Writeup

Choosing a dataset to explore

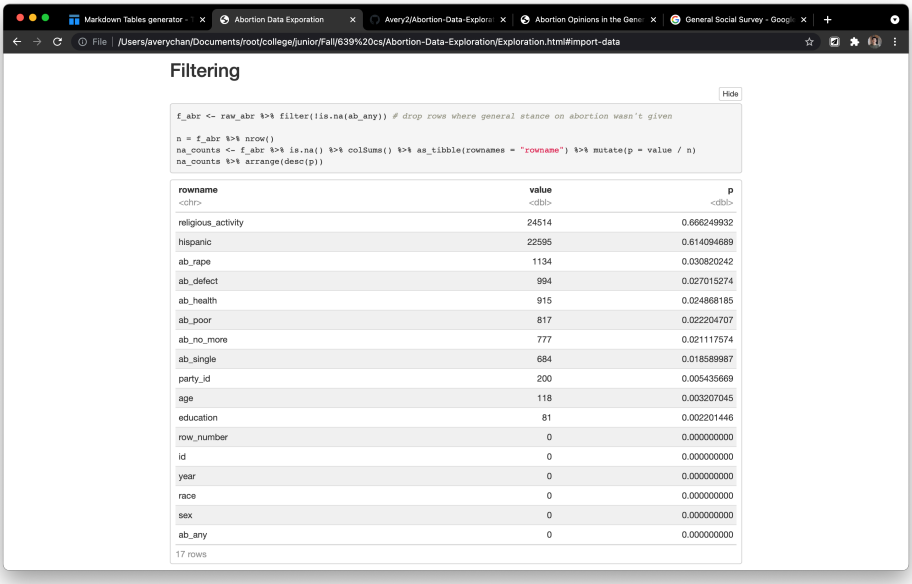
The data I explored was a subset of The General Society Survey [1] focused on stances around abortion. I found the dataset from Rdatasets [2] but it was originally at stevedata [3]. I explored a variety of datasets but eventually settled on this dataset because it was a large dataset with a variety of variables in a context I could still understand. Much of the data I avoided because it would involve more data cleaning than exploration (e.g. Unicef data [4]) or from a context I didn't fully understand (e.g. health data [5]). I found this dataset to be both usable and comprehensible. For brevity, I have described each variable in a table at the bottom of this document.

Initial Question

Given this survey data, what different demographics tend to be in favor of or against abortion?

Why am I having parsing errors?

The first step was to clean the data by specifying data types, renaming columns, and filling in missing values. I also viewed the raw data and explored it with the (non-visual) R methods `dim`, `view`, `summary`, and `str`. This step was helpful because all the parsing errors pushed me to become familiar with the structure of the data.



Filtering

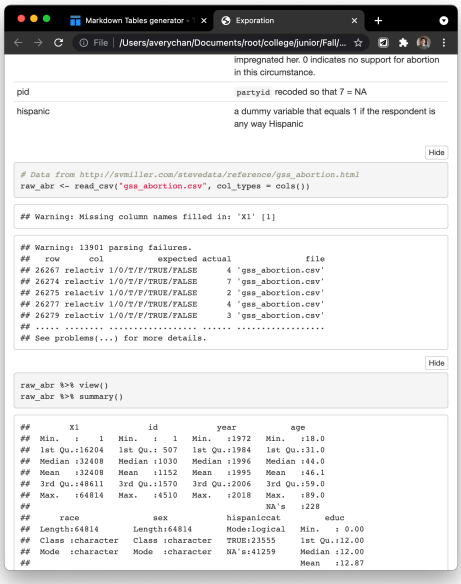
Hide

```
f_abr <- raw_abr %>% filter(!is.na(ab_any)) # drop rows where general stance on abortion wasn't given

n = f_abr %>% nrow()
na_counts <- f_abr %>% is.na() %>% colSums() %>% as_tibble(rownames = "rowname") %>% mutate(p = value / n)
na_counts %>% arrange(desc(p))
```

rowname <chr>	value <dbl>	p <dbl>
religious_activity	24514	0.666249932
hispanic	22595	0.614094689
ab_rape	1134	0.030820242
ab_defect	994	0.027015274
ab_health	915	0.024868185
ab_poor	817	0.022204707
ab_no_more	777	0.021117574
ab_single	684	0.018589987
party_id	200	0.005435669
age	118	0.003207045
education	81	0.002201446
row_number	0	0.000000000
id	0	0.000000000
year	0	0.000000000
race	0	0.000000000
sex	0	0.000000000
ab_any	0	0.000000000

17 rows



impregnated her. 0 indicates no support for abortion in this circumstance.

pid	`partyid` recoded so that 7 = NA
hispanic	a dummy variable that equals 1 if the respondent is any way Hispanic

