CMPT 353 Final Project Report

Stefan Pricope Okasha Huda

Overview

For the final project, we worked with financial data given from the financial independence surveys on Reddit in order to investigate patterns and see whether annual expenses or annual income has increased at a more drastic rate since the pandemic.

To start working towards this goal, we had to get data from previous years, in specific 2016 to 2018 & 2020 to 2023 (2019 data does not exist as the surveyor tried to let someone else run it that year and they did not do the survey and thus we will have one missing year). We used statistical analysis techniques taught in the course as well as a few other statistical methods in order to provide a clear view at answering our question.

Null Hypothesis (H_0):

- There is no significant difference between the averages of annual income (or annual expenses) between the pre-pandemic and post-pandemic periods.

Alternative Hypothesis (H_1):

- There is a significant difference between the averages of annual income (or annual expenses) between the pre-pandemic and post-pandemic periods.

Data Collection

As mentioned in the overview, we collected data from the Reddit financial independence survey that is run on the r/financialindependence subreddit on an annual basis. Specifically, we got the results from 2016 to 2023, with the exclusion of 2019, for the reason given in the Overview section. Retrieving the annual income is easy for all years of the data, as there is a column that sums all income factors together, however for expenses it is only summed up properly for data from 2021 to 2023. The rest of the years had individual expenses which needed to be summed together in order to get an accurate annual expenses column. We removed some expenses columns that we didn't think were relevant to our analysis such as the expenses related to investments.

The secondary set of data we cleaned was a set of data which includes individual expenses. The years 2016 and 2017 were removed from the datasets available for this part of analysis as the surveys were missing many core expense columns, such as healthcare, utility and tax expenses.

On all datasets we imposed some restrictions in order to clean the data of outliers or invalid input data. First, we checked all columns to not have the value of 0 for annual income or annual expenses, though in the individual expense files we allowed individual expenses to have the value of 0. Further, we dropped any NaN values to avoid having problems when performing statistical computation. All numerical columns were converted to floating point values for consistency. Finally, we restricted annual income values to be greater than $20,000 and less than $1M. We thought this would be a good way to remove outlier values, as the incomes below $20,000 were either incorrectly answered or the survey was left incomplete. The $1M cap was imposed to prevent having very large income outlier values which are extremely unlikely to be an actual true value.

Data Analysis

Before starting analysis on our collected data, we did some quick metrics such as mean values that immediately showed us that the data has very high annual income and annual expense values. The average values for each year are well in the hundreds of thousands which is beyond the average values expected for a regular population of people. This shouldn't change the analysis process, but it's good to keep in mind when looking at the patterns later.

The first thing we should do is find the p-value for our hypotheses and see if it is worth looking at patterns of expenses and incomes. The null hypothesis will be rejected if the p-value is less than 0.05 (default alpha value). We used two p-value tests in order to conclude the result, a standard t-test and Mann-Whitney U-test.

The Mann-Whitney U-test comes in handy to provide an extra opinion in case our data has unequal data for each year. In general, by looking at our cleaned datasets, we can see that there are less responses in the earlier years of the survey which is what we expected going in. The Mann-Whitney U-test allows us to have an extra test that will give a more accurate p-value if the standard t-test turns out to be affected too heavily by the pandemic data having more entry values.

Data Analysis Results

Running the code for statistical tests, we get the following p-values:

t-test Results:

	t-statistic	**p-value**
Annual Income	-24.47530907994939	2.560987369411892e-128
Annual Expense	-37.153499585460374	4.1077902719542276e-282

