Exploring the Interplay of Socioeconomic, Health, and Political Factors on Well-Being & Happiness:

Insights from the General Social Survey

Edie Thors, Vanessa Pasquarelli, Gwen Thompson

1. Abstract *(300 words)*

This paper explores the relationships between income, happiness, and stress using demographic, behavioral, and attitudinal variables. We applied various statistical and machine learning models to predict these outcomes, leveraging General Social Survey (GSS) data. Our analysis revealed that higher income and better self-reported health are strongly associated with greater happiness, while moderate stress levels and lifestyle factors, such as physical activity and religious attendance, also significantly contribute to well-being. A linear regression model predicted income using demographic and behavioral variables, with key predictors including age, education, job satisfaction, and race. However, the model's relatively low R^2 value (0.1859) suggests that income cannot be fully explained by individual characteristics and is influenced by external factors like occupation and region, which were not included in the analysis. For predicting happiness, we used logistic regression and random forest classification models. The random forest model outperformed the others, achieving an accuracy of 86.8%. This suggests that happiness is influenced by complex, non-linear interactions that are best captured using ensemble methods. Interestingly, our analysis also found that moderate political ideologies and religious attendance were associated with higher happiness levels. The K-Nearest Neighbors (KNN) model was used to classify happiness into binary categories, with an accuracy of 84.6%, but performed better at predicting "happy" individuals compared to "unhappy" ones. The stress prediction model showed a lower accuracy (62.5%) when excluding income and occupational variables, emphasizing the importance of socioeconomic factors in stress levels.

This study highlights the critical role of both structural factors and personal lifestyle choices in shaping well-being. Despite limitations such as multicollinearity and the absence of occupation-related variables, our results provide valuable insights into the complex factors influencing income and happiness. Future research should incorporate more granular data and explore non-linear relationships and interaction effects in greater depth.

2. Introduction *(Two to three pages)*

This paper will show in depth how various aspects of people's lives, such as their jobs, income, health, relationships, and political beliefs, are connected to their overall well-being. Using data from the General Social Survey (GSS), we sought to better understand how factors such as job satisfaction, financial stability, political ideology, and family life are related to emotional well-being. While happiness and stress are very personal, we were interested in identifying broader patterns across the population and what those patterns could reveal about society.

We started with a question that could be fairly simple: What makes people feel good or bad about their lives? But through analyzing data, we found that answering that question is complicated and multidimensional. People's well-being is influenced by many different things that often overlap. For example, someone's income might affect their stress levels, but so might their health, relationship status, or how safe they feel at home. By examining many of these variables together, we hoped to gain a better understanding of how these different aspects of life interact.

The GSS dataset provided a great starting point for that. It included detailed information about respondents' demographics (race, education, and gender), economic background (income and job status), personal life (marital status and number of children), health, political and religious views, and more. With all these variables, we were able to find valuable information and explore a variety of questions: Are people with higher incomes consistently happier? Do political views relate to stress? Does job satisfaction make a difference even if income is low? Are happiness levels different across demographics after controlling for various aspects?

To study these questions and analyze our dataset, we used several data science methods to uncover both expected and surprising relationships. We applied linear regression using a Lasso penalty to predict continuous variables like income, which helped us isolate the most relevant predictors while minimizing overfitting. We then used logistic regression and K-Nearest Neighbors (KNN) classification to model binary outcomes, like happiness, based on a wide range of demographic, lifestyle, and well-being factors.

Some of our findings confirmed our expectations, for example, income was positively associated with happiness and job satisfaction, and people in lower-income brackets or with less stable employment often reported higher levels of stress. However, other patterns were more nuanced. Our linear regression model revealed that while stress generally had a negative association with income, those who reported feeling stress generally had a negative association with income, those who reported feeling stresses "often" actuall earned more than those who were "never" or "hardly ever" stressed which suggests that moderate stress may be tied to demanding but high-paying jobs. Happiness was consistently positively associated with income, and those who opted out of the happiness question tended to have lower earnings.

We also found that lifestyle and ideological factors played important roles in shaping happiness. Our logistic regression and KNN classification models revealed that better self-reported health, moderate religious attendance, moderate political views, and financial satisfaction were all strong predictors of being "happy." Extreme political conservatism was associated with slightly lower happiness than more moderate or slightly liberal views. Some findings challenged assumptions. For example, heavy smokers showed a small positive coefficient for happiness, and the number of children showed a slight negative relationship with happiness.

To make sure our models were working well and producing meaningful results, we used several validation techniques. For regression models, we examined R-squared values to determine how much variation in happiness or stress could be explained by our independent variables. We also calculated the RMSE (Root Mean Squared Error) to measure how far our predictions were from the actual values. In cases where multicollinearity or overfitting became a concern, we considered using regularization methods, such as Lasso, if necessary. Finally, we used accuracy as a measure and an ROC curve with AUC for log regression. These checks helped us be more confident in our results.

Overall, this project gives us a better understanding of how different aspects of life affect people's feelings. Together, these methods helped us move beyond surface-level trends to identify more complex, sometimes contradictory relationships between well-being, economic status, and personal behavior. It also raised important questions about inequality and how people experience the world around them. The following sections of this paper provide more detailed explanations of our data, methods, and findings. The conclusion reflects on what we learned and where future research might go from here.

