CMSC 320

Kritak Patel

October 20, 2023

1 Data Issues

While exploring the data, I found and corrected the following issues:

1.1 Column Headers Size

I found the column headers to be too long and difficult to work with while analyzing the data. I elected to fix it by shortening the headers. I felt this was a valid way to fix it because the original data set is still available to see each header and with the fix, I can better handle the information and analyze it with less mistakes.

1.2 Handle Missing Data

I found two empty rows in Max's data set while analyzing the data. I elected to fix it by removing the rows entirely as it was not adding any information and was messing with analyzing. I felt that this was a valid way to fix it because the rows looked as though someone had accidently submitted a response with no information adding nothing to analyze.

1.3 Age Column Generalization

In reviewing the datasets for Max's and Fardina's classes, I noticed inconsistencies in the 'Age' column that needed attention. Max's dataset had missing values represented as 'NaN', which I addressed by replacing these instances with the average age, ensuring a more complete dataset for analysis. Moreover, the '50+' entries were causing ambiguity and were subsequently removed, streamlining the data for clearer insights. Similarly, in Fardina's dataset, there were missing values denoted as 'NaN', which I imputed with the mean age to ensure a comprehensive and meaningful analysis. These adjustments were crucial in maintaining the integrity of the data and ensuring accurate analysis and interpretation of the information. While the original datasets are still preserved, these modifications allowed for a more effective handling of the information, reducing potential errors, and enhancing the overall analysis process.

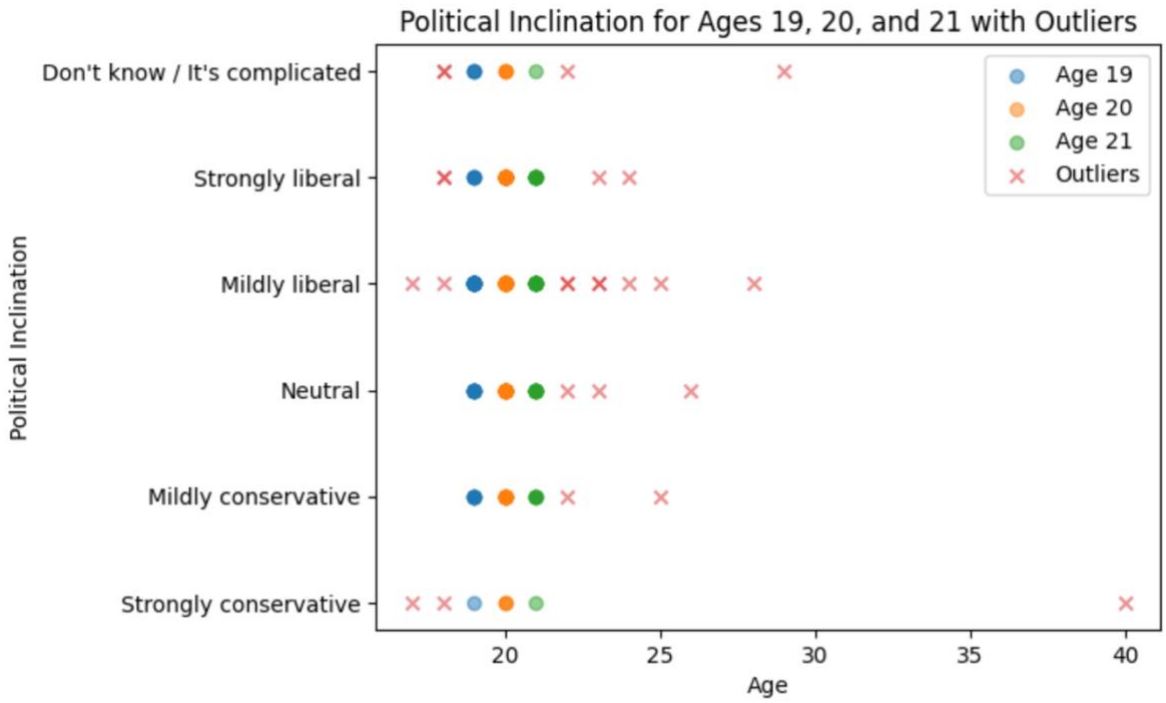
1.4 New Columns for Questions about Hypotheticals

I found that the last few columns in both datasets contained questions about hypothetical situations, each with three potential responses: 'Strongly a jerk', 'Mildly a jerk', and 'Not a jerk'. To facilitate analysis, I added three additional columns in both datasets to sum the occurrences of each response within the rows. I elected to implement this solution as it provided a simplified overview of the overall sentiment reflected in the responses without altering the original data entries. By introducing these aggregate columns, I ensured that the insights derived from the data remained unaffected, allowing for a more streamlined and comprehensive analysis of the behavioral trends within the respective datasets.

2 My Questions

2.1 Which age group has the most discrepancies about personal political inclinations?

This question was of interest to me because understanding the variations in personal political inclinations within the dataset could provide insights into the diversity of perspectives among different age groups. I investigated it by creating a scatter plot that visualized the relationship between age and political inclinations, considering all data points within the dataset. In this analysis, it became apparent that age 20 had the most discrepancies in personal political inclinations because at every level the strength shown by shade of the point is dark for age 20. This observation is noteworthy, particularly because the collected data centered around the ages 19-21, and any other ages were considered outliers in this specific investigation. Additionally, since the average age for both classes also revolved around 20, this central age might have contributed to the observed discrepancies. It's possible that this age was used as a default or placeholder for respondents who did not provide their age, leading to a relatively even distribution of responses across different political inclinations within this age group.



2.2 Does Class Year affect Religious Belief Strength?

This question was of interest to me because it aimed to explore the potential relationship between a student's academic progression and their religious belief strength. I investigated it by analyzing the combined data from both classes, creating a cross table, and conducting a chi-squared test. The analysis revealed that the p-value (0.1597) exceeded the significance threshold of α=0.01, leading to the failure of rejecting the null hypothesis. This suggests that there may not be a substantial association between Class Year and Religious Belief Strength. This outcome challenges the notion that higher class years correspond to a decline in one's religious belief, as commonly speculated considering the experiences of higher education leading individuals away from their religious convictions.

2.3 Is there a correlation between the preference for ‘Strongly a jerk’ responses and being in Max’s Class?

This question was of interest to me because I wanted to understand any potential correlations between individual inclinations towards 'Strongly a jerk' responses and their affiliation with Max's class compared to Fardina's class. At first, I viewed it through pie charts of the two classes and noticed that a slightly higher proportion of individuals in Max's class selected 'Strongly a jerk.' However, upon conducting a chi-squared test, the obtained p-value of 0.2745 suggests that there is no significant evidence supporting a connection between the preference for 'Strongly a jerk' responses and being in Max's class in contrast to Fardina's class. In other words, the observed differences may likely be due to random chance rather than a meaningful relationship. Moreover, despite this slight disparity, the difference is not substantial enough to conclude any association, as indicated by the chi-squared test results. I wanted to test this to see if Max’s class were more openly to judge and select ‘Strongly a jerk’ than Fardina’s class but this was failed to be true.