EDA LAB

The General Social Survey (GSS) is a bi-annual nationally representative survey of Americans, with almost 7000 different questions asked since the survey began in the 1970s. It has straightforward questions about respondents' demographic information, but also questions like "Does your job regularly require you to perform repetitive or forceful hand movements or involve awkward postures?" or "How often do the demands of your job interfere with your family life?" There are a variety of controversial questions. No matter what you're curious about, there's something interesting in here to check out. The codebook is 904 pages (use CTRL+F to search it).

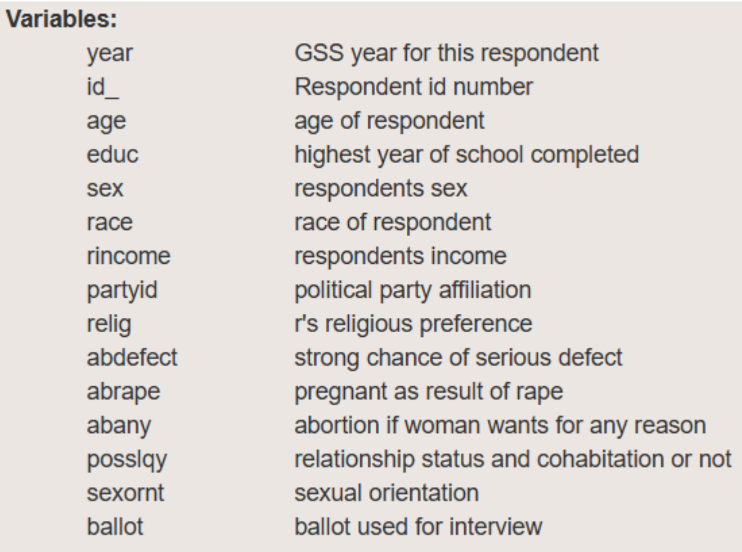
The data and codebook are available at: https://gss.norc.org/us/en/gss/get-the-data.html

The datasets are so large that it might make sense to pick the variables you want, and then download just those variables from: https://gssdataexplorer.norc.org/variables/vfilter

Here is your task:

1. Download a small (5-15) set of variables of interest.
2. Write a short description of the data you chose, and why. (1 page)
3. Load the data using Pandas. Clean them up for EDA. Do this in a notebook with comments or markdown chunks explaining your choices.
4. Produce some numeric summaries and visualizations. (1-3 pages)
5. Describe your findings in 1-2 pages.
6. If you have other content that you think absolutely must be included, you can include it in an appendix of any length.

For example, you might want to look at how aspects of a person's childhood family are correlated or not with their career or family choices as an adult. Or how political or religious affiliations correlate with drug use or sexual practices. It's an extremely wide-ranging survey.



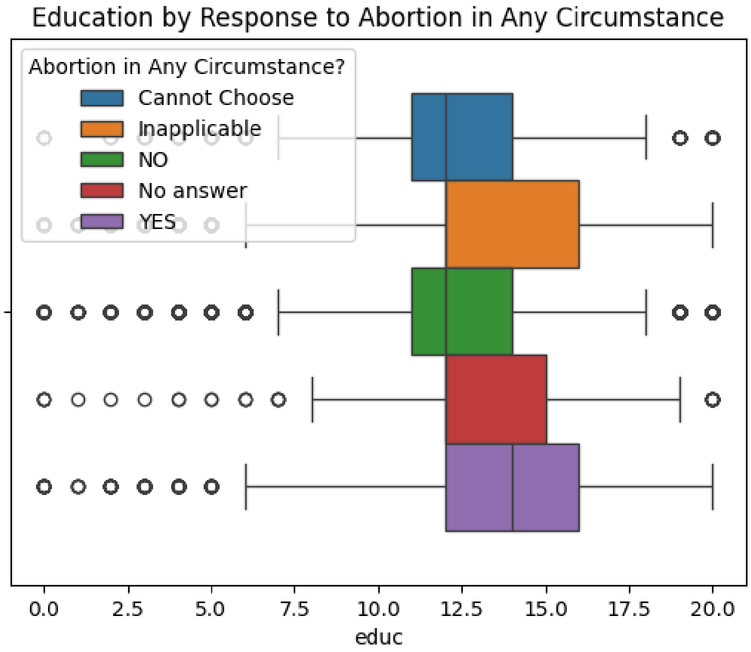
Variables:

