Happiness and Outlook on Life

Tiara Allard, Jackson Glass, Connor Rose, Kayla Sprincis

Summary

This project explores data collected in the General Social Survey (GSS) to analyze participants' perceived happiness, and determine which factors most strongly affect one's happiness. We examine a variety of data collected in the 2022 GSS, mostly relating to socio-economic status, interpersonal relationships, and participants' views of those in their community.

The project focuses on the HAPPY variable, a variable in the GSS which asks participants to rank their happiness on a scale from 1 to 3, with 1 being 'Very Happy' and 3 being 'Not Too Happy'. We explore correlations between the HAPPY variable and other variables such as age, income, happiness with partner, and variables which ask the participant about their interpersonal relationships and trust of those around them. These correlations are expressed through cross tabulations, as well as visualizations which show how happiness is distributed among different question responses. We also explore the implications of using GSS data, and the trustworthiness of social survey data for making general statements about Americans.

Our collected results show that happiness is influenced heavily by social factors, especially those related to socioeconomic status and one's relationship with their partner. Data collected through surveys such as GSS are difficult to make definitive statements with, as they depend on those surveyed, though the sample of respondents is large enough to where we can see clear trends, even if additional data would be needed to confirm these ideas. The GSS data cannot tell us underlying reasons about why these factors make someone happier, though the relationships between HAPPY and other variables do allow us to confirm and deny assumptions about what makes one happy, which could be useful in designing public policy, and in one's own interpretation of the world around them.

Data

The General Social Survey contains dozens of questions asked to participants, many of which would provide little insight into understanding happiness of participants. We narrowed the dataset down to 17 variables - including 'HAPPY' - to conduct our analysis on.

The variables we selected were based on a couple of key concepts we chose to explore. First, we included variables that give a general demographic overview of the participant, including age, inflation-adjusted income, and party ID. Second, we included variables which captured the social lives of participants, using GSS questions which ask about how often participants partake in various social events. Third, we included variables which describe the participants' outlook on life and their community, as well as their interpersonal relationships.

This grouping of variables belonging to three main themes allowed us to narrow down the dataset into 17 variables, which is enough to allow a detailed analysis, while avoiding having unnecessary or redundant variables which impede the interpretability of the data. Several of the variables were

also combined into single variables, either because the premise of the question was similar enough to warrant it, or because the question was exactly the same, but the data was collected differently. These combinations brough the overall number of variables down to 15 columns. The details of these combinations will be explained in further detail after each variable is summarized.

The relevant post-cleaning variables are briefly summarized, including the main idea of the question asked, as well as the scale it was asked on.

- AGE - The age of the participant, values range from 18 to 89.
- CONINC - A numerical variable that represents the inflation adjusted income of the participant. It has a middle value for each of the 19 income groups as defined by the GSS.
- CONINC_CAT - Modified version of the GSS data set variable. A categorical variable with 19 income groups that have varying increments and a max of $999,999.
- PARTYID - The political party of the participant. 0 means strong democrat, ranging to 6 meaning strong republican, 7 is other
- PARTYID_CAT - The categorical representation of PARYID

All of the 'SOC' variables represent social activities of the participant. Each of the range from 1 (most often) to 7 (least often)

- SOCCOMMUN - How often the participant spends social time with a friend who lives in their neighborhood.
- SOCFREND - How often the participant spends social time with a friend who lives outside their neighborhood.
- SOCBAR - How often the participant visits a bar or tavern.
- SOCREL - How often the participant spends time with relatives.

The remaining variables describe the participants happiness in their relationships with others in their community.

- HAPPY - Describes how happy the participant views themself. Ranges from 1 to 3 with 1 meaning very happy and 3 meaning not too happy.
- LIFE - Does the participant view life as exciting. Ranges from 1 to 3 with 1 meaning exciting and 3 meaning dull.
- HAPPARTNER - Is the participant happy in their relationship. Ranges from 1 to 3 with 1 being happy and 3 being not too happy.
- FAIR - Does the participant believe people are fair to them, or try to take advantage of them. Ranges from 1 to 3 with 1 meaning taken advantage of, 2 meaning fair and 3 meaning 'depends'.
- TRUST - Does the participant believe people can be trusted. Ranges from 1 to 3 with 1 meaning trustworthy, 2 meaning untrustworthy and 3 meaning 'depends'.
- HELPFUL - Does the participant believe that in general people try to be helpful or selfish. Ranges from 1 to 3 with 1 meaning helpful, 2

meaning selfish, and 3 meaning 'depends'.