Hide

```
# Data from http://svmiller.com/stevedata/reference/gss_abortion.html
raw_abr <- read_csv("gss_abortion.csv", col_types = cols())
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Warning: 13901 parsing failures.
##  row      col           expected actual               file
## 26267 relactiv 1/0/T/F/TRUE/FALSE      4 'gss_abortion.csv'
## 26274 relactiv 1/0/T/F/TRUE/FALSE      7 'gss_abortion.csv'
## 26275 relactiv 1/0/T/F/TRUE/FALSE      2 'gss_abortion.csv'
## 26277 relactiv 1/0/T/F/TRUE/FALSE      4 'gss_abortion.csv'
## 26279 relactiv 1/0/T/F/TRUE/FALSE      3 'gss_abortion.csv'
## ..... ........ ................... ...... ...................
## See problems(...) for more details.
```

Hide

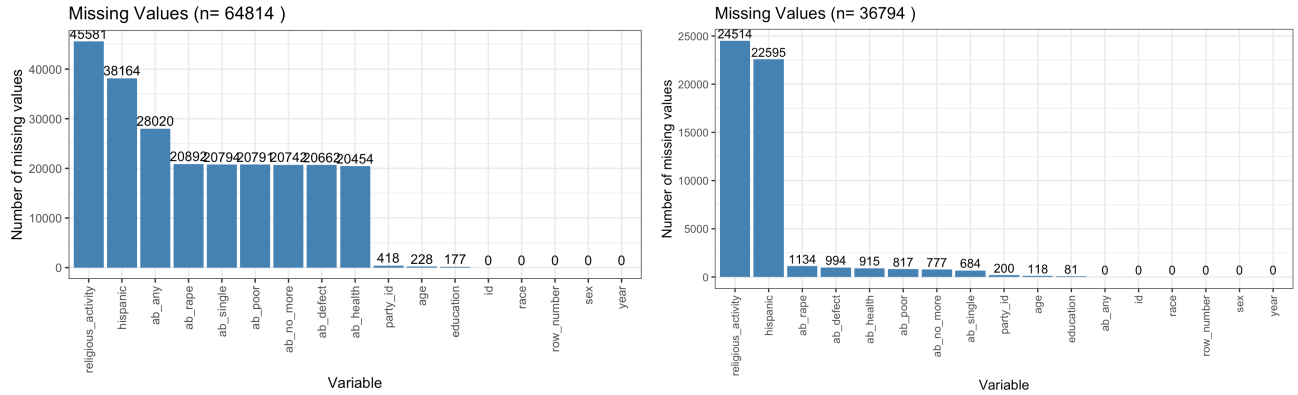
```
raw_abr %>% view()
raw_abr %>% summary()
```

```
##        X1              id             year           age
##  Min.   :    1   Min.   :   1   Min.   :1972   Min.   :18.0
##  1st Qu.:16204   1st Qu.: 507   1st Qu.:1984   1st Qu.:31.0
##  Median :32408   Median :1030   Median :1996   Median :44.0
##  Mean   :32408   Mean   :1152   Mean   :1995   Mean   :46.1
##  3rd Qu.:48611   3rd Qu.:1570   3rd Qu.:2006   3rd Qu.:59.0
##  Max.   :64814   Max.   :4510   Max.   :2018   Max.   :89.0
##                                                NA's   :228
##      race             sex          hispaniccat         educ
##  Length:64814       Length:64814       Mode:logical   Min.   : 0.00
##  Class :character   Class :character   TRUE:23555     1st Qu.:12.00
##  Mode  :character   Mode  :character   NA's:41259     Median :12.00
##                                                       Mean   :12.87
```

R giving parsing issues and missing values.

How many missing values are there?

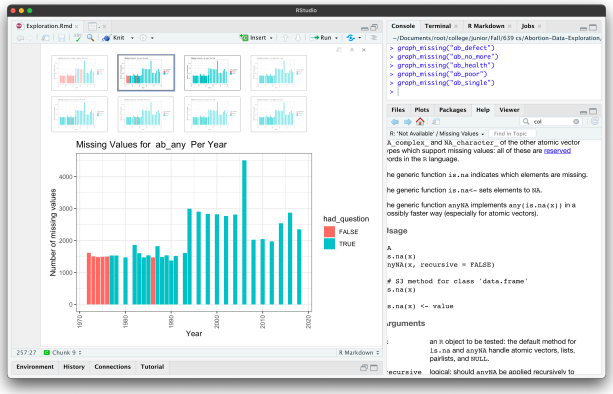
I started by visualizing the number of missing values for each column. When I ignored rows with missing values it removed too much data. You can see below when I remove rows with `NA` for `ab_any`, we go from `64814` values to `36794` values. This prompted me to investigate why it was there were so many missing values.

