**By Nurmaa Dashzeveg**

**Hypothesis**

Income varies based on job types and perhaps legal wills or inheritance. Income can indeed differ widely depending on the type of job, the industry, level of experience, location, and many other factors. Health expenses can vary depending on the healthcare provider, location, type of care, and the healthcare system in each state in United States. In the current project, we aim to see if health care expenses correlate with income level of people.

**Data collection**

To answer our hypothesis, I analyzed US\_POPULATION20.CSV for population by states throughout years 2010-2017, US\_PER\_CAPITA20.CSV for spendings by states throughout years and Income.csv for household income by states collected from United States Census Bureau (refer to attached data\_inventory.xlsx).

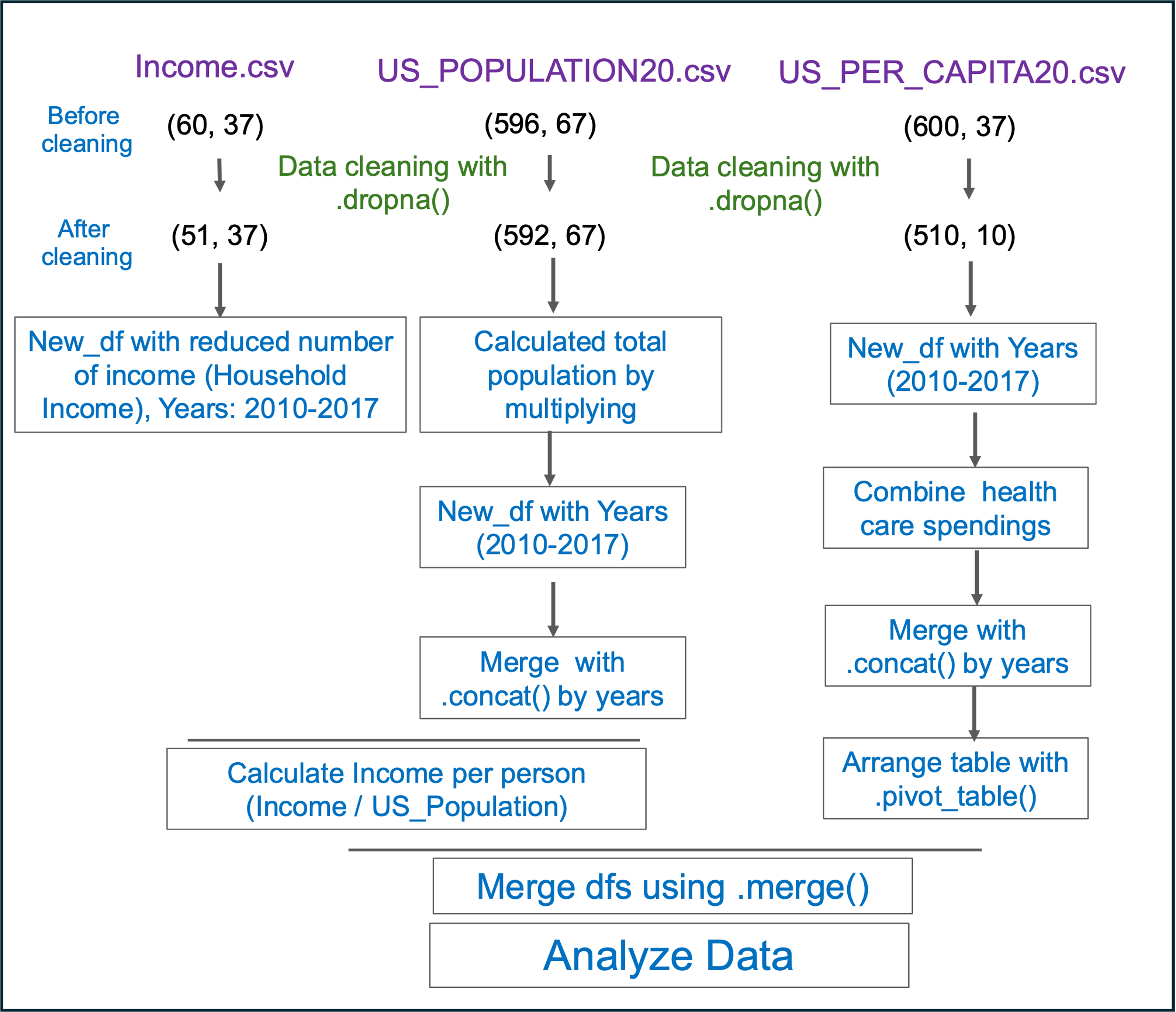


Figure 1. Strategy of data analysis

Figure 1. Three different csv files (in purple) used for the analysis. First data cleaned with .dropna() (in green). After cleaning, data were processed as indicated in steps in square box (in blue).

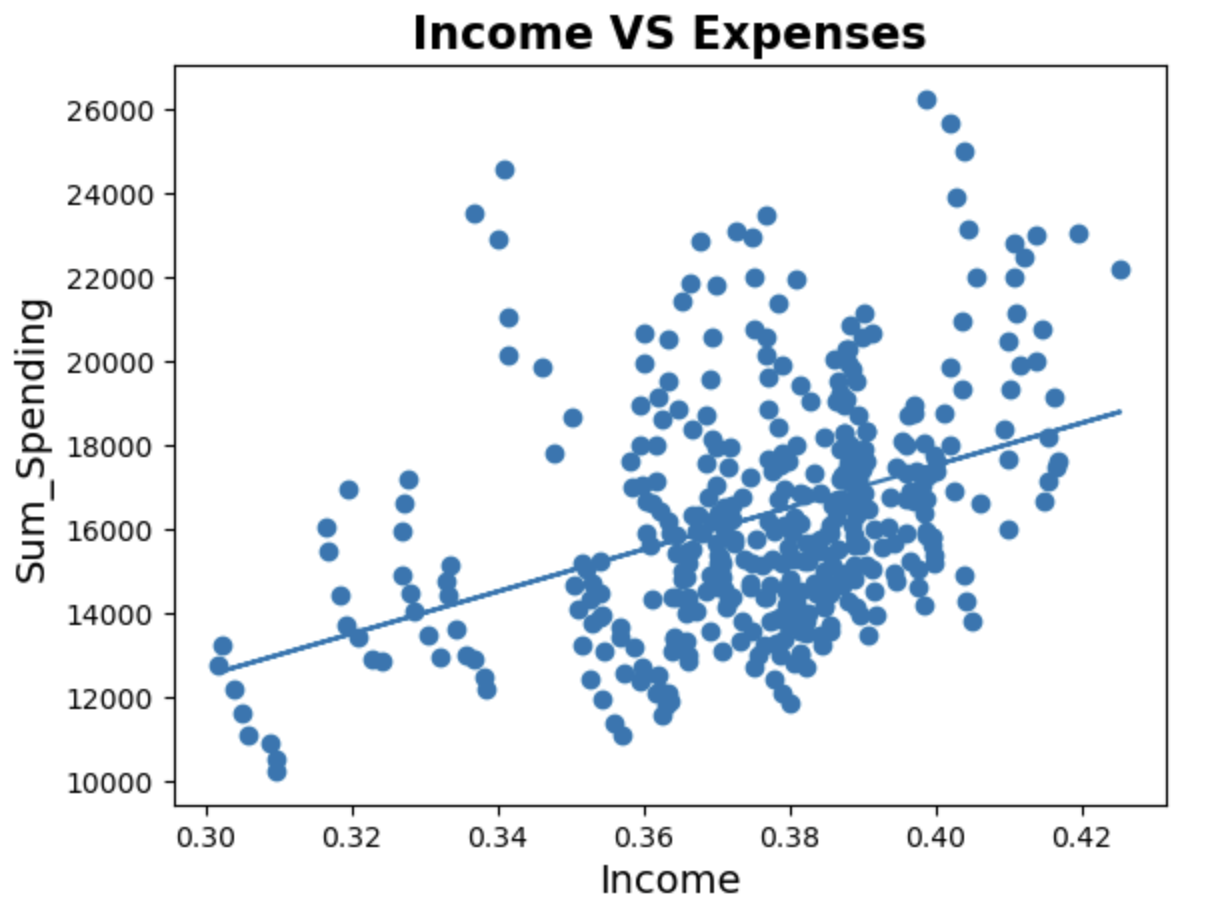
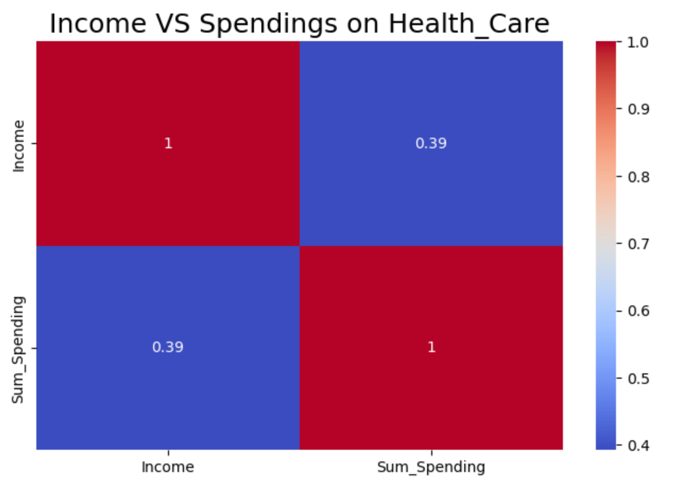
**Strategy to analyze the data**

