Financial Well-Being Survey Analysis

Introduction/Data Wrangling:

This project was completed in a group of two members. Our project involves using the National Financial Well-Being Survey in order to find correlations between certain variables related to the financial standing of the population and levels of happiness, race/ethnicity, and what generation they are from. Our dataset consisted of 217 variables and 6394 people were surveyed. Since we didn't use all 217 variables, we created a new dataframe that only included the variables we needed and used this simplified dataset. Our simplified dataset contained the following variables:

FS1_6 = I know how to keep myself from spending too much

FS1_7 = I know how to make myself save money

MANAGE1_2 = Stayed within your budget or spending plan

FWB1_1 = I could handle a major unexpected expense

FWB1_3 = Because of my money situation, I will never have the things I want in life

FWB1_4 = I can enjoy life because of the way I'm managing my money

FWB2_1 = Giving a gift would put a strain on my finances for the month

FWB2_2 = I have money left over at the end of the month

FWB2_3 = I am behind with my finances

Generation = Generation

PPETHM = Race/Ethnicity

SWB_1 = I am satisfied with my life

SWB_2 = I am optimistic about my future

Using these variables, we aimed to find some outputs that would help us answer the following questions:

1. *How financially responsible is the majority of the population?*
2. *How does general life satisfaction correlate with financial struggles?*
3. *Is there a correlation between generations, race/ethnicity, and levels of financial security?*

4. *Is there a correlation between financial status and happiness and outlook on life?*

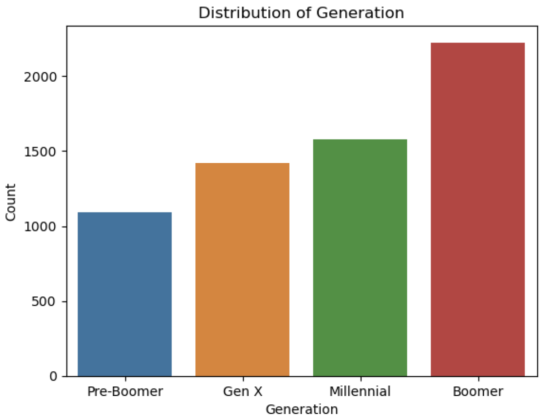
We also had to remove some negative values, which represented a refusal to respond or if the response was not written to the database. In order to clean the data, we used the “selected_columns” function and chose which variables we wanted to include. Then we used the “rows_to_remove” function to find the rows which had negative values, and then dropped them. This also helped to make our dataset slightly smaller.

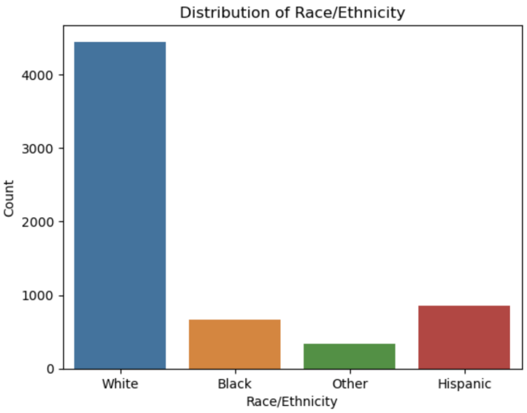
Visualizations:

