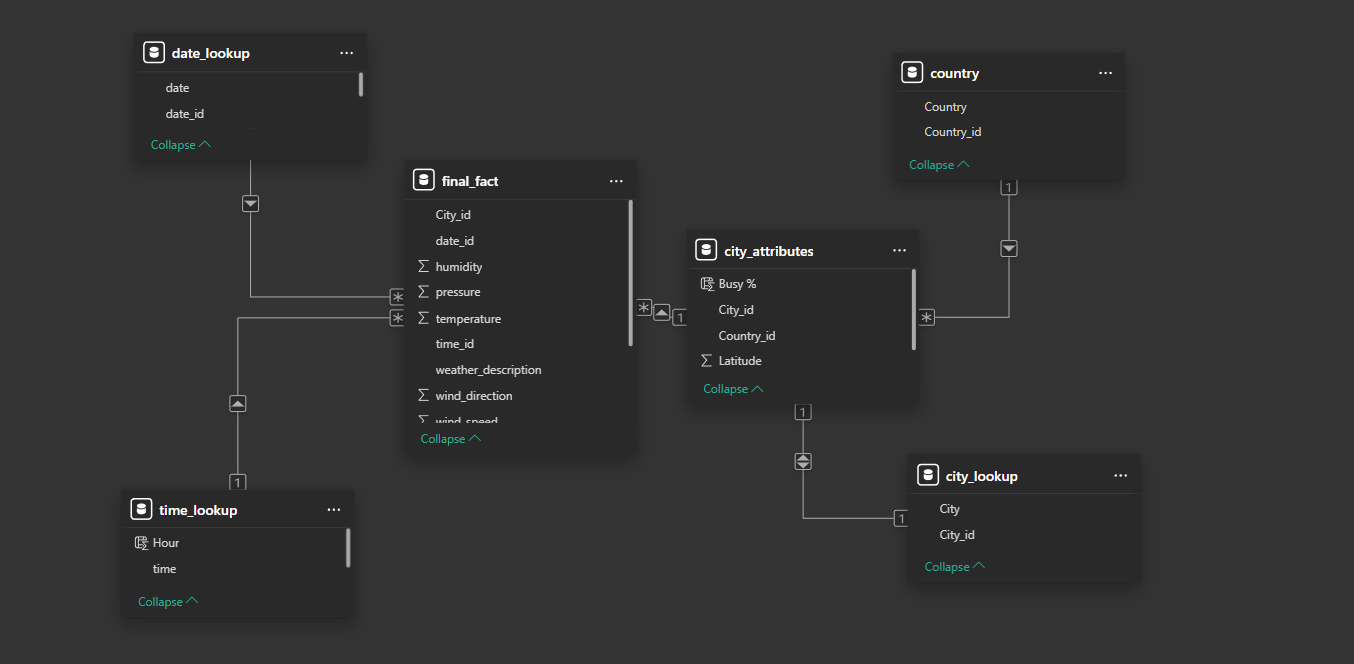
**5. country**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| country\_id | bigint | Unique country ID |
| country | text | Country name |

