Methods Of Advanced Data Egnineering

Written Report

Muhammad Arbaz

Matrikel Number: 23073711

Friedrich-Alexander-Universität
Technische Fakultät

1 Introduction

Crime rates have an impact on the economy, sociology, and public policy. High unemployment rates in a city or region significantly contribute to increased crime levels, as financial stress and limited opportunities often drive individuals toward illegal activities. This project aims to investigate this relationship by analyzing two datasets: one containing crime statistics and the other providing unemployment rates across US states, to determine whether unemployment influences crime rates or not.

2 Question

This project investigates whether there is a relationship between unemployment rates and crime rates across different states in the USA. Specifically, it aims to explore if higher unemployment rates correlate with specific types of crime in certain regions or states.

3 Data Sources and Preparation

We have chosen two datasets for this report that provide detailed insights on crime statistics and unemployment rates across different states of USA:

3.1 Data Source 1

- **Data Source 1:** US Crime Dataset
- **Metadata Url:** Link
- **Data Url:** US_Crime_DataSet.csv
- **Data Source:** Kaggle
- **Data format:** CSV
- **License:** U.S. Government Works

This dataset provides detailed crime reports from US states, including crime types, victims, perpetrators, and incident details. It allows us to analyze crime trends by state and year. The US Crime dataset is primarily based on US government data, which is typically in the public domain or under a CC0 license, allowing free use without restrictions.

Structure and Quality: The dataset is detailed and provides details about crime reports in different states of the USA since 1980. Each row consists of a crime record recorded for that time and location within the state. Several columns, particularly the victim count, perpetrator count, and incident count, are likely to have missing values. The data is not in a consistent format due to incomplete or inconsistent reporting. For instance, some records have victim counts or perpetrator counts of zero, which could lead to inaccurate aggregation results when summing these values across states, years, and months. These missing values need to be addressed to avoid skewed or inaccurate totals. Additionally, the "month" column, stored as string (e.g., "January"), requires standardization into a numeric format to facilitate proper aggregation. These data quality issues, including missing values and inconsistent formats, must be resolved to ensure reliable results from the data transformation process.

3.2 Data Source 2

- **Data Source 1:** Unemployment in America, Per US State
- **Metadata Url:** Link
- **Data Url:** Unemployment in America Per US State.csv
- **Data Source:** Kaggle
- **Data format:** CSV
- **License:** CC BY-NC-SA 4.0

This dataset contains unemployment data by state and year, including metrics and trends such as the total labor force and unemployment rate across US states. The US Unemployment dataset is licensed under a Creative Commons

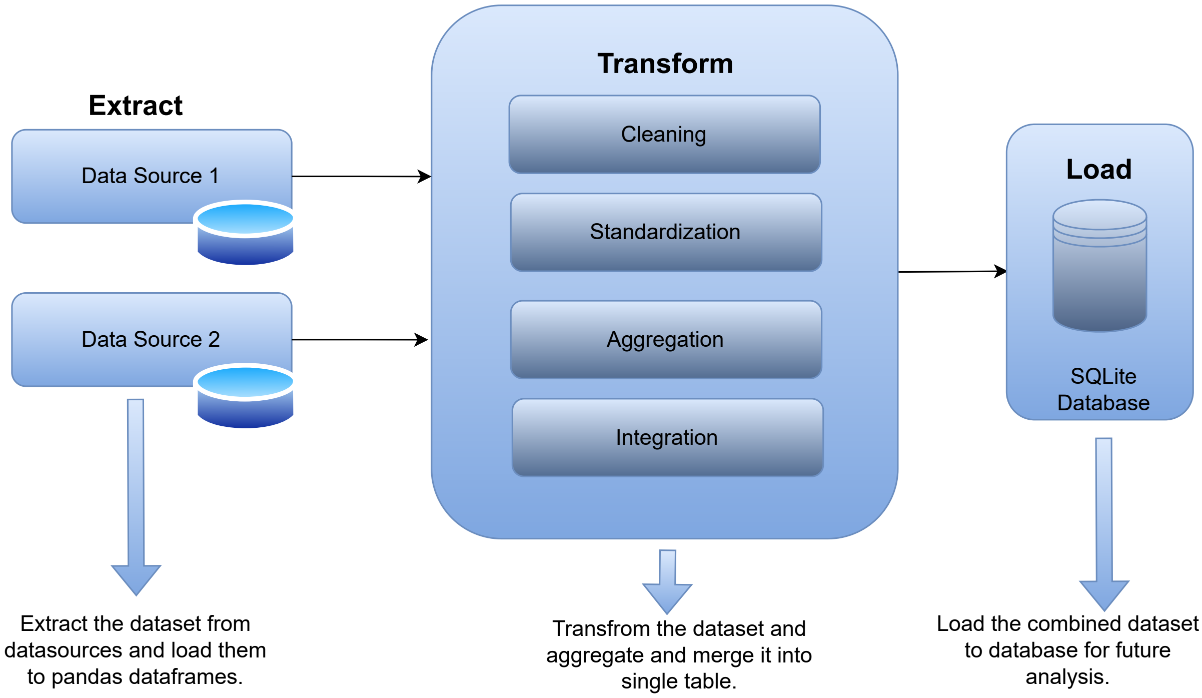
Extract
Data Source 1
Data Source 2
Transform
Cleaning
Standardization
Aggregation
Integration
Load
SQLite Database
Extract the dataset from datasources and load them to pandas dataframes.
Transfrom the dataset and aggregate and merge it into single table.
Load the combined dataset to database for future analysis.

