- Change histograms to use merged data frame (‘df’)
- ANALYZE (Data section of write up)
    - Mention the scale of graphs in the write up → 1 is the least happy, not the happiest
    - Merging the dataframes, combining for the kernel to show
    - Mention the different categories for ‘null’ responses, explain that for our research question, the difference was unimportant

- RESULTS
    - Paste the graphs
    - Because the kernel density plot is not including nulls, the fact that a different number of people responded in 2018 and 2022 is not a factor→ the graph shows the percentage out of the actual responses
        - Histograms → the number of responses for 1, 2, and 3 are shown in the bins

ADD: MERGED DATAFRAME HISTOGRAM
EXPLAIN: why histogram was less effective at communicating actual data than kernel density plot

Charlie: Results Section
Tori: Summary

The purpose of this paper was to analyze the data provided by General Social Survey that gauged a respondents’ level of happiness. The survey asked participants “Taken all together, how would you say things are these days--would you say that you are very happy, pretty happy, or not too happy?”. Respondents then rated their happiness on a three-point scale, where 1 represents "Very Happy", 2 represents "Somewhat Happy", and 3 represents "Not Happy". Specifically, this paper examined responses to the question in two distinct time periods: 2018, preceding the COVID-19 pandemic, and 2022, following the pandemic. In order to analyze the difference in answers, data was extracted from from two separate datasets, each originating from the General Social Survey from 2018 and 2022. From these two data sets, only the variable “HAPPY” was considered for analysis. This variable contained a respondent’s answer (1 through 3) or null if the

question wasn't answered for various reasons. The collected data was put into combined data sets to create multiple graphs. These visuals included histograms and kernel density plots for each separate year, as well as for the aggregate data spanning both years.

From the graphs that were created, it was found that the survey from 2022 had more respondents. Additionally, on average there were less "Very Happy" responses and more "Somewhat Happy" and "Not Happy" responses in the 2022 survey. In 2018, out of 2344 responses, 29.91% answered "Very Happy", 55.76% answered "Somewhat Happy", and 14.33% answered "Not Happy". In 2022, out of 3520 responses, 22.13% answered "Very Happy", 55.17% answered "Somewhat Happy", and 22.70% answered "Not Happy". While these findings suggest a significant shift in self-reported happiness levels, it's important to recognize that various external and internal factors may contribute to these changes. Other socio-economic, political, and personal circumstances not captured in this analysis can also play a role in how individuals rate their happiness, but it is important to acknowledge COVID-19's effect on people. COVID-19 has had a substantial impact on people's lives, which likely influenced the decline in self-reported happiness levels.

Tu-Yen: Data

- Two to five pages providing visualizations, statistics, and a discussion of your findings. If you have a lot of plots or tables, that's OK, but try to focus on a few key pieces of evidence rather than doing every single pairwise comparison of some set of variables

The data used in this report was derived from the General Social Survey, or GSS, which conducts surveys almost annually on social, economic, and political values of people in the United States. To achieve our analysis, two datasets were obtained from this survey: the data

from 2018 (before the pandemic) and the data from 2022 (after the pandemic). Each dataset was read into its respective dataframes to be cleaned and visualized in the code: df_2018, and df_2022.

After reading the data in, it needed to be cleaned. Due to the number of survey questions the GSS has, we updated these data frames to only include the variable needed for analysis: 'HAPPY'. This variable looked at people's answers to the question: "Taken all together, how would you say things are these days--would you say that you are very happy, pretty happy, or not too happy?". The data is given in numeric values of 1, 2, or 3 based on their answers, where 1 represents very happy, 2 represents pretty happy, and 3 represents not too happy. After the necessary data was extracted, the columns were renamed based on the year; 'HAPPY' in df_2018 became 'HAPPY_2018', and 'HAPPY' in df_2022 became 'HAPPY_2022'. This was done for better clarification and because later on, both datasets will need to be merged and it's important to differentiate between each column to avoid confusion. Then, the values were ensured to be numeric using the pd.to_numeric() function. Finally, null and outlier values were cleaned. The data had no outliers when checked (every value was either 1, 2, or 3), so nothing had to be handled in that case. Regarding null values, null counter variables were created to see the sum of null values in each data set (null_2018, null_2022); the 2018 data only had 4, and the 2022 data only had 24. One of the challenges faced was determining what to do with these null values, but this was solved after looking at the GSS documentation. According to the code book, the null values are caused by three different occurrences: don't know, no answer, or skipped on the web survey. Other than that, there was nothing to go off of to determine the reasoning for each of the null values. So, it was decided that the null values would not provide us much

information, nor would it skew the results drastically, so we left them as is.

The final step to preparing our data for analysis was merging the two data frames into one. As of now, the two data frames have remained separate (df_2018 and df_2022). It was important to combine the two in order to maintain clean, consistent code, and allow for easier data manipulation. In the end, we had a data frame where the rows were each individual observation/person surveyed, and the variables (columns) were the HAPPY values mentioned before, and the year that survey was conducted (2018 or 2022). The final data frame created was "df_merged".

MG: Conclusion

This project was designed to compare results to the question 'Taken all together, how would you say things are these days--would you say that you are very happy, pretty happy, or not too happy?'. This question, represented with the variable 'HAPPY' was included in both the 2018 and 2022 General Social Survey, which allowed for a fair comparison of results across the two years. Respondents could choose between '1', '2', and '3', correlating to 'very happy', 'pretty happy', or 'not too happy', in that order. We were interested in this variable for a particular reason: the Covid-19 pandemic. Covid-19, which began to spread on a global scale in 2020, was harmful for too many reasons to count. People all over the world got sick, lost family members, lost their jobs, and had no choice but to change their daily routines to protect themselves and their communities from the virus. For this reason, we predicted that respondents' overall happiness would have decreased between the years 2018 and 2022.

It is important to note that the total number of responses in 2018 and 2022 were not the same, so the following data represents a comparison of the percentages of people who answered

‘1’, ‘2’, or ‘3’ for each individual year. A straightforward response count comparison would have been inaccurate, because the total number of responses were not the same. After cleaning and modeling these data, we saw that our prediction was largely supported. More respondents in 2018 selected ‘very happy’ in 2018 than in 2022. Additionally, fewer respondents selected ‘not too happy’ in 2018 than in 2022.

While it is possible that these results were not a direct result of the Covid-19 pandemic, we would argue that many factors that would decrease a person’s general happiness were negatively impacted by the virus. For this reason, we think this comparison is worthwhile in helping understand the lasting impact of the pandemic.

This project could be expanded with the inclusion of different variables relating to the respondents. We think that ‘age’ would be a particularly interesting consideration, as it would allow us to compare how people of different ages and generations handled the pandemic in terms of their own happiness. Are young adults feeling less happy in 2022 than elderly people because of the ‘loss’ of crucial developmental time during lockdown? This question and other similar comparisons would shine light on more specific Covid-19 impacts, rather than just highlighting a nation’s general mood before and after.