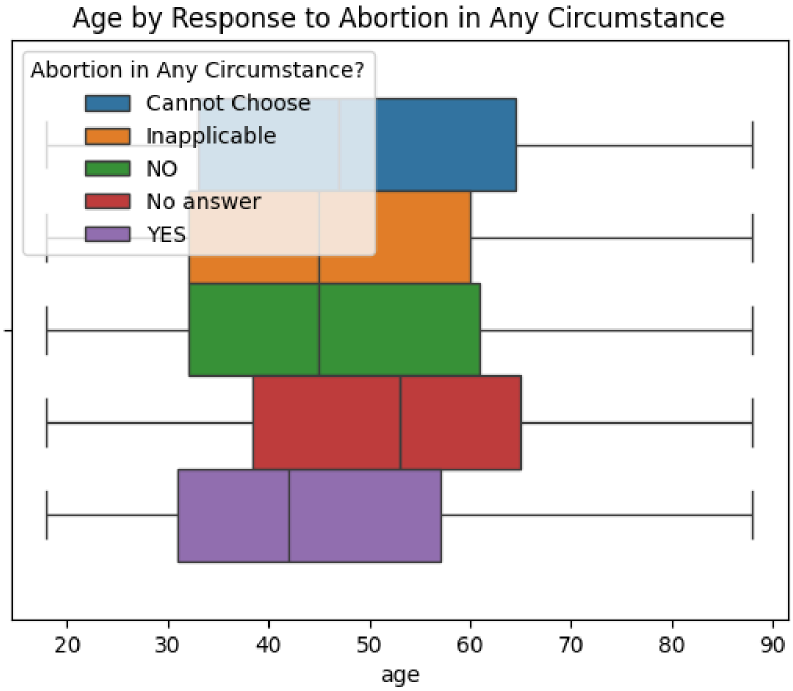
year	GSS year for this respondent
id_	Respondent id number
age	age of respondent
educ	highest year of school completed
sex	respondents sex
race	race of respondent
rincome	respondents income
partyid	political party affiliation
relig	r's religious preference
abdefect	strong chance of serious defect
abrape	pregnant as result of rape
abany	abortion if woman wants for any reason
posslqy	relationship status and cohabitation or not
sexornt	sexual orientation
ballot	ballot used for interview