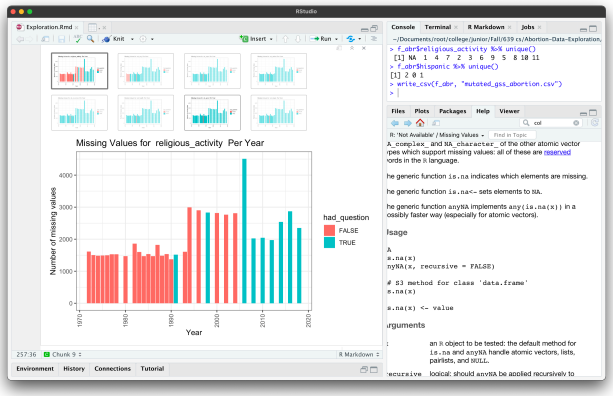


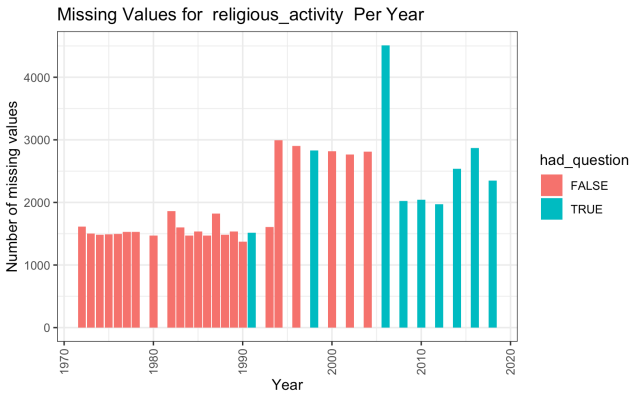
Bar charts of missing values.

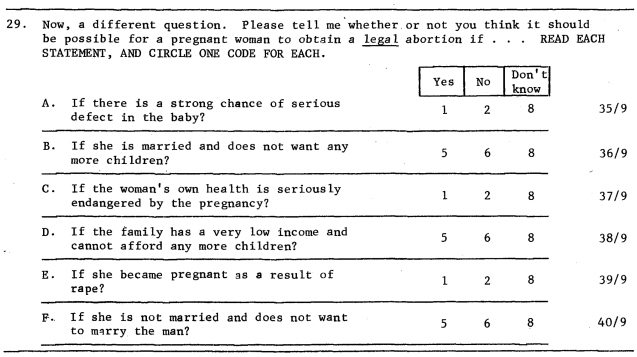
Why are there so many missing values?

Suspecting the missing data was because of changes in the survey questions I plotted the number of missing values by year and colored the bars if all of the values for that year were `NA`. Based on these graphs, I saw that some questions weren't asked until the 2000s. I was able to confirm this assumption with the archive of all GSS questionnaires [6]. This meant that I needed to find a way to clean the data without dropping rows with `NA` values or the data would be skewed towards the present.









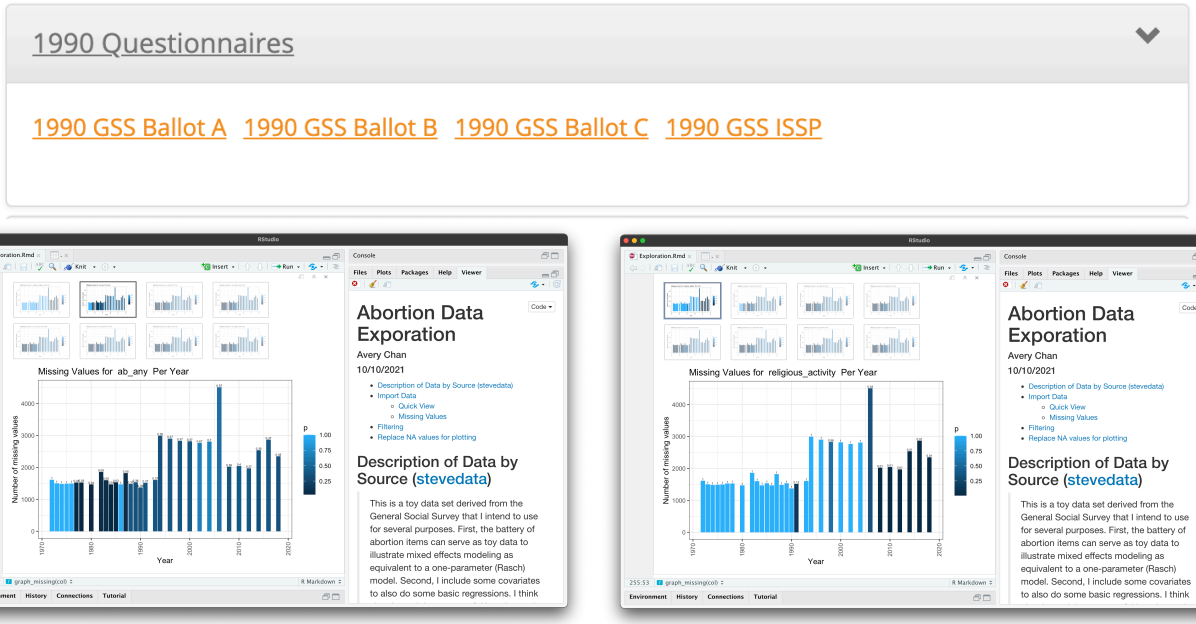
29. Now, a different question. Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if . . . READ EACH STATEMENT, AND CIRCLE ONE CODE FOR EACH.

