General Social Survey Project

Catherine Xu, Jillian Haig, Erin Moulton, Riley Tomek, Ashrita Kodali, Rushvi Patel

Summary:

In the following report, the team chose to investigate the relationship between religious affiliation and overall life satisfaction. In light of the personal belief systems and community that religion promises to its believers, it is hypothesized that maintaining a religious belief system ensures an overall more satisfactory quality of life. In order to address the question, the report dives into a data source and presents findings relating religious beliefs and general happiness at any given point in time, across a variety of age groups. Utilizing the General Social Survey (GSS) 6 variables were selected corresponding to the research question labeled as: happy, relig, attend, relig16, age, and income. GSS is an ongoing annual survey that collects various information in relation to American's social attitudes, lifestyles, economic status, and more. Each variable has continuous data covering from 1989 to 2022. After selecting the variables to represent the research question appropriately, the team began the variable cleaning process and data wrangling. For each variable, the team removed the NA's, added new columns if necessary, and combined responses to fix typos or abbreviations. Once the team had reached a concise dataset, we created visualizations to find potential relationships between the variables. Out of about 72,400 individual responses, we were able to find that the majority of respondents were religious, with only 7,500 respondents who were not. The large majority of individuals with religious beliefs provided a great foundation to answer the research question, as the sample size ensures a variety of responses from all backgrounds and beliefs.. Using a bar chart analysis, the results found no connection between an individual's specific type of belief in religion and their relative happiness. To further answer this question, we created a proportion plot that compared

the specific religious affiliations to their level of happiness. Results showed similar responses across religious groups, indicating no significant difference in happiness among them. In answering our question, we also chose to investigate the frequency of religious services attended by the respondents. Using a bar chart and violin plot to analyze the data, we found that a significant proportion of individuals who never attend religious services are “not too happy” in life, with most respondents with high frequency of attendance saying they are either “very happy” or “pretty happy.” A violin plot was also used to graph how income relates to an individual being religious or not, resulting in similar results for both kinds of people among different income levels. Lastly, to further assess the influences on religion, we created a boxplot that visualized the relationship between income levels and belief in religion. This plot showed variability in income among individuals identifying as religious, whereas those identifying as non-religious tended to have less variability and centrated in a higher income range. Via exploratory data analysis and data wrangling, the following report outlines the relationship between religious beliefs and personal quality of life across a range of demographics.

Data:

The relevant variables for analysis present in the report and notebook are happy, relig, attend, relig16, age, and income, and the data was pulled from the general social survey (GSS) which is publicly available online. The “happy” variable corresponds to the happiness level of the respondent, expressed as either “not too happy,” “pretty happy,” or “very happy.” To clean this variable, the NA’s were dropped. The relig variable represents the religion of the correspondent as a string. For conciseness, the religions were renamed into abbreviated terms; this also made the visualizations more clear, as the names were previously too long to fix clearly on a graph’s axis. Some of the respondents identified a form of Christianity that was not

Protestantism or Catholicism (such as Orthodox and Inter-nondenominational), so these instances were grouped into Christianity, under the variable “christ”. There were also observations in which the respondent recorded “relig” as their religion, so these were dropped. From the relig variable a new variable was created called relig_yn. This variable was created to define if respondents are religious or not. If the respondent put “none” as their relig response, then they’re classified as not religious and vice versa. The attend variable measures frequency of attending religious services and is expressed in terms of a string, where “never” represents the lowest frequency and “several times a week” represents the highest frequency. While the attend variable was already clean, a new variable named “attend_num” was added that created a numerical scale from 0-8 based on attendance frequency; “never” corresponds to 0 while “several times a week” corresponds to 8 which made visualizing the relationship between attendance and other variables more clear. The religion in which each survey respondent was raised in is measured in the “relig16” variable; for clarity, this variable was renamed “relig_fam.” To prepare this variable for analysis, it was cleaned in the same manner as the “relig” variable in which the religions recorded were abbreviated. We did not end up using this variable for any visualizations as we felt the relig_yn and relig variables were most useful in answering our question. The “age” variable represents the age of each survey respondent. Each observation included the age in the form of a string, so these were converted into numeric for analysis. Ultimately, we did not use age for any visualizations as it did not provide useful insights into happiness level and religion**.** Finally, the income variable includes the income group in which the respondents’ families fell into from the previous year (2021) before taxes, measured in US dollars. In the dataset, family incomes were string types, and they included dollar signs and ranges of income levels which made analysis difficult without first cleaning the variable. To do this, a new variable called “income_num” was

created that measured income as integers instead of strings. For observations that included a range of numbers (i.e. “7000 to 7999”), the average of each of the numbers taken, and this number, in integer form, replaced the string.