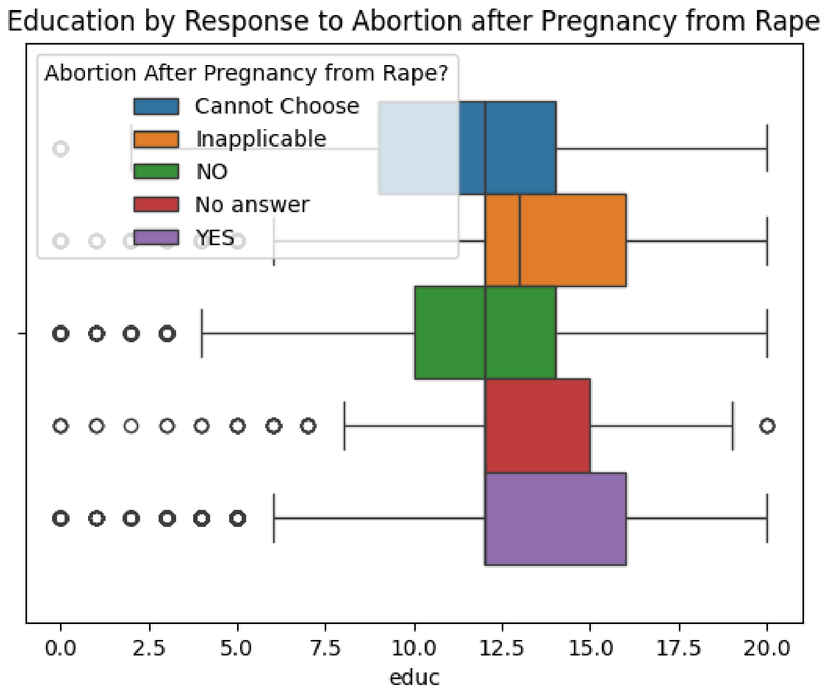
Why I chose this topic

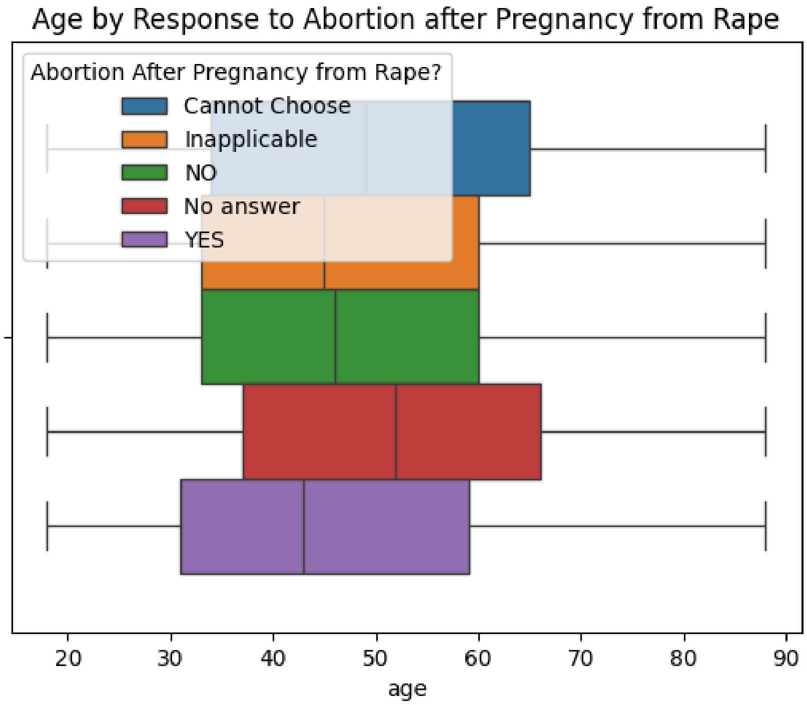
Abortion is a topic that has become very polarizing in American politics within the past few election cycles and I personally believe that a women's right to choose is something that should not be taken away. I think there is a lot of misinformation and bias surrounding how people, myself included, view people across the aisle in politics and on how our preconceptions lead us to lump people into boxes based on one or two political beliefs they may have. I was also curious to see how different demographics view a few different reasons for having an abortion including: Abortion in Any circumstance, Abortion because of High Chance of Defect, and Abortion for Pregnancy because of Rape. I chose these variables to look at because I was curious to see if age, education, sex, race, income, political party id, or religion had any relationship with responses and views on abortion. Some of my preconceptions or biases include believing that people of lower education, of right leaning political ideology, of lower income level, being a male, older, being white, and of Christian religion being less likely to support abortion. Through this analysis I will see if any of those will be proven in one way or another.

Visualizations



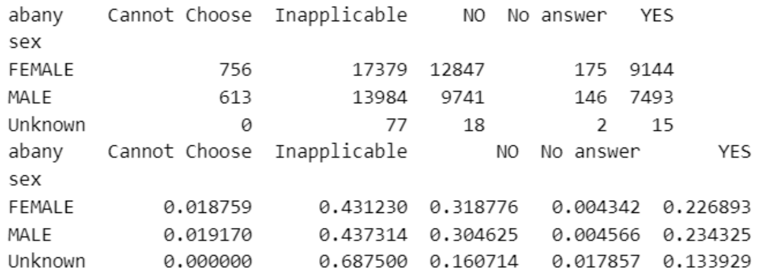



Education by Response to Abortion after Pregnancy from Rape
Abortion After Pregnancy from Rape?
Cannot Choose
Inapplicable
NO
No answer
YES
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
educ


Age by Response to Abortion after Pregnancy from Rape
Abortion After Pregnancy from Rape?
Cannot Choose
Inapplicable
NO
No answer
YES
20
30
40
50
60
70
80
90
age