---

3. Data

Our data was composed of the General Social Survey (GSS), which included a variety of variables related to the demographic, economic, social, and health-related information of the participants. It also covered areas such as job satisfaction, political views, family dynamics, and more. The dataset contained both categorical variables (e.g., marital status, race, political views) and numerical variables (e.g., income, stress level). A breakdown of the specific variables was as follows:

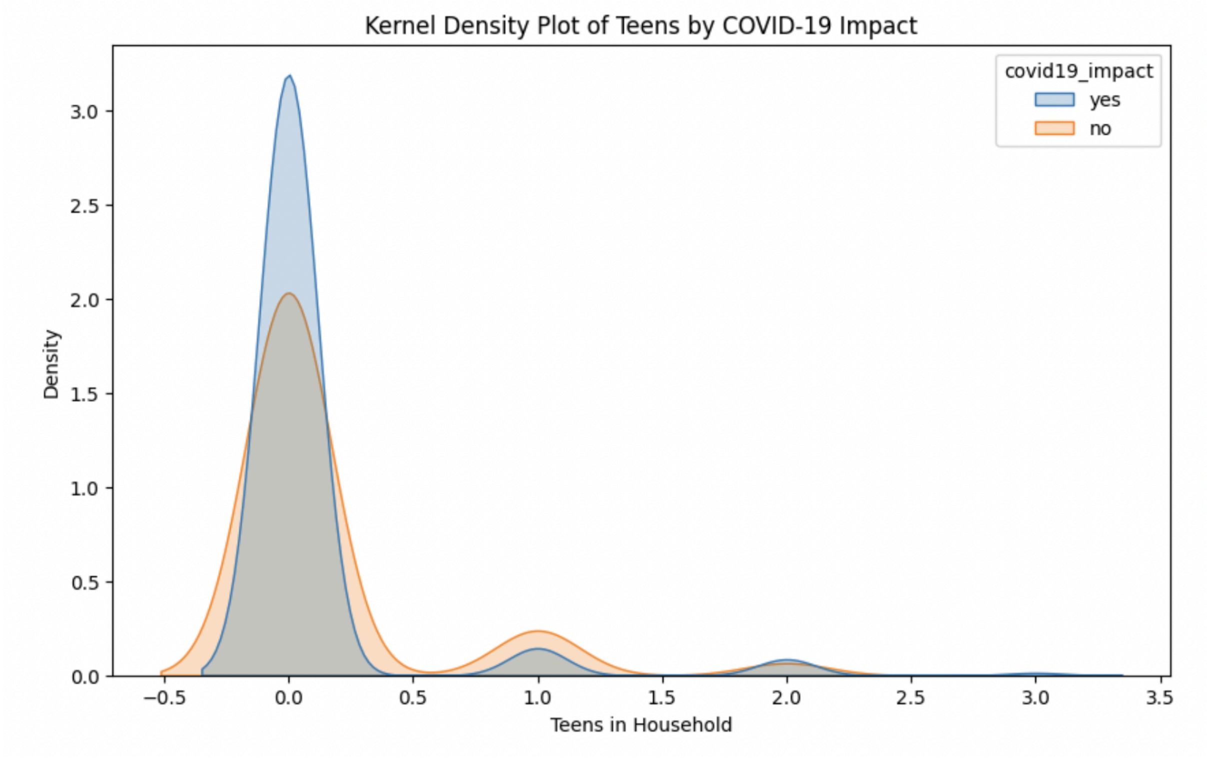
Demographic variables included marital status (e.g., married, divorced, never married), divorce status, whether the respondent had children in the household, the number of children, expected household population, gender, education level, type of degree, and parental education levels. Other demographic details included work status, race and ethnicity, gender identity, sexual orientation, and age. Economic variables included the socioeconomic index (a measure of occupational status), real income or household income adjusted for inflation, income from occupation, and social class. Family and household data consisted of the number of individuals in various age groups within the household, as well as the number of sexual partners the respondent had in the past year. Health-related variables included self-reported health status, stress levels, safety perception in the neighborhood, physical activity levels, and smoking habits. Social and political views were captured through variables such as political views, religious affiliation, frequency of prayer and religious service attendance, beliefs about life after death, fears related to safety or societal issues, opinions on gun laws, and the level of trust in government and people. Job and work-related data included whether the respondent had lost a job, whether they found a job after losing one, work status of the spouse/partner and coworkers, and job satisfaction. Well-being and happiness variables included overall happiness, happiness in marriage, happiness in cohabitation, and satisfaction with financial situation. Safety and security variables encompassed perceptions of vaccine safety, the impact of COVID-19, views on educational or social issues, perceptions of helpfulness in society, and whether the respondent had been arrested. Other variables included the type of health insurance, use of contraception (specifically condoms), and job finding after job loss.

These variables from the GSS allowed us to examine how social structures, personal well-being, cultural beliefs, and economic conditions interacted to shape the state of the world, particularly in the context of global political and social dynamics. By analyzing relationships between demographics, economic factors, and emotional well-being, we aimed to uncover how societal issues such as job insecurity, income inequality, and family life impacted overall happiness and stress levels. Additionally, we sought to explore how political polarization and social divisions in the current era might be reflected in this data. The goal was to understand how political and religious views shaped perceptions of trust, safety, and societal issues. By studying these factors, we hoped to gain insights into how they influenced well-being and societal outcomes, which could inform more targeted interventions and strategies for policymakers to promote a more equitable and less polarized society.