Results:

The first visualization created is a bar chart representing the religion variable, relig (Figure 1). Of the 10 levels for religious groups, the most prominent religion is Protestantism, followed by Catholicism. Most individuals represented seem to follow a religion, with about 7,500 individuals stating that they are not affiliated with any religion.

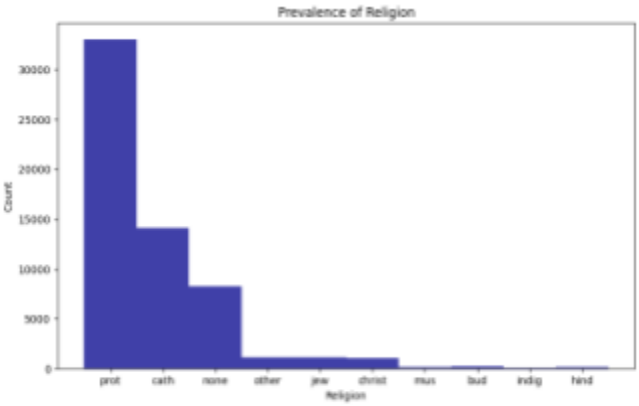

Figure 1: Prevalence of Religion

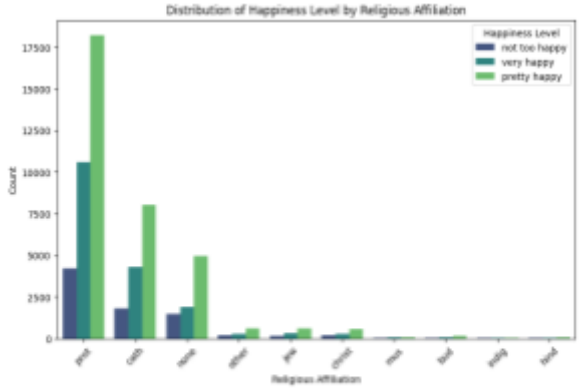

Figure 2: Distribution of Happiness Level by Religious Affiliation

After getting an understanding of the respondents religious affiliations, a bar chart (Figure 2) was created comparing religious affiliation to their relative level of happiness. The results of this visualization were difficult to evaluate since the majority of the religious affiliations had significantly less respondents and varying group sizes.

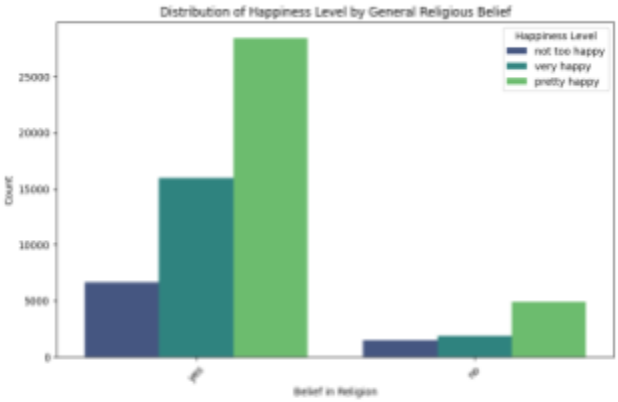

Figure 3: Distribution of Happiness Level by General Religious Belief

The previous visualization inspired the creation of a new variable, relig_yn. This variable was created to define if respondents are religious or not; this new variable was then used to compare an individual's belief

in religion and their relative happiness. While there are nearly 27000 individuals and 15000 individuals who are ‘pretty happy’ and ‘very happy’ respectively, the relative rates of happiness between those who are religious or have belief in religion and those who do not are similar. It appears that belief in religion and happiness level seem to be largely unrelated. This new bar chart, Figure 3, created a more concise visualization that provided a stronger answer to our research questions.

While Figure 3 provided us with a good visualization based on an individual’s belief in religion, we felt it was still necessary to see how specific religious affiliation groups scored their happiness. We chose to create a bar plot represented in proportions rather than the previous bar plot using counts. This proportion barplot takes into account the varying sizes of the different groups and allows for easier comparison. From this visualization (Figure 4) we can see that there is no drastic difference in happiness level among religious affiliation. The only group that seems to have lower levels of “very happy” responses, is the indigenous religion group. While there are slight differences, we can conclude that different religious affiliations have little to no influence on happiness levels.