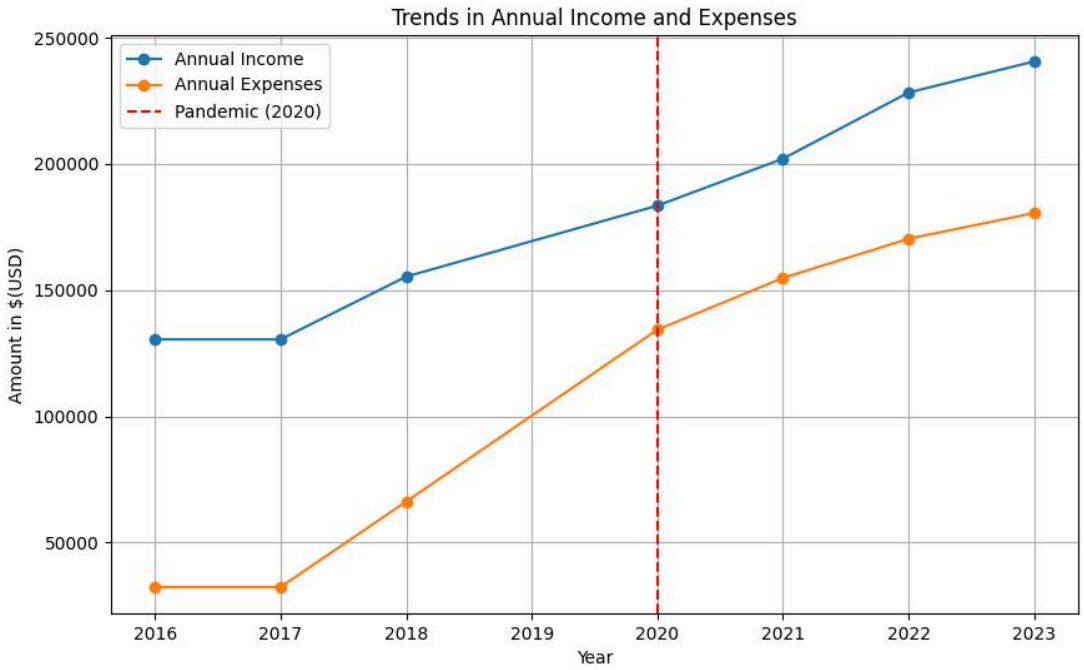
Mann-Whitney U-Test Results:

	U-statistic	p-value
Annual Income	7288972.5	4.832275882066054e-176
Annual Expense	3336800.0	0.0

As we can see, for both tests and on both annual income and annual expenses, our p values are significantly smaller than our alpha value of 0.05. This allows us to reject the null hypothesis and conclude that there is a statistically significant difference between pre-pandemic and post-pandemic data.