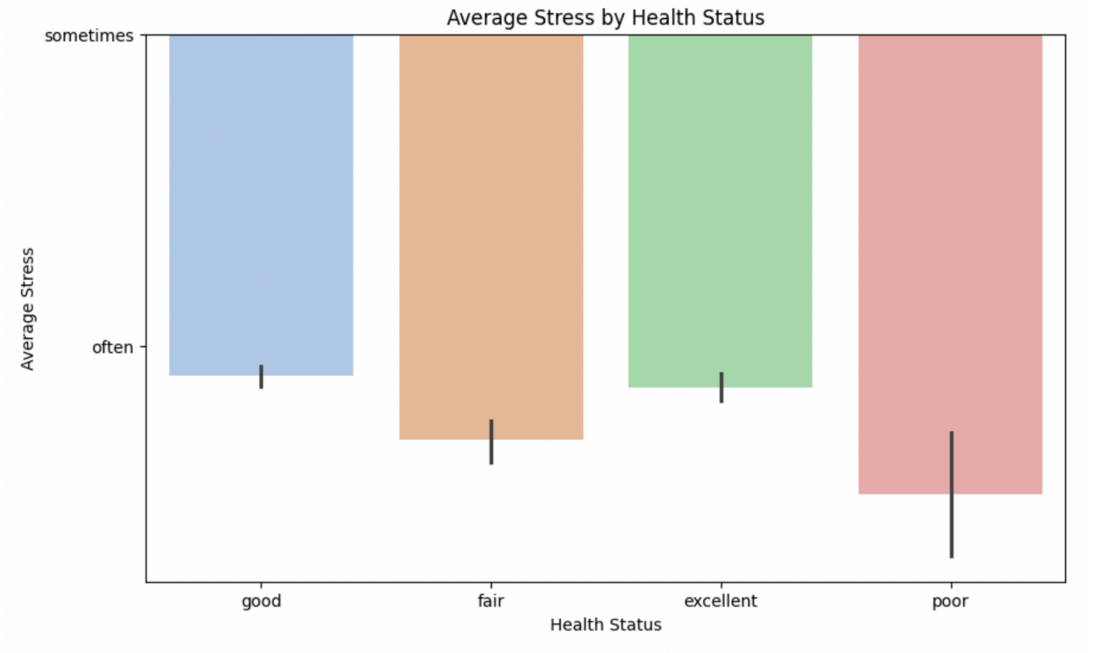
While the dataset appeared relatively clean, several challenges were anticipated. One challenge involved data completeness, as some variables had missing, inconsistent, or incomplete values. Decisions were made about whether to impute these values or drop rows with missing data. Data standardization was also necessary, particularly for variables like political views, religious beliefs, and income, which may have been recorded in different formats or categories over time. Another challenge was multicollinearity, where certain variables were highly correlated with one another, leading to redundancy in our analysis. This was mitigated by dropping variables that represented the same information. Additionally, some relationships between variables were non-linear, making it difficult to visually interpret how they interacted. Careful attention was given to how graphs and trends were read to ensure the correct interpretation of the relationships between variables and the broader picture they represented.

Below are some of the visualizations we created, inspecting the relationships between education level and vaccine safety/political views, the degree of COVID-19's impact on teens, stress level's relationship to health status, the relationship between whether people thought the COVID-19 vaccines were safe and whether COVID had an impact on the participants, and the relationship between health status and race/ethnicity.

