Project Report

TASK 5: ANALYZING INDIAN GOVERNMENT OPEN DATA

Introduction:

This project aims to analyze the annual minimum temperature data in India from 1901 to 2021, obtained from the Open Government Data Platform of India. The dataset contains monthly minimum temperature data for each year, which was used to calculate the annual minimum temperature. The analysis includes data cleaning, preprocessing, visualization, and interpretation of the results.

Data Cleaning and Preprocessing:

The dataset was initially loaded into a pandas DataFrame, and it was found that all the values were numeric, and there were no missing values. The YEAR column was converted to datetime format, and the month and year were extracted as separate columns. The monthly columns were then pivoted to create a matrix for further analysis.

Data Analysis:

The average temperature for each month was calculated, which showed that the temperature is lowest in December and highest in June. The annual average temperature was also calculated, which showed a slight increase over the years. The temperature anomaly for each year was calculated by subtracting the mean annual temperature from the annual temperature. It was found that the temperature anomaly fluctuated over the years, with some years showing higher temperatures than the mean and some showing lower temperatures.

Data Visualization:

The annual average temperature was visualized using a line chart, which showed a slight increasing trend over the years. The temperature anomaly was visualized using a line plot, which showed the fluctuations in temperature over the years. The monthly average temperature was visualized using a line chart and a bar chart, which showed the trend of temperature over the months and the average temperature for each month, respectively.

The temperature distribution by month was visualized using a boxplot, which showed the range and quartiles of temperature for each month.

Principal Component Analysis (PCA):

PCA was applied to the monthly average temperature data to reduce the dimensionality of the data. It was found that the first two principal components explained 60.87% of the variance in the data.

t-Distributed Stochastic Neighbor Embedding (t-SNE):

t-SNE was applied to the monthly average temperature data to visualize the high-dimensional data in a two-dimensional space. It was found that the data points were clustered together, indicating similar temperature patterns.

K-Means Clustering:

K-Means clustering was applied to the monthly average temperature data to group similar temperature patterns together. It was found that the optimal number of clusters was 5, which explained 72.1% of the variance in the data.

Standardization:

The monthly average temperature data was standardized to have zero mean and unit variance. This was done to ensure that all the variables were on the same scale and to avoid any bias towards variables with larger values.

Conclusion:

The analysis of annual minimum temperature data in India from 1901 to 2021 showed a slight increasing trend over the years. The temperature anomaly fluctuated over the years, indicating the variability in temperature patterns. The data visualization and dimensionality reduction techniques used in this project provided insights into the temperature patterns and variations in India. The K-Means clustering results can be used to identify regions with similar temperature patterns and to develop targeted climate change mitigation strategies.

Limitations:

The analysis in this project was based on annual minimum temperature data, which may not provide a complete picture of the temperature patterns in India. Other factors such as humidity, precipitation, and wind speed may also affect the temperature patterns and should be considered in future studies.

Recommendations:

The analysis in this project can be extended to include other temperature variables and to consider the spatial variation in temperature patterns. The K-Means clustering results can be used to develop targeted climate change mitigation strategies for different regions in India. The analysis can also be updated with more recent data to provide a more up-to-date picture of the temperature patterns in India.