Healthcare Spending and Income in the United States.

Marcus Jundt & Mitchell Stapelman

Summary of questions and results:

Research Questions.

> **Question 1**: Which race is spending the highest percentage of income on healthcare?
>
> Results: The Black population spends the highest percentage of their income on healthcare while Asian, Native Hawaiian and Pacific Islander spend the least.
>
> **Question 2**: Which age range is spending the highest percentage of income on healthcare?
>
> Results: The 75+ age range is by far spending the highest percentage of income on healthcare while ages 35 to 44 spend the least.
>
> **Question 3**: What does healthcare spending look like in the future?
>
> Results: The percentage of income spent on healthcare in the future is uncertain. Looking at different ranges of years may yield different results. However, it is clear over the past couple decades we are spending an increasing percentage of income on healthcare per year.

Motivation:

Our proposed questions are significant in answering racial disparities between healthcare coverage and costs as well as finding a clear measurement for the cost of healthcare in the US based on household income. We find this important because having a proper understanding of health and financial data for minority and marginalized populations provides insight into the root causes behind structural racism, poverty and differences in opportunity for all people in the US.

Plotting the spending disparities between different age groups shows how impactful healthcare spending is proportional to the income of that age range. This would show how rising healthcare costs and needs increase the cost burden on the households affected.

Finding the current and future projections for health spending will assist in measuring the impact that health spending has on household income and the potential results that an increase or decrease in health spending may have on families and across different populations in the US.

These results can be used to draw conclusions on impact of spending, future health coverage and also income trends for different US populations.

Dataset:

1. United States Healthcare Spending by Race and Ethnicity 2002-2016 | GHDx

2. Historical Income Tables: People (Table P-1, P-10)

Our first dataset comes from the Global Health Data Exchange, it contains estimates on total health care spending for six major race/ethnicity groups and across types of care, sex, age, health conditions for 2002-2016.

Our second dataset we use comes from the US Census. It contains US census survey data for the current US population and the per capita income across different races and age groups.

Method:

Question 1: Which race is spending the highest percentage of income on healthcare?

- Filter the health spending dataset to show spending per capita for each race (Asian, Native Hawaiian, Pacific Islander, Black, Hispanic, White) in each year (2002-2016), ignoring irrelevant factors for plotting like sex, cause, type of care, and age range
- Create individual dataframes for each race
- Join each health spending dataframe with the associated Census per capita income dataset (Table P-1, each race is separated into different datasets)
    - We will use the 'combined' race Census data rather than 'alone' where appropriate
    - Columns will be renamed for clarity in plots and data processing
    - Census data has additional years outside of the 2002-2016 range which will be ignored for this question
- Compute the mean percentage of income spent on healthcare using per capita health spending and per capita income for each race/ethnicity.
    - Store each calculation in a new column of the dataframe

- Create a bar, line and scatter plots with *plotly* to show distribution for mean percentage of income spent on health for different race groups.

Question 2: Which age range is spending the highest percentage of income on healthcare?

- Filter the health spending dataset to show spending per capita for each age group (25-34, 35-44, 45-54, 55-64, 65-74, 75+) in each year (2002-2016), ignoring unimportant factors like sex, cause, type of care, and race
    - The dataset is separated into 5 year age ranges meaning we will average across each pair of ranges to get the desired 10 year range
- Preprocess the Census data by creating an additional column with age range and separating the two different median and mean income dollar amounts
- Join the health spending dataframe with the Census per capita income dataset
    - Census data has additional years outside of the 2002-2016 range, median income, and population number which will be ignored
- Compute the mean percentage of income spent on healthcare using per capita health spending and per capita income for each age group.
    - Store each calculation in a new column of the dataframe
- Created bar, line and scatter plots with *plotly* to show distribution for mean percentage of income spent on health for different age groups.

Question 3: What does healthcare spending look like in the future?

- Using *scikit-learn*, fit a linear regression model to the age and race percentage of income spent on health that we calculated in Questions 2 and 3
    - We will have a separate model for each subgroup within age and race
    - To analyze each model we will test and calculate its mean square error on parts of the data set (excluded from training set)
- Create new scatter plots with *plotly* with projected future percentage of income spent on health

Results:

Question 1: The Black population has the highest percentage of income spent on healthcare with 28.6% in 2016. The Asian, Native Hawaiian, and Pacific Islander population has the lowest percentage of income spent with 12.2% in 2016. The Hispanic and White populations are between them; both with 22.9% in 2016 (**Fig. 1**).