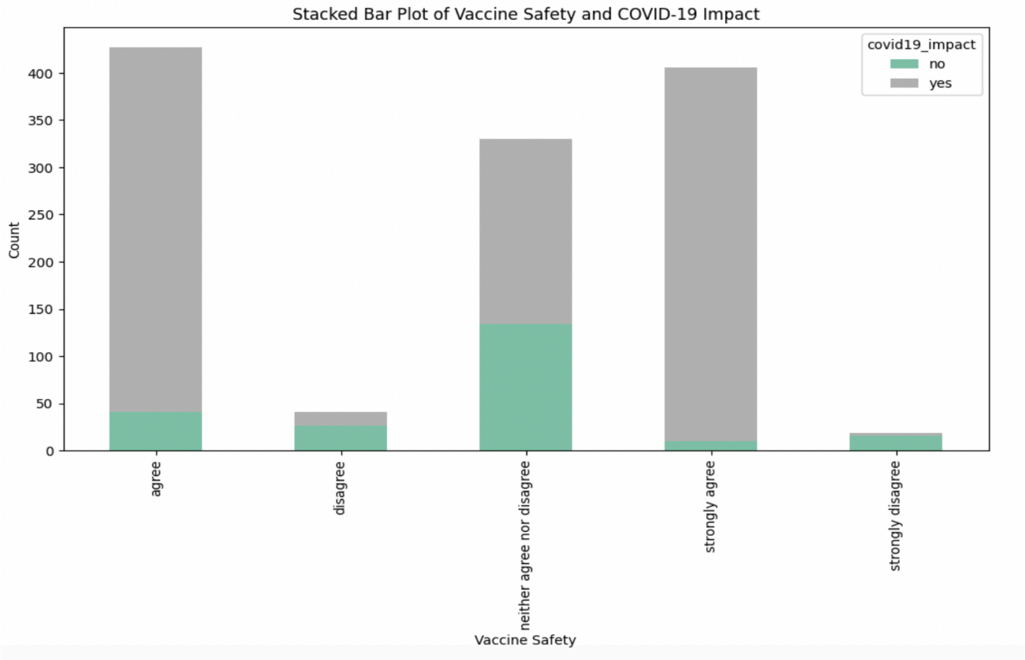
Graph 1



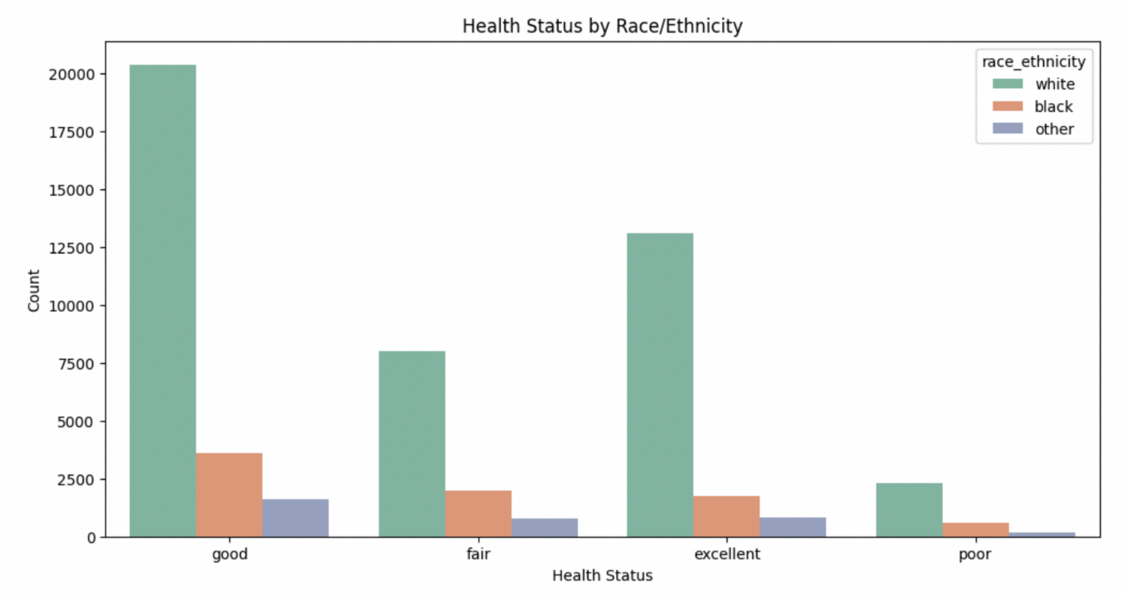
Graph 2



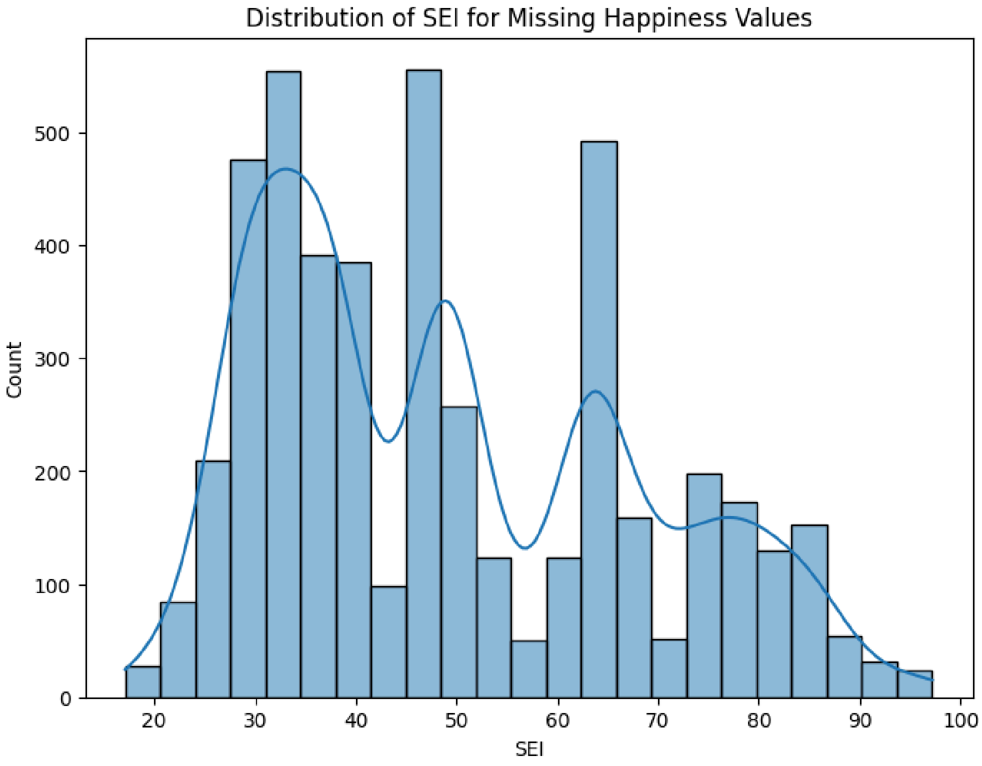
Graph 3



Graph 4



Graph 5



4. Methods

The goal of this analysis is to predict subjective well-being (happiness and stress) based on a variety of demographic, economic, health, and lifestyle factors. Specifically, we aim to understand how income, job satisfaction, stress levels, and other individual characteristics influence happiness and stress levels within the population.

