Economic Well-Being Project Writeup

Anna Le

This is my personal project and I was responsible for all aspects. I processed the database, wrote and organized the code.

I. Project Goals:

The primary objective of this project was to explore the relationships between economic well-being, demographic factors and life expectancy. Specifically, I aimed to address the following:

1. **Understand Income Trends Over Time:** I wanted to examine how per capita income has changed over time and across geographic regions in the United States. This included analyzing growth rates and identifying potential disparities between states or demographic groups.
2. **Investigate Correlations Between Income and Population Demographics:** By analyzing population data segmented by age groups, I sought to understand how demographic changes (e.g., shifts in age distribution) are associated with changes in economic metrics such as per capita income.
3. **Assess the Relationship Between Income and Life Expectancy:** A key focus was to evaluate whether economic well-being, represented by household income, correlates with improvements in race-adjusted and unadjusted life expectancy. This was done to explore the extent to which economic factors impact public health outcomes.
4. **Identify Gender Disparities:** I aimed to uncover any differences in the relationships between income and life expectancy when stratified by gender. This was to better understand how economic benefits and health outcomes might vary between males and females.

I hope to gain a deeper understanding of how economic and demographic factors interact and influence health outcomes. This knowledge could be a foundation for policymaking or further research into reducing economic and health inequalities.

II. Data Description:

1. **Population by Age and Sex Dataset (from StatsAmerica):**
   This dataset provides population estimates by age and sex for U.S. states and counties from 2000 to 2019. It includes detailed breakdowns of the population across various age groups as well as male and female population counts. This dataset was instrumental in analyzing demographic trends and their relationship with economic metrics.

   For certain sections, I preprocess the data by focusing on relevant columns such as age groups, total population and gender-specific populations. I also filter the dataset to include only rows that represent the U.S. as a whole and ensure that state-level data is distinguished from county-level data to prevent overrepresentation. Furthermore, I merge income data with population data using geographic identifiers and years to create a comprehensive dataset.

2. **BEA Per Capita Personal Income Dataset (from StatsAmerica):**
   This dataset details annual estimates of per capita personal income for U.S. states and counties from 2001 to 2023. It served as the primary metric for assessing economic well-being over time and across geographic regions.

I preprocess the data by merging income data with population data using geographic identifiers and years. I then calculate the growth rates of per capita income to evaluate changes over time.

3. **Life Expectancy Estimates Dataset (from Opportunity Insights):**
   This dataset represents life expectancy estimates for men and women at age 40, reported for each percentile of the national income distribution and separated by year. It includes both race-adjusted and unadjusted life expectancy measures, along with associated standard errors.

   For the preprocessing steps, I renamed columns for clarity. Moreover, I filtered and grouped data by demographic attributes such as gender and income percentiles to facilitate targeted analysis. For the regression model, I also encode the "Gender" variable to numeric to be able to perform regression analysis and predictive modeling.

4. **Data Task - Data Wrangling: Joining Demographics and Income datasets**
   To analyze income by demographics, I merged the Population by Age and Sex dataset with the BEA Per Capita Personal Income dataset. The datasets were joined on the common columns Statefips, Countyfips and Year. Before merging, I ensured that these key columns had consistent data types in both datasets to avoid mismatches. An inner join was performed using *pandas.merge()* to align demographic and income data at the state and county levels across the specified time periods. After inspecting the merged dataset, I verified and handled duplicated columns by standardizing names.

III. Analysis:

1. **Analyze Income Differences Across Gender, Age and Geographic Regions**
   (Methods: Visualization and Geocoding)

   a. **Income by Gender - *Visualization*:** I calculated and visualized average income for males and females over time (2001–2019) using a line chart. Results revealed that female income consistently exceeded male income with both genders experiencing steady growth and a relatively stable income gap over the years.
   b. **Income by Age Groups - *Visualization*:** I grouped data by year and calculated income for various age groups, visualizing trends with a line chart. Middle-aged groups (45–64 and 25–44) contributed the most to income while older groups (65+) showed a notable rise. Unexpectedly, the 5–17 age group had higher income levels than the 18–24 group which suggests an area for further exploration.
   c. **Income by Geographic Region (based on State) - *Visualization & Geocoding***: I created an interactive bar plot and a scatter mapbox plot using geocoded state locations, sized and colored by income. Findings highlighted the Northeast as having the highest income. This trend could be due to the Northeast states having the highest education levels in the country. Further exploration could investigate the relationship between education and income levels for deeper insights.