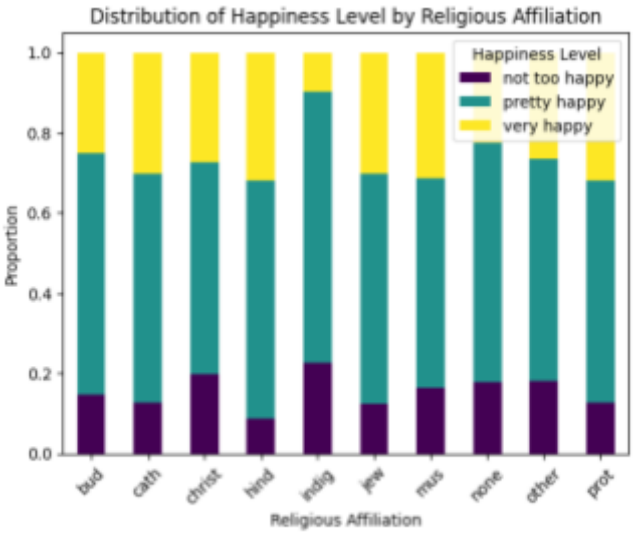

Figure 4: Distribution of Happiness Level by Religious Affiliation

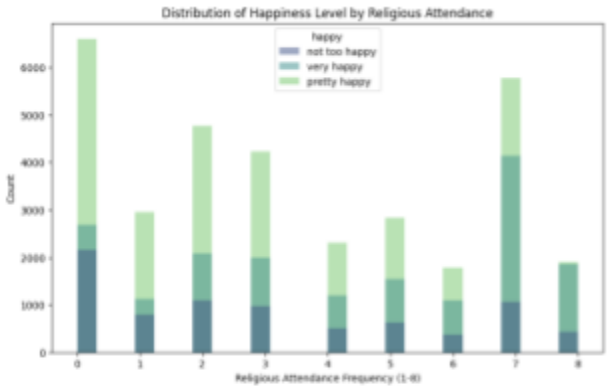

Figure 5: Distribution of Happiness Level by Religious Attendance

Figure 6: Violin Plots of the Distribution of Attendance by Happiness Level

To see how frequency of religious attendance relates to happiness, we created the two graphs shown above. The leftmost bar graph above (Figure 5) measures the happiness level of individuals, varying with how often they attend religious services on a scale from 0-8, with 0 being individuals who never attend religious services and 8 being those who attend several times per week. Based on this visualization, about a third of those who never attend services state they are “not too happy” in life. Those who attend religious services the most often have a proportionally lower response level of being “not too happy,” and about three-quarters of these respondents claimed they were “very happy” in their lives. The level of respondents that stated they were “not too happy” appears to proportionally decrease as the scale increases, indicating that those who attend religious services more often seem to be happier than those who do not attend and those who attend infrequently. This is further justified by the violin plot above on the right (Figure 6) in which we see that those who are “not too happy” also do not attend religious services (the area near the value 0 is the largest). However, the individuals who reported that they were “very happy” or “pretty happy” had a more of their distribution skewed to the left and had higher frequencies of attending religious services.

In determining if religious affiliation affects overall life satisfaction, we also believed

income should be examined to study whether this variable affects if someone is religious. This boxplot (Figure 7) displays the spread of income by an individual's belief in religion. The income variable was initially categorical, but was modified to become quantitative. We wanted to see if there was a particular range of individuals within similar income ranges that were more religious than others. The graph indicates that there is a larger variance of income with people that believed in religion. This suggests that those who are religious span across a more diverse range of income. On the other hand, individuals without religious affiliations predominantly concentrated in the higher income range of around $17,000 and above. By factoring in income and its role in religion and life satisfaction, we are able to see an example of the complexities in answering our question. There are lots of factors that affect religious affiliation and life satisfaction, and by putting multiple variables together, we are able to get a more comprehensive picture.