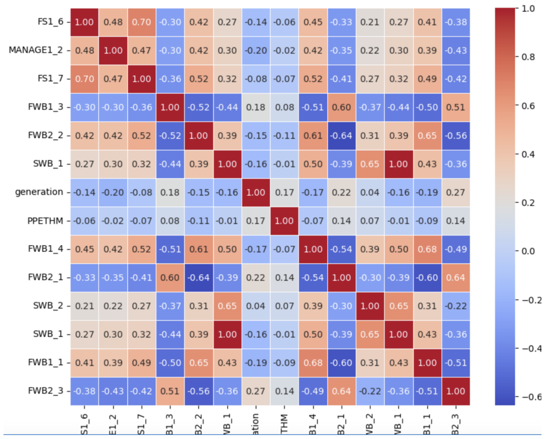
For some exploratory analysis, we made some plots and graphs in order to show visualizations of which race/ethnicity and generation made up the most of the population. We did this by mapping the generation and race/ethnicity variables using the “mapping” and “.map” functions. Once this was done, we plotted some distribution plots for each variable by using the “sns.countplot” function and adding axis labels and a title. Our distribution plots showed that the majority of the population consisted of White people and those belonging to the “Boomer” generation. In order to further understand whether there was a correlation between these variables and levels of financial security, we created a grouped bar-plot using the mapped data and the “sns.barplot” function. Our grouped bar-plot shows us that the majority of the population, regardless of their race or what generation they belong to, have a relatively high level of financial security. However, people of black ethnicities tend to have a lower financial security level compared to white and other ethnicities due to their money management habits. We can also see that people from the “Pre-Boomer” generation tend to have higher financial security compared to other generations, and those in the “Gen X” generation have lower financial security.

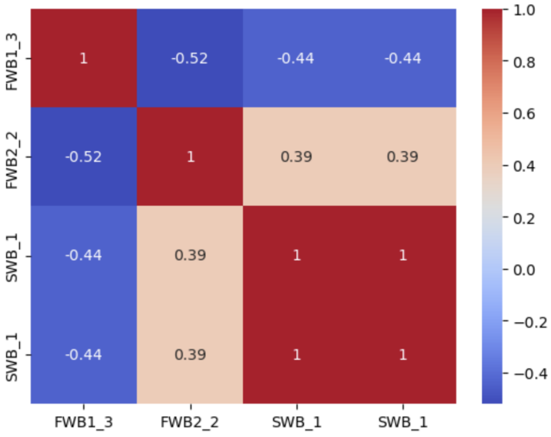
We then proceeded to create a heat map using the “sns.heatmap” function for all the variables we were interested in to see if any of them correlated with each other. Using a distinct color combination, we saw that those blocks colored blue indicated very little or no correlation and those that were orange showed higher correlation. Just by looking at the heatmap, we can immediately see that there are more variables that are colored blue or in tones of blue compared to orange, which suggests that there is little correlation between variables. This assumption was somewhat expected, since we assumed that variables like FWB2_3 and FWB2_2 should be negatively correlated as being behind with finances is more likely to mean that one does not have money left at the end of the month. Our heatmap shows this correlation in a dark blue color, and the correlation coefficient is -0.56, which indicates a negative correlation.

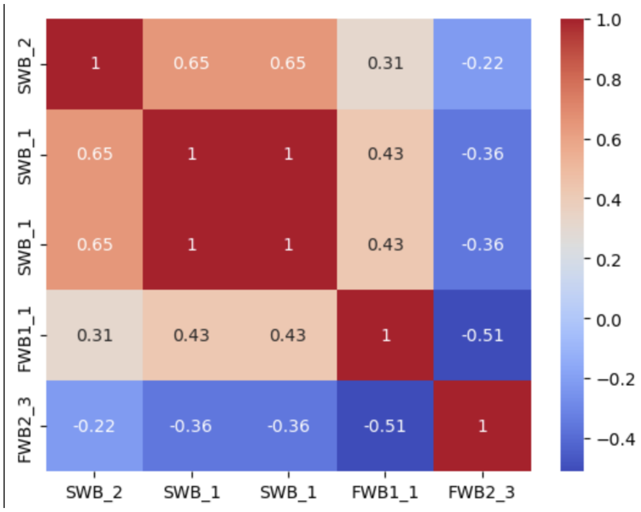
In order to further understand the correlation between similar variables, we conducted separate heatmaps using specific variables related to our research questions. The results we found from these heatmaps show a negative correlation between variables such as FWB1_3 and FWB2_2, suggesting that people do not have the things they want because they do not have money by the end of the month. We also see a positive correlation between variables like SWB_1 and FWB1_1, suggesting that people are satisfied with their life and can handle a major unexpected expense. So far, the results we found follow through with our expectations.