	Yes	No	Don't know	
A. If there is a strong chance of serious defect in the baby?	1	2	8	35/9
B. If she is married and does not want any more children?	5	6	8	36/9
C. If the woman's own health is seriously endangered by the pregnancy?	1	2	8	37/9
D. If the family has a very low income and cannot afford any more children?	5	6	8	38/9
E. If she became pregnant as a result of rape?	1	2	8	39/9
F. If she is not married and does not want to marry the man?	5	6	8	40/9

Missing values by year and example questionnaire from 1990.

Why are there so many missing values? Pt. 2

I was decided to color by the proportion of `NA` values and noticed while some years would have `0.37` or `0.67` missing values. This prompted me to investigate the questionnaires themselves [6], and I found that the GSS had started to give multiple versions of the questionnaire. Some years only 1 of 3 of the questionnaires would have questions about abortion, so there would be `1/3 = 0.37` missing values.



Proportion of missing values by year and multiple questionnaires for 2000.

How can I fix these missing values?

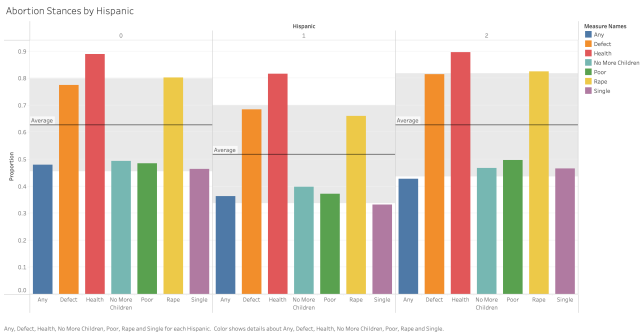
Now that I knew the reason for the missing values, I had a better idea of how to replace them without misrepresenting the data. I will skip the details for brevity, but besides `religious_activity` I was able to replace the `NA` values with something to represent that a question was not asked.

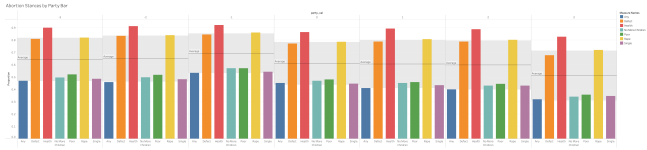
Old data on left, cleaned data on right.

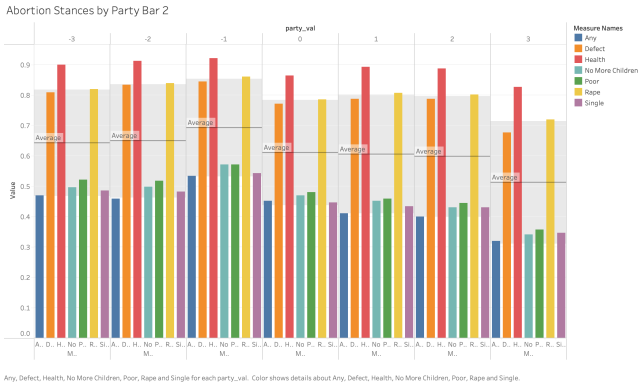
With this long step of cleaning data over, I could start plotting in Tableau.

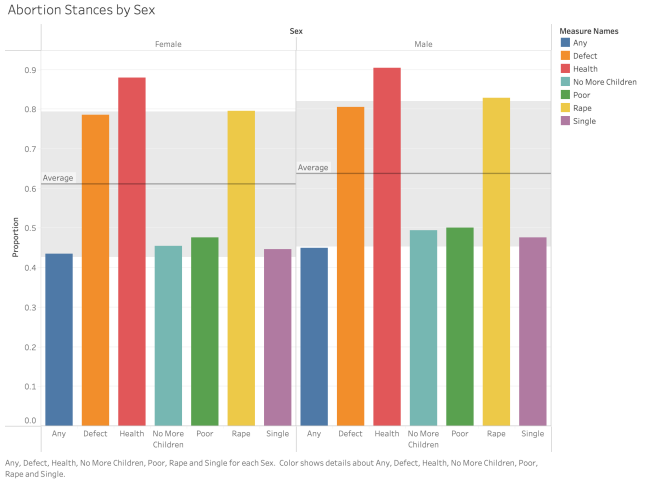
Are there any obvious differences between demographics?

