Capstone Project Proposal:

Analyzing Drug Adverse Event Reports from OpenFDA

1. Introduction and Problem Statement

An adverse drug event (ADE) refers to any undesirable symptom or experience that occurs while a patient is taking a medication. Reporting ADEs is critical for improving patient safety, identifying potential risks associated with medications, and informing regulatory decisions.

To support this effort, the U.S. Food and Drug Administration (FDA) maintains the Adverse Event Reporting System (FAERS), a centralized database that collects information on adverse events and medication errors reported to the agency. Adverse event reports can come directly from healthcare professionals—such as physicians, pharmacists, and nurses—as well as from consumers, including patients, family members, and legal representatives. In addition, manufacturers who receive adverse event reports are typically required to submit them to the FDA. Together, these reports contribute to a growing body of data that helps monitor the safety of approved medications.

The problem I aim to address is:

- Can we identify patterns in drug adverse events?
- Which drugs are most frequently associated with serious outcomes (e.g., hospitalization, death)?
- Are certain adverse events more common among specific demographics (age, gender)?
- Can we detect possible underreported or overlooked drug safety signals?

2. Data Source

The openFDA drug adverse event dataset which has been collected from the FDA Adverse Event Reporting System (FAERS), a database that contains information on adverse event and medication error reports submitted to FDA.

Key features of the dataset:

- Information about drugs involved in adverse events
- Patient demographics (age, gender, weight)
- Reported adverse outcomes (hospitalization, death, etc.)
- Report submission date
- Event seriousness (e.g., serious, non-serious)

Dataset documentation: OpenFDA Drug Event API Documentation

3. Data Cleaning and Joining Plan

Since the raw data is nested JSON and spans multiple fields, the expected data preparation tasks include:

- **Data Extraction**:
    - Pull event data via API or download bulk datasets.
    - Flatten nested JSON structures into relational tables.
- **Data Cleaning**:
    - Handle missing values (e.g., missing age or gender).
    - Standardize drug names (e.g., normalize brand names and generic names).
    - Remove duplicate reports (same event reported multiple times).
- **Data Joining**:

    - Link related tables, such as patient demographics with drug information and outcomes.

- **Data Transformation**:
    - Convert dates into proper formats for time series analysis.
    - Categorize continuous values (e.g., age groups).

4. Tools and Technologies

I plan to use the following tools:

- **Data Extraction**:
    - Python (requests library for API interaction)
- **Data Processing and Cleaning**:
    - Python (pandas, json libraries)
    - SQL (for structuring and querying the cleaned dataset)
- **Data Storage**:
- **Data Analysis and Visualization**:
- **Optional (for advanced exploration)**:

5. Project Goals

The goals of the project are:

- Build a clean, queryable relational database from the OpenFDA drug adverse events dataset.
- Create analytical queries and reports to:
    - Identify drugs most associated with serious adverse outcomes.
    - Analyze demographic trends (e.g., age, gender) in adverse events.

    - Visualize the frequency of reports over time to detect possible spikes (signals).
    - Group and classify adverse events by drug class or type.
- Discuss possible biases or limitations in the dataset (e.g., voluntary reporting bias).

Conclusion

This project will highlight my ability to perform key data engineering tasks such as extracting, cleaning, transforming, modeling, and analyzing data using a publicly available healthcare dataset.
It also provides an opportunity to apply technical skills to a meaningful public health challenge, demonstrating both engineering expertise and analytical problem-solving.