Each observation in the study will represent an individual respondent from the General Social Survey (GSS) dataset. The dataset consists of 72,390 survey responses with 52 variables, which include both categorical and numerical variables, such as marital status, race, income, stress levels, and political views. These variables provide a comprehensive picture of the factors that might influence well-being and societal outcomes.

For this analysis, we plan to use supervised learning techniques to predict continuous (income) and categorical (happiness and stress) outcomes. Specifically, we will employ regression models for continuous outcomes and classification models for categorical ones. For income prediction, we will apply linear regression with Lasso regularization, which will help us address multicollinearity and focus on the most relevant predictors such as education, job satisfaction, and stress levels. To predict happiness, we will use logistic regression to categorize respondents as "happy" or "not happy," based on their subjective well-being, while controlling for income and demographic factors. Additionally, we will apply the K-Nearest Neighbors (KNN) classifier, which is effective at handling complex, non-linear relationships, to predict binary happiness status. Finally, we will use a random forest classifier, which is particularly good at capturing non-linear interactions and relationships between predictors.

Feature Engineering

We plan to preprocess the data in several steps. Missing data will be imputed by replacing missing numerical values with the mean of the respective column and categorical variables with the label "missing." Variables such as income and stress levels, which are on different scales, will be standardized to ensure consistency across the dataset. Categorical variables, including political views and marital status, will be one-hot encoded to make them suitable for machine learning algorithms. These preprocessing steps will ensure that the data is properly formatted for the models we intend to use.

To evaluate the models, we will use several key metrics. For the regression models, R-squared (R^2) will be used to assess how much of the variance in happiness and stress can be explained by the predictors. For the classification models, we will evaluate performance using accuracy, sensitivity, and specificity. We will also use k-fold cross-validation to ensure that the models generalize well to unseen data. Additionally, we will compute the Area Under the Curve (AUC) for binary happiness classification to assess how well the models distinguish between happy and unhappy individuals. These metrics will help us gauge the effectiveness of each model.

Several challenges are anticipated during the analysis. Missing data, particularly in variables like income and health status, will be addressed using imputation techniques. If the missing data is not missing at random, it could introduce bias, but we will examine patterns in missingness and adjust accordingly. Multicollinearity is another concern, especially given the high correlation between some predictors like income and socioeconomic status. We will address this by using Lasso regression, which penalizes irrelevant variables and shrinks their coefficients to zero. Non-linear relationships between variables, such as those between income and happiness or stress and well-being, will be explored using non-linear models like random forests, and transformations like logarithmic scaling will be applied where necessary.

The results will be presented through several key outputs. For linear regression, a regression coefficients table will be provided, including coefficients, p-values, and confidence intervals to help us understand the impact of each predictor on income and happiness. For classification models, confusion matrices will be used to illustrate model performance, showing true positives, true negatives, false positives, and false negatives. We will summarize performance metrics such as R^2, RMSE, accuracy, and AUC scores in a comparison table. Additionally, we will use visualizations, including scatter plots and heatmaps, to present the relationships between key variables, such as happiness and income, or stress and well-being.

5. Results

We performed many exploratory data analyses in order to understand initial trends and relationships in our data. We looked at summary statistics, distributions for numerical variables, and crosstabs and groupings for categorical variables. After understanding these dynamics, we built and tested various models for predicting income and happiness indicators.

Table 1: Crosstab of Happiness by Income Group - Percentages

Happy	**Missing**	**Not too happy**	**Pretty happy**	**Very happy**
Income group				
Low	6.530612	13.654917	54.594929	25.219542
Medium	6.534541	14.665351	49.968613	28.831495
High	7.029015	7.982564	53.500885	31.487536

Table 2: Crosstab of Political Group by Gender - Percentages

Political Group	**Conservative**	**Liberal**	**Moderate**
Sex			
Female	31.231934	28.358770	40.409296
Male	36.694138	27.721289	35.584573
Missing	34.408602	24.731183	40.860215

Table 3: Crosstab of Health by Income Group - Percentages

Health	**Excellent**	**Good**	**Fair**	**Poor**	**Missing**
Income Group					

Low	21.587302	37.423212	13.951763	2.176871	24.860853
Medium	18.560368	32.807222	18.222581	7.159298	23.250531
High	29.062798	38.135131	8.561504	0.892249	23.348318

From Table 1, we see that “pretty happy” is the most common answer in the survey and that people in the highest income bracket tend to be happier overall. Table 2 reveals that most people identify as politically moderate and that men tend to be more conservative, and women tend to be more liberal. Table 3 shows that most people believe that their health status is “good”. Overall, people in the higher income bracket tend to self-report that their health is “excellent” or “good” compared to lower income levels. Interestingly, people in the middle income bracket have higher rates of “poor” health and lower rates of “excellent” health than those in the lowest income bracket.