I started by creating bar charts of some nominal variables (`Hispanic`, `Party`, `Sex`) against Abortion stances. I had expected to see more difference in these stances between groups, especially between the sexes, but it didn't seem there was a significant difference. One unexpected thing this did highlight was the differences between the different reasons for abortion. Some reasons like health were more accepted than a reason like poverty across these demographics.





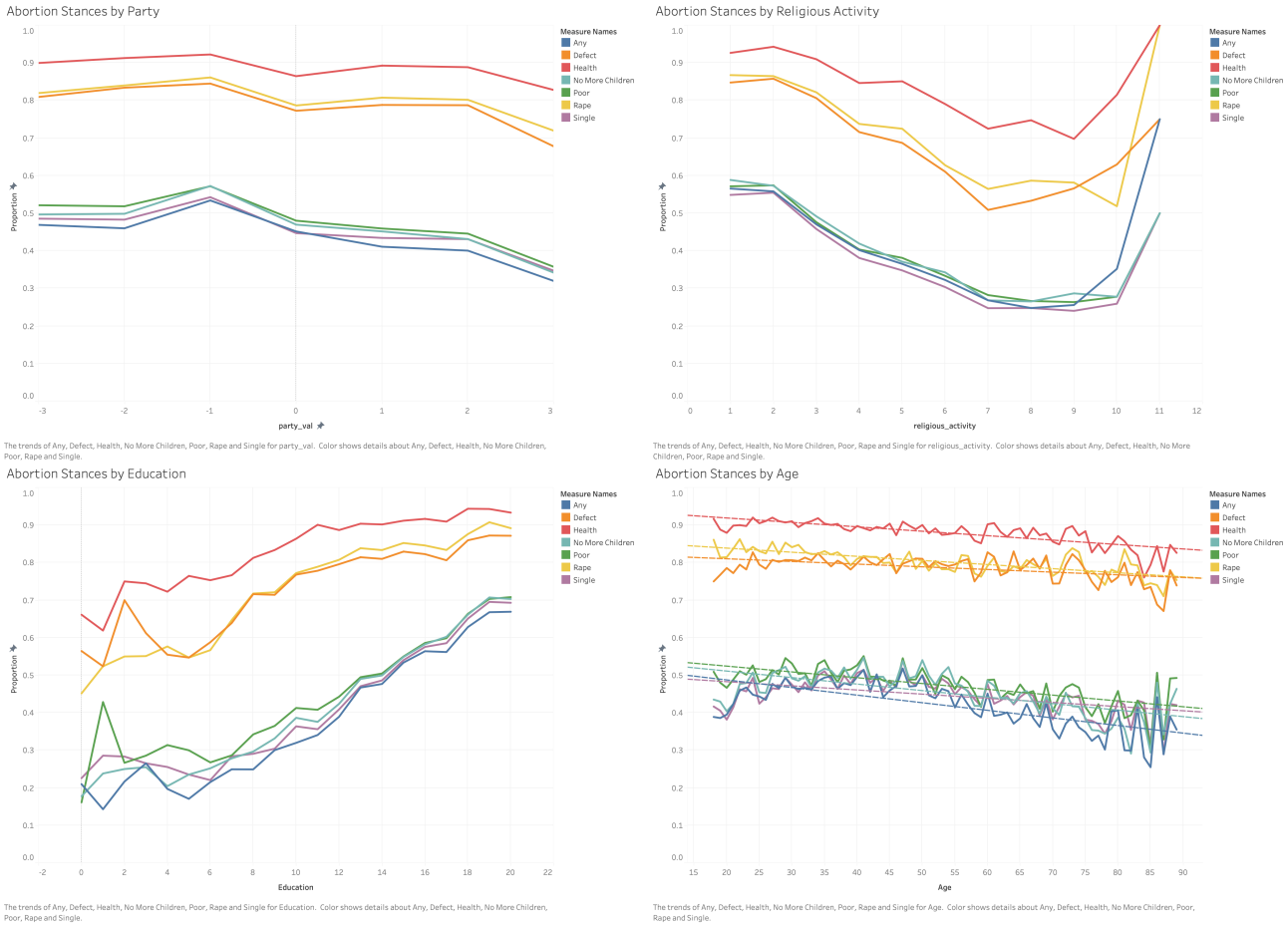




Bar charts for abortion stances against different variables with average and 95% ci.

Are there any correlations?

