CSE 519 DATA SCIENCE PROJECT PROPOSAL

POLITICAL POLARIZATION AND MARRIAGE

Synopsis:

Given a dataset of electorates for state elections spanning with a period of 4 years, we need to identify the political affiliations of each of the members. We then try to analyze from this dataset as to which of the members are cohabitating and with some probability assert whether they are married.

Having identified these couples from the dataset, we then focus on their political affiliations, more specifically whether the members in a couple have different political affiliations (Democrat, Republican, or some other). Their cohabitation is an important factor. With data from the next sets of elections over the years, we try to identify whether these couples are cohabitating or living separately.

The difference in their cohabitation over election sessions is meant to give some insight on whether their political affiliation played a part in their separation.

Objective:

To assert the probability of a married couple with different political affiliations cohabitating in the next election cycle, and calculate how much of a factor does alternate political affiliations account in the strength of the marriage.

Background Research:

As we all know that the United States is a divided country. In these few decades, people are not divided based on race, religion or on their economical status but on the political party to which they affiliate. Research shows us that two-third of the electorate affiliated with either the Democratic or Republican Party. Research done by Eitan Hersh (assistant professor of political science at Yale University) shows us that most people don't like people from the other political party. It is also found that they avoid dating one another, hate to live near each other and refuse to let their kids marry someone outside of their party.

Eitan Hersh(Professor at Yale), and Yair Ghitza(a chief scientist at political data firm Catalist), by studying a database of more than 18 million couples drawn from voter registration records found that of the couples who are married, 70% of them are the people from the same Political party. It is shown that of those married couples of the same political party, Republican-Republican marriages are 30%, Democrat-Democrat marriages are 25% and Independent-Independent marriages are 15%. The remaining couples (different party couples) consist of 30%. Democrat(Male)-Republican marriages are 3%, Republican(Male)-Democrat marriages are 6%. From this we can say that out of the 10 married couples, 1 couple contains both a Democrat and a Republican.

They also found that men are more likely to be Republican than women. By seeing the above analysis, we can easily say that households with (Republican(M)-Democrat) (6%) are twice that of the (Democrat-Republican(F)) (3%) households. In their research, they also found that older couples are more likely to be of the same political party affiliation when compared to younger couples. They said that it may be because of them being married for years, one spouse may have changed their political views because of the other and joined their party.

In one survey, spouses were asked to assess presidential candidates Hillary Clinton and Donald Trump on a 100-point scale in 2015. Married couples with both partners on the same party gave more scores to Hillary. Trump was more likely to win among participants with different political affiliations. So, we can say that a Republican partner in (Republican-Democrat) relationship was likely to rate Trump higher than a Republican partner in (Republican-Republican) relationship.

If we take a look at the participation of voters, we see huge effects of household composition on voter turnout. Those who are married to the partner of the same party voted at much higher rates than those who are married to the independents or those of the opposite party. Based on the research done in the 2012 and 2014 general election of New York state, a (Republican-Republic) married couple is 10% likely to vote than a (Republican- Independent) or a (Republican- Democrat).

If we take the closed primaries, where the independent voters are not eligible to vote, then the voters who are married to the independents have very low turnout when compared to the (Republican- Republic) or (Democrat-Democrat). In 2012 and 2014 closed primaries, it is recorded that Democrats and Republicans were 17 to 18 percentage points less likely to vote if they were married to an independent, which is enormous considering that overall turnout in these elections is only 30 to 40 percent among registered partisans. From the above statistics, we can say that political affiliations play a large role in the lives of people. It would be interesting to know what effect does the political affiliation play in the marriage of a couple in this rising trend of political polarisation.

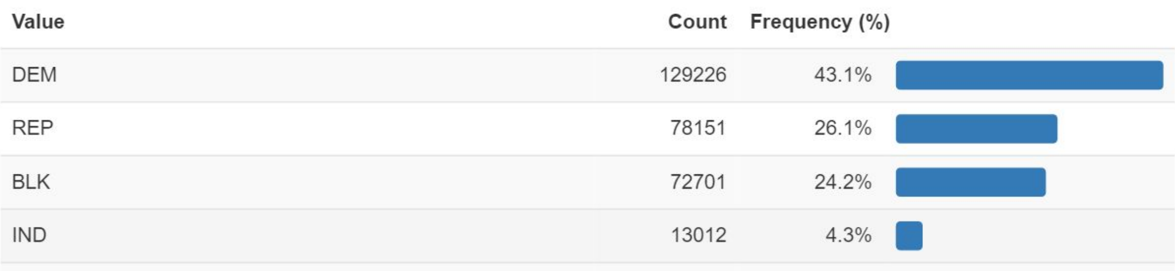
Dataset:

At present we have the Voter registration data records for the state of New York. The data was collected by Professor Jason Jones from the Department of Sociology at Stony Brook University for the election years 2012, 2016, 2017 and 2019. The exploratory analysis was done for the year 2017 only.

