Data Exploration of Responses to AITA Posts

Shi Feng Wu

October 21, 2023

1 Significance

A significance level of $\alpha = 0.05$ was selected for this analysis. This means that results with a p-value less than 0.05 were considered statistically significant.

2 Data Issues

2.1 Misnamed Columns

The last three questions were written slightly differently, which led to pandas treating them as two different questions when combining the datasets. I fixed this issue by renaming the columns to the same thing before combining the datasets. This will not hurt data quality because the questions were exactly the same, just with different spacing, so these differences would not affect responses between groups.

2.2 Missing Data

Many columns have missing data in seemingly random places. If demographic data important to my question was missing, I dropped the entire row. If response data was missing, I dropped only that entry. This does not hurt data quality because the missing data was not intentionally left blank; it was just a software error, so omitting an entry is just lowering population size arbitrarily. Since I omitted data differently for each question, I will note how I did it.

2.3 Spelling Errors

"Female" was spelled as "Famale" for some entries, adding another category. I just fixed it by correcting the typo. This does not hurt data quality since it was just a typo, and the intended answer is clear.

3 Questions

3.1 Is there a significant association between political leaning and response to the LGBT wedding question?

3.1.1 Motivation

As political leaning is heavily correlated with opinion on gay marriage and LGBT people in general, I was curious if it would also correlate to opinion on a question where an LGBT couple was involved.

3.1.2 Missing Data Handling

Any rows with missing political leaning or response to the wedding question were dropped.

3.1.3 Methodology

1. Visualize response distribution.
2. Assign a ranking to all responses.
   "Not a jerk" = 1, "Mildly a jerk" = 2, "Strongly a jerk" = 3.
3. For each political leaning, add all the responses to the wedding question from that political leaning to a list.
4. Pass all lists into Kruskal-Wallis test.
5. Run Dunn's test if Kruskal-Wallis returns $p < \alpha$.

The Kruskal-Wallis test was used since comparing data from > 2 groups, and the responses are ordinal data formal a non-normal distribution (ruling out a parametric test like ANOVA).

3.1.4 Hypotheses

H_0 : There is no significant association between political leaning and responses to the wedding question; the medians of "Not a jerk," "Mildly a jerk," and "Strongly a jerk" responses are equal across all political leaning groups.

H_A : There is a significant association between political leaning and responses to the wedding question; at least one political leaning group has a different median response ("Not a jerk," "Mildly a jerk," or "Strongly a jerk") compared to the other groups.

If the null hypothesis is false, post-hoc tests will be run on the data to determine which groups are different.