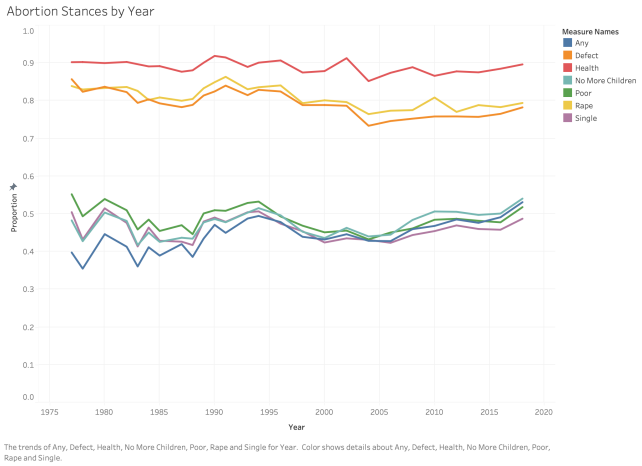
For variables that I thought could be treated as ordinal, like `Religious Activity`, `Age`, `Education`, or `Party` (adjusted to a numerical scale from Democratic to Republican). By plotting on a simple line chart I was able to see some simple linear correlations for `Education`, `Age`, and `Party`. `Religious Activity` (on a scale from 1:11) had some interesting behavior once it got to `11` but I was unable to find the original scale from the questionnaires to find out what this meant because the questionnaire had become computerized which made it much harder to search for a specific question.

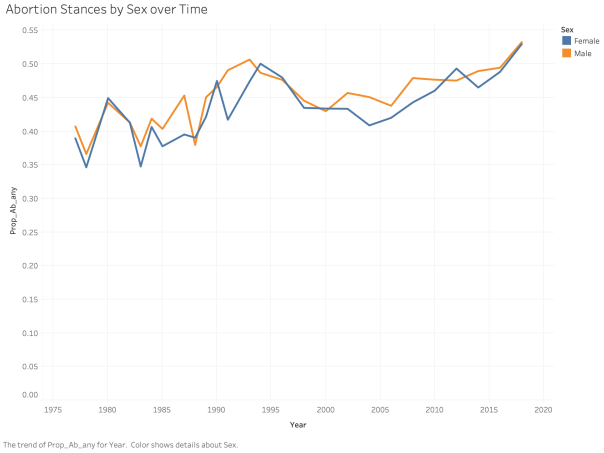


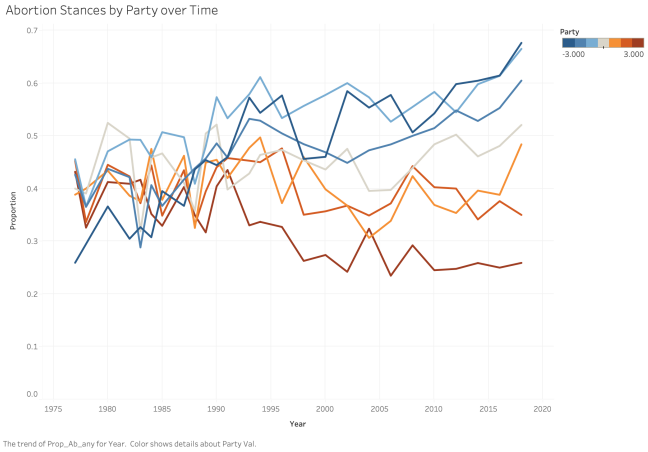
Line plots to show correlations between abortion stances and `Party`, `Religious Activity`, `Education`, and `Age`.

What effect has time had on abortion stances?

I plotted against time next because of the correlation with `Age` shown above and I just suspected there would be a pattern. I was surprised that overall there wasn't much change in the proportion of answers to questions on abortion. I did however notice the distinction between reasons for abortion even more clearly than in the bar charts. The chart by `Party` over time was most interesting because you can see the stances on abortion diverge among party lines with the most Democratic `Party` encoded as `-3` and the most Republican encoded as `3`.







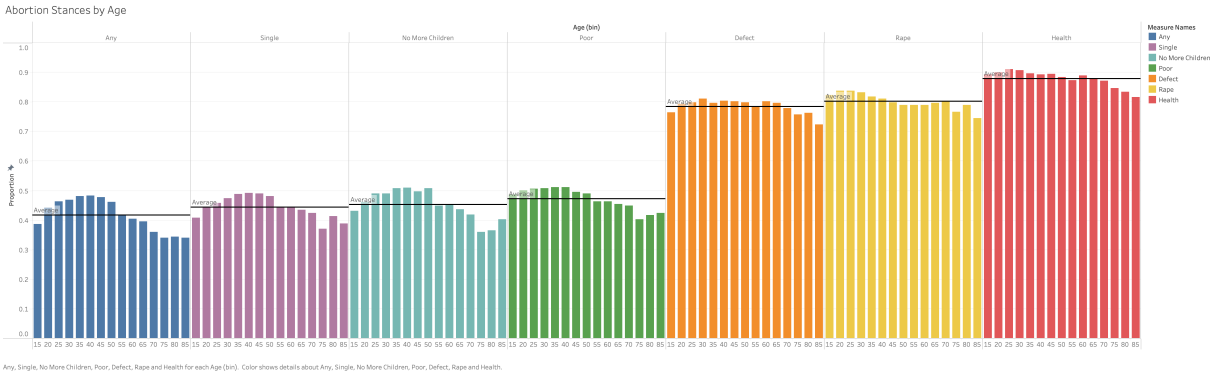
Abortion stances by year overall and for both `Sex` and `Party`.