I. Linear Regression on Real Income using Lasso

For our first model, we decided to do a linear regression using many demographic and behavioral variables to predict real income, a continuous variable. We aimed to understand how happiness and stress levels impact income, controlling for many confounding and relevant variables. We used the following regression equation to create the predictive model.

$$\hat{y} = \beta_0 + \beta_1 \cdot happy + \beta_2 \cdot marital + \beta_3 \cdot educ + \beta_4 \cdot sex + \beta_5 \cdot hours\_worked + \beta_6 \cdot attend + \beta_7 \cdot age + \beta_8 \cdot job\_satisfaction + \beta_9 \cdot mother\_educ + \beta_{10} \cdot father\_educ + \beta_{11} \cdot degree + \beta_{12} \cdot stress + \beta_{13} \cdot physact + \beta_{14} \cdot work\_status + \beta_{15} \cdot sexornt + \beta_{16} \cdot class + \beta_{17} \cdot race\_detailed$$

$$Minimize: \ \Sigma(yi - \hat{y}i)^2 + \lambda \cdot \Sigma|\beta j|$$

$$Where \ \lambda = 5.0$$

We used a Lasso regression method in order to narrow down the relevant predictor variables. With an alpha value of 5, our model was relatively more strict at penalizing irrelevant variables and highlighting the most significant relationships. Even with this high alpha value, most of our variables had non-zero coefficients. The direction of our coefficients for our variables mostly matched our hypotheses, especially for the confounding variables. As we expected, age, education, working hours, being male, degree, job satisfaction, class, and being white all had a positive relationship to real income.

Table 4: Happy and Stress Coefficients:

Stress	**Coefficient**	**Happy**	**Coefficient**
Missing	- 497.9129	Missing	*Base category*
Never	- 426.3183	Not too happy	1.6557
Hardly Ever	- 680.6843	Pretty happy	342.9206
Sometimes	- 251.8080	Very happy	508.3346
Often	530.3346		
Always	*Base category*		

The “happy” variable also has a positive relationship: as happiness increases, so does income. Interestingly, all coefficients for happiness are positive compared to the base category “missing”, indicating that opting out of the happiness survey question is associated with lower income. Our stress variable coefficients show that being stressed “often” is associated with higher levels of income. This might be attributed to the fact that higher-paying jobs and high-achieving lifestyles might come with a moderate amount of stress. Likewise, rarely feeling stressed has a negative relationship with income, compared to the base category of always feeling stressed. It appears that too much stress and too little stress have a negative association with income. In this regression, we also see that being married, moderate physical exercise, and moderate religious engagement are positively associated with higher levels of income. This first regression gives us some introductory insight into the main factors affecting income and reveals a correlation between financial health and overall well-being and happiness.

The model used an 80/20 train-test split to ensure that the model is robust, properly fit, and can predict outside of the training dataset. The R^2 value is 0.185930, and the MSE is 357,070,353.53. These statistics reveal that our model has relatively low predictive power, likely due to omitted variable bias, as occupation, industry, job level, and region directly affect income levels. Many of the variables are correlated, introducing bias due to multicollinearity, and the model currently neglects possible non-linear relationships - age and education have peak income ranges, and interaction or polynomial terms might capture these phenomena better.

II. Logistic Regression on Happy_Binary

$$\mathit{logit}(\hat{Y}) = \beta_0 + \beta_1 \cdot \mathit{marital} + \beta_2 \cdot \mathit{educ} + \beta_3 \cdot \mathit{sex} + \beta_4 \cdot \mathit{hours\_worked} + \beta_5 \cdot \mathit{attend} + \beta_6 \cdot \mathit{age} + \beta_7 \cdot \mathit{number\_of\_children} + \beta_8 \cdot \mathit{health} + \beta_9 \cdot \mathit{smokeday} + \beta_{10} \cdot \mathit{polviews} + \beta_{11} \cdot \mathit{number\_partners} + \beta_{12} \cdot \mathit{inj\_drugs} + \beta_{13} \cdot \mathit{physact} + \beta_{14} \cdot \mathit{stress} + \beta_{15} \cdot \mathit{race\_ethnicity} + \beta_{16} \cdot \mathit{household\_income} + \beta_{17} \cdot \mathit{sexornt} + \beta_{18} \cdot \mathit{financial\_satis} + \beta_{19} \cdot \mathit{class} + \beta_{20} \cdot \mathit{sei} + \beta_{21} \cdot \mathit{neighborhood\_safety}$$

For our next model, we used a logistic regression to predict happiness using various lifestyle factors, while controlling for demographic variables and income-related measures. During our analysis, we determined that most of the missing values for happiness were distributed among lower SEI values, indicating that people with a lower income might be more likely to have skipped this question in the survey (see Graph 5). Based on this data and our own assumptions regarding non-response bias (questions on stigmatized or uncomfortable topics), we determined that we could reasonably categorize the missing values with the “not too happy” observations when making the happy variable binary for analysis purposes. Therefore, we combined “very happy” and “pretty happy” into one category and “missing” and “not too happy” into another category for this analysis and used the “happy_binary” variable as our categorical dependent variable. From this analysis, we see that financial well-being is positively correlated with happiness. While controlling for income-related variables and general demographics, we can see the effect of lifestyle choices and ideological differences on happiness. Since this model includes many complex categorical variables, we selected a few interesting variables and their relationships with the coefficients in Tables 5 and 6 below.