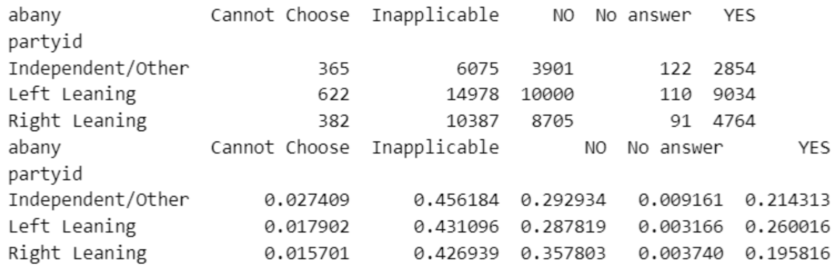
Intersection Tables and Proportion Calculations

‘Sex’ x ‘abany’



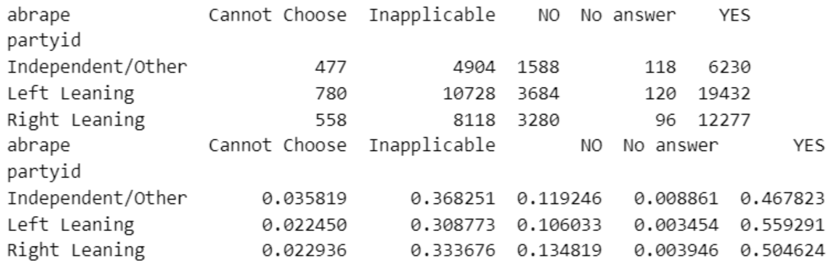
```
abany    Cannot Choose  Inapplicable     NO  No answer   YES
sex
FEMALE             756         17379  12847        175  9144
MALE               613         13984   9741        146  7493
Unknown              0            77     18          2    15
abany    Cannot Choose  Inapplicable        NO  No answer       YES
sex
FEMALE        0.018759      0.431230  0.318776   0.004342  0.226893
MALE          0.019170      0.437314  0.304625   0.004566  0.234325
Unknown       0.000000      0.687500  0.160714   0.017857  0.133929
```

‘partyid’ x ‘abany’



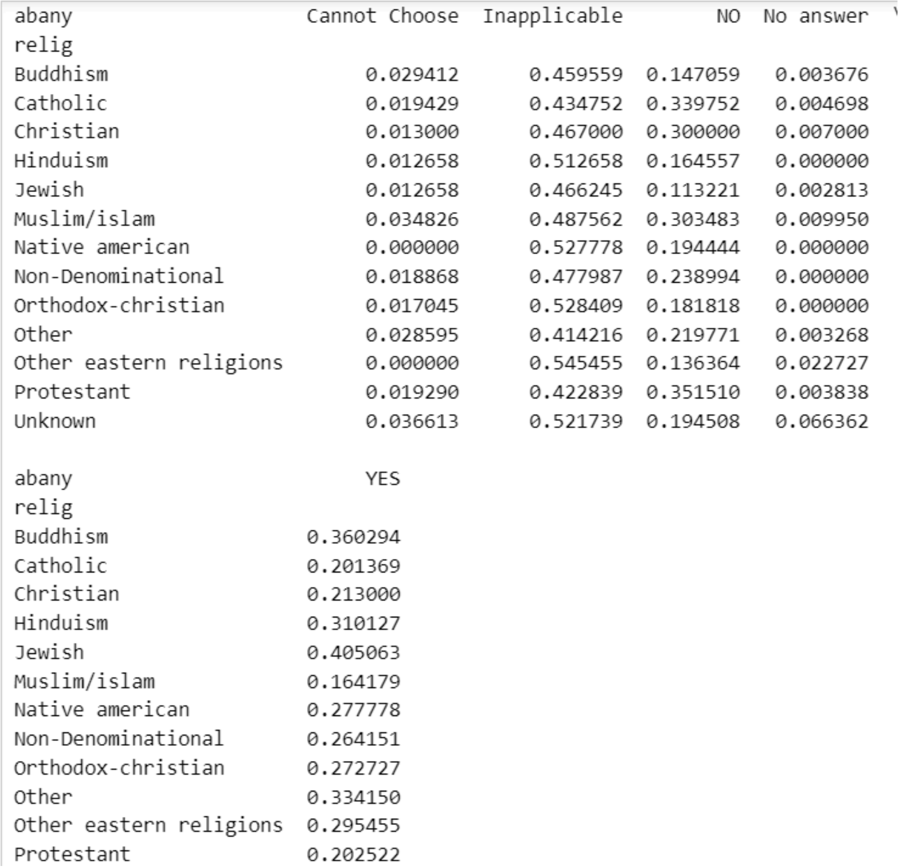
```
abany              Cannot Choose  Inapplicable     NO  No answer   YES
partyid
Independent/Other            365          6075   3901        122  2854
Left Leaning                 622         14978  10000        110  9034
Right Leaning                382         10387   8705         91  4764
abany              Cannot Choose  Inapplicable        NO  No answer       YES
partyid
Independent/Other       0.027409      0.456184  0.292934   0.009161  0.214313
Left Leaning            0.017902      0.431096  0.287819   0.003166  0.260016
Right Leaning           0.015701      0.426939  0.357803   0.003740  0.195816
```