Of the set of 13 columns which we used in the analysis and visualization, a few were found to be very useful in analysis. The variables HAPPARTNER, AGE, CONINC, and LIFE are of particular interest, as will be seen in the results section.

Cleaning the data from the original GSS format proved challenging in many ways. The original data file for the GSS is in the .sas7bdat format, which we had never previously encountered. To read the file in to our notebook, a pandas function for reading SAS files had to be used.

Working with social survey data also presents unique challenges. The participants are not required to answer every question, and may choose to refrain from answering any question, which leads to larger numbers of missing values in the dataset.

To deal with these missing values, we developed a system for what rows should be kept, and which should be removed. Since our analysis was focused on understanding the HAPPY variable, any row which had a NaN value for HAPPY was removed from the dataframe, as it provided little to no insight into the initial question. 24 Rows had NaN values for the HAPPY variable and were dropped accordingly.

As previously mentioned, several of the initial 17 variables were combined together because they either asked the same question, or similar enough of a question to warrant combination. Initially, the variables FAIR, TRUST, and HELPFUL all had V and NV counterparts, originally being two separate variables, of the form FAIRV, FAIRNV. The variables with the suffix -V/-NV represent a different, more updated, data collection method. To keep consistency across the data set only the variables with the -V/-NV suffix were considered. Since the question was the same in both the V and NV columns, we combined them into a single column to make analysis of the question easier. Though there would have been more data available for analysis if rows with the older data collection method were used, we believe that it is important to keep situations involving the data as consistent as possible. This prevents results that are incorrectly screwed.

Additionally, rows which had a NaN value for the FAIR_C variable were also removed, as those rows represented participants that were interviewed with the old data collection system, and did not provide any useful data for the analysis.

The variable HAPPARTNER is also a combination of two variables. The original data had variables HAPMAR (for happy marriage) and HAPCOHAB (for happy relationship). The question asked in both variables is very similar, and only refers to those with a romantic partner that they are currently living with. As such, combining them into a single variable simplifies our ability to analyze the impact of being in a happy relationship, regardless of marriage.

Overall, the data were cleaned with the intention to narrow down the initial large GSS dataset to a manageable number of variables, and to ensure that those variables were not redundant. The remaining 15 variables allow us to look at several factors relating to happiness while not having too many variables to work with.

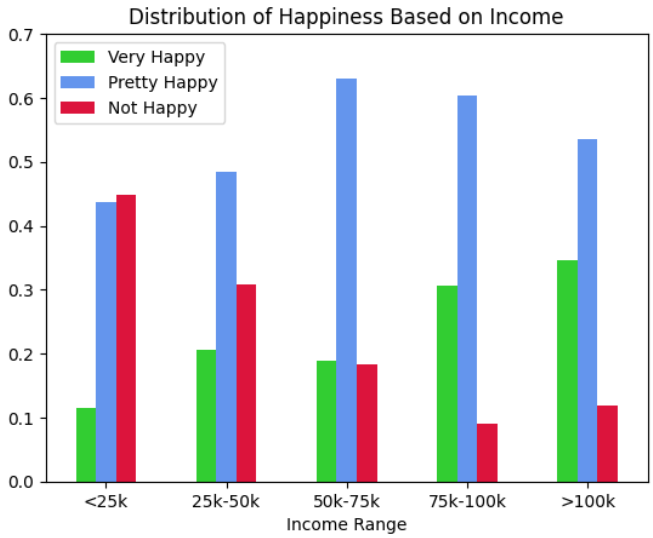
Results

Using the cleaned dataset, we conducted exploratory data analysis to isolate variables which had strong positive or negative correlations with the HAPPY variable. To properly examine these correlations, we created a correlation matrix between all of the relevant variables in the dataset. The correlation matrix gave us a good starting place for interesting visualizations, particularly as it showed strong positive correlations between HAPPY and HAPPARTNER, with a 0.46 correlation, as well as HAPPY and LIFE, with a 0.30 correlation. The full correlation matrix of the dataset can be found in the Jupyter Notebook. Additionally, we examined the distributions of each variable, to get a better idea of the nature of the dataset.