Tables 5 and 6: A Summary of Variable Coefficients

Demographic Variables			**Health/Wellness Variables**		
Variable	*Coefficient*	*Base Category*	*Variable*	*Coefficient*	*Base Category*
Male	0.0023024	Female	Health Poor	-0.6065529	Health Excellent
Heterosexual/Straight	0.2735624	Homosexual/Gay	Stress Often	-0.157383	Stress Always
White	0.2823569	Black	Physical Activity Never	-0.014266	Physical Activity Daily

Neighborhood Very Safe	0.2408572	Neighborhood Very Unsafe	Smoke 40+ Cigarettes/Day	0.0015228	Never Smoked

Lifestyle/Ideology Variables		
Variable	*Coefficient*	*Base Category*
Household Income	1.678216	NA
Number of Children	-0.0690705	NA
Married	0.611649	Divorced
Never Attend Religious Service	0.2417678	Attend Religious Service 2-3 times Per Month
Extremely Conservative	-0.0239223	Conservative

The majority of the variables show coefficients that support our hypotheses. However, the smoking variable and the number of children variable show an opposite relationship than we were expecting. Interestingly, extremely conservative political views are the only category with a negative coefficient (compared to the base category “conservative”), and “slightly liberal” and “slightly conservative” have the highest positive coefficients. This might suggest that relatively moderate political views are associated with higher happiness levels. Surprisingly, smoking 40+ cigarettes per day is associated with higher happiness than being a non-smoker. After controlling for demographic variables and economic factors, we see that lifestyle choices, health and stress, family formation, religiosity, and political views are correlated with happiness in the hypothesized ways, except for a few variables with unexpected coefficients.

For this model, we used accuracy and an ROC curve to measure the predictive power. The accuracy was 0.853847216, and the AUC score from the ROC curve was 0.801. The AUC score evaluates the ability of a classification model to distinguish between classes, with scores over 0.5 indicating a relatively effective model.

III. Classification for Happy and Stress Variables

We applied a K-Nearest Neighbors (KNN) classification algorithm to predict binary happiness status (happy_binary) using the following features, the same as used previously in our logistic regression.

- *Demographics:* household_income, educ, marital, sex, race_ethnicity, age, sexornt, class, sei
- *Health & Well-being:* health, stress, physact, smokeday, inj_drugs
- *Social Behavior:* attend (religious service attendance), polviews, number_partners
- *Family & Work:* number_of_children, hours_worked
- *Other:* financial_satis, neighborhood_safety

We used a classification model in order to avoid the issues of non-linearity and be able to handle complex interactions better than a simple logistic regression. By using the same predictor variables, we can compare the performance of the two models. The accuracy of this classification model is 0.84576598, slightly lower than the accuracy of the logistic regression. The F-1 scores for happy and unhappy are 0.91 and 0.49, suggesting that the model performs better at predicting “happy” than predicting “unhappy” - this is also evident in the confusion matrix results in Table 7 below.

Table 7: Confusion Matrix of Happy_Binary Classification

	Predicted: Not Happy (0)	**Predicted: Happy (1)**
Actual: Not Happy (0)	True Negative (TN) = 11,173	False Positive (FP) = 499
Actual: Happy (1)	False Negative (FN) = 1,734	True Positive (TP) = 1,072

We also built a classification model for the stress variable, which we recoded into three categories: “always”, “moderately”, and “rarely”. In this model, we intentionally excluded variables related to employment and income to determine if stress levels can be predicted somewhat accurately without these factors. The accuracy of the model is 0.625, relatively low compared to our other models. Including financial variables as controls is likely the more effective method of determining how lifestyle and health choices shape stress levels.

IV. Decision Tree and Random Forest Algorithm For Happy Variable

Finally, we decided to use a random forest algorithm to make predictions using the happy variable. We used the happy_binary variable again for interpretability. We used the same variables as previously used for our logistic regression and classification with happy_binary. The accuracy for this new model was 0.8677, higher than the classification accuracy of 0.8457 and the logistic regression accuracy of 0.85384, making this our best-performing model of the analysis. The random forest model handles nonlinear relationships well, can process multiple different types of variables, and is resistant to outliers. Therefore, a random forest outperforms our logistic regression and simple KNN classification model.

Table 8: Confusion Matrix of Happy_Binary Random Forest

	Predicted: Not Happy (0)	**Predicted: Happy (1)**
Actual: Not Happy (0)	True Negative (TN) = 11,530	False Positive (FP) = 142
Actual: Happy (1)	False Negative (FN) = 1,772	True Positive (TP) = 1,034

Our analysis explored a range of models to understand factors influencing happiness, stress, and income. We found consistent evidence that higher income and better self-reported health are associated with higher levels of happiness, while moderate stress and lifestyle factors such as physical activity and religious attendance also play significant roles. The linear regression for income prediction showed limited explanatory power, likely due to omitted variables and multicollinearity, but still highlighted key predictors like education, job satisfaction, and class. The logistic regression and classification models for predicting happiness both performed well, with the random forest model achieving the highest accuracy at 86.8%, confirming its strength in handling nonlinear and complex relationships. Overall, our results suggest that both structural factors (like income and education) and personal lifestyle choices (like health behaviors, family formation, and community engagement) significantly shape individual well-being.