3.1.5 Results

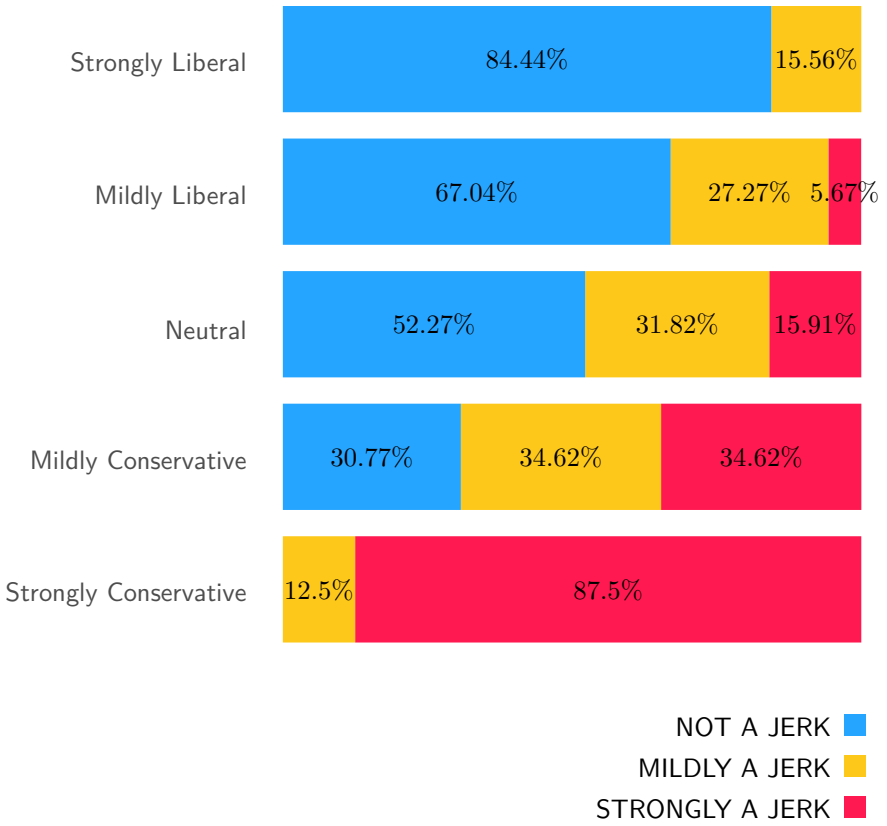

Figure 1: The percentages of responses from each political leaning. ($n = 211$)

At first glance, it looks like there are differences in distribution for every group. To check if they are actually significant, we run Kruskal-Wallis.

Kruskal-Wallis returned a p-value of 0.000000000644. Since this is less than $\alpha = 0.05$, we reject the null hypothesis.

Since there are significant differences between one or more groups, we perform post-hoc analysis using Dunn's test to determine which groups are different:

	Str. Lib.	Mild Lib.	Neutral	Mild Cons.	Str. Cons.
Str. Lib.	1.000000	0.463672	0.013631	0.000012	0.000000
Mild Lib.	0.463672	1.000000	0.516578	0.001374	0.000012
Neutral	0.013631	0.516578	1.000000	0.312310	0.001455
Mild Cons.	0.000012	0.001374	0.312310	1.000000	0.179903
Str. Cons	0.000000	0.000012	0.001455	0.179903	1.000000

Table 1: p-values for all pairs of groups from Dunn's test.

The table altered to show which pairs were significantly different:

	Str. Lib.	Mild Lib.	Neutral	Mild Cons.	Str. Cons.
Str. Lib.	No	No	Yes	Yes	Yes
Mild Lib.	No	No	No	Yes	Yes
Neutral	Yes	No	No	No	Yes
Mild Cons.	Yes	Yes	No	No	No
Str. Cons.	Yes	Yes	Yes	No	No

Table 2: Whether p-values for each pair were $< \alpha = 0.05$.

Those strongly liberal and mildly liberal did not show significant differences in responses. The same goes for strongly conservative and mildly conservative. However, all the liberal groups showed significant differences from all the conservative groups. The neutral group did not show significant differences from the mild groups, but did for the strong groups. It seems that for all groups, their similarity of their responses with another group directly correlated with their proximity on the political spectrum.

This might be because feelings about gay marriage and LGBT people in general can contribute to more or less sympathy for the LGBT person that made the original post.

3.2 Is there an association between political partisanship and partisanship in responses?

3.2.1 Motivation

I was curious if tendency to be on the political extremes would also have a tendency for more extreme answers and if people who are politically neutral would also be more mild in their responses.

3.2.2 Missing Data Handling

Any rows with missing political leaning were dropped. If a response was missing, only that response was dropped while keeping the rest of the responses in that

row.

3.2.3 Methodology

1. Visualize response distribution.
2. Assign a ranking to all responses. Extreme responses were "not a jerk" and "strongly a jerk"; a mild response is "mildly a jerk". Rankings were assigned accordingly, so that all extreme responses were treated the same, and all mild the same.
   "Not a jerk" = 2, "Mildly a jerk" = 1, "Strongly a jerk" = 2.
3. For each political leaning, add all the responses from that political leaning to a list.
4. Pass all lists into Kruskal-Wallis test.
5. Run Dunn's test if Kruskal-Wallis returns $p < \alpha$.

The Kruskal-Wallis test was used since comparing data from > 2 groups, and the responses are ordinal data formal a non-normal distribution (ruling out a parametric test like ANOVA).

3.2.4 Hypotheses

H_0 : There is no significant association between political leaning and extremity of responses to questions; the medians of extreme and mild responses are equal across all political leaning groups.

