Correlation Between Unemployment Rates and Crime Incidents in USA(2004-2014)

November 27, 2024

Question

The primary question of this project is to investigate the correlation between unemployment rates and crime incidents in the USA from 2004 to 2014. Specifically, the project aims to identify states with the highest correlations, examine temporal changes, and explore overall trends.

Data Sources

Unemployment Data

- **Source1**: Kaggle (Dataset: "Unemployment in America Per US State")
- **Reason for Choice**: This dataset provides comprehensive state-wise unemployment rates across different years and months, which is essential for analyzing economic conditions.
- **Content**: The dataset contains monthly unemployment rates for each state in the USA.
- **License**: Standard open-data license, allowing for analysis and sharing with proper attribution.
- **Link**: https://www.kaggle.com/datasets/justin2028/unemployment-in-america-per-us-state

Crime Data

- **Source2**: Kaggle (Dataset: "US Crime DataSet")
- **Reason for Choice**: This dataset offers detailed records of crime incidents across various states, crucial for understanding crime patterns.
- **Content**: The dataset includes records of crime incidents with details about the nature and frequency of crimes.
- **Link**:https://www.kaggle.com/datasets/mrayushagrawal/us-crime-dataset

.

Data Pipeline

Technology Used

- **Programming Language**: Python
- **Libraries**: Pandas, NumPy, SQLite3, Kagglehub

Steps

1. **Data Extraction**: Download datasets from Kaggle using `kagglehub`.
2. **Data Cleaning**:
   - Remove duplicates and handle missing values.
   - Convert month names to numeric values for consistency.
   - Filter data for the years 2004 to 2014.
   - Ensure columns have appropriate data types.
3. **Data Transformation**:
   - Rename columns for consistency.
   - Merge unemployment and crime datasets on 'State', 'Year', and 'Month'.
   - Aggregate crime incidents by state and year, calculating the mean unemployment rate.
4. **Data Storage**: Save the cleaned and transformed data to a CSV file and SQLite database.

Problems Encountered and Solutions

- **Inconsistent Data Formats**: Resolved by standardizing month formats and ensuring consistent data types across datasets.
- **Missing Values**: Addressed by implementing appropriate imputation techniques or excluding incomplete records.

Meta-Quality Measures

- Error handling routines to manage invalid data entries.
- Regular checks to maintain data integrity and quality.

Results and Limitations

Output Data

The final dataset is a cleaned and aggregated table with columns for State, Year, 'Percent (%) of Labor Force Unemployed', and the total number of crime incidents. The data is saved in CSV format and an SQLite database for easy querying and analysis.

Data Structure and Quality

- **Structure**: Well-organized table suitable for analysis.
- **Quality**: Verified for completeness and accuracy through cleaning and transformation steps.

Format

CSV and SQLite were chosen for their simplicity and compatibility with data analysis tools.

Critical Reflection

- **Potential Issues**: Possible biases in data collection or reporting discrepancies in crime data across different states.
- **Limitations**: The analysis is limited to the available data and selected timeframe, which may not capture all influencing factors.
- **Future Work**: Additional data sources and longer timeframes could provide more comprehensive insights. Including other socioeconomic factors could also enrich the analysis.

Visualizations

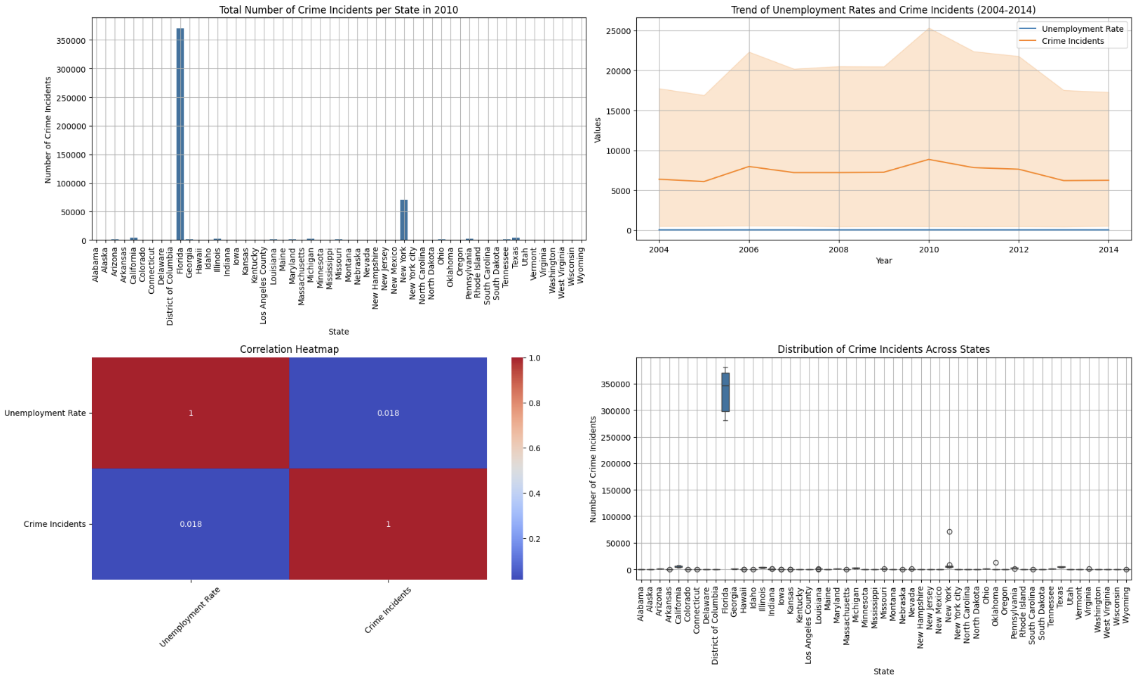

Figure 1: Bar Graph, Historical Line Graph, Heatmap