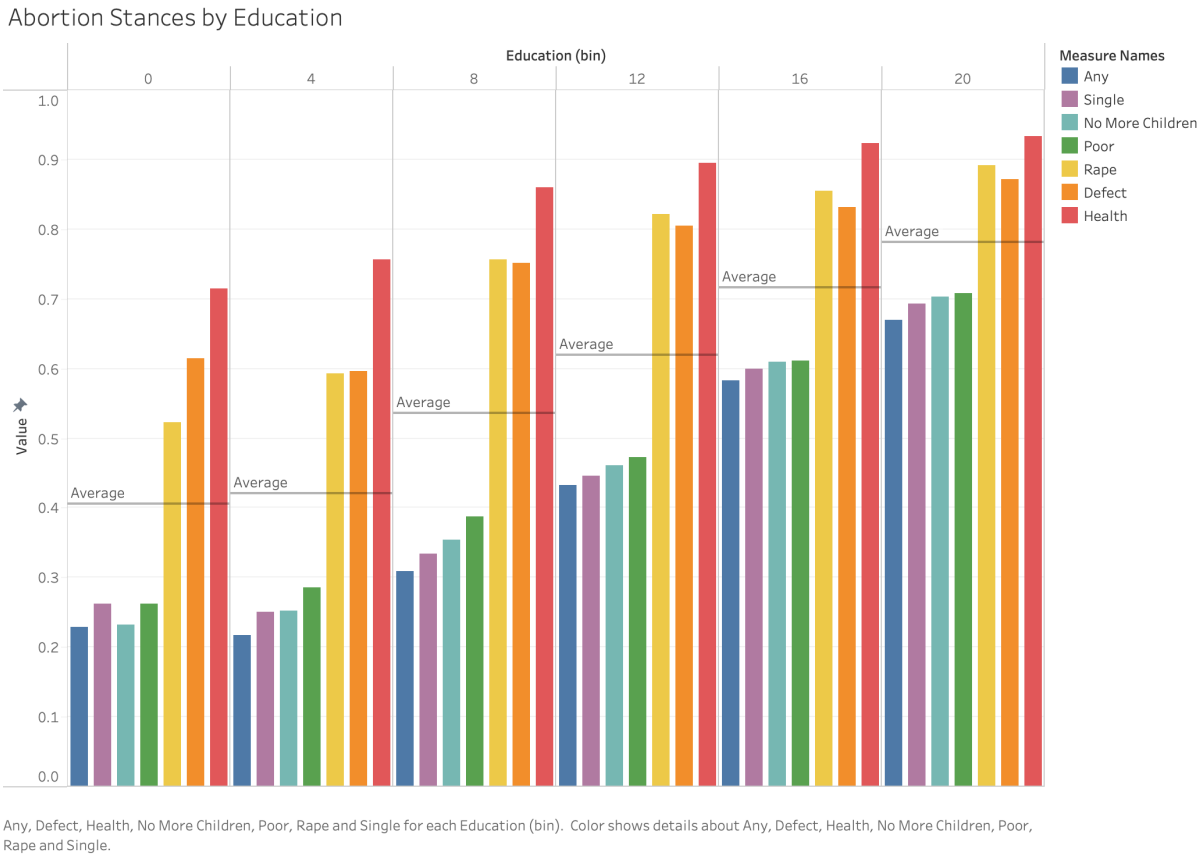
What insights do I find most interesting?

Based on previous visualizations, I found these insights the most interesting:

1. The correlation between `Education` and abortion stances
2. The correlation between `Age` and abortion stances
3. The divergence of abortion stances by `Party` over time
4. The difference between reasons for abortion
5. The reason for missing values (survey question changes)

I created new visuals to highlight each of these besides the last point (because it was part of data cleaning). I combined points 2 and 4 by grouping by reason before age.