**6. date\_lookup**

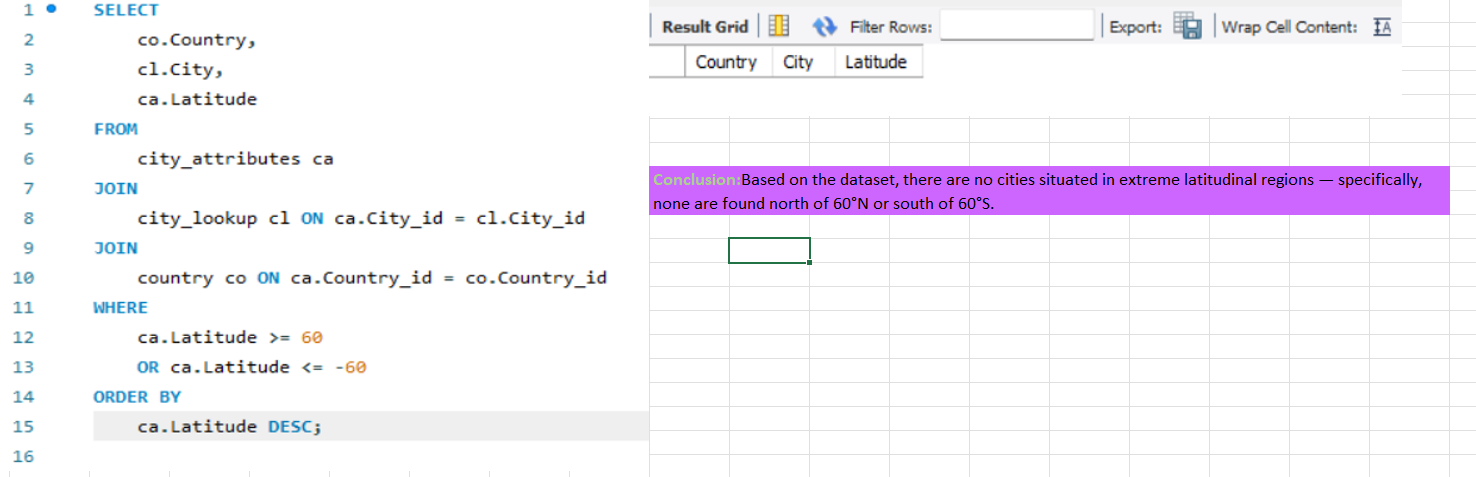
| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| date\_id | bigint | Unique date ID |
| date | text | Calendar date string |

**ER Diagram**



**EDA Question**

**1.** **Are there any countries with cities located at extreme latitudes, and how might this impact their climate?**

****

To determine whether any cities in the dataset are located at extreme latitudes — specifically above 60° North or below 60° South — a SQL query was written to extract relevant data. The query:

* Joined three related tables: city\_attributes, city\_lookup, and country using their respective IDs.
* Filtered the records to include only those cities where the latitude is either greater than or equal to 60, or less than or equal to -60.
* Selected key columns (Country, City, and Latitude) to display the results clearly.
* Sorted the output in descending order by latitude to highlight the most extreme values first.

