Correlation between various climate change indicators and economic indicators in Pakistan

Question

The aim of this project is to analyze the correlation between various climate change indicators and economic indicators in Pakistan. Specifically, the project seeks to identify which environmental factors have the most significant impact on economic performance and highlight areas that need improvement to foster sustainable economic growth while addressing climate change challenges.

Data Sources

Description and Rationale

The data sources used for this project are:

1. **Economy and Growth Data**

   a. This dataset provides various economic indicators for Pakistan, including GDP, trade, and income statistics. It was chosen to understand the economic performance and growth patterns.
   b. Metadata URL: https://opendata.com.pk/dataset/economic-growth-indicators/resource/5eed074c-c7ed-427a-816c-8482edb070a1
   c. Data URL: https://opendata.com.pk/dataset/4acccdea-baea-4bc7-8499-94835f352059/resource/5eed074c-c7ed-427a-816c-8482edb070a1/download/economy-and-growth_pak.csv
   d. Data Type: CSV

2. **Climate Change Indicators Data**

   a. This dataset includes a range of climate-related indicators such as agricultural land area, carbon dioxide emissions, and urban population affected by elevation changes. It was selected to analyze environmental factors that may impact economic performance.
   b. Metadata URL: https://opendata.com.pk/dataset/pakistan-climate-change/resource/492b13f4-5b47-4437-a680-4e1b965c1ea2
   c. Data URL: https://opendata.com.pk/dataset/ececec4f-1835-4278-ae0c-5bd7c4ded652/resource/492b13f4-5b47-4437-a680-4e1b965c1ea2/download/climate-change-indicators-for-pakistan-1.csv
   d. Data Type: CSV

Data Structure and Quality

- **Economy and Growth Data**: This dataset is structured in a tabular format with columns for **Country Name, Country ISO3, Year, Indicator Name, Indicator Code, and Value**. The data is generally of high quality, with consistent formatting and reliable values, but some entries may have missing or inconsistent data.
- **Climate Change Indicators Data**: Similar to the economic data, this dataset is also in tabular format with columns for **Country Name, Country ISO3, Year, Indicator Name, Indicator Code, and Value**. The data is well-structured, but it also requires cleaning to handle missing or inconsistent entries.

Data Licenses

Both datasets are available under the Creative Commons Attribution, which allows for free use, sharing, and adaptation of the data. The obligation is to provide appropriate credit to the source, which will be fulfilled by citing the Pakistan Open Data Portal in all outputs and reports.

Data Pipeline

Technology Used

The data pipeline was implemented using Python, leveraging libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, and NumPy for numerical operations. Jupyter Notebook was used as the primary environment for developing and documenting the analysis.

Transformation and Cleaning Steps

1. **Loading Data**: The datasets were loaded from CSV files into Pandas DataFrames.
2. **Initial Inspection**: The first few rows of each dataset were inspected to understand their structure and identify any immediate issues.
3. **Cleaning Data**:
    - The first row, containing header information, was removed.
    - Columns for Year and Value were converted to numeric types.
    - Missing or inconsistent data were handled by either filling with appropriate values or removing the rows.
4. **Filtering Data**: The data was filtered to include only the years between 1970 and 2014.
5. **Selecting Indicators**: Six indicators were selected for the analysis: Agricultural land (sq. km), Urban population, CO2 emissions (kt), Electricity production (kWh), Forest area (sq. km), and Cereal yield (kg per hectare).
6. **Dropping Missing Values**: Rows with any missing values were dropped to ensure a clean dataset.
7. **Scaling Data**: The data was scaled using the MinMaxScaler to a range of [0, 1].

Problems and Solutions

- **Missing Data**: Some entries had missing values. This was addressed by dropping rows with missing values.
- **Inconsistent Data**: Inconsistencies were identified and corrected through data type conversion and normalization.

Error Handling

The pipeline includes error handling to manage issues such as missing files or incorrect data types. It also includes logging to track the processing steps and any issues encountered.