We also have an external dataset from the CDC (Centers for Disease Control and Prevention) which provides the statistics for the marriage and divorce rates through NVSR (National Vital Statistics Reports), this dataset will play an important role in validating the results of our prediction.

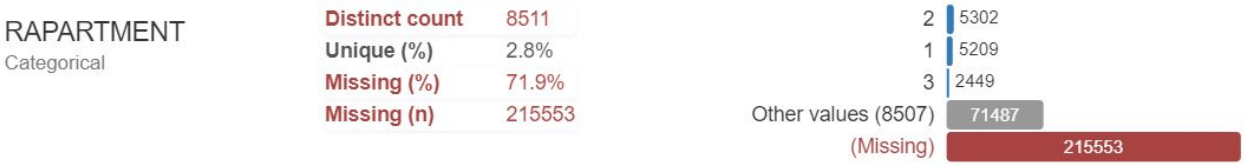
We are planning to take datasets from other states and do a similar procedure across all the datasets from the states we find the dataset of. We will take the dataset of the state of Florida next month as they roll out the data only on a monthly basis and the next extract date is in November.

Exploratory Analysis:



Value	Count	Frequency (%)
DEM	129226	43.1%
REP	78151	26.1%
BLK	72701	24.2%
IND	13012	4.3%

Figure 1 : Party affiliation statistics



RAPARTMENT
Categorical

Distinct count	8511
Unique (%)	2.8%
Missing (%)	71.9%
Missing (n)	215553

2	5302
1	5209
3	2449
Other values (8507)	71487
(Missing)	215553

Figure 2 : Apartment information

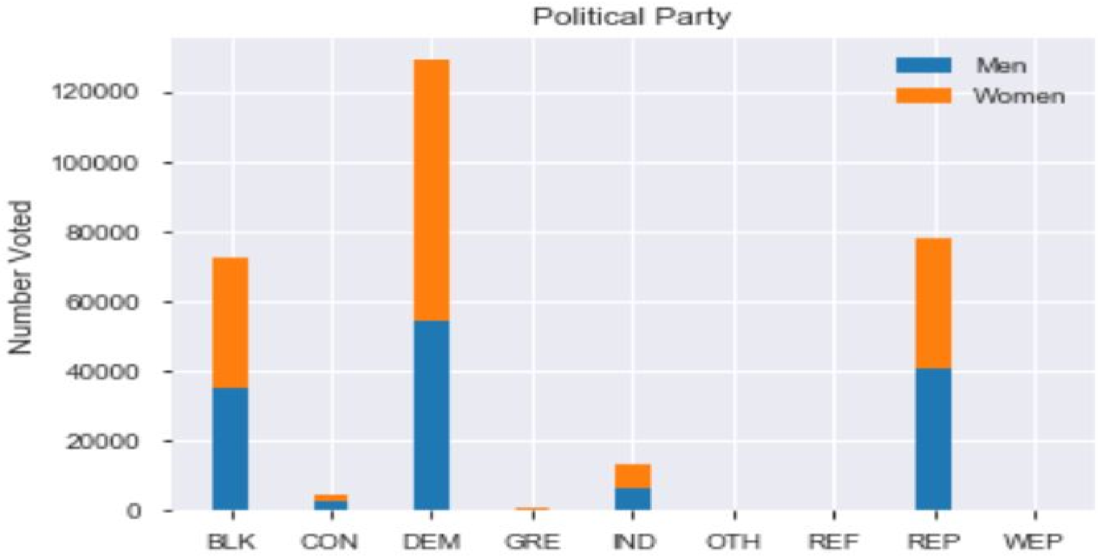


Figure 3: Political Party Affiliations by Gender

For extracting some meaning out of the data, we analyzed the data provided by Professor Jones. It is the election data of the New York state for the year 2017. It contains around 17 million data points. We extracted 3,00,000 data points sequentially to get a gist of the variables used and to find any pattern within the data. We did this process multiple times on our computer to reduce the sampling size bias. We found out the proportion of party affiliations of people which can be seen in figure 1. We found that majorly people belonged to the Democratic party at 47%, Republicans were the next at 26%.