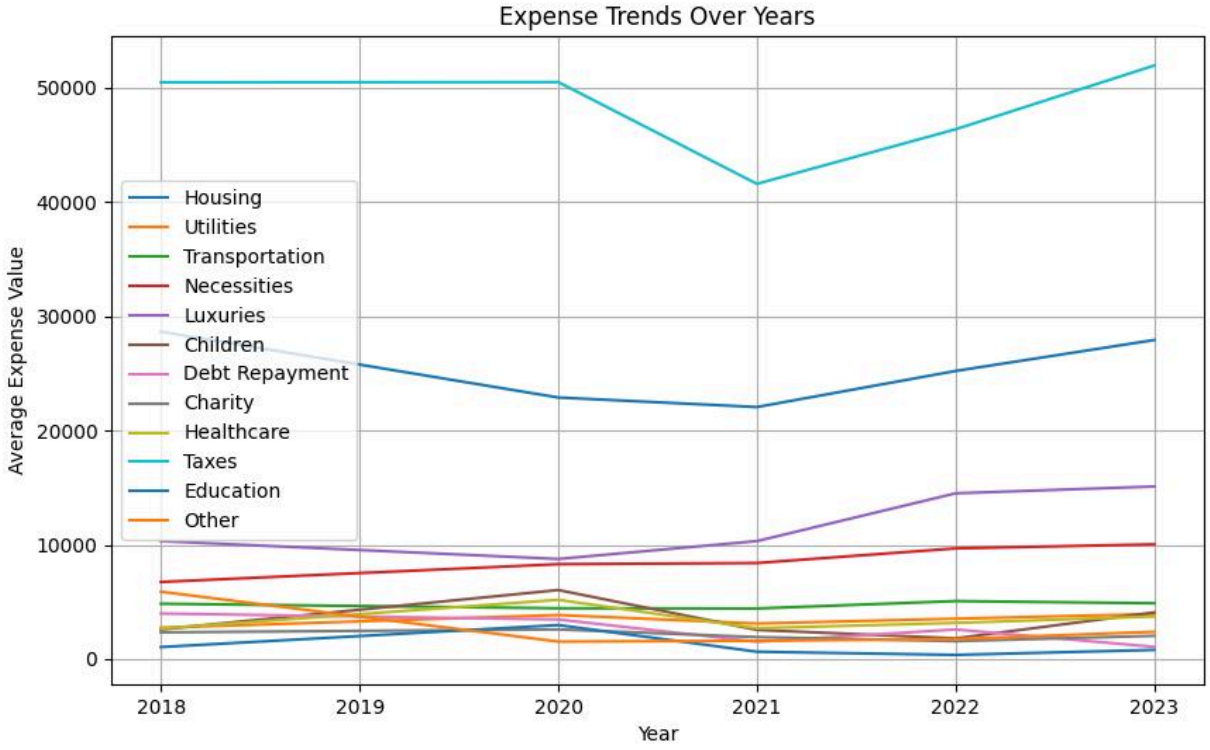
Further Analysis

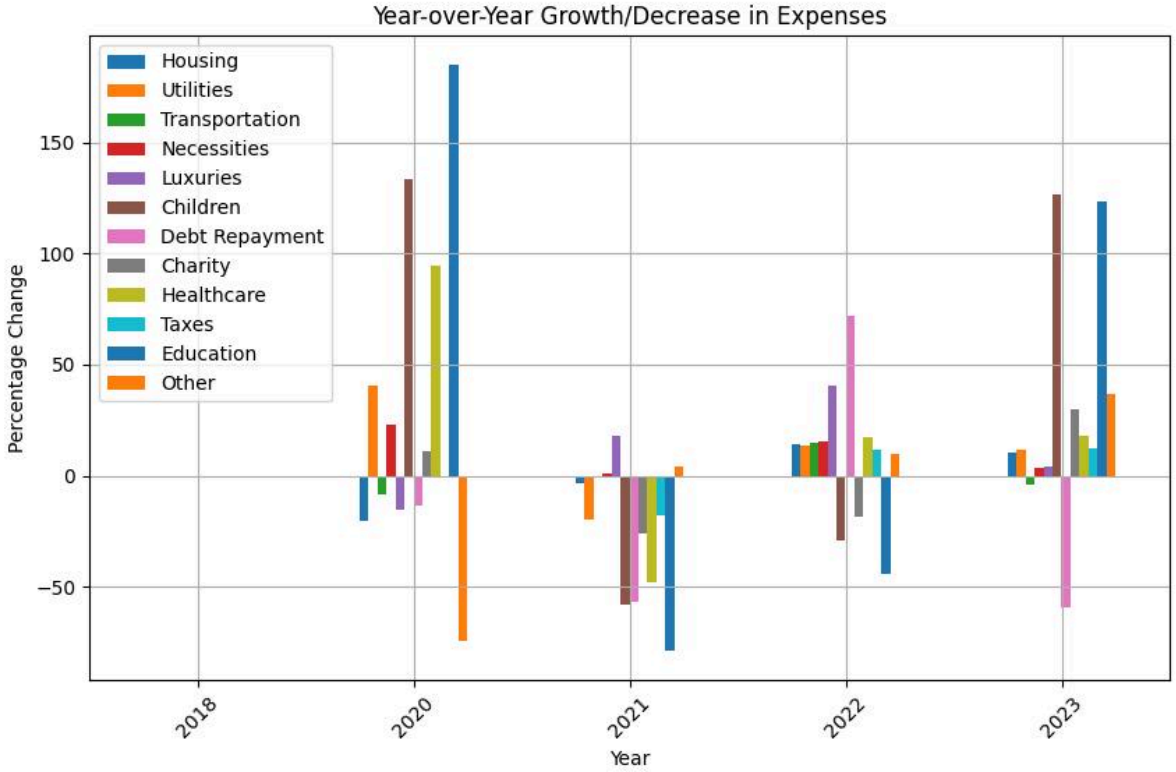
Now we know that there is a significant difference between pre-pandemic and post-pandemic data but this does not tell us any information about what this difference actually is. To gain a deeper understanding of what is going on in our data, we combined all the datasets together and assigned them a year column. We then grouped our data by those years and aggregated them with the mean values of the annual income and expense columns. By plotting the data on a trend plot using our annual income and expenses data, we obtain the graph below. We can see the graph has a very distinct increasing pattern over the years. This is consistent with our statistical conclusion, as the graph showcases a very clear increasing relationship for the growth of each of our annual factors.