First, all data were cleaned using .dropna() (Fig.1). After cleaning, data showed equal number of contents when checking with .count, and reduced number of columns and rows when use .shape. Later, useful columns for further analysis were selected and saved in new dataframes. In US\_POPULATION.csv, population were multiplied by thousand to create total population and the years in Year-columns were shifted to the rows by using looping and .concat(). Income per person in different states were calculated by dividing Income.csv by number of total populations. Next, using looping and .concat(), columns\_Year is shifted to the rows and using .pivot\_table() contents in the Items column were listed as columns to analyse US\_Per\_Capita20.csv to create same arrangement of the table as in Income.csv. After that, the values of the spendings in different health care were combined to estimate total expenses of health care. Finally, all data were merged by utilizing .merge() for the further analysis.

**Results**

Figure 2. Correlation between Income and health care spendings

Figure 2. Total mount of income estimates from US Census Bureau (Income) were compared with total health care expenses and shown in heat map (A) and scatter plot (B, R=0.3921, p=1.88e-16



A

B

In comparing US income throughout the US States and Years (2010-2017), the correlation of Income with spendings in health care showed mild positive correlation (R=0.39, P=1.188e-16) (Fig.2) indicating wealthy population spends more in health care. According to the analysis the average income of the states was almost the same. States with highest income was District of Columbia, Maine, and Vermont, however, the states with lowest income were Utah, Hawai and California (Fig. 3).

**Alternative potential strategy:** In the current analysis, I combined all spendings together and compared with income level. It might be better to categorize spendings into health serious illness-related and healthcare services for pleasure will estimate better correlation with income.