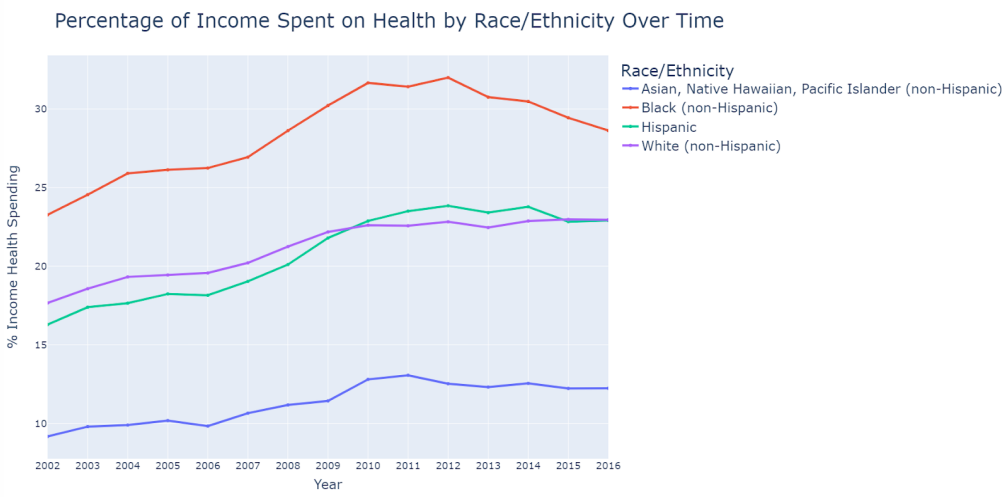

Fig. 1

The Black population spent an average of $7,900 per capita on health care per year; around $2,900 less than the White population. However, the Black population spent 6% more of their income on average on healthcare due to the much lower per capita income (at $27,500 in 2016 vs. $46,900 for the White population, **Fig. 2**). Despite having less per capita income to use on health, the Black population still spends a very high percentage of their income on health.

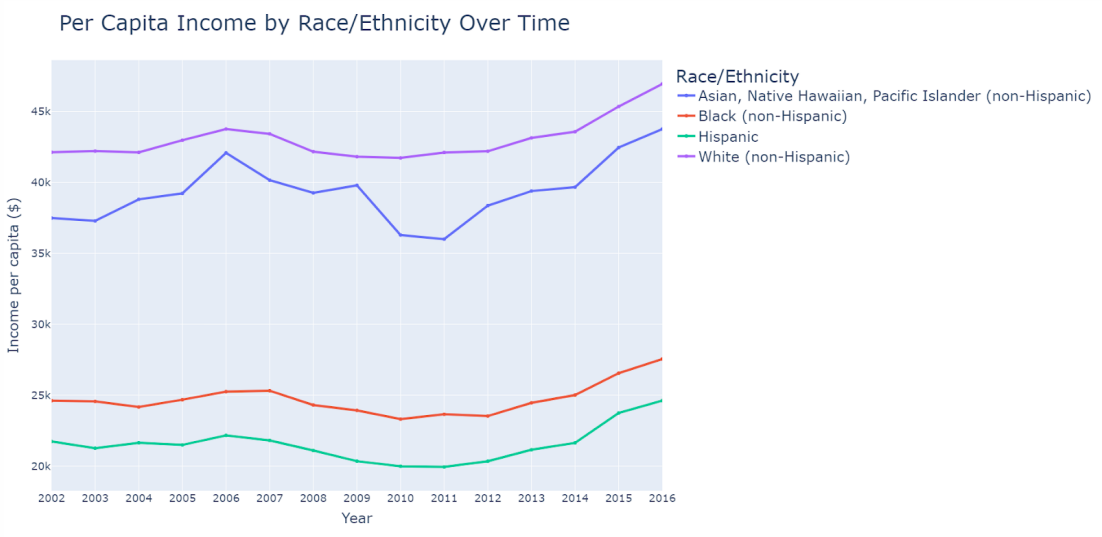

Fig. 2

Another interesting result is the similarity between the Hispanic and White population's percentage of income spent on healthcare. Despite their similarity, the Hispanic population has a lower per capita income ($24,600 in 2016) and lower health spending per capita ($5,640 in 2016). Unfortunately, this data does not tell us the impact this percentage has on each respective population. It is possible the 22.9% of income spent on healthcare is much more impactful for the Hispanic population compared to the White population depending on geographical and environmental factors like proximity to health facilities, food deserts and other forms of structural violence that may adversely increase health risks and cost of health services within a region.

The Asian population has the lowest percentage of income spent on healthcare at 12.2% (in 2016). The population is unlike the others with its high per capita income and low per capita health spending. The low per capita health spending is particularly strange. Are there major underlying barriers of access to healthcare that are not sufficiently addressed? Is the Asian population less likely to develop health problems? Or are they less likely to pay for traditional health care options?

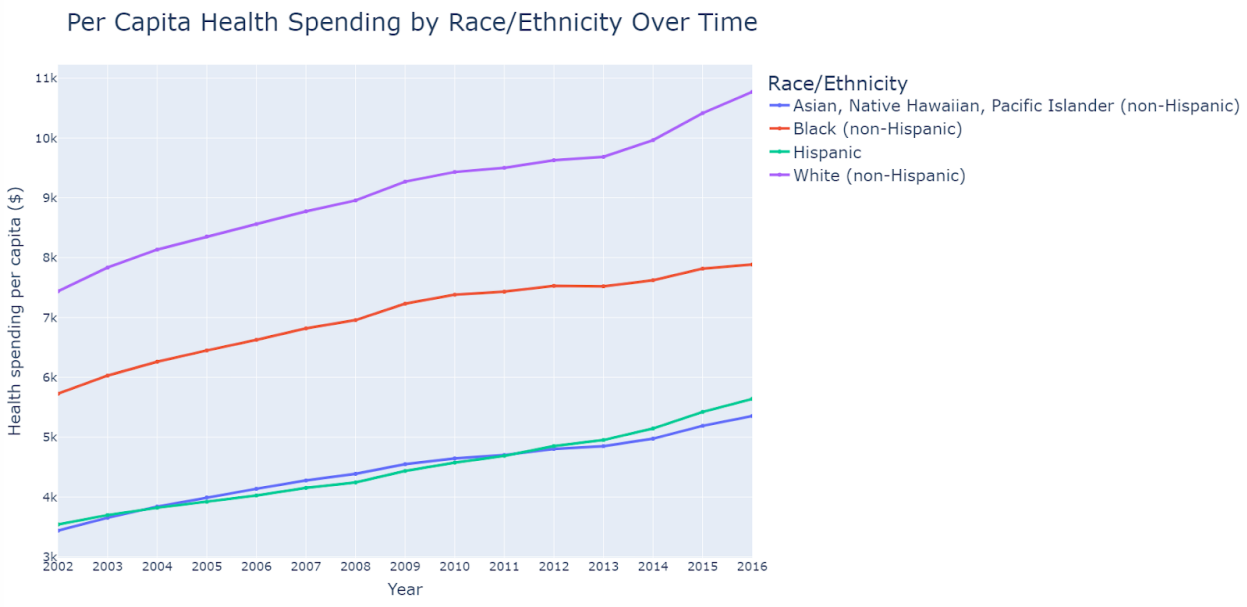

Fig. 3

Question 2: Looking at **fig. 4**, we can observe an obvious trend of higher percentage of income spent on healthcare as the population gets older. Ages from 35 to 64 on average have around the same income; while ages past 65 income drops significantly. In addition, health care spending increases significantly as one gets older. Both of these factors contribute to the increase in percentage of income spent on healthcare as the population ages.

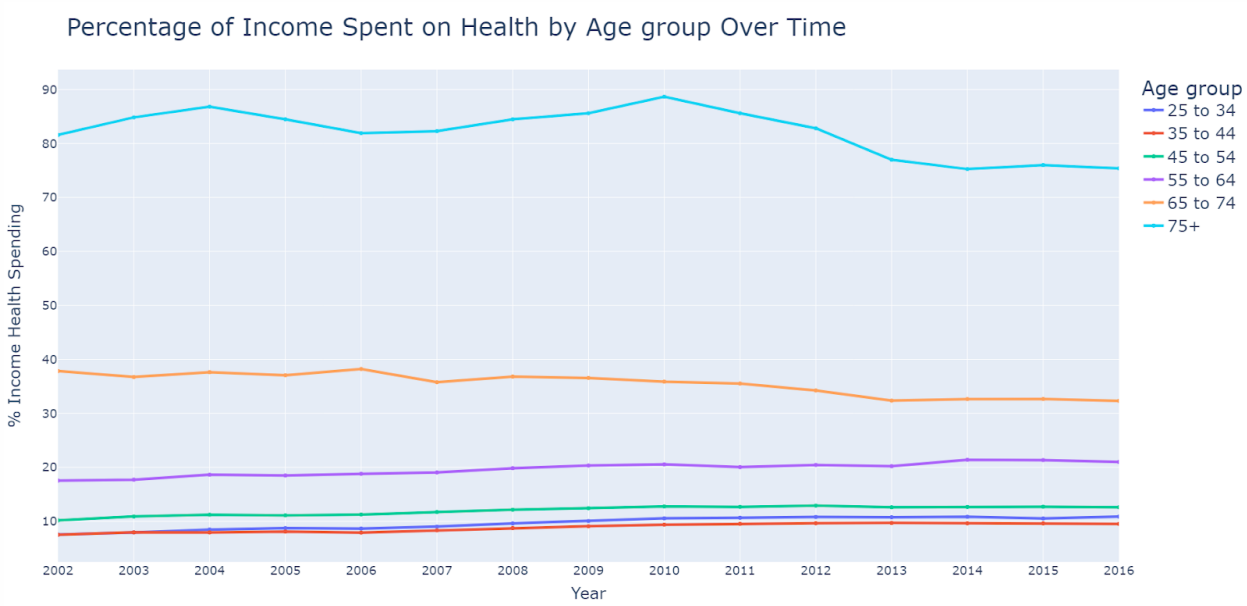

Fig. 4

Outside of 2002 and 2003, ages 25 to 34 have a higher percentage of income spent on health care than 35 to 44. Thus 25 to 34 is the only age range to have a higher percentage than an age range older than it. The increased risk of health problems (and health spending) is outweighed by the increase in income.

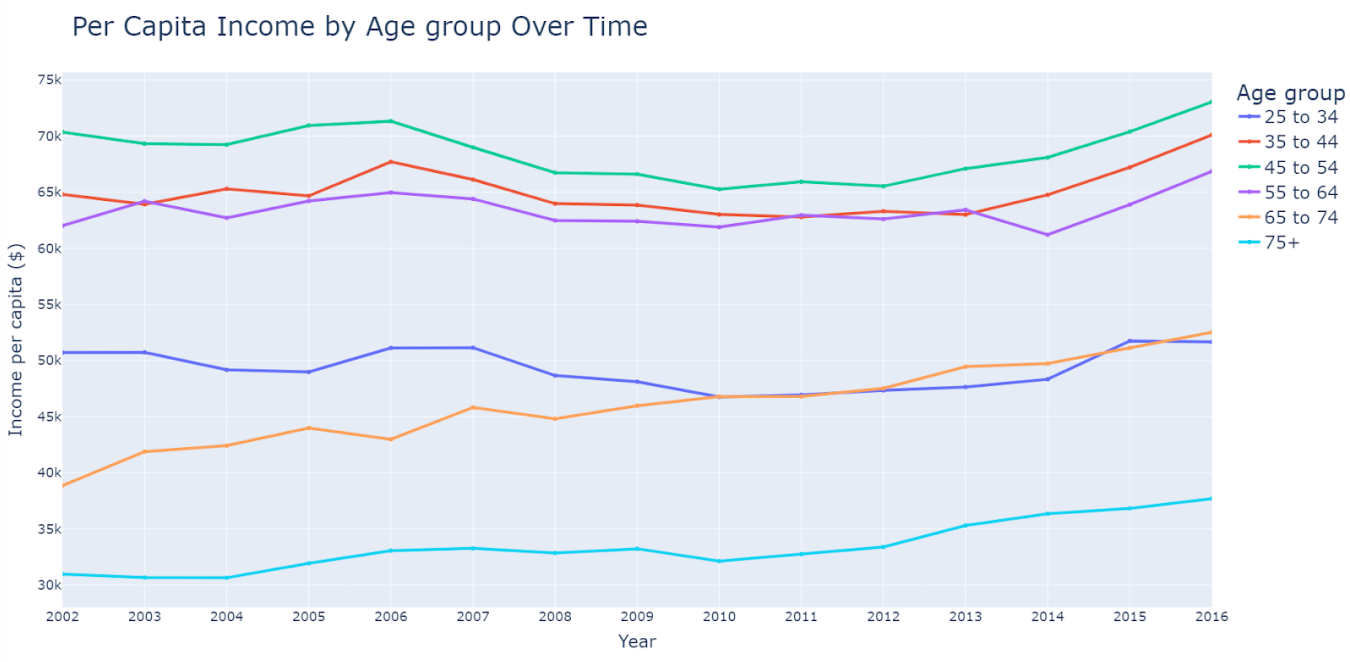

Fig. 5

Question 3: *The future of healthcare spending and income is largely a reflection of greater economic and world affairs. The results here should be taken with a grain of salt as by no means a simple linear regression model could predict how the global or U.S economy will unfold in the upcoming years.*

Race/ethnicity. The linear regression model of 2002-16 predicts the percentage of income spent on healthcare will continue to increase after 2016 (**Fig. 6**).

- **Asian:** 2.65 increase in percent from 2017-28
- **Black:** 4.9 increase in percent from 2017-28
- **Hispanic:** 6.95 increase in percent from 2017-28
- **White:** 4.1 increase in percent from 2017-28

However, looking at a more recent range of years 2012-2016, the percentage for each race is trending down or stagnating. It's difficult to say which trend in the data will be accurate, especially with such a limited amount of data points.