‘partyid’ x ‘abrape’



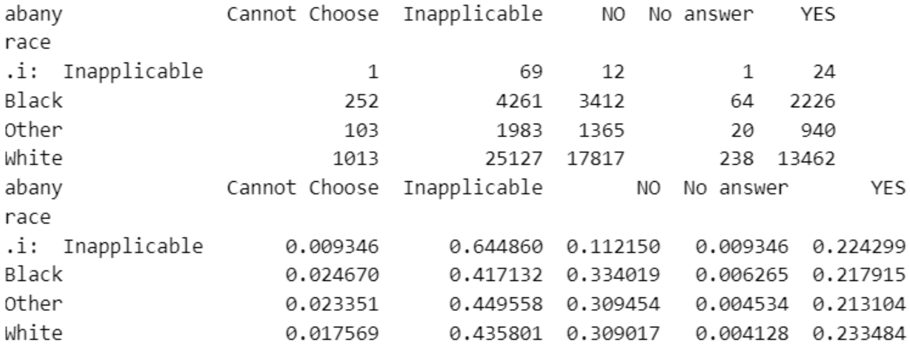
```
abrape             Cannot Choose  Inapplicable    NO  No answer    YES
partyid
Independent/Other            477          4904  1588        118   6230
Left Leaning                 780         10728  3684        120  19432
Right Leaning                558          8118  3280         96  12277
abrape             Cannot Choose  Inapplicable        NO  No answer       YES
partyid
Independent/Other       0.035819      0.368251  0.119246   0.008861  0.467823
Left Leaning            0.022450      0.308773  0.106033   0.003454  0.559291
Right Leaning           0.022936      0.333676  0.134819   0.003946  0.504624
```

‘relig’ x ‘abany’



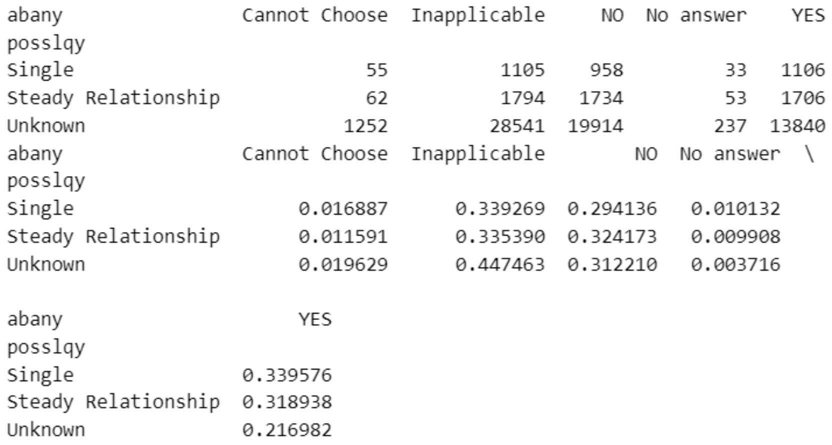
```
abany                    Cannot Choose  Inapplicable        NO  No answer  \
relig
Buddhism                      0.029412      0.459559  0.147059   0.003676
Catholic                      0.019429      0.434752  0.339752   0.004698
Christian                     0.013000      0.467000  0.300000   0.007000
Hinduism                      0.012658      0.512658  0.164557   0.000000
Jewish                        0.012658      0.466245  0.113221   0.002813
Muslim/islam                  0.034826      0.487562  0.303483   0.009950
Native american               0.000000      0.527778  0.194444   0.000000
Non-Denominational            0.018868      0.477987  0.238994   0.000000
Orthodox-christian            0.017045      0.528409  0.181818   0.000000
Other                         0.028595      0.414216  0.219771   0.003268
Other eastern religions       0.000000      0.545455  0.136364   0.022727
Protestant                    0.019290      0.422839  0.351510   0.003838
Unknown                       0.036613      0.521739  0.194508   0.066362

abany                         YES
relig
Buddhism                 0.360294
Catholic                 0.201369
Christian                0.213000
Hinduism                 0.310127
Jewish                   0.405063
Muslim/islam             0.164179
Native american          0.277778
Non-Denominational       0.264151
Orthodox-christian       0.272727
Other                    0.334150
Other eastern religions  0.295455
Protestant               0.202522
```