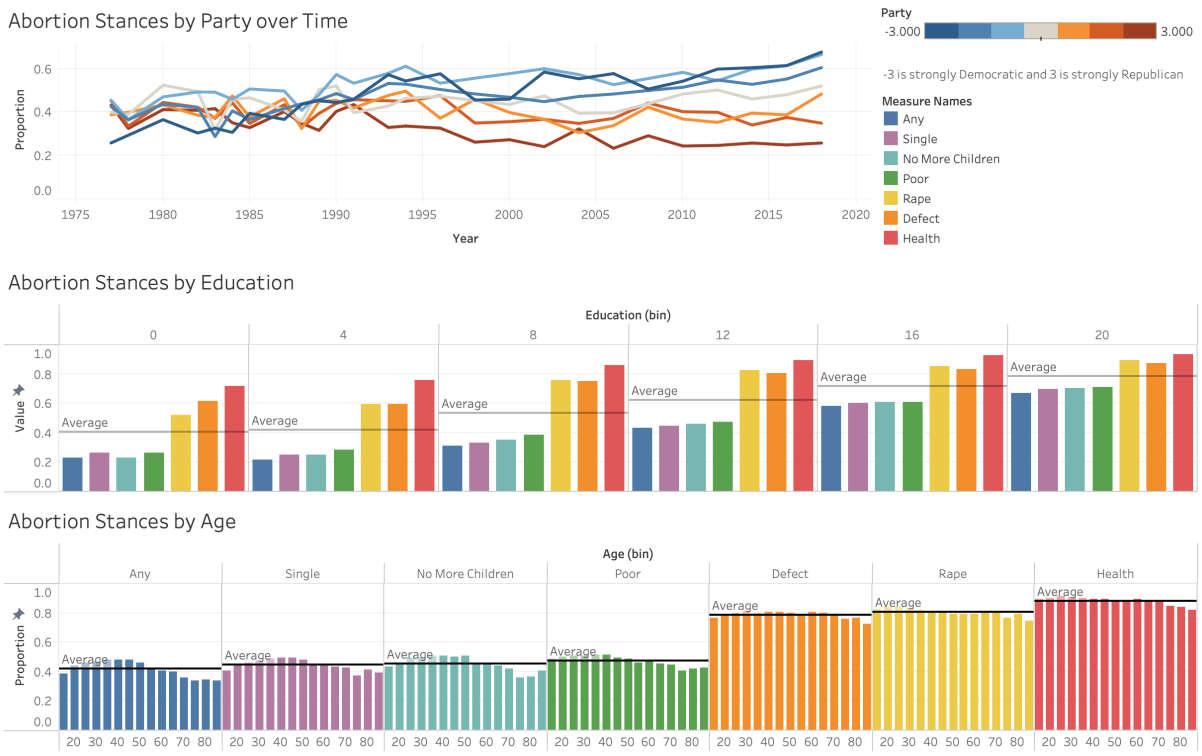
Newly created visuals.

Summary of Lessons Learned

Now let us return to the original question:

> Given this survey data, what different demographics tend to be in favor of or against abortion?

I found age, political party, and education to be demographic information relevant to stances on abortion. The summary visual below attempts to show these tendencies. During this data exploration, I learned to pay attention to missing values because they can tell you something about the data itself. I also learned that particular groupings can yield interesting insights, as was the case with seeing stances on abortion diverge along party lines.



Final visualization emphasizing `Age`, `Education`, and `Party` (over time)

Appendix

source code

Variable Description Table

Variable	Description
id	a unique respondent identifier
year	the survey year
age	the respondent's age in years
race	the respondent's race, as character variable
sex	the respondent's gender, as character variable
education	how many years the respondent spent in school
party	the respondent's party identification, as character variable
religious_activity	the self-reported religious activity of the respondent on a 1:11 scale
ab_any	a binary variable that equals 1 if the respondent thinks abortion should be legal for any reason. 0 indicates no support for abortion for any reason.
ab_defect	a numeric vector that equals 1 if the respondent thinks abortion should be legal if there is a serious defect in the fetus. 0 indicates no support for abortion in this circumstance.
ab_nomore	a numeric vector that equals 1 if the respondent thinks abortion should be legal if a woman is pregnant but wants no more children. 0 indicates no support for abortion in this circumstance.
ab_hlth	a numeric vector that equals 1 if the respondent thinks abortion should be legal if a pregnant woman's health is in danger. 0 indicates no support for abortion in this circumstance.
ab_poor	a numeric vector that equals 1 if the respondent thinks abortion should be legal if a pregnant woman is poor and cannot afford more children. 0 indicates no support for abortion in this circumstance.
ab_rape	a numeric vector that equals 1 if the respondent thinks abortion should be legal if the woman became pregnant because of a rape. 0 indicates no support for abortion in this circumstance.
ab_single	a numeric vector that equals 1 if the respondent thinks abortion should be legal if a pregnant woman is single and does not want to marry the man who impregnated her. 0 indicates no support for abortion in this circumstance.
hispanic	a dummy variable that equals 1 if the respondent is any way Hispanic

1. https://gss.norc.org/About-The-GSS ↩
2. https://vincentarelbundock.github.io/Rdatasets/articles/data.html ↩
3. http://svmiller.com/stevedata/reference/gss_abortion.html#details ↩
4. https://data.unicef.org/resources/dataset/learning-and-skills/ ↩
5. https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?select=heart.csv ↩
6. https://gss.norc.org/get-documentation/questionnaires ↩ ↩