**Summary:**

Our primary research question when approaching this project was “what are the general trends of Americans’ confidence in medical institutions, and are there demographic factors that influence this?”. In short, we were hoping to gain a deeper understanding into whether Americans' trust in these institutions oscillates, which could provide insight into whether this correlates with economic or social events. We included other institutions to compare these confidence levels. We chose to focus on medical institutions because if they can correlate a smaller confidence with certain economic/social events,  they will be able to predict and account for this in the future, which can help to prevent us from economic devastation, since medicine is a large part of our economy. The strategy to answer this question included cleaning the data, then creating visualizations to represent the fluctuations throughout the years, as well as visualizations of the confidence overall. Our first action after developing this question was collecting and cleaning the data. We used the General Social Survey (GSS) for all of the data in this project. We imported the raw data that we gathered from this survey, and cleaned it with several methods. We dropped several rows that contained no information, as well as adjusting all of the columns so they had the correct mappings to correlate to the codebook, and  checked for nans throughout the data. Once we were satisfied that our data had been sufficiently cleaned, we moved forward to the visualization and summarization processes. We created a variety of visualizations instead of one main visualization as a way to confirm our conclusions and examine the data from multiple angles. The visualizations that provided us the best insights were the density plot of ‘conhlth’ by race and the line plot of other institutions compared to the medical institutions. These visualizations provided us with interesting results, the first being that race is likely not a factor in whether an individual has a low or high confidence level. We also found it interesting that

compared to other institutions, medicine rose and fell through the years in a similar fashion. Overall, we found that Americans' generally have 'some confidence' in these institutions, and this does not seem to be skewed by demographic information.

**Data:**

In the summary above, it was stated that our primary research question is "what are the general trends of Americans' confidence in medical institutions, and are there demographic factors that influence this?". To answer this question, we had to collect multiple variables of interest from the General Social Survey (GSS). The primary variables that we are interested in are 'conmedic' and 'conhlth'. The former represents the confidence each subject had in medicine. The answer choices were: great deal of confidence, only some confidence, or hardly any confidence at all. We chose this specific variable to focus most of our research on, because we were interested in seeing how confidence in medicine fluctuated since there have been several widespread diseases in the last 50 years (HIV, COVID-19, etc.). The second variable, 'conhlth' represents confidence in health care provided. The responses for this variable were: complete confidence, a great deal of confidence, some confidence, very little confidence, and no confidence at all. This variable is important for similar reasons to 'conmedic'. We included other variables in the data that we cleaned as well, totaling 23 columns with 73297 rows. Some notable variables include year, age, race, and polviews (where the subject would place themselves from liberal to conservative).

The first challenge with this data arose almost immediately, as we quickly realized the data was incredibly large and hard to read into Google Collab. We had to select the variables that we wanted from the GSS website, which were then fortunately combined into a much smaller CSV for us to read in. Once we were able to read this in, we ran into our second obstacle. Seven of our rows showed just column names, there was no data in these rows. We had to find these seven rows, and remove them from the dataframe. We now have a dataset with 23 columns, 73290 rows, and strings in each column saying things such as “only some” and “a great deal”. The answers that were given to the survey questions were recorded as these phrases, and each correlates to a number in the GSS (typically one through 3, although some have more responses). In order to sort these and create visualizations easier, we mapped the phrases to their corresponding number. We could not use dummy variables, since there were more than two options, and dummy variables are used for variables that are either 0 or 1. The first column that we mapped was ‘polview’. There were seven options for this {'extremely liberal': 1, 'liberal': 2, 'slightly liberal': 3, 'moderate, middle of the road':4, 'slightly conservative':5, 'conservative':6, 'extremely conservative':7}. By using the “.map()” function, we were able to switch these strings into their corresponding numbers. We did the same technique for ‘conhlth’, which had five responses that we needed to map. The remaining variables (excluding sex, age, race, and year) had three responses that were the same, so we were able to use the same .map() function for all of them. Finally, we coerced year and age to become numeric rather than strings, and checked our columns for NaNs.