An interesting fact to look into is which of the two had a higher growth rate throughout the years. On an initial look at the graph, expenses appear to have a larger growth especially from the 2018 to 2020 portion, but we need to do a calculation to be certain. The method we chose for this is to look at the compound annual growth rate (CAGR) stats for the annual income and annual expenses. It is a statistic that represents the smooth growth of a metric over a number of years. Calculating CAGR values for annual incomes and expenses, we get a CAGR value of 9.13% for annual income, and 27.82% for annual expenses. Notice that both are growing (because they are positive) but the expense value in particular has grown significantly faster than the income value.

We also created trend plots and bar graphs showcasing the growth of individual expense factors over the years. The growth plot shows that most factors have remained fairly consistent throughout the years, but we can see that standouts include a large drop in tax expenses for between 2020 and 2021, and a sizable increase in education costs. The barplot showcases growth between individual years in percentage, and interestingly we can see extremely large spikes in 2020 for education, healthcare and children costs. These factors make sense along with the other factors having increased, as the pandemic is a likely factor to the increase of these values. In 2023 we can also see larger increases with education once again and children costs. Interestingly, 2021 and 2022 had sizable education value drop offs, but it’s a possibility this is because of different demographics answering for a specific year.


Expense Trends Over Years
Average Expense Value
Year
Housing
Utilities
Transportation
Necessities
Luxuries
Children
Debt Repayment
Charity
Healthcare
Taxes
Education
Other
0
10000
20000
30000
40000
50000
2018
2019
2020
2021
2022
2023


Year-over-Year Growth/Decrease in Expenses
Percentage Change
Year
Housing
Utilities
Transportation
Necessities
Luxuries
Children
Debt Repayment
Charity
Healthcare
Taxes
Education
Other
−50
0
50
100
150
2018
2020
2021
2022
2023

Limitations

There were several limitations for this data analysis project. For one, the missing data from 2019 could be seen as a detriment to the analysis. The data from that particular year would add nicely to the pre-pandemic analysis; helping us understand the financial trends leading up to the onset of the COVID-19 pandemic. The data from 2016 and 2017 were available, but were less comprehensive in comparison to data from 2018 and later. The earlier data lacked several key variables that we excluded due to insufficient data quality. As a result, we were left with only one year of pre-pandemic data for certain analyses, particularly on individual expenses. This imbalance — one year of pre-pandemic data compared to three years of post-pandemic data — posed challenges in forming a baseline and limited the scope of our comparative analysis.

Another limitation was the bias in the dataset toward high-income individuals. Public financial data often face this issue, as comprehensive data on low- or average-income populations is rare to find. While the data we used provided valuable insights, this skew limits the generalizability of our findings to the broader population. A more representative dataset could have allowed for a more impactful analysis, particularly around the financial behaviors of diverse income groups.

Although we initially planned to incorporate machine learning techniques, this approach was quickly turned down by weak correlations between most variables. The few features that did show hints of a stronger correlation yielded predictable results. This limited the potential for meaningful predictive modeling or deeper insights, narrowing the scope of our project. If we had extra time, we would have liked to do more research and find a more meaningful method of performing machine learning on our data.

Project Experience Summary

Okasha Huda

- Cleaned and consolidated seven annual financial datasets, totaling over 100,000 records, by standardizing columns, resolving missing data issues, and filtering outliers, resulting in a 30% improvement in data completeness
- Performed statistical tests, t-test and Mann-Whitney U-tests, revealing significant changes in income and expenses pre- and post-pandemic, with income increasing more than 30% during the pandemic
- Addressed data biases, including overrepresentation of high-income individuals, to provide insight into financial behaviour despite limitations in data coverage

Stefan Pricope

- Segmented datasets into pre- and post-pandemic periods to measure significant changes in financial

- Analyzed expense trends and savings rate, identifying a 15% decline during the pandemic compared to pre-pandemic levels
- Visualized trends in income, expenses and individual spending categories using Matplotlib and Seaborn, creating [insert number of viz created] detailed visualizations, including CAGR calculations
- Developed a predictive model using linear regression, estimating annual income based on expense patterns for 2023 data