Some of the more important distributions of variables were the distributions for HAPPY, CONINC, and AGE. Histograms overlaid with KDE plots can be seen in the Jupyter Notebook. The distribution for HAPPY shows that 591 responses are moderately happy, 304 responses are unhappy, and 241 responses are very happy. There is a peak towards moderate happiness, with a similar number of happy and unhappy responses on either side. The CONINC distribution shows a steady decrease as income increases, with a spike in higher income ranges. Finally, the AGE distribution is roughly constant, with small peaks at about 30 and 65 years. Confirming that age is well distributed and in line with the population tells us that the correlations among the dataset are more likely to be applicable to the broader American population.

One important feature of the data that greatly impacted our chosen visualizations was the categorical nature of many of the data. Though the entries are represented as numbers, each number represents a category (such as 1 being very happy in the HAPPY variable). As such, we focused more on visualizations which work well with categorical comparisons, such as bar charts and heatmaps, rather than other visualizations such as scatterplots. We ultimately settled on 4 visualizations that display the most important features of the dataset well, with respect to the HAPPY variable

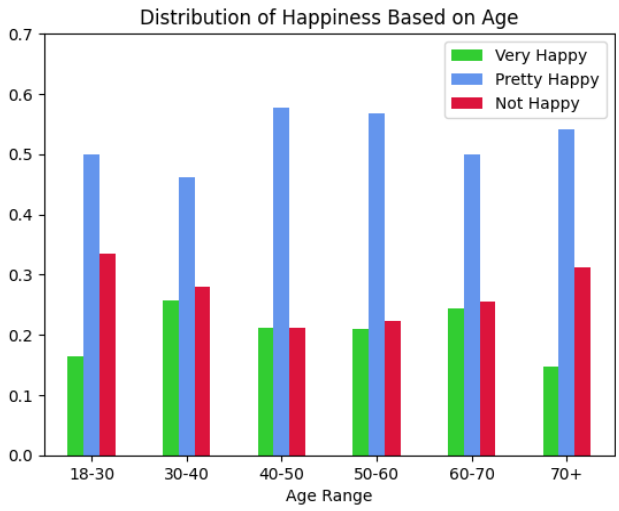
The distribution of happiness in relation to the binned inflation-adjusted income variable shows the clear relationship between income and happiness, as pictured below. Each income bin is normalized, such that the sum of the density of happiness responses in each income bin is the same, making it easier to compare across incomes.



There are a few notable features of this visualization, which tell us about the relationship between income and happiness.

As expected, the number of participants who self reported as very happy increases with increasing income, and the number of reported not happy responses decreases generally with income (Except from 75-100k to >100k, which potentially tells us that if someone is unhappy while making a comfortable amount of money, more money is unlikely to make them happier). The middling happiness category peaks at middling income ranges, and is slightly higher at higher incomes than lower incomes. From this, we also see that the magnitude of 'not happy' responses is greater at the <25k range than the magnitude of 'very happy' responses is at the >100k range. This discrepancy could mean that the unhappiness that comes from having not enough money is greater than the happiness from having more money, potentially due to the stresses induced by lower income. This visualization shows the not-unexpected result that having a higher income, and the ability to live comfortably generally positively correlates with happiness.

We created another similar bar chart, comparing age and happiness, grouped by the three possible responses for happiness.



The relationship between age and happiness is less obvious than that of happiness and income, though there are notable features of the plot which give insight into happiness by age. Similarly to the previous bar chart, each age range bin was normalized to plot with density instead of count.

