Data Pipeline Report

Abdul Basit

23460819

1 Question

What is the Relationship Between Crime Rate and Unemployment Rates Across the US from 1990 to 2010.

2 Data Sources

2.1 Description of Data Sources

2.1.1 US Unemployment Rate by County, 1990-2016

This data represents the Local Area Unemployment Statistics from 1990-2016, broken down by state and month.
URL: `https://www.kaggle.com/datasets/jayrav13/unemployment-by-county-us`
Data Type: CSV Format
License Type: CC0: Public Domain

2.1.2 US Crime DataSet

The Dataset contains the record of all the crimes in the US from 1980. It has been further broken down by state.
URL: `https://www.kaggle.com/datasets/mrayushagrawal/us-crime-dataset`
Data Type: CSV Format
License Type: U.S. Government Works

According to **17 U.S.C. § 105** [`https://www.copyright.gov/title17/92chap1.html#105`], works created by U.S. federal government employees as part of their official duties are not eligible for copyright protection. As a result, such works are in the public domain.

2.2 Structure and Quality of Data

2.2.1 US Unemployment Rate by County, 1990-2016

This dataset is in **CSV** format. The data quality is pretty good and there are no missing values. Each row represents the Unemployment Rate for a specific month in a specific County. The dataset contains the following columns:

`'Year', 'Month', 'State', 'County', 'Rate'`

2.2.2 US Crime DataSet

Same as the Unemployment Dataset, this dataset is also **CSV** format, clean and complete. There are no missing values. Both datasets are well-maintained and well-kept. Each row represents a single instance of a crime in the US. Following columns are available in this particular dataset:

```
'Record ID', 'Agency Code', 'Agency Name', 'Agency Type', 'City',
'State', 'Year', 'Month', 'Incident', 'Crime Type', 'Crime Solved',
'Victim Sex', 'Victim Age', 'Victim Race', 'Victim Ethnicity',
'Perpetrator Sex', 'Perpetrator Age', 'Perpetrator Race',
'Perpetrator Ethnicity', 'Relationship', 'Weapon', 'Victim Count',
'Perpetrator Count', 'Record Source'
```

3 Data Pipeline

3.1 Pipeline Overview

The data pipeline is implemented in Python. The library **Pandas** was heavily utilized to clean and transform datasets. I have used a slightly modified version of the **Medallion Architecture** [`https://learn.microsoft.com/en-us/azure/databricks/lakehouse/medallion`] to store my data.

3.2 Transformations and Cleaning Steps

- Filtered data from 1990-2010 on both datasets to make it fit our problem domain.
- Validated the Month Names, Year, and Rate data types on both datasets.
- Added a Month Number column in both datasets to aggregate or use in various graphs or calculations easily.
- Added an Unemployment Level column by binning the unemployment rates to visualize as Low, Medium, and High easily.
- For the unemployment dataset, aggregated at the State level instead of the County level to make it comparable.
- For the Crime Rate dataset aggregated on Year, Month, and State to make it comparable.
- The Crime Rate dataset contains a lot of unused columns, so we filter out only the columns that we need
- Merged the two datasets on Year, Month, and State level to get the final, analysis-ready data.

3.3 Problems Encountered and Solutions

- The Crime Rate dataset was quite large and difficult to aggregate. Because each row represented a crime instance. I had to aggregate it down to calculate the number of crimes per month for each state.

3.4 Meta-Quality Measures

- Validation scripts to check data consistency after each step.
- In the next steps, I will add logging of all errors and dropped columns

4 Result and Limitations

4.1 Output Data

The final output data is stored in an SQLite Database called **'crime_unemployment_analysis'** As mentioned above, I have used a slightly modified version of the Medallion Architecture. There is no bronze layer. Transformed and Cleaned data will be stored in the Silver layer in two tables. The unemployment dataset in its own table (**unemployments**) and the Crime Rate dataset in its own (**crimes**). This data is not yet analysis-ready or tailored to any business use case. The merged dataset will be stored in the Golden layer and is ready for the final analysis. It is stored in the table **unemployment_crime_merged**

4.2 Data Structure and Quality

The final data is of high quality with very minimal missing values. The data is in tabular form with each row representing the unemployment rate and crime count for a month, year, and state in the US.

4.3 Output Format

SQLite Database for efficient querying and storage of larger datasets.

4.4 Potential Issues

- Economic changes like unemployment may take time to affect crime rates, which the analysis may not fully capture.
- Imputation and mapping processes could introduce biases, especially in regions with sparse data.