Zeb Moffat
Individual Project
10/14/2025

INFO 511: Final Project Milestone 2 Reporting Update

Data Acquisition

The data used in this project is from the U.S. Chronic Disease Indicators dataset from the U.S. Department of Health & Human Services. It is publicly available from the data.gov website (https://catalog.data.gov/dataset/u-s-chronic-disease-indicators). The dataset is under the Open Data Commons Open Database License (ODbL) v1.0, which allows use, modification, and distribution of the data for research purposes with appropriate attribution. The dataset contains 115 chronic disease indicators collected across U.S. states and territories from 2015-2022. For this project, the data has been filtered down to focus exclusively on cancer-related indicators.

EDA

The original dataset contained multiple measurement types for each chronic disease indicator. The cancer data had three measurement types: raw count, crude rates, and age-adjusted rates. Age-adjusted rates was selected because it standardizes for population age distributions, which makes it the best way to compare across demographic groups at different ages. There are 5480 records for age-adjusted cancer data.

Issues

There were three major issues with the age-adjusted cancer data.

- **Suppressed and missing values** - Some of the records contained flags in the DataValueFootnoteSymbol column (e.g. ~, #, *). These flags represent suppressed data due to small sample sizes, data quality issues, or privacy issues. Any records with these flags were dropped to ensure reliability in the data. Also, any records with missing DataValue entries were dropped.
- **Stratification Structure** - The dataset uses a single "Stratification1" column that contains mixed demographic categories (sex, race/ethnicity, and Overall aggregates). Each record has just one of these dimensions, so records cannot contain a combination of sex and race/ethnicity. Due to this, the cancer data will be split into two smaller datasets, sex (1118 records) and race/ethnicity data (2702). These two new datasets will be used to train two separate models for predictions and answer the research question, since they cannot be reasonably used with each other.
- **Geographic Coverage** - The dataset includes geographic data from all 50 states, Washington D.C., Puerto Rico, and aggregated U.S. data. Aggregated U.S. data has been dropped as it doesn't allow for comparison between locations. Puerto Rican data has also been dropped because it only had 6 records, which is not nearly enough for data analysis.

Final Analysis

After filtering and cleaning the original data, there are 3820 records left over which have been split into two smaller datasets.

- **Sex Dataset** - 1118 records covering Males and Females for 7 cancer indicators

- **Race/Ethnicity Dataset** - 2702 records covering White non-Hispanic, Black non-Hispanic, Hispanic, Asian or Pacific Islander non-Hispanic, and American Indian or Alaska Native non-Hispanic populations across 50 states and 7 cancer indicators

Future Analysis

Next steps involve training two separate linear regression models to examine demographic differences in cancer rates. Model 1 will examine sex differences in cancer rates, while Model 2 will examine race/ethnicity differences in cancer rates. This two-model approach is necessary because the original data is stratified only once by sex or by race/ethnicity, not simultaneously. By training these two models, the research question “Do cancer incidence rates differ significantly between demographic groups (sex, race/ethnicity) across US states?” will hopefully be answered with meaningful significance.