Data cleaning took us longer than expected to finish, primarily because we ran into the difficulties loading the data. We also had to spend time looking through the codebook in order to correctly map our variables. It was somewhat challenging that some of the variables had five categories of responses, while some only had three. This made mapping more tedious, and seemed to be an inconsistency that the GSS should look further into.

**Visualizations:**

It is important to note that as confidence levels increase, the confidence is getting lower. For example, level 3 means that there is less confidence in the institution than level 1. We started off our visualizations by creating two histograms to showcase our two highlighted variables, ‘conhlth’ and ‘conmedic’. We chose to do this because histograms are a great representation when looking at discrete variables. Since we used ‘count’ as our y-axis, we were able to clearly see the number of responses for each confidence level. Looking at the ‘conhlth’ variable first, the highest count is clearly level 3 (‘some confidence’). Levels 2 and 4 (‘a great deal of confidence’ and ‘very little confidence’ respectively) were the two next highest, with about equal counts. The next histogram deals with the variable ‘conmedic’. This histogram looks much different than the prior one, due to an incredibly low count of the confidence level 3 (‘hardly any’). We had previously assumed that confidence levels for medicine would be similar to the confidence levels for health care, but were somewhat proven wrong when comparing these two histograms. There seems to be much greater confidence in medicine in general, which could be due to underlying factors such as race or gender of the subjects surveyed.

Once we noticed that the histograms of ‘conhlth’ and ‘conmedic’ were different, we decided the next visualization we should create is a kernel density plot of confidence in health systems grouped by race. The thought process of this was that a kernel density plot grouped by race would help to remove some of the underlying factors that may be causing the histograms to be different. We quickly identified that the primary race represented in the surveyed data is white, which could be an indicator as to why the confidence levels may be high, as America has disproportionate health care equity. However, looking at the lines that represent the other two races, black and other, we notice a similar pattern as the line for white. They all peak with the third level, which shows us that our data is likely not skewed by race. Despite this, it is always a good idea to collect more data that is increasingly representative of all races, instead of primarily white.

Next, we have a boxplot with the statistical summary for ‘conhlth’. The statistical summary tells us that the count of responses is 1147, the mean is 3.05 and the standard deviation is .958. These are all represented in the box plot. It reinforces the idea that most people chose 3 as their level of confidence, with 1 being shown as an outlier on the boxplot. This is interesting since the histogram did not do a great job of showing that 1 was low enough to be an outlier, we now have an idea of how low the number of responses with confidence level 1 were.

We appreciated the information that our first boxplot gave us, so we decided to create another one that was slightly more advanced. We used the same variable, ‘conhlth’, but grouped this by race. Once again, we are interested in seeing the role that race plays on the subjects’ answers. All three race categories had a median of 3, which shows that the majority of people chose this, regardless of race. However, the ‘black’ race and ‘white’ race both had the same standard deviation and interquartiles, while ‘other race’ had a large standard deviation. There were no outliers for ‘other race’, but ‘black’ and ‘white’ both had an outlier at 1, which corresponds to the first boxplot. Overall, the data in this boxplot seems to be fairly symmetric, with ‘black’ and ‘white’ slightly skewed to the right. This tells us that subject’s who identified with those races felt less confident about healthcare than the median. Even though we have previously noted that there were more subjects who identified as white, their confidence distributions have remained similar, once again discounting that this could be a confounding variable.

After deciding that race is likely not a variable that is causing differences in confidence levels, we decided to take a look at sex. This data only includes ‘male’ and ‘female’ as their answers for sex, which we note could be problematic. We created another kernel density plot to examine the ‘conhlth’ variable grouped by sex. There were more women than men who took this survey, but the densities are close enough that this does not seem to be an issue. The patterns of both the male and female lines in this density plot are very similar, they follow almost the exact same pattern. Confidence level 3 is the highest for both, while level 1 is the smallest. This provides us with evidence that men and women both have similar viewpoints on their level of confidence in health care, but we decided to look at a boxplot in order to confirm this.