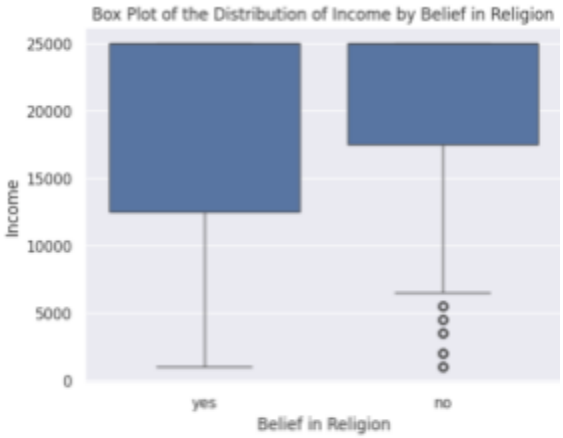

Figure 7: Box Plot of the Distribution of Income by Belief in Religion

Conclusion:

In all, we aimed to determine whether variables such as religious affiliation, attendance to religious services in conjunction with variables such income and age have a relationship with overall life satisfaction. Using the general society survey, we were able to find 6 variables that were important to our analysis. We proceeded to remove observations that had NA's and fixed typos or abbreviations whenever it was necessary. Following the data cleaning, we proceeded to create numerous visualizations and tables to determine whether some of the variables we were looking at had a significant relationship with our response variable, life satisfaction. First, we

created a bar chart to determine the spread and prevalence of religion and saw that Protestantism was the most prominent. We proceeded to make bar graphs of the happiness variable by an individual's religious affiliation and belief in religion and saw that the relative happiness levels were the same regardless of affiliation and belief. However, it was difficult to come to a conclusion as the counts were not similar across each level for the religious affiliation and religious belief variable. In order to alleviate the counts issue, we proceeded to create a bar chart using proportion instead and saw that among all the religions the proportion of individuals who were happy and those that were not were relatively the same. Afterwards, we determined that those who attend religious services had a greater overall life satisfaction in comparison to those that did not by creating a proportional bar chart and a violin plot. After plotting the income variable by happiness level, we saw that the centers for each boxplot for each level of happiness were quite different from one another.

In all, there are potentially significant relationships between variables such as income level, attendance of religious services, and belief in religion with overall life satisfaction. Criticism of the exploratory analysis might include that categorical variables were forced to become quantitative and missing observations were simply dropped. Even though categorical variables such as the variable happy and income were forced to become numeric, it was easier to visualize and understand trends when these variables became numerical quantities as we could now compute statistics (although biased) such as the mean, median, etc. Furthermore, while we were cleaning the data, it was much easier to remove missing values rather than trying to fill them with our best estimates. If we attempted to estimate or impute missing values we could have potentially caused biased estimates and results in the future when we created graphs or calculated statistical quantities such as the median or mean. Furthermore, since the sample size

was quite large, removing the NA's still resulted in variables having over 1000 observations, which was adequate for analysis. While the exploratory data analysis is still quite rudimentary, it provides important insights on how life satisfaction may be correlated with certain variables. Further analysis should be carried out by determining whether there are any interactions between the explanatory variables that are strongly correlated with life satisfaction. For instance, it might be useful to see how income level and age might influence overall life satisfaction. Additionally, it is important to conduct further analysis on other variables. As an example, it might be important to test variables such as marital status/relationship status, gender, employment, and more against overall life satisfaction.

Appendix:

Table 1: Summary of Age by Happiness Level

	age							
	count	mean	std	min	25%	50%	75%	max
happy_num								
0.0	8017.0	46.942622	17.501210	18.0	32.0	45.0	60.0	89.0
2.0	33174.0	45.396545	17.170852	18.0	31.0	43.0	58.0	89.0
3.0	17725.0	47.026629	17.358704	18.0	32.0	45.0	61.0	89.0

Table 2: Summary of Happiness Levels by Income

	happy_num							
	count	mean	std	min	25%	50%	75%	max
income_num								
1000	830.0	1.603614	1.145921	0.0	0.0	2.0	2.0	3.0
2000	1359.0	1.705666	1.073261	0.0	0.0	2.0	2.0	3.0
3500	1267.0	1.737964	1.057959	0.0	2.0	2.0	2.0	3.0
4500	1149.0	1.816362	1.033520	0.0	2.0	2.0	2.0	3.0
5500	1270.0	1.870866	1.001505	0.0	2.0	2.0	3.0	3.0
6500	1178.0	1.822581	1.002088	0.0	2.0	2.0	2.0	3.0
7500	1263.0	1.871734	1.021644	0.0	2.0	2.0	3.0	3.0
9000	2192.0	1.840785	1.010089	0.0	2.0	2.0	2.0	3.0
12500	6565.0	1.938005	0.963199	0.0	2.0	2.0	3.0	3.0
17500	5030.0	1.993241	0.922506	0.0	2.0	2.0	3.0	3.0
22500	5226.0	2.010333	0.926117	0.0	2.0	2.0	3.0	3.0
25000	31963.0	2.130182	0.847840	0.0	2.0	2.0	3.0	3.0