Project 1 – GSS

Author(s):

Gabriel Jackson

Naad Kundu

Thao Nguyen

Kayla Nguyen

Halbert Nguyen

Ben Willoughby

Omar Zeineddine

DS 3001: Foundations of Machine Learning

March 12, 2024

SUMMARY

For this project, our research question was: "Can we determine the impact of age, religion, happiness, and education on political views?" In order to answer this question, we used a lot of different methods, involving different types of graphs and visualizations to come closer to our goal. In order to have normalized, un-skewed graphs, we first needed to clean our data. We first removed all of our unclean data, and changed our year range to only include data from the years 2012-2022. We also changed some column names to make our data more readable, as well as cleaning all of our important variables. Our important variables included religion, age, degree, education, happiness, and political_view. We cleaned these variables by converting the necessary ones to numeric, removing NAs and replacing them with unknown, removing .0s or unnecessary data from our dataframes. Also, for political_view, to make it easier to visualize our data in graphs, we created the political_view_id variable, which is a scale between 1 and 7 representing someone's political view. 1 stands for very liberal, and 7 stands for very conservative.

DATA

For our project, the data was originally very unclean and a lot of work had to be done to clean it and prepare it for our data visualization. Firstly, in terms of our 'year' variable, we had to make it numeric. Also, we renamed the columns of our key variables to make them more readable to prepare for our visualization. Before we even cleaned up our other key variables though, a challenge arose. The dataset stretched all the way back into the 20th century in terms of years, and we didn't want older data to skew all of our results. This is why we decided to only include data from between the years 2012 to 2022. The reasoning is pretty simple- we wanted both data for at least a decade and the most recent data possible to show more relevant political

data. Since 2022 was the most recent year available in our dataset, it only made sense to go back ten years from that year to 2012. This allows for the most recent data to be shown.

For our other key variables, a lot of cleaning had to be done for most of them. In terms of our first key variable, 'religion', the first thing we did is capitalize the names. For example, we changed "catholic" to "Catholic". It is a simple, yet effective change for making our data more aesthetically pleasing. Next, we replaced all NAs with "Unknown" to keep our data consistent for our categorical variables, and prevent our data from being skewed by the NA values.

Our next key variable, 'age', involved some similar cleaning. Firstly, we converted the age variable to numeric values. For age, we decided to keep the NAs as is so that we could eventually graph it as a numerical type. Another challenge that arose is that a lot of the age data points had a .0 at the end of their values. We fixed this by converting them to the integer type (if they were not NAN).

Our next variable, degree, involved a similar process to cleaning our religion variable. We firstly capitalized the names (for example, bachelor's to Bachelor's). Then, we replaced NAs with unknowns, as degree is a categorical variable.

'Education', a numerical variable, was our next key variable to clean. This means that we converted it to numeric values, and left the NAs as is to allow for graphing later.

The 'happy' variable, a categorical variable, was cleaned the same way as the other categorical variables- we capitalized the values, and replaced any NAs with "Unknown".