Statistical Associations/Analyses:

We then proceeded to perform some statistical tests to see the relationship between these variables. Our first step was to perform an ANOVA test to see how financially responsible the population was. The p-value we got was an extremely small number, suggesting that our null hypothesis that the majority of the population is not financially responsible is rejected. We then did a Tukey test using the "from statsmodels.stats.multicomp import pairwise_tukeyhsd" function and we saw that there was a significant difference between FS1_6 and FS1_7, suggesting that the population may find it more challenging to save money compared to avoiding excessive spending. We then conducted some statistical analysis tests and the first thing we did was calculate the mean of all variables. We saw that the highest mean value was seen in the SWB_2 variable, suggesting that most people were optimistic about their future. We also saw that the variable with the lowest mean was FWB2_3, which tells us that not many people struggle with finances. We then conducted multiple frequency distribution tables to see which rating was voted for the most for each of our research questions and found that for each question, the majority of the population has a positive outlook on their life, is not financially struggling, and can handle and manage expenses well.

Our final statistical analysis test was to create a correlation matrix for each research question using the “correlation_matrix = df[].corr()” function. By identifying what each correlation coefficient means, we were able to gain a clearer understanding of which variable is correlated with what and we could also see the relationship of the correlation.

Predictive modeling:

The final part of our project includes predictive modeling. In order to test our research questions, we decided to use Linear Regression models and Random Forest classification models. We decided to use a Linear Regression model for the research questions “How does life satisfaction correlate with financial struggles” and “What is the correlation between financial status and happiness and a positive outlook on life” because we expected to find a linear relationship between these variables. We split our data into a training and testing set and created a Linear Regression model using the “model_variable = LinearRegression()” and “model_variable.fit(X_train_variable, y_train_variable)” functions. The results we obtained from these models were a bit confusing, since our Mean Squared Error value was very large (greater than 1), which suggests that our model lacks accuracy and had some errors. We assumed this was because our data was collected on a scale of 1-5, which doesn’t necessarily work well when trying to find a linear relationship. Our R-squared value was very low, which means that the model did not explain much of the variability in the data. We concluded that this was because the data was very concisely obtained with exact values of 1-5.

We used a Random Forest classification model for the research questions “How financially responsible is the majority of the population” and “Is there a correlation between generations, race, and financial security level”. The results from both these models were the same: perfect scores of 1 for recall, precision, and f1-scores. We suggested that this may be due to the fact that the results are a whole number on a scale of 1-5. We also saw that in both confusion matrices, there were 0 false positives and false negatives. This was sort of expected, since the population voted an exact number that they felt, so there was no scope for false reports. We also assumed that the model gave us perfect scores due to overfitting or underfitting.

Conclusion:

Overall, the statistical analysis part of our project gave us the desired results. However, our predictive modeling techniques didn’t seem to work the way we wanted them to. We believe that this occurred because of the nature of our data and that because it was recorded on the basis of a scale, the use of the models didn’t work out right. In order to conduct these analyses and models, we used a variety of functions and tools such as pandas, seaborn, matplotlib.pyplot, sklearn.preprocessing, and many more. We can conclude that based on our statistical results and visual plots, there seems to be a correlation between financial management and happiness. There is also a correlation between financial struggles and optimism, and we noticed a relation between race/generation and financial security.

The results of our project can provide us with further scope on this topic and can allow researchers to focus on more aspects in order to improve financial management and decision making. Seeing that race and generation also plays a role in financial standing, further research can explore on how to improve financial decision making in certain races/generations, and how to maintain high financial standing in certain people. Although some aspects of our project did not work the way we expected, our analyses show us that there is indeed a correlation between life satisfaction and financial struggles, levels of financial status and happiness, generation/race and financial security, and if the majority of the population is financially responsible.