‘race’ x ‘abany’



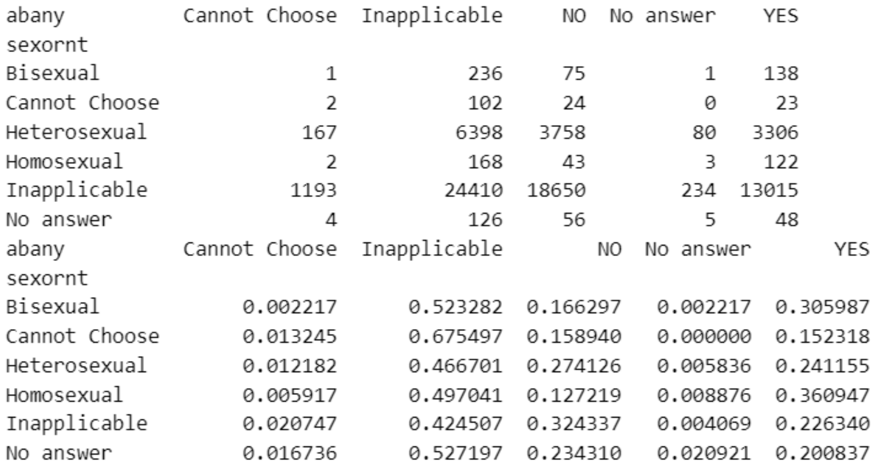
```
abany              Cannot Choose  Inapplicable     NO  No answer    YES
race
.i:  Inapplicable              1            69     12          1     24
Black                        252          4261   3412         64   2226
Other                        103          1983   1365         20    940
White                       1013         25127  17817        238  13462
abany              Cannot Choose  Inapplicable        NO  No answer       YES
race
.i:  Inapplicable       0.009346      0.644860  0.112150   0.009346  0.224299
Black                   0.024670      0.417132  0.334019   0.006265  0.217915
Other                   0.023351      0.449558  0.309454   0.004534  0.213104
White                   0.017569      0.435801  0.309017   0.004128  0.233484
```

‘posslqy’ x ‘abany’



```
abany               Cannot Choose  Inapplicable     NO  No answer    YES
posslqy
Single                         55          1105    958         33   1106
Steady Relationship            62          1794   1734         53   1706
Unknown                      1252         28541  19914        237  13840
abany               Cannot Choose  Inapplicable        NO  No answer  \
posslqy
Single                   0.016887      0.339269  0.294136   0.010132
Steady Relationship      0.011591      0.335390  0.324173   0.009908
Unknown                  0.019629      0.447463  0.312210   0.003716

abany                    YES
posslqy
Single              0.339576
Steady Relationship 0.318938
Unknown             0.216982
```