H_A : There is a significant association between political leaning and extremity of responses; at least one political leaning group has a different median response (extreme response, mild response) compared to the other groups.

If the null hypothesis is false, post-hoc tests will be run on the data to determine which groups are different.

3.2.5 Results

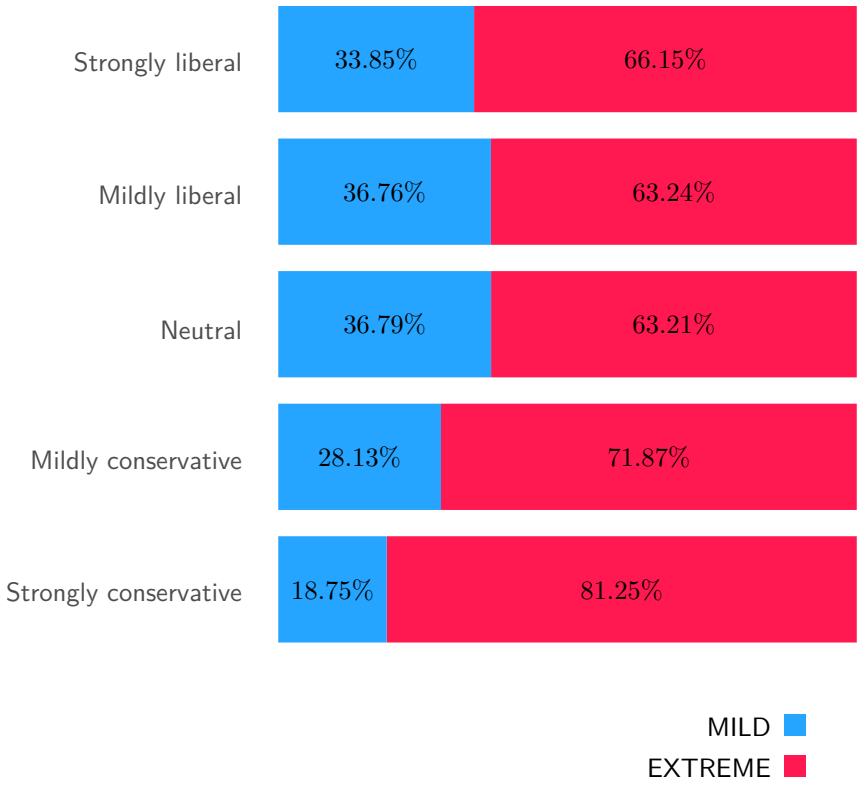

Figure 2: The percentages of responses from each political leaning. ($n = 2973$)

Most replies for all distributions were extreme, and the distributions look fairly similar. To check if there were significant differences, we run Kruskal-Wallis.

Kruskal-Wallis returned a p-value of 0.0001212696. Since this is less than $\alpha = 0.05$, we reject the null hypothesis.

Since there are significant differences between one or more groups, we perform post-hoc analysis using Dunn's test to determine which groups are different:

	Str. Lib.	Mild Lib.	Neutral	Mild Cons.	Str. Cons.
Str. Lib.	1.000000	0.899796	0.957135	0.502873	0.018839
Mild Lib.	0.899796	1.000000	1.000000	0.024331	0.001224
Neutral	0.957135	1.000000	1.000000	0.058925	0.002182
Mild Cons.	0.502873	0.024331	0.058925	1.000000	0.505769
Str. Cons.	0.018839	0.001224	0.002182	0.505769	1.000000

Table 3: p-values for all pairs of groups from Dunn's test.

The table altered to show which pairs were significantly different:

	Str. Lib.	Mild Lib.	Neutral	Mild Cons.	Str. Cons.
Str. Lib.	No	No	No	No	Yes
Mild Lib.	No	No	No	Yes	Yes
Neutral	No	No	No	Yes	Yes
Mild Cons.	No	Yes	Yes	No	No
Str. Cons.	Yes	Yes	Yes	No	No

Table 4: Whether p-values for each pair were $< \alpha = 0.05$.

Both liberal groups and the neutral group answered similarly extreme. The conservative groups answered similarly to each other, but different from the other three. Looking at Figure 2, this means that conservative groups were significantly more likely to answer extremely than the other three groups. Political partisanship does not correlate to extremity of answers, but being conservative directly correlates.

