Analyzing the Consumption and Health Risk Factors of Alcohol in the United States

1 Introduction

1.1 Question addressed in the Project

To investigate the relationship between alcohol consumption and its health-related impacts in the United States, this project focuses on two key questions:

"What are the patterns of alcohol consumption across states, and how do these patterns vary geographically?" To address this question, the report will explore critical metrics such as per capita ethanol consumption, excessive drinking rates, and state-level differences in alcohol use behaviors. Insights will include identifying states with the highest alcohol consumption per capita and those with the most prevalent excessive drinking behaviors.

"How does alcohol consumption correlate with health and safety outcomes in the United States?" This question will examine the association between excessive drinking rates and health-related risks, such as the percentage of driving fatalities involving alcohol. The analysis aims to reveal trends that link high consumption patterns to increased risks, providing a foundation for understanding broader public health implications.

2 Dataset

Two datasets were selected to support this research: the "Alcohol Consumption by State 2024" dataset from Kaggle and the "U.S. Chronic Disease Indicators (CDI)" dataset from Data.gov. These datasets provide comprehensive information on state-level alcohol consumption patterns and related health indicators. Upon review, both datasets were deemed suitable for analysis as they include meaningful metrics such as per capita ethanol consumption, excessive drinking rates, and alcohol-related driving fatalities. They are provided in CSV format, which ensures compatibility with data integration and analysis tools. Python was chosen for processing and analysis due to its rich ecosystem of libraries, including support for accessing the Kaggle API, handling CSV files, and performing data transformations seamlessly within an ETL workflow.

1. **Alcohol Consumption by State 2024:**
   The dataset is available to the audience through the following source:

   - Metadata URL: https://www.kaggle.com/datasets/annafabris/alcohol-consumption-by-state-2024
   - Data URL: https://www.kaggle.com/datasets/annafabris/alcohol-consumption-by-state-2024
   - Data Type: CSV
   - License: MIT

   The dataset consists of the Alcohol Consumption and driving fatalities involving alcohol by State in 2024. The dataset consists of the 5 columns as given below :
   State Name: names of the states of the USA.
   State Abbreviations (USPS): the United States Postal Service (USPS) abbreviations for each state. These two-letter codes are commonly used to represent states in various contexts.
   Gallons of Ethanol per Capita (Gallons Consumed): amount of ethanol consumed per capita in each state, measured in gallons.
   Driving Fatalities Involving Alcohol (Percentage): the percentage of driving fatalities in each state that involve alcohol.
   Excessive Drinking Rate (Percentage): the percentage of the population engaging in excessive drinking behaviors in each state. Excessive alcohol consumption includes heavy drinking and binge drinking. Heavy drinking is eight or more drinks per week for women and 15 or more drinks per week for men. Binge drinking is four or more drinks during a single occasion for women and five or more for men.

2. **U.S. Chronic Disease Indicators (CDI):** This dataset is based on indicators described in MMWR "Indicators for Chronic Disease Surveillance — United States, 2013" https://www.cdc.gov/mmwr/preview/mmwrhtml/rr6401a1.htm before 2024 CDI refresh. It provided cross-cutting set of 124 indicators that were developed by consensus and that allows states and territories to uniformly define, collect, and report chronic disease data that are important to public health practice and available for states, and territories. In addition to providing access to state-specific indicator data, the CDI web site www.cdc.gov/cdi provides current release and additional information and data resources.

   - Metadata URL: https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi
   - Data URL: https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi
   - Data Type: CSV
   - License: Open Database License (ODbL)

Data Pipeline

The data pipeline for this project integrates two datasets to analyze alcohol consumption and its health-related impacts across the United States. The pipeline follows the Extract, Transform, Load (ETL) framework, implemented using Python. Data is first extracted from Kaggle and Data.gov sources. Then, specific cleaning and transformation steps are applied to standardize and integrate.

Technologies Used

- **Programming Language:** Python
- **Libraries:** Pandas (for data manipulation), NumPy (for data cleaning), SQLite (for database storage), Matplotlib, Seaborn, plotly (for visualizations)
- **Data Storage:** SQLite database
- **Data Formats:** CSV for input, SQLite for intermediate and structured output.

Transformation and Cleaning Steps

- **Standardization:** Unified column names across datasets.
- **Missing Value Handling:** Imputed missing values in numerical fields using mean and median methods.
- **Filtering:** Focused on state-specific data, excluding irrelevant indicators.

Problems Encountered and Solutions

- **Inconsistent state names between the datasets:** Used a mapping function to align state names and abbreviations across both datasets.

Meta-Quality Measures

- Validity checks for column data types
- Logging errors and anomalies during the transformation process.
- Automated schema validation to handle structural changes in the source data.

Result and Limitation

For this project, I utilized a SQLite database for its lightweight nature and minimal configuration requirements. The ETL pipeline processed the data into a SQLite database named database.sqlite, containing integrated and cleaned tables. These tables include state-level metrics such as per capita alcohol consumption, excessive drinking rates, and alcohol-related driving fatalities. The data is accessible for querying through SQL commands, and exploratory visualizations have been created to highlight trends and insights.

The pipeline produced several outputs:

1. Choropleth maps displaying per capita alcohol consumption by state, visually highlighting geographic patterns.
2. Bar plots comparing key metrics such as alcohol consumption and health impacts (e.g., renal failure) across states over time.
3. Correlation visualizations showing the relationship between excessive drinking and health-related metrics.

The dataset is structured in a relational format. Compared to unstructured or semi-structured formats, the structured nature ensures consistency and accuracy. The data reflects real-world metrics (e.g., state-reported health statistics and alcohol consumption data), ensuring a high degree of reliability. Unified column names, data types, and standardized state abbreviations ensured consistency across datasets. The dataset aligns with the research objectives by focusing on alcohol consumption and its health-related effects. The final output is stored in SQLite format for its advantages in querying and managing structured data. Additionally, CSV exports are provided for interoperability with visualization and reporting tools.

The data pipeline successfully integrates and processes the datasets into a usable format. However, future work should address data granularity by incorporating more localized or demographic-level data. Additionally, expanding the temporal range could provide deeper insights into trends. Maintaining completeness and representativeness is crucial for reducing biases and improving the reliability of findings.