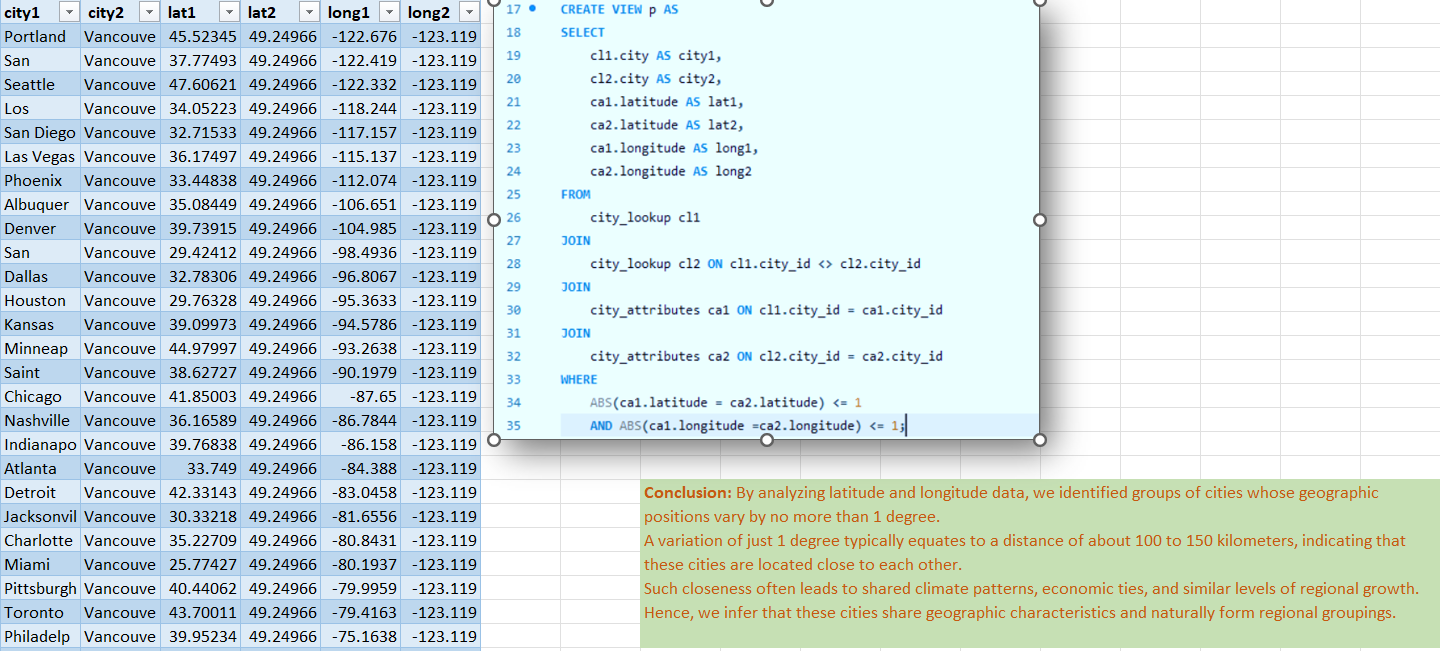
After executing the query, the result grid returned no records, indicating that there are no cities in the dataset located in these extreme latitudinal zones.

This means that:

* No cities are situated in the Arctic (above 60°N).
* No cities are situated in the Antarctic (below 60°S).

The absence of such cities may be due to the nature or limitations of the dataset, such as regional focus or data availability.

**2.Can you identify any clusters of cities with similar latitude and longitude values? What factors might explain these clusters?**



To identify geographically close cities in the dataset, a SQL view was created to analyze differences in both latitude and longitude. The query aimed to detect pairs of cities that are positioned within a narrow geographic range. The approach included:

* Joining the city\_lookup and city\_attributes tables twice to compare each city against every other city.
* Using the ABS() function to calculate the absolute difference in both latitude and longitude between each pair.
* Filtering results to include only those pairs where the latitude and longitude differences are less than or equal to 1 degree.
* Selecting city names and their corresponding geographic coordinates for comparison.

After executing the query, the result table returned multiple city pairs where the variation in coordinates is minimal — indicating proximity within approximately 100 to 150 kilometers.