**What I learnt:**

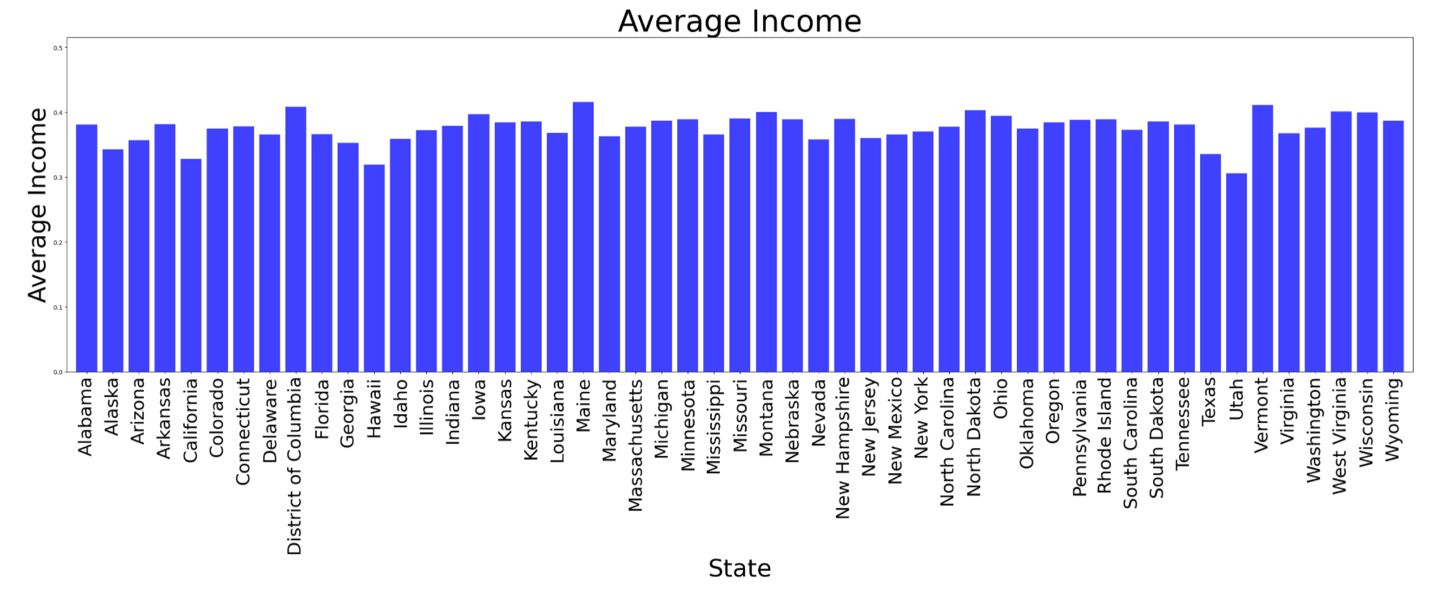


Figure 3. Average Income by States

Figure 3. Average Income were calculated with the mean of values in years (2010-2017).

1. Understanding the data and understanding the questions to be answered is the most important.

2. Exploring in data deeply and make a detailed strategy for the analysis and collect the information (codes) that is useful for this analysis to answer the specific questions are also required. Data are broad that is enough to create numerous of output, focusing on the specific question will save time and reproducibility.

3. Checking data quality and the resources of data are essential to be considered before starting analysis.

**What should be done:** (1) In the current analysis, I have combined different expenses of health cares to compare Income level and mild correlation of expenses with Income. If expenses were categorized, it might lead better outcome. (2) The income values of per person was less than one which was less than expenses on health care. The original value in Income might be in thousands, however it was not indicated in the resource. It traced back in the future.

**Conclusion**

"Analyses of income levels versus healthcare expenses provide a mild positive correlation, indicating that people spend, and purchase healthcare products based on their income level.

**Discussion**

(1) People with higher incomes tend to invest more in preventive care and healthier lifestyle choices (e.g., gym memberships, organic foods, wellness programs), which can result in higher overall healthcare spending. These expenditures reflect the ability to prioritize long-term health.

(2) Although higher-income people can afford routine healthcare and prevention, lower-income people tend to spend a higher percentage of their income on healthcare in emergency situations. But the overall spending remains lower due to limited access to services. (3) Higher-income groups tend to purchase more expensive, high-quality healthcare products—such as prescription drugs, supplements, medical devices, or advanced treatments—which increase total healthcare spending.

(4) Wealthier individuals are more likely to have **comprehensive health insurance** that covers a wider range of medical services, often allowing them to seek specialized care or advanced treatments, which they can afford either through insurance premiums or out-of-pocket spending.

(5) The results of Average Income analysis of States showed California and Utah as a lowest income states, however, these states are known to have high income earnings indicating the variation of the income is high in those states.