The boxplot that we created to look at ‘conhlth’ grouped by sex shows us a deeper story than the kernel density plot. Both male and female have the same median, but female is skewed to the right, while male is symmetric. Female also includes an outlier at 1, while male does not. Due to this, we can conclude that there may be a slight difference between male and women answers to the survey. Since females had a higher number of answers than men, this means that the data is potentially skewed to have confidence levels closer to 4 and 5 for the ‘conhlth’ variable. Since the higher confidence levels mean lower confidence in this data, there may be less confidence in ‘conhlth’ due to more women answering the survey who do not have large amounts of confidence in the healthcare system.

Finally, as a last bit of analysis, we were interested in comparing the confidence in medicine to the confidence in other institutions. We thought that ‘conmedic’ would be an interesting variable to compare since we wanted to get a more in depth look into one of the aspects of healthcare. Healthcare is difficult to generalize, so looking specifically into the medicine aspect versus other institutions could give us clearer insight. We created a line plot that represents the fluctuations in confidence intervals throughout several institutions between 1972 and 2022. Immediately, it’s shown that Congress has had the most drastic increases in confidence levels. If you remember that confidence levels are worse as the levels get higher, this is not a good thing for congress. In addition, we see a fairly steady rise in the lack of confidence for medicine throughout the years. There is a rise in the confidence level at 2020, which is likely due to the COVID-19 pandemic and a lack of trust for the vaccine. Other institutions follow similar trends, most notably confidence in science. Science and medicine both intersect in terms of policies and regulations, so it is not surprising to see that their levels have similar trends.

**Conclusion:**

Overall, we are able to answer our research question with confidence. A few critiques and concerns that typically arise with research are data collection, missing data, and bias in questions. The data must be collected from a reputable source, it must be correctly cleaned in order to be accurate, and the question must not be misleading the visualizations to show something that isn't really there. We collected our data from the General Social Survey, which is reputable and uses standard, fair surveying techniques. We cleaned our data to ensure there were no missing or odd variables that could skew our data and cause issues in our findings. Finally, we used statistical methods to create visualizations that provided accurate representations of the data. Our research question "what are the general trends of Americans' confidence in medical institutions, and are there demographic factors that influence this?" was open-ended and did not have implicit bias that could lead to us attempting to create or interpret visualizations in a misleading manner. We had several versions of our visualizations, and chose the ones that provided the most interesting and representative takeaways.

Through analyzing these multiple graphical and numerical summaries, we have found that Americans' tend to have 'some confidence' in healthcare systems and medicine. We do not believe that this is being influenced by demographics, although we would preferably like to look further into the ‘male’ and ‘female’ answers to confirm this. There is not an abundance of trust in medical institutions, yet there is also not a lack of it. Most subjects that answered this survey are middle of the road, with just ‘some’ confidence in these institutions.

Additional work that was outside the scope of this specific project includes combining more datasets with this one, and looking further into the other institutions to see more of the ways that they interact with confidence in healthcare systems/medicine. If we were able to find more datasets with similar surveys, we could ensure that there is no bias in our data. This is due to the fact that the data would have been collected in different ways, so we would have multiple variations of data to look through and compare. If we had more time, we would have loved to investigate further into other institutions. We were intrigued by the way confidence in science had a similar trend to confidence in medicine, and we all wanted to dive deeper into the confidence in Congress. Congress typically sets regulations for medical institutions, so this would be a fascinating relationship to explore.

In conclusion, there is neither high nor low confidence in medical institutions in America. The confidence has fluctuated some throughout the years, but remains fairly steady. This could be used to find ways and target demographics that may not have as much trust in the institutions, to help create more credibility for these organizations.