Fig. 1: Overview of the ETL (Extract, Transform, Load) pipeline showing the pipeline flow from raw data extraction from online sources to data transformation and final loading into the target database.

Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license. It complies with the terms for public use, provided appropriate attribution is given, any changes are indicated, and the dataset is not used for commercial purposes.

Structure and Quality: It includes both monthly and yearly data, but some fields are recorded as string and require conversion before being saved into the database. While there are no missing values, several columns contain inconsistent data. One issue is the presence of numeric values stored as strings in columns such as "total civilian non-institutional population in state" and "total civilian labor force in state," which need to be converted to proper numeric types for accurate analysis and database storage.

These two data sources are merged based on the "state," "year," and "month" columns to enable more effective analysis on a single records of table. The final output of the dataset is shown in Table.

In the following section, we will discuss the ETL pipeline for our project.

4 Data Pipeline

The ETL (Extract, Transform, Load) pipeline for this project is developed in Python!Van Rossum and Drake Jr [1995] using the Pandas pandas development team [2020] and SQLAlchemy Bayer [2012] for data manipulation and database operations. This can automate the task of data extraction, transformation, and loading into database. The ETL pipeline diagram is shown in Figure 1.

4.1 Extract

The pipeline begins by downloading datasets from online data sources mentioned in above section. Kaggle API is used to download the required CSV files, which are then loaded into Pandas DataFrames.

4.2 Transform

In the transformation phase, the following steps are applied to clean and merge the datasets:

Data Cleaning: Missing values are dropped to ensure data consistency. Column names are standardized by converting them to lowercase and replacing spaces and special characters with underscores. This ensures compatibility with database operations and future analysis.

Data Standardization: The US crime dataset's month names are mapped to numeric values for consistency with the US unemployment dataset.

Data Aggregation: The crime dataset is aggregated by state, year, and month to compute total incidents, victims, and perpetrators, creating a summary dataset of crimes in order to merge it with US unemployment dataset.

Data Integration: The aggregated crime dataset is merged with the US unemployment dataset on state, year, and month using an inner join. This integration ensures that only records with matching keys in both datasets are included.

4.3 Load

The transformed data is loaded into an SQLite database using SQLAlchemy. The merged dataset, which contains both unemployment and aggregated crime statistics of US states, is saved as a single table named US_crime_unemployment. The database provides a structured format for exploratary data analysis and enables efficient querying of the data.

5 Result and Limitations

The combined dataset can be used to calculate the correlation between unemployment rates and crime rates across different US states. The output data is aggregated at the state and year level, providing a clear picture of regional trends.

5.1 Data Structure and Quality

The final data is in a structured format, with cleaned columns and consistent data types. The dataset is in CSV format, which was chosen for its simplicity and compatibility with various tools. Also, It is interesting to observe the trends and patterns from the data, as they provide valuable insights into the relationship between unemployment rates and crime rates across different US states.

5.2 Limitations

There may be some inherent biases due to missing or incomplete data, particularly in crime statistics. The correlation analysis may not cover all external factors influencing crime rates, such as social policies or economic changes as they are not reflected in datasets.

References

Michael Bayer. Sqlalchemy. In Amy Brown and Greg Wilson, editors, *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*. aosabook.org, 2012. URL `http://aosabook.org/en/sqlalchemy.html`.

The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL `https://doi.org/10.5281/zenodo.3509134`.

Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.