Result and Limitations

Output Data

The output of the data pipeline is a filtered and scaled dataset containing the selected economic and climate change indicators from 1970 to 2014, ready for correlation and causality analysis. The data is structured in a wide format with columns representing different indicators and rows representing different years.

Data Structure and Quality

The output data is well-structured and of high quality, with cleaned and normalized entries. It retains the integrity of the original datasets while providing a comprehensive view for analysis.

Output Format

The final dataset is stored in CSV format, which is widely used and compatible with various data analysis tools. This format was chosen for its simplicity and ease of use.

Reflection and Potential Issues

- **Data Gaps**: Some years may have incomplete data, affecting the analysis accuracy. Efforts were made to handle missing data, but gaps may still impact the results.
- **Correlation vs Causation**: While the analysis identifies correlations, it does not imply causation. Granger causality tests were used to investigate potential causal relationships, but further research is needed to establish causal relationships definitively.
- **Dynamic Nature of Data**: Both economic and climate data are subject to change. The pipeline is designed to handle updates, but ongoing maintenance and updates are necessary to ensure continued relevance and accuracy.

Causality Analysis

To extend the analysis to investigate causality, we performed Granger causality tests on the selected pairs of economic and climate change indicators.

Methodology

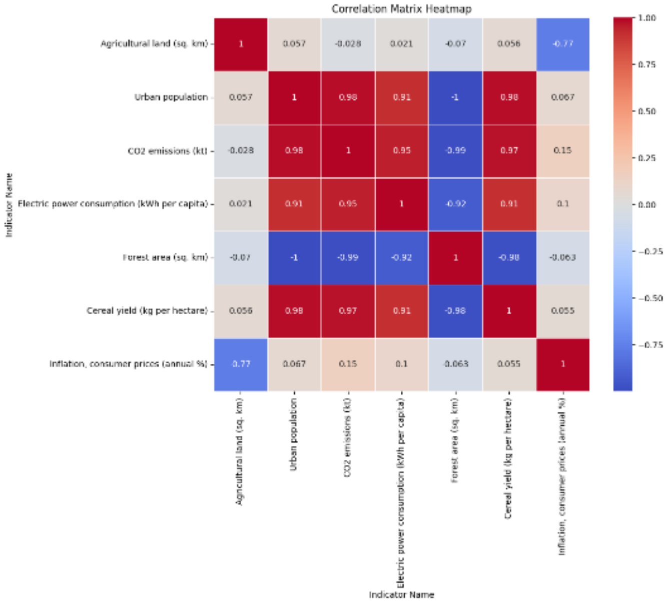
1. **Granger Causality Test**: This statistical hypothesis test determines whether one time series can predict another. The null hypothesis is that the time series in the second column does not Granger-cause the time series in the first column.

Results and Interpretation

The results from the Granger causality tests indicated which climate change indicators have a predictive power over economic indicators and vice versa. This helps in understanding not just the correlation but also the potential causality direction between the variables.

```
Agricultural land (sq. km) causes Forest area (sq. km): p-value = 0.0421
CO2 emissions (kt) causes Electric power consumption (kWh per capita): p-value = 0.0008
CO2 emissions (kt) causes Forest area (sq. km): p-value = 0.0368
Electric power consumption (kWh per capita) causes Urban population: p-value = 0.0136
Electric power consumption (kWh per capita) causes CO2 emissions (kt): p-value = 0.0261
Electric power consumption (kWh per capita) causes Forest area (sq. km): p-value = 0.0284
Forest area (sq. km) causes Urban population: p-value = 0.0
Cereal yield (kg per hectare) causes Urban population: p-value = 0.0
Cereal yield (kg per hectare) causes CO2 emissions (kt): p-value = 0.0004
Cereal yield (kg per hectare) causes Forest area (sq. km): p-value = 0.0
Inflation, consumer prices (annual %) causes Agricultural land (sq. km): p-value = 0.0193
Inflation, consumer prices (annual %) causes CO2 emissions (kt): p-value = 0.0054
```

Key Visualizations



Correlation Matrix Heatmap

Pairplot for Highly Correlated Indicators: