What is the relationship between Income and debt if any?

As the title suggests, this paper aims to look at a single question. What is the relationship between Income and Debt? The topic at hand was designed purposefully to be open ended as there are many directions one could broach the topic from. One such direction would have been to look at the earnings and debt of different age groups, generations, socio economic backgrounds, etc. Another direction would have been to look at the data in regards to specific professions. Yet another direction could have looked at it from an international scope. We could've taken the average or total debt/earnings of a country's population and compared the result of different countries. Alas, I chose to look at the average income and debt per household, per state in the US.

To put it simply, I was interested in knowing if there was some sort of correlation between Income and debt. Be it positive or negative, I was also curious to know if this correlation would be a weak or a strong one.

To tackle the issue, I chose to use three data sets. The first dataset is titled "EFA: Household Debt" and is from the Board of Governors of the Federal Reserve System. This data set was collected between 1999 and 2022. The data covers each US State. The data shows the lower bound of Debt-to-Income ratio and the upper bound of Debt-to-Income ratio for each quarter of the years between 1999 and 2022. There are five variables in this dataset. The second dataset is "Household Debt Statistic by State" and is gathered by the Federal Reserve Bank of New York. The period that this data set covers is between 2003 and 2022. We only get information for the fourth quarter of each of the years covered. Within the dataset, there are multiple tabs that cover credit card debt, mortgage debt, student loan debt, total debt, auto delinquencies, credit card delinquencies, mortgage delinquencies and student loan delinquencies. I will only be using data from the tab that covers total debt. The total debt adds up credit card, mortgage and student loan debt. The last dataset comes from the national non-profit known as SSTI and gives us the "Median Household Income by State" from 1984 to 2018. The data in this dataset covers all the US States and is adjusted to 2018 dollars. Eventually I replaced the "EFA: Household Debt" data set with "Credit Card" data set that was from the list of Federal Bank of New York data sets.

I imported all the data into R-Studio to clean them up and transform them. There really wasn't that much cleaning to be done with the data itself in any of the data sets. All the data sets imported into R-Studio with the column names included as rows in the data frames. So one of the first things that I had to do was to turn the rows that contained the name of the columns into column names. After that it was simply a fact of removing other rows and columns that I wasn't going to be using in the analysis. The "Total Debt" and "Credit Card Debt" data sets only had data starting from 2003. Also, PR(Puerto Rico) did not have any data after 2016. As a result of this I decided to make the range of the analysis between 2003 and 2016. So, for all three data sets, all columns outside of the years between 2003 and 2016 were excluded. The final data sets and the process taken to achieve them can be found in the supplemental document provided along with this document.

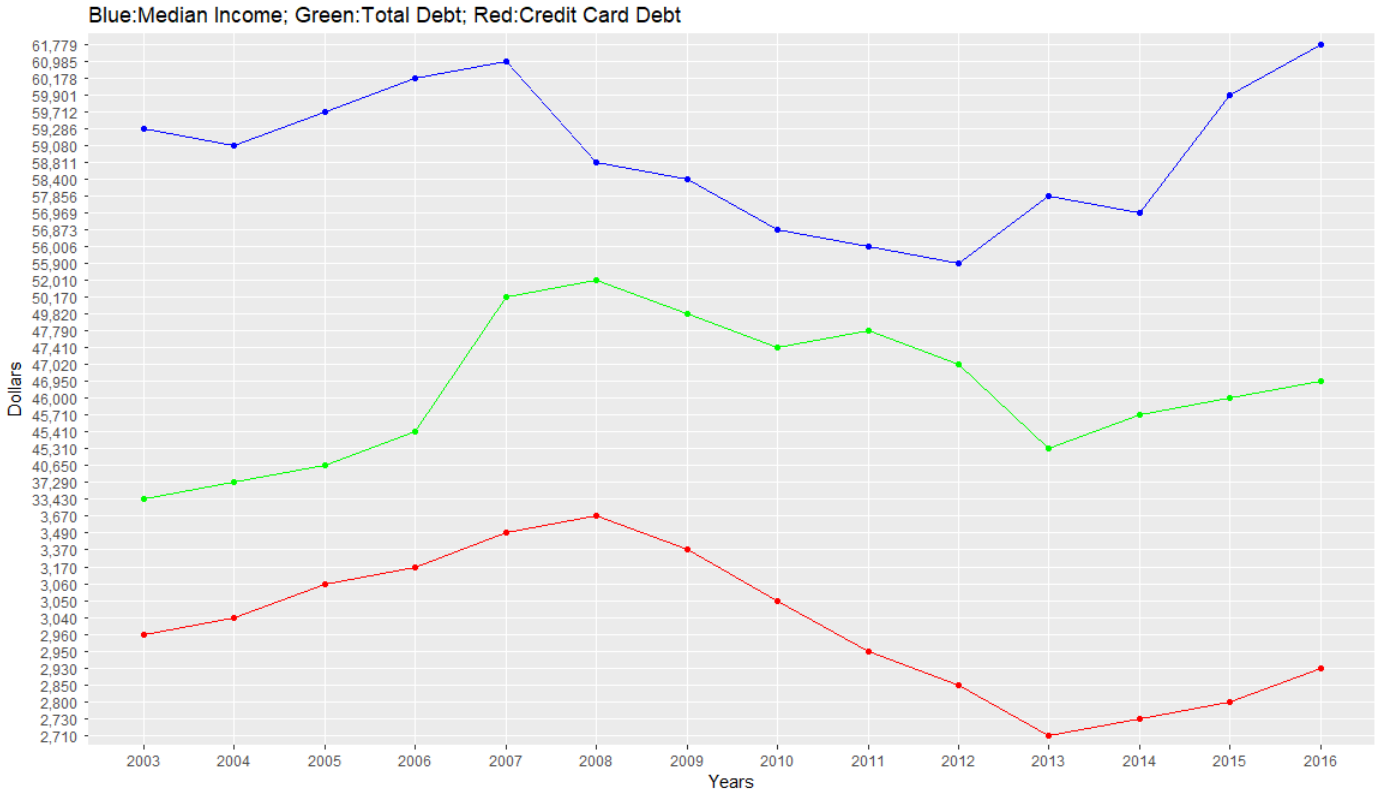
Later on, I decided to focus on the US overall as opposed to the different US States. As such I pulled out from each of the three data sets, the row that contained the information for the US in general. I then combined all three of these individual rows into a data frame of its own. This data frame contained the median income in the US, the average US household credit card debt and the average US

total debt between 2003 and 2016. See Fig 1. Once this had been done I created a combined geom_line and geom_point graphs with all of the data in the same graph. See Fig 2.

Fig 1

	Median_Income	Total_Debt	Credit_Card_Debt
2003	59,286	33,430	2,960
2004	59,080	37,290	3,040
2005	59,712	40,650	3,060
2006	60,178	45,410	3,170
2007	60,985	50,170	3,490
2008	58,811	52,010	3,670
2009	58,400	49,820	3,370
2010	56,873	47,410	3,050
2011	56,006	47,790	2,950
2012	55,900	47,020	2,850
2013	57,856	45,310	2,710
2014	56,969	45,710	2,730
2015	59,901	46,000	2,800
2016	61,779	46,950	2,930

Fig 2



DSC 520 Christian Campbell 8/12/2023
Project Milestone 3

As I couldn't get the code for the multiple regression to work, all my analysis comes from the graph visual. In addition, for some reason I couldn't get the labs() function to create a legend for my graph. So I put the legend in the title. To reiterate, however, the blue dot/line graph gives us the median income. The green dot/line graph gives the total debt, and the red dot/line graph gives us the credit card debt. To no one's surprise the credit card debt and total debt follow a similar trend. This is to be expected as credit card debt is a subset of total debt. The real analysis comes from looking at both debts in comparison to the median income. As the median income rises, so do the debts and as median income falls, so do the debts along with it. They seem to be in lockstep with each other. This to me denotes a positive correlation between median income and debt.

In all honesty, the analysis from the graph is a complete surprise to me. My personal hypothesis was that there would be a negative correlation between income and debt. It just made sense to me that we would see a decrease in debt if median income increased and an increase in debt if the median income decreased. I assumed that if a household income increased then that meant that that household now had more money to pay off their debts, as such I believed their debt would decrease. Similarly, I assumed that if a household income decreased, that household might not have as much money to cover all their necessary expenses, which means that they might need to take out more money. So it's a real shocker for me to see the complete opposite happening from the graph analysis.

What the analysis implies is that the wealthier people get, the more debt they take on. After having digested that information it no longer seems all that shocking to me anymore. People with more money feel more confident in their abilities to pay off debt as such they take out more debt.

One major limitation to this project it could be said is my skill level with R and data analysis/manipulation. I'm clearly no expert, so I tried to simplify things as much as I could so as not to make it overly complex for me. It is likely that in one way or another this could have had an effect on the final product. For example, even though the datasets I chose looked at all 50 US States individually, I didn't feel comfortable tackling all 50 US States individually due to me not feeling confident in my R abilities. So I simplified it and just looked at the combined data for all the States. Another limitation has to do with the scope of the data in terms of years. 14 years is a decent amount of time to carry out an analysis, however I feel like a larger timeframe would have yielded an even more accurate result.

In conclusion, I set out to find a relationship between two-ish variables; and I did. Just not the relationship that I was expecting to find. I would argue that this project was a very eye-opening experience for me. I gathered data sets, cleaned them, sliced, diced and combined them. I eventually created a visual from the final data which I then used to answer the problem statement. This experience was the first real look at what it would mean for me to be a data scientist. Having done this, I'm aware that I still have a lot more to learn and I still have a long way to go yet. That said it's made me more confident in this path that I've set myself on.

Citations

- Useful stats: Median household income by state, 1984-2018. SSTI. (2019, October 24). https://ssti.org/blog/useful-stats-median-household-income-state-1984-2018
- Board of governors of the Federal Reserve System. The Fed - Table: State Debt-to-Income Ratio, 1999 - 2022. (2023, June 16). https://www.federalreserve.gov/releases/z1/dataviz/household_debt/state/table/
- Center for Microeconomic Data: Data Bank. CMD Data Bank - FEDERAL RESERVE BANK of NEW YORK. (2023). https://www.newyorkfed.org/microeconomics/databank.html