2. **Growth Rate of Income Over Time and Its Correlation with Population Demographics**
   (Methods: Statistical Summarization, Statistical Associations and Visualization)

Income growth rates show a generally positive trend over time with occasional disruptions that may reflect major economic events or demographic shifts. Correlation analysis reveals that middle-aged populations, as primary earners, are positively associated with income growth, while regions with younger populations experience slower growth. Visualizations including a line chart of income growth over time and a correlation heatmap provide a clearer understanding of these patterns and their demographic associations.

3. **Analyze the Correlation between Income and Life Expectancy**
   (Methods: Statistical Associations and Visualization)
   a. **Relationship between Income and Unadjusted Life Expectancy / Race-Adjusted Life Expectancy:** Using Pearson's r and scatterplots, the analysis reveals a statistically significant but weak positive correlation (~0.3) between income and both unadjusted and race-adjusted life expectancy. Scatterplots show most observations with low income and life expectancy clustered in the lower-left corner while higher income levels exhibit greater variability in life expectancy.
   b. **Group by Income Percentile and Compare Life Expectancy:** A line chart comparing life expectancy across income percentiles shows both unadjusted and race-adjusted life expectancy increasing with higher income percentiles. There is a steep increase in lower-income groups which indicates more pronounced health improvements for these populations. Race-adjusted and unadjusted measures show minimal differences.

   While higher income is associated with longer life expectancy, other factors such as access to healthcare, lifestyle and education may also play significant roles. Further exploration into these variables could provide a more comprehensive understanding of the complex relationship between income and life expectancy.

4. **Analyze the Correlation between Income and Race-Adjusted Life Expectancy by Gender**
   (Methods: Statistical Associations and Visualization)
   A boxplot reveals that females generally have a slightly higher median life expectancy than males. Using Pearson's r and a scatterplot, gender was found to strengthen the correlation between income and race-adjusted life expectancy, with correlation coefficients of 0.44 for males and 0.42 for females, both statistically significant. This suggests that higher income is associated with increased life expectancy for both genders. However, females show more consistent and higher life expectancy, warranting further investigation into biological, social and behavioral factors.

5. **Predicting Race-Adjusted Life Expectancy based on Gender and Income**
   (Methods: Statistical Associations, Visualization and Predictive modeling using scikit-learn)
   The correlation between race-adjusted life expectancy, gender, mean household income and income percentile was analyzed using Pearson's r and ordinary least squares regression. A heatmap was generated to visualize the correlation matrix. For predictive modeling, a linear regression model was built using scikit-learn to predict race-adjusted life expectancy based on gender and income-related variables. The scatter plot of true and predicted values showed a strong positive correlation with the points closely following the perfect prediction line.The model

shows good predictive power with high accuracy but its simplicity limits its potential. Future work could explore more complex models to capture non-linear relationships and improve prediction accuracy.

IV. Challenges Encountered and How They Were Addressed:

Throughout the analysis, several challenges arose, particularly with data integration and consistency. One significant challenge was merging the demographics dataset with the income dataset due to duplicated columns and inconsistencies in geographic naming conventions. For example, “Bedford + Bedford City, VA” and “Bedford, VA” represented the same area. To resolve this, I standardized geographic names to ensure consistency by replacing mismatched entries with a consistent identifier such as “Bedford, VA.”

Another challenge was extracting meaningful trends at the national level as the dataset was duplicated for each state and county. Initially, I was unsure how to focus on national-level trends without redundant data. Upon closer inspection, I realized the need to filter the dataset to include only rows that represented the U.S. as a whole, removing duplicates caused by state- or county-level entries.

V. Conclusion and Further Exploration:

Through the analysis tasks, I uncovered several key insights on socio-economic patterns in income and life expectancy. Female income consistently exceeded male income, with both genders experiencing steady growth and maintaining a stable income gap over time. Middle-aged groups (45–64 and 25–44) were the largest contributors to income and the Northeast emerged as the region with the highest income levels. An interesting observation was the positive correlation between higher income and longer life expectancy. Females generally had a slightly higher and more consistent median life expectancy than males. Additionally, gender appeared to strengthen the relationship between income and race-adjusted life expectancy.

The analysis raised several intriguing questions and opportunities for further investigation. For example, the higher income levels in the 5–17 age group compared to the 18–24 group suggest further investigation into factors like family income or child labor laws. The Northeast’s higher income levels could be linked to its high education levels. Future analysis could investigate the relationship between education and income to validate this hypothesis. Moreover, the consistently higher and more stable life expectancy of females calls for deeper exploration into biological, social and behavioral factors that may contribute to this disparity.

Lastly, there is potential for applying predictive modeling to better understand the complex relationships between socio-economic factors and life expectancy. Exploring more advanced machine learning models could reveal nuanced insights that policymakers and public health officials could use to design targeted interventions. These models could help identify at-risk populations, optimize resource allocation and even tailor health programs to communities that need them the most.