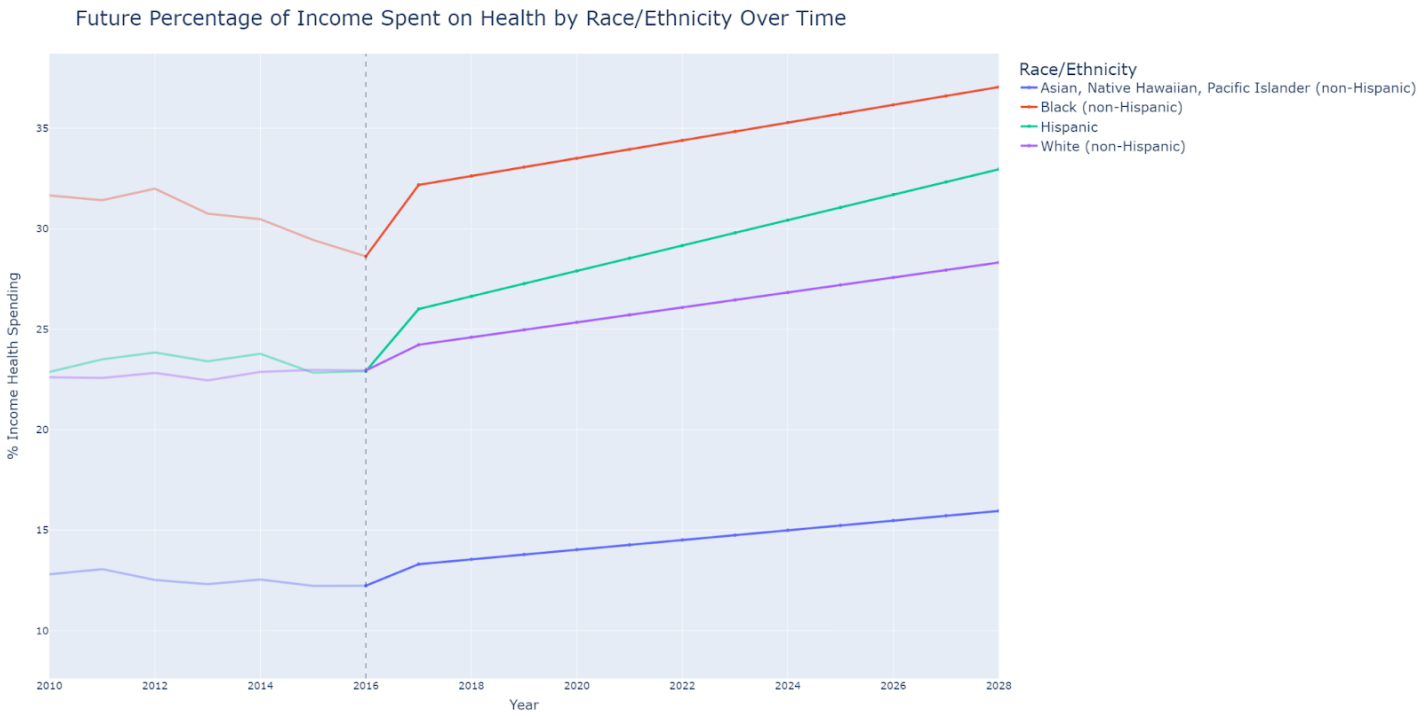

Fig. 6

Age. The linear regression model of 2002-16 predicts the percentage of income spent on healthcare will increase for ages below 65 and decrease for ages above 65 (**Fig. 7**).

- **25 to 34:** 2.9 increase in percent from 2017-28
- **35 to 44:** 1.9 increase in percent from 2017-28

- **45 to 54:** 2.2 increase in percent from 2017-28
- **55 to 64:** 2.9 increase in percent from 2017-28
- **65 to 74:** 4.5 *decrease* in percent from 2017-28
- **75+:** 5.6 *decrease* in percent from 2017-28

Although the decrease in percent from ages over 65 is much more significant than ages below 65, the population of people over 65 is much less. Thus the result still aligns with the race data with an overall increase in percent. Once again looking at a more recent range, 2013-16, the percentage seems to stagnate.

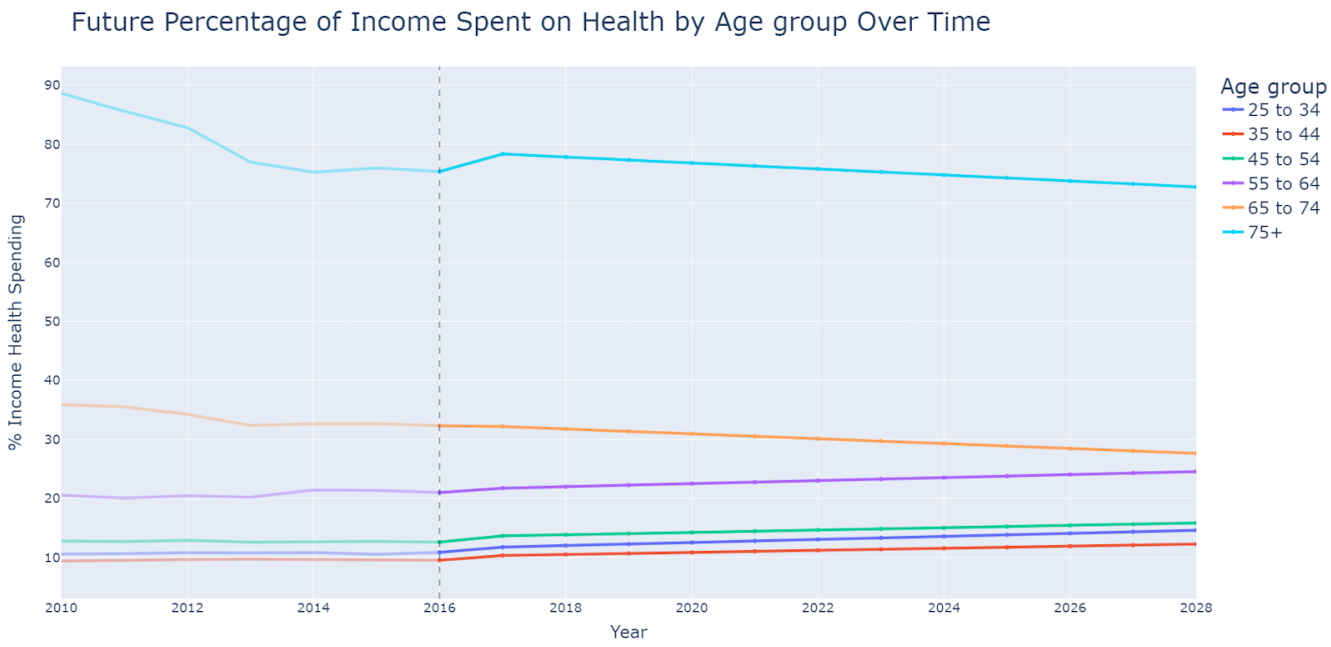

Fig. 7

Although it is uncertain what trend the percentage of income spent on health will take, it still is a reflection of how it has changed from the past. As time goes on, we are spending more and more on healthcare relative to our income.