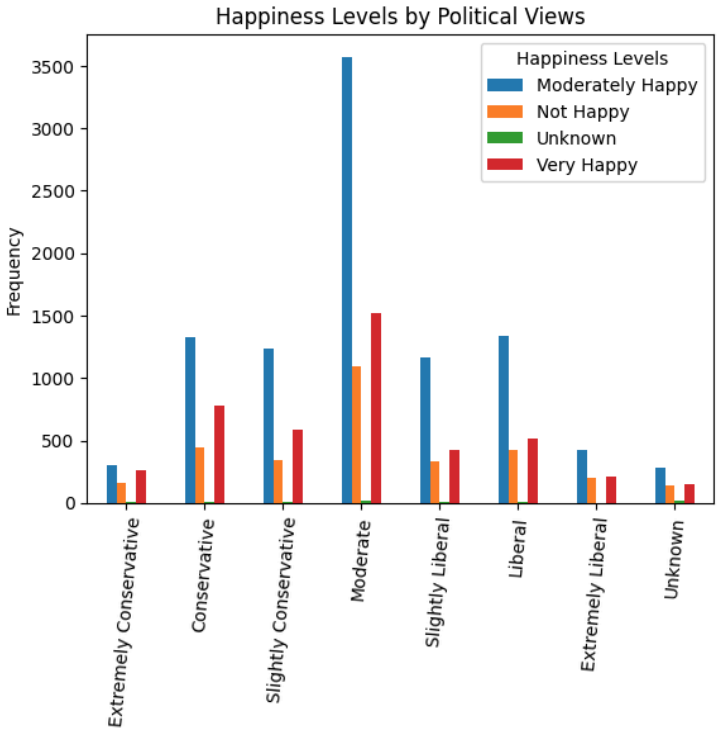
Our last key variable to clean was the 'political_view' variable. The first thing we did was capitalize the values, like our other categorical variables. We also replaced NAs with 'Unknown'. Interestingly though, in order to best show this variable and represent it in an aesthetic and interesting manner, we decided to create a new variable, 'political_view_id'. Since the 'political_view' variable had 7 values ranging from "Extremely liberal" to "Extremely Conservative", we decided to make 'political_view_id' a numerical variable ranging from 1-7, and we mapped 'political_view' to 'political_view_id'. This decision allowed us to easily graph and visualize different political affiliations with other variables.

RESULTS

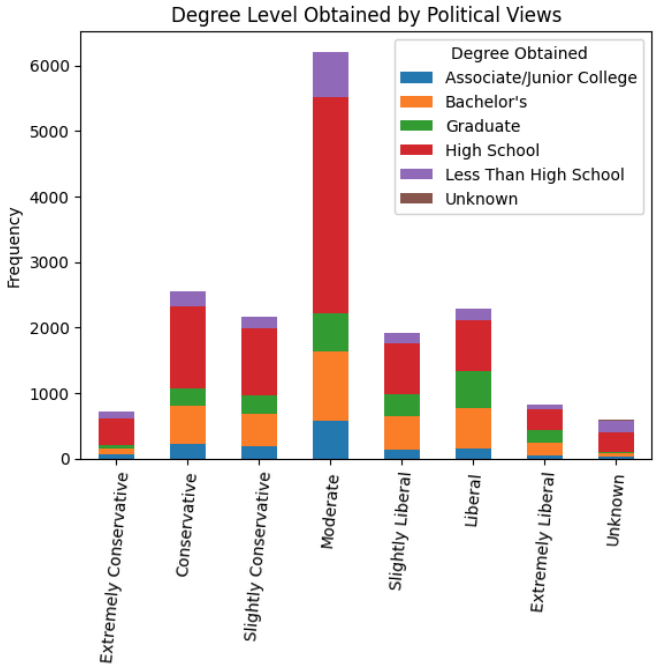
In our question aiming to determine the impact of age, religion, happiness, and education on political views, we set out on a comprehensive analysis, utilizing various data wrangling and visualization techniques to convey our findings. Through the meticulous cleaning and normalization of our data from years 2012-2022, we ensured the comprehensibility of our data, focusing on key variables such as religion, age, years of education, degree, happiness, and political views. As mentioned before in the summary, the newly created “political_view_id” variable is on a scale from 1 to 7, liberal to conservative respectfully.

Our visualization efforts revealed a fascinating pattern regarding happiness and political views. Conservatives report the highest levels of happiness across all three levels; however, the amount of unhappy individuals are similar among both sides, suggesting an unclear pattern

