**Data Engineering Project Write-Up**

**Abstract**

The goal of this project was to engineer a data storage and processing pipeline for an interactive web application visualizing the rates and trends of hypertension-related cardiovascular disease (CVD) mortality in the US from 2000 to 2019. The dataset was provided by the Centers for Disease Control and Prevention (CDC) with over 1 million rows of data with information on the counties, states, gender, race and ethnicity, and age groups, and the number of deaths per 100,000 residents.

**Design**

The design of the web app on Streamlit is to highlight the rates and trends of CVD mortality in the US. Some key takeaways from the dataset show that CVD mortality is steadily increasing throughout the years, affecting vastly more states in the Midwest and south. Surprisingly, it also shows that the rate of CVD deaths between genders has a narrower gap than one would expect.

**Data**

The dataset was provided by the CDC and consisted of over 1 million rows of data on one table. Each row contained information on the year (from 2000 to 2019), location (county, state), gender, race and ethnicity, age groups, data values (number of deaths per 100,000), confidence levels, and latitude and longitude coordinates. The dataset was downloaded as a CSV file and stored in SQLite.

**Algorithms**

Preprocessing: Column names had to be cleaned up for readability, and numeric datatypes converted appropriately.

Visualization: Web application displays pie charts that group the average number of cases by gender, age groups, and race and ethnicity; bar graphs that group cases by states with an animation bar that automatically runs through the values in the years from 2000 to 2019; and a chloropeth map that shows the most current data values of the dataset by county in the US.

**Tools**

* SQLite and SQLAlchemy for data storage and extraction
* Python, pandas, Pandasql for data processing
* Matplotlib, Plotly for Visualizations
* Streamlit, Github for app deployment
* Canva for presentation slides

**Communication**

Compressed file of the full dataset, CSV files of queried results, pipeline scripts and parts, and project slides are included in the repo. Web app can be accessed at streamlit here: https://snazzys1026-data-pipeline-cv-mortality-app-jrv7g0.streamlitapp.com/