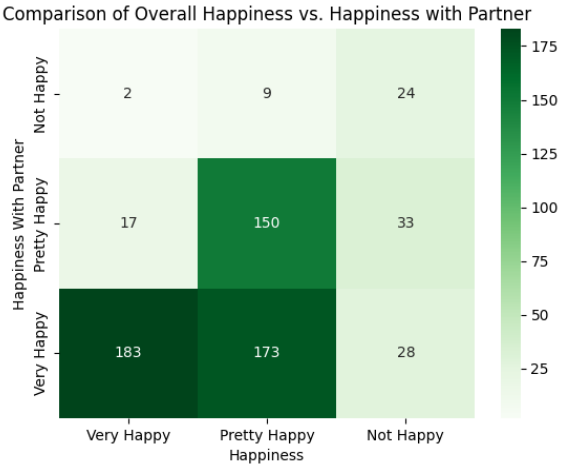
Unlike with income, the relationship between age and happiness is not monotonically increasing or decreasing. Instead, the very happy category has two distinct low points in the 18-30 and 70+ range. For 18-30 year olds, they are likely dealing with the stress of establishing a new career, which also means a likely lower income, which correlates with a smaller very happy category. For the 70+ category, many respondents are likely beginning to deal with problems associated with aging, such as chronic illnesses and separation from their community, leading to a higher proportion of not happy respondents.

There are high points for 'very happy' responses in the 30-40 and 60-70 age ranges. For those in the 30-40 range, it is possible that they are finding greater career stability, or settling down with a partner (which we will see positively correlates with happiness as well). Those in the 60-70 age range are likely beginning to reach retirement age, allowing for greater freedom.

These relative low and high points of happiness correlate with the approximate ages of major life changes, which suggest that when one is in a period of change, their happiness is more likely to be variable, relative to average. Though, we cannot definitively say that specific life events are

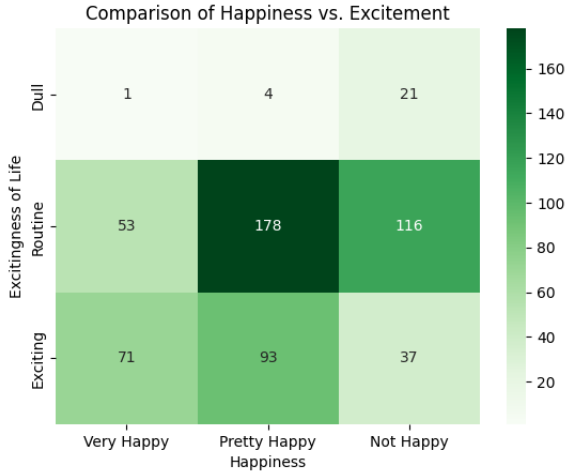
causing happiness or unhappiness, as that information is outside the scope of the GSS.

To examine the relationship between happiness and the respondents happiness in their relationship, we used a Heatmap, which allowed us to visualize the crosstabulation between these two variables. The heatmap takes happiness on the x axis, and happiness with partner on the y axis. From this heatmap, we can see a clear connection between the happiness of one's relationship and their overall happiness. Of responses we consider, only 2 listed them-



selves as being very happy despite not being happy with their partner. Of those who were not happy with their partner, 68.6% were also not happy in general, whereas 47.7% of people who were very happy in their relationship were very happy in general, compared to an average of 32.6% of the sample being very happy. Being in a happy relationship increases the likelihood that the participant is very happy in general. Another interesting result of this heatmap is that those who list themselves as pretty happy with their partner, largely list themselves as pretty happy in general, with 75% of entries falling into the pretty happy category.

We similarly used a heatmap to look at the relationship between the HAPPY and LIFE variables. LIFE is a measure of how exciting the participant views their daily life to be. The correlation matrix we created during our EDA suggested that there is a correlation between excitement and happiness, with a correlation of 0.30. To further explore this correlation, we turned the crosstabulation of LIFE and HAPPY in



to a heatmap. This heatmap gives us several interesting results. First, similar to the previous heatmap, there is a large amount of data which fall into the very center of the grid. One potential reason for this is that both the distributions for HAPPY and LIFE have peaks at the center, and so the most likely value in their cross tabulation will be the center of the grid. Those that described their life as routine are less likely to be very happy, but likely to be moderately happy. Among those that are very happy, a small majority of them describe their life as routine. One might expect there to be a larger clustering of very happy results in line with exciting life, but instead we see a cluster more towards the center, with those who have exciting lives predominantly describing themselves as moderately happy.