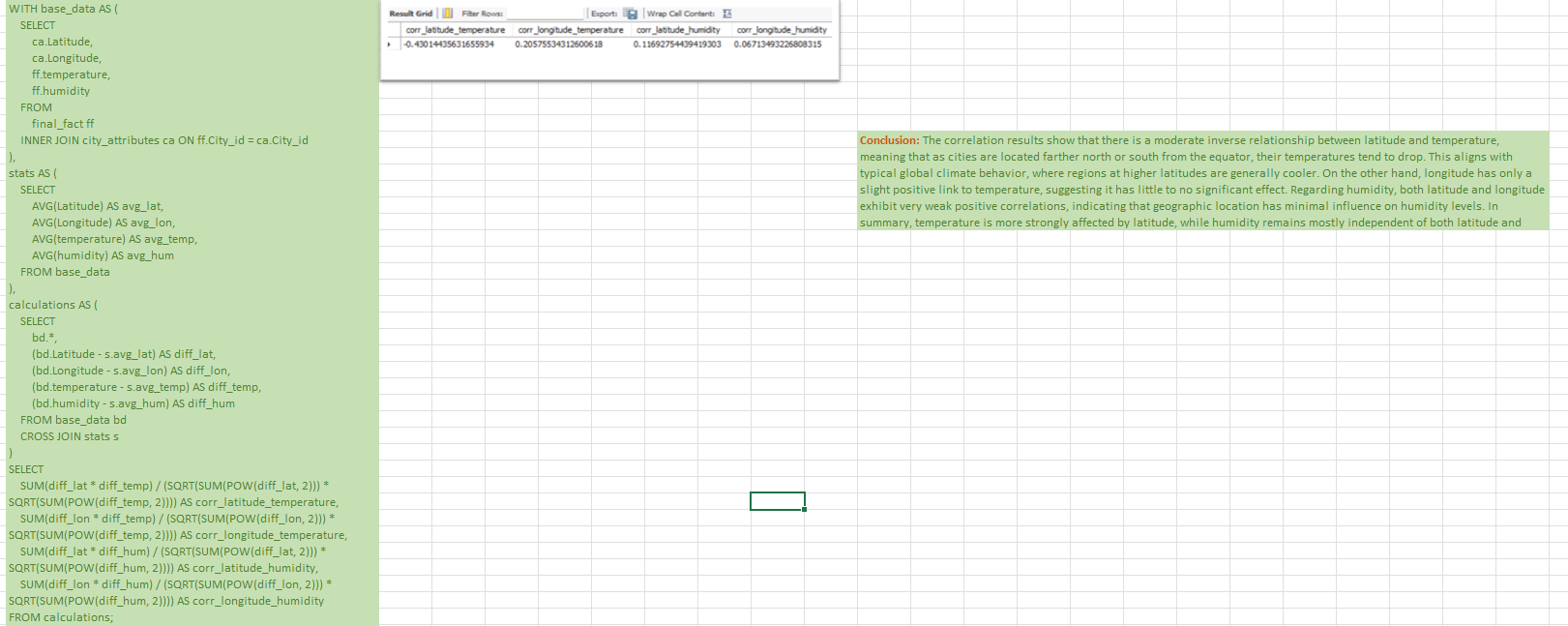
This suggests that:

* These cities are geographically near each other.
* Such proximity often implies shared climate conditions, regional transportation networks, economic interactions, and sociocultural connections.
* These city pairs can be considered regional clusters, naturally grouped based on their spatial location.

The analysis demonstrates how simple geographic thresholds can be used to uncover meaningful patterns of urban distribution and regional association.

**3.Are there any correlations between a city's geographical location (latitude and longitude) and its weather attributes, such as temperature or humidity?**

**Here is your detailed conclusion in the same style as the previous ones, based on the correlation analysis between geographic coordinates and weather attributes:**



To explore the relationship between geographic location (latitude and longitude) and weather variables (temperature and humidity), a SQL query was developed to compute correlation coefficients. The process involved:

* Creating a base\_data CTE (Common Table Expression) that joined city weather attributes with their geographic coordinates from related tables.
* Calculating average values for latitude, longitude, temperature, and humidity using the stats CTE.
* Computing the deviation of each value from its mean using a calculations CTE.
* Applying the Pearson correlation formula to measure the linear relationship between:
  + Latitude and temperature
  + Longitude and temperature
  + Latitude and humidity
  + Longitude and humidity

The final output showed the following correlation values:

* Latitude vs. Temperature: ~-0.43 — a moderate negative correlation
* Longitude vs. Temperature: ~0.20 — a weak positive correlation
* Latitude vs. Humidity: ~0.12 — very weak positive correlation
* Longitude vs. Humidity: ~0.07 — negligible correlation

These results imply that:

* Temperature tends to decrease as latitude increases (i.e., moving farther from the equator), which reflects well-known global climate patterns.
* Longitude has minimal influence on temperature, indicating it’s not a significant factor.
* Humidity is largely unaffected by either latitude or longitude, as indicated by the near-zero correlation values.

This analysis confirms that latitude has a notable impact on temperature, while humidity appears to be influenced by other non-geographical factors, such as local terrain, vegetation, or atmospheric conditions.