6. Conclusion

In this paper, we explored the complex relationships between demographic, behavioral, and attitudinal variables and three major outcomes: income (via linear regression), happiness (via logistic regression, KNN, and random forest classification), and stress (via multiclass classification). Using General Social Survey (GSS) data, we applied a variety of statistical and machine learning models to extract insights while also addressing common challenges in social science research such as missing data, variable multicollinearity, and the limits of linear modeling.

Key Findings

Across all models and exploratory analyses, financial well-being and self-reported health consistently emerged as the strongest predictors of happiness. The logistic regression, KNN classification, and random forest models all revealed that income, health, and marital status are robust correlates of subjective well-being. Interestingly, some lifestyle variables—like frequency of religious attendance and political views—also had statistically meaningful relationships with happiness. For example, we found that moderate religious attendance and moderate political ideology were both associated with higher happiness, suggesting a potential "balance" effect that could warrant deeper psychological or sociological investigation.

The linear regression model for predicting income using Lasso regression helped highlight key drivers such as age, education, hours worked, job satisfaction, and race, while penalizing less informative predictors. Although this model's R^2 value was modest (0.1859), this relatively low explained variance is not unexpected in income prediction without direct occupational controls. In fact, this limitation helps underline the model's value: it makes clear that income cannot be fully explained by individual demographics and behavior alone; it is shaped by institutional, regional, and occupational factors not present in the dataset.

When predicting happiness using a binary classification model, we found that the random forest algorithm outperformed logistic regression and KNN in terms of accuracy (0.8677). Random forest models are particularly well-suited for this type of problem because they handle complex, non-linear interactions and mixed data types effectively. This suggests that the relationship between happiness and other social variables is not easily reducible to linear or additive terms and supports future use of ensemble methods in sociological modeling. Our classification of stress using a simplified three-category model ("always," "moderately," and "rarely" stressed) produced lower predictive accuracy (0.625), particularly after removing financial and occupational variables. This result itself is insightful: it reinforces how integral socioeconomic status is to self-reported stress and shows that lifestyle factors alone provide an incomplete picture of psychological distress.

Limitations and Future Directions

Several methodological and data challenges emerged during this study, many of which point toward rich directions for future work. One of the more difficult challenges involved interpreting missing values—especially for subjective variables like happiness. We found that happiness non-responses were disproportionately clustered among low-SEI (socioeconomic index) individuals, leading us to classify missing values as "unhappy" in our binary model. While this decision is defensible and helped improve model consistency, it raises broader questions about how nonresponse can be systematically related to well-being. Future work could use imputation methods, latent class analysis, or qualitative approaches to explore what missingness itself might signify.

Many demographic variables, such as education, income, and occupational status, are highly correlated, which introduces multicollinearity into regression-based models. Although Lasso regression helped mitigate this by penalizing unnecessary complexity, there remains value in explicitly modeling the hierarchical or nested nature of these variables (e.g., structural equation modeling or multilevel regression). Further exploration of interaction effects—for example, how the impact of stress on happiness varies by income—could also reveal nuanced patterns masked by simpler additive models.

The modest R^2 in our income model is a reflection of missing variables that are typically strong predictors of earnings (e.g., occupation, industry, geographic location). This limitation points toward a clear path for improvement: future studies should incorporate occupation and job-level data where available. Alternatively, researchers might explore more contextual models (e.g., regional economic indicators, policy environment) to account for external drivers of income inequality.

Some variables had counterintuitive effects. For example, people who smoked 40+ cigarettes per day reported higher happiness than non-smokers, and those who felt stressed “often” had higher income than those who felt stressed “never.” These results may reflect complex causal pathways (e.g., high-income individuals in demanding careers) or unmeasured mediating variables. They also highlight the limitations of purely statistical interpretation without richer context. Future research could triangulate these findings with qualitative data or use causal inference methods (like instrumental variables) to probe mechanisms more deeply. Several variables likely follow nonlinear relationships with outcomes (e.g., income peaks with age, education has diminishing returns). Our use of tree-based models (random forest) partially addressed this, but more explicit modeling of nonlinearity would allow deeper insight. Interaction effects, like the compounded effect of being low-income and in poor health, should also be explored in future analyses.

This project provides a comprehensive, data-driven look at how demographic, lifestyle, and psychological factors interrelate in shaping income, stress, and happiness. While our models demonstrate solid predictive performance, particularly for happiness, they also reveal the limitations of survey-based, cross-sectional data in fully explaining complex human outcomes. Future work should build on these insights, incorporating longitudinal data, deeper causal modeling, and broader contextual variables to develop a more holistic understanding of social well-being. Through continued iteration and model refinement, we can move closer to identifying not just who is happy or wealthy, but why.

---

7. References/Bibliography

General Social Survey 2022 Data
https://gss.norc.org/us/en/gss/get-the-data.html
General Social Survey 2022 Codebook
https://gss.norc.org/content/dam/gss/get-documentation/pdf/codebook/GSS%202022%20Codebook.pdf

Acknowledgments: We used ChatGPT to help us with parts of this assignment.