Weather Data Analysis Report

By Aarush Puri
Date: 01-05-2024

Introduction

This report details a methodical approach to analyzing a comprehensive weather dataset. The process involves initial data cleansing, followed by detailed analysis and visualization using Python and Tableau, respectively. The data includes various key weather parameters such as temperature, humidity, wind speed, and pressure, which are crucial for understanding atmospheric conditions. This document aims to guide beginners through the essential data science processes, from manipulating raw data to creating meaningful visual representations and extracting actionable insights.

Data Preparation (Phase 1: Python)

The first phase of the project focused on preparing the dataset for analysis using Python. This involved several critical steps to ensure data quality and usability.

Detailed Steps Performed:

Loading and Examining the Data: The dataset was imported into a Pandas DataFrame, enabling a preliminary examination to identify missing values, anomalies, and general data characteristics.

Handling Missing Values:

Missing entries in the 'Precip Type' column were replaced with the most frequently occurring value, assuming that the prevailing weather conditions would persist.

Outlier Identification and Correction:

For the 'Temperature (C)' and 'Pressure (millibars) etc' variables, outliers were identified using Z-scores. Records with Z-scores beyond the thresholds of 3 or -3 were considered extreme and excluded to normalize the data.

Zero values in 'Pressure (millibars)', which were unrealistic, were substituted with the median pressure from the dataset to correct these inaccuracies.

Data Type Adjustments:

The 'Formatted Date' column was transformed from string format to datetime to better support subsequent time-series analysis.

Data Analysis and Visualization (Phase 2: Tableau)

The cleaned data was then analyzed visually using Tableau, which provided a platform for creating interactive and intuitive visualizations.

Visualizations Developed:

Temperature Fluctuations Over Time:

A line graph was generated to depict the seasonal and yearly temperature variations, highlighting both expected weather patterns and unexpected temperature anomalies.

Monthly Average Humidity Trends:

A bar chart displayed how average humidity levels fluctuated monthly, offering insights useful for sectors dependent on moisture levels, such as agriculture.

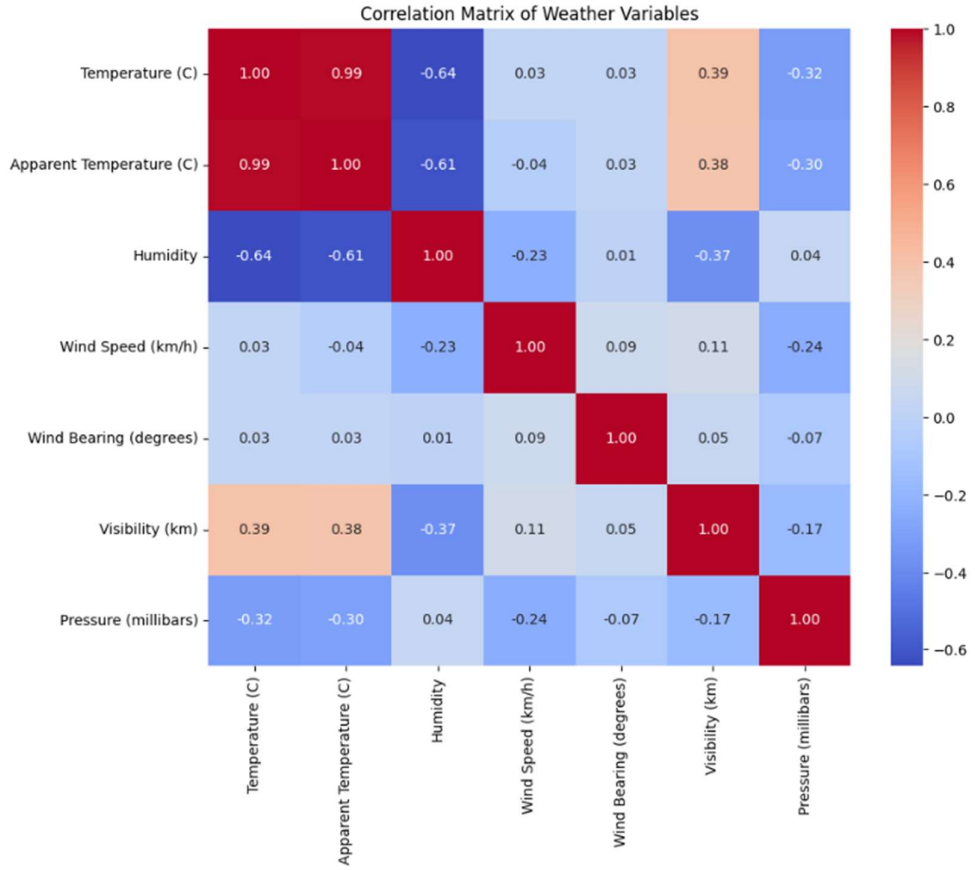
Correlation and Regression Analysis (Phase 3: Python)

In the final phase, deeper statistical techniques were applied to explore relationships between weather elements and to predict future conditions.

Analysis Techniques Employed:

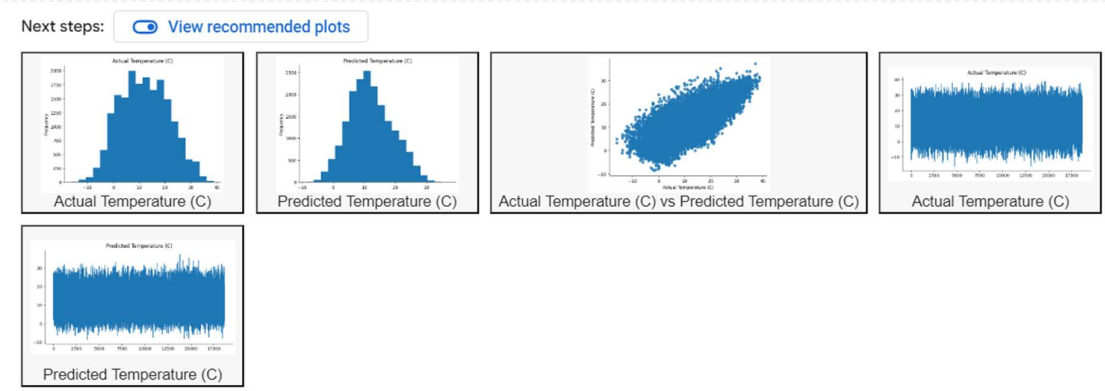
Correlation Analysis:

This analysis helped uncover the relationships between different meteorological variables, identifying which factors most significantly impact temperature variations.



Regression Modeling:

A linear regression model was created to forecast temperatures based on variables like humidity and wind speed. The effectiveness of this model was evaluated using statistical metrics such as the R-squared value and Mean Squared Error (MSE), ensuring the model's accuracy and reliability.




```
[109] # Evaluating the model
      mse = mean_squared_error(y_test, predictions)
      r2 = r2_score(y_test, predictions)
      print("Mean Squared Error:", mse)
      print("R^2 Score:", r2)

      Mean Squared Error: 40.865641963733175
      R^2 Score: 0.5470752425622818
```

Conclusion:

The project showcased a fundamental approach to weather data analysis suitable for novices, from thorough data cleansing in Python to dynamic visualizations in Tableau and advanced statistical evaluations. The findings highlight significant seasonal and weather-related trends within the dataset, demonstrating how data science can be applied to real-world issues, aiding decision-making in weather-dependent industries. This exercise provides a robust foundation for future analytical projects that require a methodical and detailed approach to data analysis.