Another major point we noticed in the data was that the RAPARTMENT column which gives the apartment number of a person. One critical observation was that 71.9% of the data was missing in that column. Hence, we need another dataset that will be used to extract the exact address of people that we can join with this data to assign addresses to the missing values. **data.ny.gov** is a resource that we found where we can find address related details. One dataset which we found was Property Address Directory, containing addresses of all properties in New York state and which is a potential candidate for finding the addresses.

The data shown in figure 3 shows the political affiliations by gender and gives a relative comparison between them. The Y axis is the count of the number of votes, the X axis being the political party affiliated with. What is evident from this dataset is there are more people affiliated with the Democratic party than the Republican, however within these two datasets we can see that there are a number of female Democrats than male Democrats. As for the Republicans, the number of female and male voters is nearly the same. What we can see is a disparity between Male and Female voters for the Democratic party and this disparity could possibly give us an interesting insight.

Challenges:

1) Identifying whether two people of different genders cohabitating are married.
    a) We are checking by last names. But, even fathers and daughters can have the same last name. Even, mothers and sons can have the same last name. Brothers and sisters can have the same last name.
    b) It's become a common practice now for women to not change the last name after marriage at all. So, it would further add to our challenge of identifying a marriage between 2 individuals.
2) Identifying whether two married people who are not cohabitating in the recent elections but were cohabitating previously are actually separated or migrated to other states or countries for a different job or any other reason.
3) We found that the rapartment field was missing in most of the locations of the dataset, this would make it difficult in precisely locating the address of a person and to compare it with other voters to identify whether they were cohabitating.
4) Working with such a large dataset (~6GB per election year per state) is also cumbersome, we would need access to high-performance clusters to process our data quicker.

Steps:

1. Identify which fields are relevant in identifying the address accurately, this will involve multiple columns merged.
2. We would also need to identify external datasets that could be merged with our data, this is in order to increase the accuracy with which we identify the address of a user. With this, we could also fetch the father's name of the electorates to better identify which pair is a married couple.
3. Sanitation and cleansing of the data would have to be done and irrelevant columns could be removed that would affect our analysis and training.

4. We then train our data using a Logistic regression model, this would be our baseline model that would give a probability of a couple getting divorced.
5. To get better results from our dataset, we will use a more complex model, it would possibly involve building upon either RandomForest Regression, Ridge Regression techniques to get a better score of our results.
6. Lastly, we would compare our data with real-world data either from the CDC or from actual election data for the next upcoming election.

Validation:

We are finding the probability of a married couple not cohabitating in the next elections because of a difference in their political party affiliations. For that, we are classifying the data into two datasets. Married couples with the same political party affiliation and married couples with different political party affiliations. For each dataset, we then will calculate the overall divorce rate. We will then validate the divorce rates of married couples with different political affiliations and divorce rates of married couples with the same political affiliations against the average divorce rates of that location. We have the divorce rate available to us from the dataset by the Centers for Disease Control and Prevention via the National Vital Statistics System Report (NVSR)

We will then validate our probabilities of each married couple cohabitating in the next election cycle by checking the election data for the next election. Precisely, we will come up with a probability threshold above which we will classify whether they will be cohabitating or not. We will calculate the threshold by training the data. We will validate our model by calculating various metrics like accuracy, precision, recall, and F1.

Reference:

[1]https://www.cdc.gov/nchs/nvss/marriage-divorce.htm?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fnchs%2Fmardiv.htm

[2]https://pdfs.semanticscholar.org/6a28/a5ffe7119fe2475c25ce868c99fa960869cc.pdf

[3]https://fivethirtyeight.com/features/how-many-republicans-marry-democrats/

[4]https://www.washingtonpost.com/news/wonk/wp/2016/07/01/the-interesting-thing-that-happens-when-a-republican-marries-a-democrat/

[5]https://qz.com/1410962/political-partisanship-in-the-us-is-now-deciding-love-and-marriage/

[6]http://www.inquiriesjournal.com/articles/127/the-effect-of-marriage-on-political-identification