TOPIC: Exploring Patterns and Trends in Weather Data

Introduction

Weather patterns significantly impact our daily lives, agriculture, and overall climate health. Analyzing weather data allows us to understand these patterns, identify trends, and make predictions. In this project, I want to explore patterns and trends in various weather datasets. By using machine learning models, we can uncover valuable insights that can inform weather forecasting and climate research.

Dataset Description

I have ten sub-datasets, each containing specific weather related variables, each sub-dataset contains 372,059 rows and 6 columns. The subsets are:

1. Annual mean temperature
2. Max temperature of warmest month
3. Annual precipitation
4. Temperature annual range
5. Min temperature of coldest month
6. Max temperature of warmest month
7. Annual precipitation

Eah dataset has a unique feature and can not be combined or augmented to make one dataset. The source of this dataset is **kaggle.** Initial observations indicate variation in temperature and precipitation, highlighting the diverse climate conditions across different regions.

Data Cleaning

Data cleaning is crucial for ensuring the accuracy of our analysis. I have encountered several issues during this process,including filling missing values, removing outliers and making the dataset to be consistent.

1. **Missing Values:** are blank values in columns of dataset, i had used isnull() function to and sum() function to find missing values in each column and to fill them i have filled them using mean of each column and using the fillna() function
2. **Outliers :** firstly I have used boxplot to visualize outliers then i used z-score to find lower limit and upper limit so that I can replace the values in column that is greater than upper limit with column mean and applies the same with number that is less than lower limit is also replaced with column mean.

3. **Inconsistent Data Formats**: I standardized the data formats across all sub-datasets to ensure consistency.

Data transformation

Changing numerical columns so that they can have mean of o and standard deviation of 1 using z-score normalization.

Creating new variables

1. creating variable by calculating the ratio of "COUNT" to "AREA" and the name of the variable is named "COUNT_AREA_RATIO"
2. creating variable by calculating the ratio of"MEAN" to "AREA" and the name of the variable will be "MEAN_AREA_RATIO"

Encoding Categorical Variables

This is changing variable values or strings to represent numbers that increase generalization of the model. I have used labelencoder class from sklearn to do column encoding

Extrapolatory Data Analysis

1. **Finding summary statistics**

   I have found summary of statistics using describe() function in pandas it give the minimum value, maximum value, mean, 25 percent, 75 percent and standard deviation of dataset

2. **Plotting histogram**

   Histograms are used for understanding distribution of individual columns, the histogram of annual mean temperature is showing high distribution and it is left skewed meaning that most temperatures are higher with occasional lower values.

   The histogram of area is showing nothing meaning area have small distribution that can be viewed

3. **Scatterplot**

   They help to visualize relationship between pair columns

   1. **Relationship between area and mean**

      The relationship is showing a 90 degrees upward points line suggesting that there's no any relationship and it is also showing that areas have constant values.

   2. **Relationship between Mean and Count**

      The plot is showing no relationship because all the points are scattered but points are showing a little downward points meaning there's little negative relationship between these two variables

4. **Pair plot**

It is used to show relationship of variable pairs in the dataset but only numerical pair

The pair plot is not showing any relationship of pair variables

6**. correlation heatmap**

Correlation heatmap shows correlation between different variables in the dataset, here it is showing that all the numerical column have no any relationship

Saving dataset to csv

The I saved my transformed dataset

Challenges faced

1. **Dealing with missing values**

   Deciding the best approach to handle missing values was challenging. Replacing missing value with mean can sometime change the dataset
2. **Outlier detection and removing**

   Removing outliers was challenging, outliers can affect analysis.
3. **Data consistency**

   Ensuring data format consistency across all columns was difficult.
4. Interpreting graphs was also challenging