**Data Sources and Relationships (Credit Card Fraud Detection)**

Akshay Sharma

Department of Data Science, Bellevue University

DSC540-T303: Data Preparation

Catherine Williams

May 28, 2025

## **Introduction**

Credit card fraud poses significant challenges to consumers, financial institutions, and merchants, leading to substantial financial losses and undermining trust in payment systems. The dynamic nature of fraudulent activities necessitates advanced data science techniques to detect and prevent unauthorized transactions effectively. This research focuses on detecting and analyzing **patterns in credit card fraud** by integrating transaction-level data, geolocation-based insights, and regional economic/demographic data. The goal is to uncover environmental and behavioral signals associated with fraudulent activity.



## **Data Sources**

**Flat File (CSV)**

* **Description**: A dataset of anonymized credit card transactions labeled as fraudulent or legitimate. Each row includes transaction amount, time, and engineered features.
* **Source**: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>
* **File Format**: CSV
* **Data Sample**: 284,807 transactions, 31 columns
* **Note**: Features are PCA-transformed, so additional location-based context will be joined in from other sources.

**API**

* **Description**: Geolocation API that converts IP addresses or other location info into city/region/country data, which can help link anonymized transaction points with geography.
* **Source**: https://ipapi.co/
* **Use**: Simulate location context for transactions based on anonymized fields or generated IP-like inputs.

**Website (HTML Table)**

* **Description**: U.S. Census Bureau or similar source with economic/demographic info (e.g., median income, education level, urban/rural designation) by zip code or region.
* **Source**:https://en.wikipedia.org/wiki/List\_of\_U.S.\_states\_and\_territories\_by\_median\_household\_income
* **Format**: HTML table
* **Use**: Match regions from the API with income level or demographic data to analyze socioeconomic impact on fraud risk.

## **Relationships Between the Data**

* The **flat file** contains anonymized transaction records with labeled outcomes (Fraud = 0/1), amounts, and timing.
* The **API** data will simulate or enhance regional tagging of transactions using transaction metadata (such as timestamps or IDs) to assign a location.
* The **Wikipedia** demographic data (scraped as a table) provides income-level data by region. This can be linked to the geolocation info from the API.

**Key Relationship Approach:**

* Create a synthetic "Location ID" or mapping field to simulate IP/Region from transaction data.
* Use that to pull API location metadata, which in turn links to state/region info scraped from the website.
* This creates a connected web where:

**Transactions ↔ Location (via API) ↔ Region demographics (via Website)**

## **Final Project Summary & Ethical Reflection**

For this final project, I explored the relationship between credit card transactions and state-level socioeconomic indicators such as median income and income growth. My goal was to examine patterns in fraud occurrence by enriching transactional data with geographic and economic dimensions. I worked with three datasets: anonymized credit card transactions, median household income by state (web-scraped and cleaned), and IP-based location data retrieved from an API. These datasets were loaded into a SQLite database and merged using SQL joins to create a unified analytical view.

In preparing the data, several **transformations** were necessary. The most notable change was the addition of a simulated State column in the credit card dataset to enable joining with state-level income data, since no geographic information was originally present. I also performed type normalization, column renaming, income growth calculation, and standardized inconsistent formatting.

**Legal and regulatory** concerns with this type of project would be significant if the credit card data were not anonymized. Thankfully, the dataset used was publicly available and devoid of personally identifiable information (PII). However, in real-world applications, any geolocation or behavioral profiling must comply with laws such as GDPR, CCPA, and PCI-DSS standards.

The **primary risk** in transforming this data lies in over-interpreting patterns from simulated or assumed joins. For instance, since the State assignments were synthetic, any trends between fraud and income are illustrative, not definitive. Additionally, incorrect data merging or fuzzy matching can lead to biased insights and potential harm if decisions were to be made from such data in practice.

Several **assumptions** were made, particularly that simulated joins could mimic real geographic relationships. Income growth was computed as a simple percentage change over time, which assumes uniform economic influence across states—a simplification that may not hold true in real-world settings.

The **data sources** included government-reported income tables from Wikipedia and an open credit card fraud dataset, both widely cited in academic settings. API data was collected responsibly without scraping or violating terms of use. All datasets were used strictly for educational, non-commercial purposes, aligning with ethical research standards.

To mitigate **ethical risks**, future iterations should avoid simulated geography and instead rely on verifiable geolocation tied to consented user data. Transparency about assumptions and limitations, as done in this report, is also crucial. Finally, ethical AI practices—such as fairness audits and human oversight—would help ensure that any predictive models using such data are used responsibly.