Impact and Limitations:

Some potential implications of these results are the disproportionate percentages of income that certain race and age populations spend on healthcare. Some that may both be harmed or benefit from this analysis are the populations analyzed in the data. Given findings that the Black population and the older age range from 75+ years spend the highest percentage of income on health may lead these populations towards reducing their healthcare expenditure to save money, but at the cost of reduced coverage and higher risks when they fall ill or are injured and in need of proper care. Without larger societal and systemic shifts to reduce the burden that health spending has on these populations in particular, puts all the risk on these people themselves, rather than the institutions that enforce it. This leaves them with the option either to pay more out-of-pocket for health coverage, putting many in debt and poverty or to risk their health long term to remain financially stable, a decision that should not have to be considered by anyone in our society.

Some major limitations present in our analysis is the limited range of cases for our dataset. To perform our analysis on the different populations, we standardized the year range to 2002-2016 for consistency, but this also limits the long term applicability of our findings. Another limitation is that it does not account for the current COVID-19 pandemic data or its significant effects on the global economy, healthcare systems and aid. The impractical task of predicting future economic trends also limits our results as economic downturn and other major economic or health crises can greatly skew spending or household income for the majority of the measured populations.

Challenge Goals:

One challenge goal we implemented was combining **multiple datasets** to perform a deeper analysis using both health spending and total income in the US. We worked towards this goal because we can use our findings from the total per capita income to gain a deeper understanding into the impact that high healthcare costs and spending will have on various populations. This also allowed us to perform a stronger analysis by finding the percentage of income used on health spending across different race and age groups. The Census tables have somewhat **messy data** that is not presented in CSV format which requires a great deal of preprocessing with filtering out rows that aren't valid data.

We learned a **new library** to create more interactive and descriptive data visualizations for our finding using *plotly*. Learning new libraries allowed us to be more flexible in our implementation of our work and also in presenting our findings with different visualization tools with more information.

Work Plan Evaluation:

Adjusted Work Plan:

- **Task 1 (5 hours):** Preprocess and create a working CSV for the Census data
- **Task 2 (3 hours):** Use pandas to filter and combine our two datasets.
- **Task 3 (1.5 hours):** Use pandas to calculate the mean percentage of income spent on healthcare by race and separately by different age ranges.
- **Task 4 (3 hours):** Use plotly to visualize our results and get a baseline for our machine learning models.
- **Task 5 (7 hours):** Use scikit-learn to fit our data to a model and predict the future. Involves exploring various model types and finding one that works best.
- **Task 6 (3 hours):** Analyze our findings and make conclusions based on the data, models and visualizations that we created.

Workflow evaluation:

We underestimated the work required to preprocess the Census data across all of the different tables. A large amount of time was invested into formatting and filtering the datasets to be usable across the project and work for all questions we wanted answered. Splitting up each population then merging them into a proper dataset that could be used for all our questions was a large hurdle to overcome but it also helped us find more innovative ways to make the merged datasets as applicable to our project as possible. We were effective in compartmentalizing our code and adapting it so methods could be reused with different parameters rather than having repetitive and overlapping code.

Testing:

We tested our project using assert statements to ensure that our calculated means of income and that filtered values in each method are accurate. Assert statements from the cse163_utils file are also used to assess the accuracy of our linear regression model used for predicting future spending trends. It also accounts for underlying variance between tests and checks for error. A smaller dataset created from our original health spending dataset was used to test individual functions more efficiently.

Collaboration:

Library documentation: **Plotly**, **Pandas**, **NumPy**.
Uses cse163_utils file from CSE 163 to test expected vs received values to ensure that calculations for our functions are accurate.