Mortality rates for people in the US

Understanding trends in mortality across the United States is a useful exercise because it can provide clues about the ways we can improve people's lives. This project is useful for any government agencies and nonprofit organizations that want to understand factors affecting mortality rates.

The **main goal** for the project is to understand **what socioeconomic and environmental factors affect the mortality rate of people in the US.**

Data collection

We collected data from different governmental sites. We used the Center for Disease Control and Prevention (CDC) to obtain the **number of deaths** reported for each US county during 2002 to 2016. This data also included demographic information including race, sex, age group and Hispanic origin. The United States Census Bureau provided **yearly poverty estimates** for each county during the 15-year interval. The Environmental Protection Agency (EPA) provided several **air quality markers** such as ozone, particulate matter and lead for several counties in the same time interval. Due to privacy concerns, the CDC data suppresses population and deaths numbers when there are fewer than 10 people, so the number of deaths should be treated as an estimate.

Modeling approach

The data was divided into three 5-year intervals: 2002-2006, 2007-2011 and 2012-2016, to better understand any temporal trends. **We created a major dataset (401389 rows, 39 columns)** to store all the data collected from all three governmental agencies. We calculated death rates, poverty rates and air quality safety index, and plotted them in a **US choropleth map** to help illustrate geospatial trends using plotly.express.

Furthermore, we used scattering and log scale plots, histograms and corner plots to investigate the possible trends and correlations associated with geospatial locations (US regions), population demographics (age, race, ethnicity, gender) and environmental factors (e.g. particulate matter). **We built linear regression models to best fit** the death rate vs. poverty rate for each one of the factors aforementioned.

Conclusions and Future directions

Our findings show that there's a **nationwide trend of poverty reduction in the last 15 years in the US, as well as an air quality pollution reduction for counties that reported**. This is a positive result for current and future generations, and it can be an indication that self-reporting pollution numbers can help counties remain accountable and reduce numbers.

The analysis on the US regions show that the Midwest has a steeper correlation than any other region between death rate and poverty rate. The Northeast shows the lowest poverty and death rates. Analysis on age group showed that infants (less than 1 year old) have unusually high death rates, while older people have the highest death rates. Deaths of younger people looked to be the most affected by poverty. Analysis on environmental factors show little to no correlation with death rates given the limited data.

In summary, **the factors that affect death rates the most are: age, county population, and poverty.** Thus, the most impactful action item to reduce the mortality rate is to reduce poverty rates nationwide. The next steps for this project are: 1. exploit the geographical connections between counties, 2. use predictive models to estimate missing values for counties in the data, and 3. include the next time interval (2017-present) which would include information of the COVID-19 pandemic and its effects on the factors considered.