We also utilized cross tabulations between Happy and a few variables related to one's feelings about those around them, using variables such as Trust, Helpful, and Fair. Each of these cross tabulations can be found in the Jupyter Notebook. From these cross tabulations we see that in general, for Helpful and Trust, happy respondents are generally split between believing a positive (i.e. people are helpful or people are trustworthy) response and believing the negative response. For respondents who were unhappy though, the majority chose the negative response in the associated Helpful and Trust variables. From this, we see that the participants' outlook on Helpful and Trust don't necessarily mean they are happier, but they can tell us about unhappiness. Another interesting result can be found in the cross tabulation between Happy and Fair. A large number of respondents who listed themselves as unhappy also reported that they believe people would take advantage of them given the chance. These cross tabulations tell us about the relationship between participants' social interactions and their happiness.

Conclusion

Our analysis of happiness provides us with several key insights into the relationship between one's happiness, their socio-economic status, and their social connections to those around them.

Based on our analysis, we believe that our claim that happiness is influenced by social and economic factors is correct. Our analysis and visualizations clearly show that happiness is not static, and that the participants' relationship to those around them, as well as their socio-economic status have an influence on their overall happiness.

To support this claim, we can point to the clear correlations between happiness and the variables happy partner, income, life excitement, and fair, among others. With these variables there is a clear positive trend between happiness and the respondents' relationship with their community.

Though we see this positive result for our claim, it is important to note that there are variables which we expected to have different relationships to happiness. For example, as mentioned in Results, cross tabulations between Happy and other variables show a similar number of happy respondents choosing the positive and negative answers for the Fair, Helpful, and Trust variables.

Despite these variables which have a different relationship to happiness than expected, we believe that overall our result is a useful one. The usefulness of our analysis is based on the idea that we can use GSS data to statistically confirm or deny the correctness of our assumptions about factors which are correlated with happiness.

Many of the results we present in this project may seem trivial or obvious, such as the correlation between happiness and higher income, though being able to confirm this assumption provides a useful basis for further explorations of happiness. Similarly, as previously mentioned, certain variables which one may expect to correlate strongly with happiness have a much smaller effect.

When working with social survey data, the biases of the one working with the data can often be overlooked, which can

lead to inaccurate and harmful results, so despite the expected nature of many of our results, we believe that the utility of our analysis is in confirming or denying assumptions about what factors correlate most strongly with happiness.

Despite what we believe is a useful result, there are many related questions not covered by our project, or by the GSS. One difficulty we had in working with the GSS happiness data is that they rank happiness on a 1 to 3 scale. This ranking means 3 fairly broad and ill-defined categories that are highly subjective. Though we can make broad statements about happiness based on these categories, the specificity of our claims is limited by the small scale of the data, as where a given respondent draws the line between each category is vague. A more in depth analysis of happiness, in which we collected our own data, would have a more empirical understanding of happiness.

We also cannot make definitive statements about whether happiness is caused by these factors, or being happy causes these factors. Since the GSS does not specifically ask respondents about the sources of their happiness, it would be against the goal of the project to make assumptions about whether a given factor causes happiness, or whether it is a result of happiness. A more in depth project should analyze respondents' views on what they believe causes their happiness or unhappiness.

These limitations are a result of the format of the GSS, which is intended to give a broad overview of Americans' opinions on a variety of topics. A more in depth study focused solely on happiness would need questions designed to analyze whether these relationships are causal, as well as questions which provide a less-vague interpretation of happiness, potentially asking multiple questions about different interpretations of happiness.

Analyzing a concept such as happiness presents many challenges, such as dealing with large numbers of missing values due to survey data, the categorical nature of many variables in the dataset, and the openness to interpretation of many of the survey questions. Despite these challenges, we believe that our analysis successfully provides a foundation for confirming or denying assumptions about how various social and economic factors correlate with happiness.

Appendix

1. Correlation Matrix for our Cleaned Dataset. A larger version is also present in our associated python notebook (Group16Final.ipynb)
2. Histograms/KDE plots for HAPPY, AGE, and CONINC, as well as additional plots for each variable are present in our notebook.
3. Cross Tabulations between Happy and the variables Fair, Trust, and Helpful can be found in our final notebook.
4. In our data cleaning, we cite the following links which were used to read in the SAS format, and replace NaNs with 0s.
    a. https://www.marsja.se/how-to-read-sas-files-in-python-with-pandas/

b. https://sparkbyexamples.com/pandas/pandas-replace-nan-values-by-zero-in-a-column/#:~:text=Use%20the%20DataFrame.,but%20returns%20a%20new%20DataFrame.