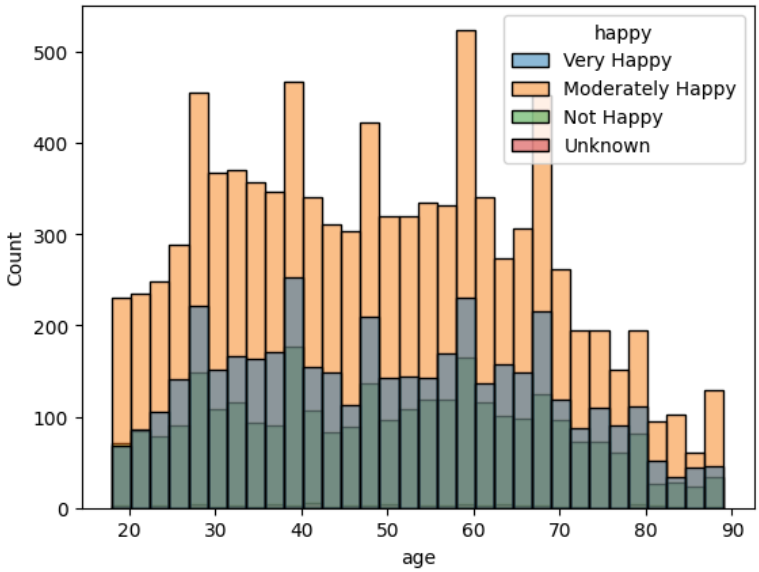
between political views and overall life satisfaction. As we dive deeper into the extremes, individuals who identify themselves as “extremely conservative” tend to report as “very happy” in comparison to being “extremely liberal,” with most being only “moderately happy.” This disparity becomes less prominent if we compare those who are only slightly forward with their position, signifying a convergence of happiness levels among less polarized individuals.



As for educational background, most of the surveyed held up to a high school diploma, with the highest reaching up to graduate school. When we further examine the data, we can see that liberals tended to have higher levels of education, particularly graduate school. This pattern highlights the potential influence that education can have on political views, suggesting that educational background can shape political views. By knowing this data, we shouldn’t jump to conclusions, as correlation does not mean causation. Beyond the level of education,



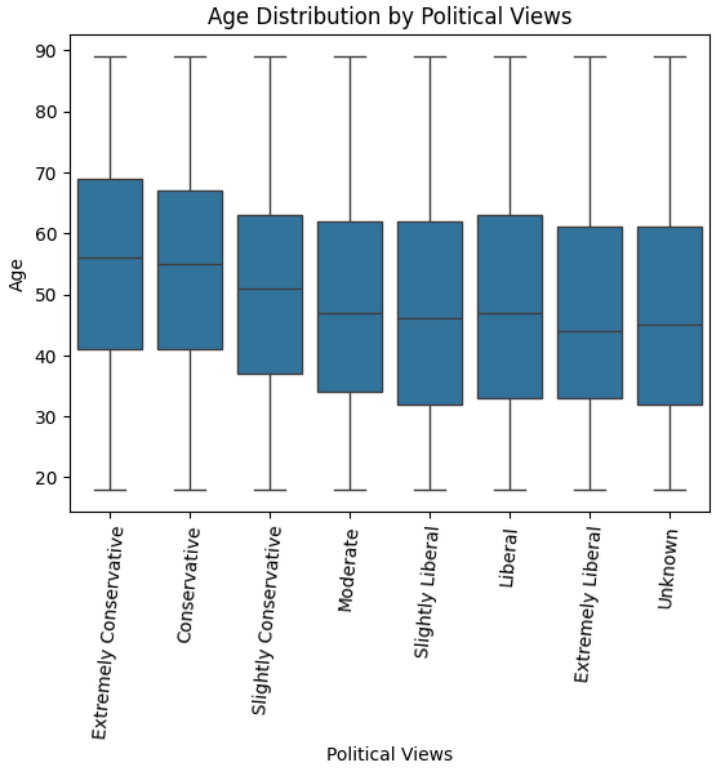
their fields of studies could also influence political views. For example, those with higher education might lean liberal due to the general environment in being exposed to liberal qualities such as inclusivity and equality.