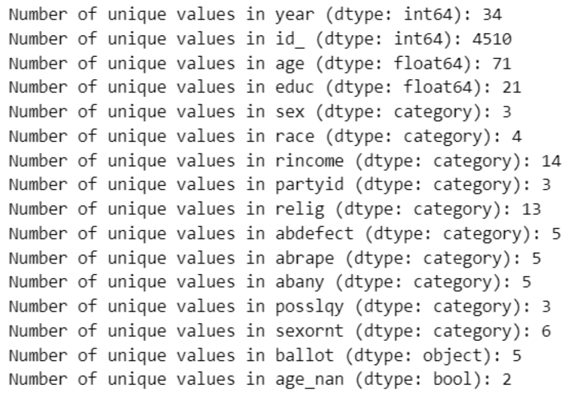
‘sexornt’ x ‘abany’



```
abany         Cannot Choose  Inapplicable     NO  No answer    YES
sexornt
Bisexual                  1           236     75          1    138
Cannot Choose             2           102     24          0     23
Heterosexual            167          6398   3758         80   3306
Homosexual                2           168     43          3    122
Inapplicable           1193         24410  18650        234  13015
No answer                 4           126     56          5     48
abany         Cannot Choose  Inapplicable        NO  No answer       YES
sexornt
Bisexual           0.002217      0.523282  0.166297   0.002217  0.305987
Cannot Choose      0.013245      0.675497  0.158940   0.000000  0.152318
Heterosexual       0.012182      0.466701  0.274126   0.005836  0.241155
Homosexual         0.005917      0.497041  0.127219   0.008876  0.360947
Inapplicable       0.020747      0.424507  0.324337   0.004069  0.226340
No answer          0.016736      0.527197  0.234310   0.020921  0.200837
```

Findings/Discussion

There was definitely a bit of a learning curve with cleaning the data and making the best decisions for what type of data each variable should be. The resulting variables and types were:



```
Number of unique values in year (dtype: int64): 34
Number of unique values in id_ (dtype: int64): 4510
Number of unique values in age (dtype: float64): 71
Number of unique values in educ (dtype: float64): 21
Number of unique values in sex (dtype: category): 3
Number of unique values in race (dtype: category): 4
Number of unique values in rincome (dtype: category): 14
Number of unique values in partyid (dtype: category): 3
Number of unique values in relig (dtype: category): 13
Number of unique values in abdefect (dtype: category): 5
Number of unique values in abrape (dtype: category): 5
Number of unique values in abany (dtype: category): 5
Number of unique values in posslqy (dtype: category): 3
Number of unique values in sexornt (dtype: category): 6
Number of unique values in ballot (dtype: object): 5
Number of unique values in age_nan (dtype: bool): 2
```

I was a bit unsure of whether I should use One-Hot Encoding for variables like the abortion questions. The main variables that were numeric ended up being age and education, with race, sex, income, partyid, religion, relationship status (posslqy), and sexornt being categorical. I struggled at first during visualizations because most of the variables were categorical, but figured out I could do intersections tables in the case where I couldn't do scatter plots of boxplots. I found were that the proportion of people who think abortion should be possible in any circumstance is higher among people with on average higher education. I also found that older people are less likely to says yes to abortion after pregnancy from rape or in any circumstance. One thing that surprised me was that there was a slightly higher proportion of women that answered NO to the question of abortion in any circumstance than men. One statistic that I felt was not surprising was that the proportion of people with Right Leaning political ideology was higher for NO to any abortion than Left Leaning people. A trend I noticed throughout the different factors was that support for abortion in the case of pregnancy from rape was much higher across the board, no matter the demographic, but tended to increase proportionally. Religions more likely to be in support of abortion in any circumstance were Buddhism, Jewish, Hinduism, Other, and Other Eastern Religions while Catholic, Christian, Muslim/Imam, and Protestant had higher proportions who said NO. A statistic I would like to look into more is that Black people had higher proportions of NOs to abortion in any circumstance than White people. A couple concerning notes about the survey were that there were only 2 possible races (white and black) to choose from and all others being under the umbrella of Other. Also, the income ranges were so low, that the highest range was $25,000 or greater which I feel does not account for representation or enough difference between low-income and very high income stats that I would like to pinpoint. There was a higher proportion of people saying NO who were in steady relationships than who were single. Bisexual and Homosexual people were also more likely to answer YES on abortion in any case.