There could be many reasons for these results. One could be that the questions were generally ones that conservatives might have strong feelings on, leading to more black and white situations for them, which then show in their responses.

3.3 Is there a significant association between age and responses to the question about drinking?

3.3.1 Motivation

Since the age range on the survey was wider than I thought it would be, $18-50+$, and may have different frequencies of alcohol use, I wondered if this would affect opinion on whether it was okay for someone to not allow a partner to drink if they could not.

3.3.2 Missing Data Handling

Any rows with missing age or response to the alcohol question were dropped.

3.3.3 Methodology

1. Visualize response distribution.
2. For each type of response, put all ages into a list.
3. Pass all lists into ANOVA test.
4. Run Bonferroni's test if $p < \alpha$.

The ANOVA test was used since comparing data from > 2 groups, and the distribution of ages would follow a normal distribution.

3.3.4 Hypotheses

H_0 : There is no significant association between age and responses to the alcohol question; the means of "Not a jerk," "Mildly a jerk," and "Strongly a jerk" responses are equal across all age groups.

H_A : There is a significant association between age and responses to the alcohol question; at least one age group has a different mean response ("Not a jerk," "Mildly a jerk," or "Strongly a jerk") compared to the other groups.

If the null hypothesis is false, post-hoc tests will be run on the data to determine which groups are different.

3.3.5 Results

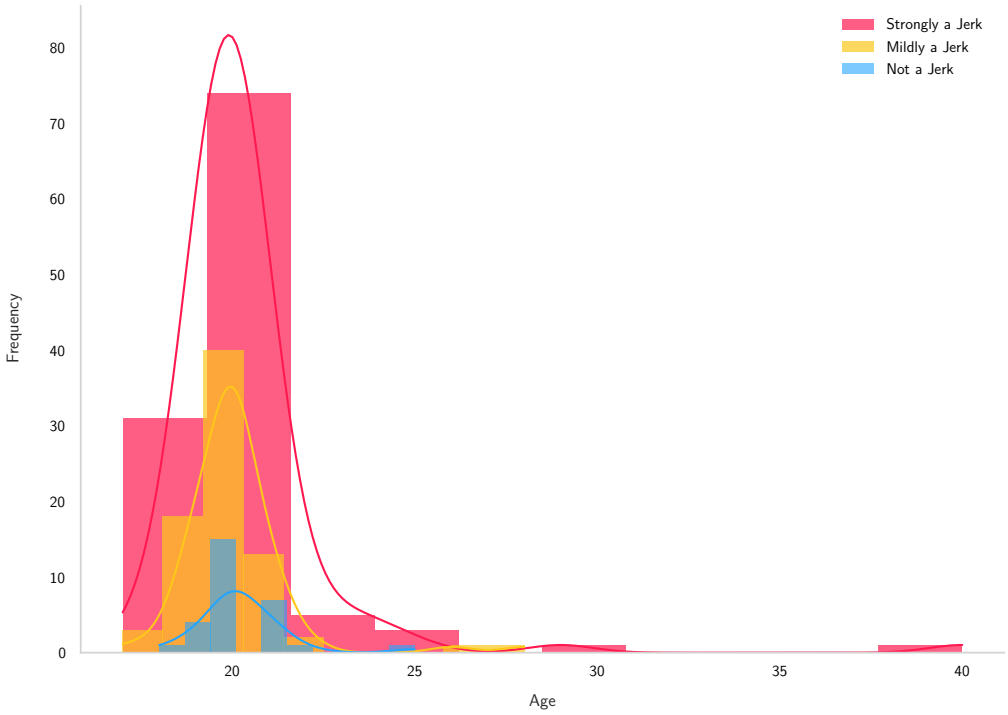

Figure 3: Age distributions for each type of response. ($n = 222$)

ANOVA returned a p-value of 0.667984. Since this is greater than $\alpha = 0.05$, we accept the null hypothesis.

There was no significant association between age and the answer to the alcohol question.

This could be because the question had more to do with opinion on relationship boundaries than alcohol, or because age just is not a large factor in opinions about alcohol.