Ages from most of the surveyed lies between 30-70, with the amount of participants greatly decreased above 70 years old. With the wide range of ages, our findings indicate a relatively even distribution of happiness across the various age groups, suggesting that there are other factors than age that can impact an individual's happiness. As we can see on the graph above, the proportion of the orange bars (moderately happy) to green (not happy) are similar across all age groups. This means that no matter the age, we are still in touch with our emotions. One little thing to think about is how education can affect happiness. Higher education allows for a higher ceiling in salary, employment opportunities, and social status. All of these can greatly improve the happiness as an individual as they move up in the workforce. On the other hand, the stressful lifestyles of being a student can adversely affect happiness. All of this suggests a complex relationship between education and happiness.

Although results from happiness vs. age were underwhelming, a more pronounced trend can be seen with political views and age. As we can see from the box and whisker plot below, older individuals are more likely to lean conservative, where the liberal party contains individuals around 30 years old. This age related trend highlights the generational differences in

political ideas, as different generations have lived through different political climates.



One thing to mention was the overwhelming prevalence of the "moderate" category, as we can see in the bar graphs above. This makes it difficult to interpret the data, going against the role of a data scientist whose goal is to be able to present data in a way that everyone can understand. This suggests that there are many neutral views among the surveyed population. Humans are naturally followers, this means that many will gravitate towards the moderate option in order to conform to the socially acceptable views. Additionally, there are individuals who don't fully believe either side's views, so they probably chose "moderate" in order to find a way to most accurately capture their stance. By not totally adhering to a single stance but supporting certain ideas from both sides, these people aim to find a middle ground. The recurrence of selecting the "moderate" option could also signal a general discontent with the current political offerings. These individuals feel as if none of the sides fully convey their full beliefs, resulting in them to select "moderate" as a form of ambiguity. The prominence of the moderate category is a reflection of the evolving nature of political identity in society. It signifies a shift towards a more neutral and individualized approach to politics, where the current traditional labels no longer suffice to capture the complex and personalized view of

these individuals. This trend in individualism challenges policy makers and political parties to rethink how they interact with the public, suggesting a need for change.

Through our comprehensive analysis of the General Social Survey, we now have an informative understanding between personal attributes and political views. The bias towards selecting moderates across our dataset challenges the current political sides, advocating for a refined image and understanding of politics. As we continue observing the relationships between certain variables, we can see that the subject of political orientation is composed from a diverse platter of life experiences, education, and morals.

CONCLUSION

In conclusion, our analysis of the General Social Survey revealed significant trends between voters belonging to more extreme sides of the political spectrum and their age, happiness, and education. Liberals tend to have higher levels of education, particularly at the graduate level, while conservatives tend to report higher levels of happiness. Liberals also tend to be younger while conservatives are typically older.

Ultimately, we believe that our findings should serve as a warning sign for conservative politicians. With higher education becoming more readily available and younger people opting to vote blue, the Democratic Party seems to possess an advantage in the demographic that will make up the future of American citizens. While higher happiness levels seem like a benefit conservatives could cite to attract voters to their cause, the self-reported nature of these scores make it difficult to determine whether those happiness rates were due to an actual better quality of life or if they were skewed upwards by a hidden, confounding variable such as religion which

can positively affect worldview. Obviously, there are many other factors that determine a presidential election such as the candidates, the current state of the country, and campaign finances, however, voter demographics can serve as a solid indicator of which candidate's platform may be more appealing. Another important issue of using certain demographic qualities to predict political allegiance, and by extension an election, is the presence of an overwhelming moderate population. Just like our study shows, it is difficult to categorize these voters because they share both conservative and liberal characteristics. This concern, however, only serves to reinforce our conclusion that liberals may have an advantage in elections in the future. If neither party can consistently rely on moderates during elections, then the party with more aligned voters should have the advantage.

One behavior our group would like to investigate further after conducting this analysis is if people that identify with a certain political party maintain those beliefs as they age. Political scientists often refer to the tendency for older people lean conservative and younger people to lean liberal as the "generation gap". We would like to know whether that phenomenon is true over time and people switch from being liberal to conservative as they age or if older people are holding on to their beliefs from a significantly different political climate. If people do not have strong political party allegiance over time, our conclusion of Democrats having an advantage in demographics will not be as future-proof.