Political Polarization and Marriage

CSE 519: Data Science Fundamentals Progress Report[1]

I. Abstract

A number of features of the contemporary environment affect voters' proclivity to divide the world into a liked-in group (one's own party) and a disliked-out group (the opposing party). Several studies show that many couples find themselves experiencing marital problems over their polarizing political views. As of late, breakups caused by political differences are becoming more and more public. Voter registration data is effective for identifying cohabiting couples and tracking them over each electoral snapshot to see whether they are still cohabiting. In this project, we try to find out if couples with alternate party affiliations (one Democrat and one Republican) ceased cohabiting at a greater than expected rate after President Trump was elected in 2016.

II. Overview

The primary objective of our study is to effectively analyze the effect of political affiliations of couples on their marriage post the election of 2016. We try to explore the voting registration data from two different states in US : Florida and New York where we first do an in-depth study of the couples from three different dimensions, the age group of couples, the race distribution of couples and the effect of demographics like overall popularity of a party in the area where couple resides. We explain our results for Florida and NY by presenting them in the same categories as above and discuss on the comparison between the results obtained from the two states together.

In order to identify the two most important features for our analysis, the set of couples and the separated couples, we designed two rule based algorithms that use multiple features including identification information and demographic information of each person in the dataset to find couples that fulfill maximum number of constraints. We analyzed over 20 million rows of data from Florida and 15 million rows from New York voter data to generate our two derived datasets. This helped us to perform aggregate level analysis over the entire states as well as deep-dive analysis on specific county ensuring generalized results that have minimum skew as if encountered in case of only county level analysis. All the analysis has been performed by identifying couples from a year and half prior to the 2016 elections where we estimate that the election campaigns don't start at least 1.5 years prior to the next elections and so the party affiliations of the couples are least affected. Similarly we identify the separated couples within 5 months of post elections with the hypothesis that political activities and discourse is the highest during the post months of elections and hence effect on the marriage of the couples will be most prominent during this time frame.

III. Experimental Setup

In this section we discuss our development setup and details about the datasets that we used in generating our results. We provide a summary about the end-to-end pipeline that we developed to find insights from the voter registration data.

A. *Development Environment*

One of the most crucial tasks in our project work was to have a solid development setup that could adapt to our requirements and modifications over time. Our top 4 candidates were Google Colaboratory, Google Cloud Compute (Limited free credits), Seawulf Cluster and running a Local Run-Time on one of our systems with decent specifications. Google colaboratory and Seawulf cluster stood out as two great choices for our project as they provide excellent resources in terms of hardware for running heavy data manipulation and analysis tasks. However we

[1]Professor Steven Skiena, Stony Brook University, NY - 11794.

went ahead with Google Colaboratory as our main development environment due to its flexibility and features over bare metal Seawulf cluster instances. Google Colaboratory is an excellent platform designed for data analysis which integrates wide array of tools for complex task that one would need for analysis. It provides a cloud based services for running Jupyter notebooks which are easy to use and shareable with other members. It runs an enhanced version of the Jupyter environment with advanced features like code completion, active member collaboration and wide number of storage integration options.

B. Dataset

Identification of right datasets for the project forms an important step in achieving right results. Voter registration data served as the primary dataset for our project that contains most of the information required for generating results. We also looked at external datasets that could aid our primary dataset such as records of married couples. However such data is personal identification information and wasn't accessible to everyone.

*1) **Florida Voter Registration Data**:* We choose Florida voter registration dataset as our main data source for building an analysis pipeline. The dataset contains voter registration information for all the 67 counties in the state in separate files. Each file contains information about multiple descriptors like Personal Information, Residential and Mailing Information along with Precinct level data.

*2) **New York Voter Registration Data**:* In addition to Florida data, we decided to analyze New York state data to understand and compare trends across the two states. This analysis helps us validate our results about the effect of elections across states. The New York voter registration data is different from Florida data in terms of features provided in the dataset. We found that fields like person's race and other identity fields were missing in NY data at the same time fields like previous address, previous vote county and similar fields were present. We modified our pipeline for NY data in order to accommodate new changes and extract as much useful information as possible. We have tried to analyze NY data in the similar way to Florida data as much as possible in order to provide accurate comparison between the results.

*3) **Output Dataset - Couples Identification**:* This is a derived dataset generated by our couple identification algorithm. The algorithm was run over all the counties in Florida and was able to identify 3,110,231 couples across the state. We generate a record with voter registration number along with other features for each couple we identify. Below is a list of features we include in the dataset:

- **County Code:** Field to identify the county where the couple resides.
- **Male Voter ID:** Field to identify male partner in the dataset with O(1) lookup.
- **Female Voter ID:** Field to identify female partner in the dataset with O(1) lookup.
- **House Address:** Field to keep track of previous residential address of the couple when performing analysis.
- **Male Party Affiliation:** Represents the electoral party supported by the male partner in the couple.
- **Female Party Affiliation:** Represents the electoral party supported by the female partner in the couple.
- **Race:** Represents the combined race of the couple, If the race of the two partners don't match, we quote this as Interracial marriage.
- **Male Age:** Field to save age of the male partner at the time of registration for voting.
- **Female Age:** Field to save age of the female partner at the time of registration for voting.

*4) **Output Dataset - Separated Couples Identification**:* This is another derived dataset generated by the separated couples identification algorithm. This dataset is the basis for our analysis on number of separated couples due to electoral reasons. The dataset includes all the same fields as the couples dataset above with three additional fields as below:

- **Previous House Address:** Field to keep track of previous residential address of the couple when performing analysis.

- **Male New House Address:** Field to save changed address of the male partner assuming the couple separated.
- **Female New House Address:** Field to save changed address of the female partner assuming the couple separated.

C. *Data Preprocessing*

Significant efforts are required to handle inconsistent records in the datasets that contribute to incorrect observations and results. We developed a detailed pipeline of steps to identify and eliminate such records. Below are few of the many challenges we faced while working with the data and the steps we too to handle them.

1) A null value analysis reveals fields that have majority of null values. We eliminated features that had more than 75 percent values as null and provided redundant information that was accessible from some other field. We use the interpolation method for null value treatment in the dataset wherever applicable.
2) Some fields had data values as '*'. This is the case when the voters deny to share their PII information publicly. We removed these records as sufficient data was missing to identify couple pairs.
3) We identify people living in the same household by matching their addresses using Levenstein distance metric which provides an excellent way to check if strings are close match to one another. We match and fix all the addresses that are at most 3 edit-distance away.
4) The data provides multiple fields for address identification. We ignore fields that are redundant or mostly null. Residence Address Line fields are our primary features for address identification which are augmented by city, zipcode and county fields.
5) The dataset provides date of birth of each voter, we calculate the age of each voter based on the date they registered with the Florida government for voting rights.
6) We performed basic lowercase operation and whitespace character removal from string fields.

D. *Algorithms*

1) ***Couples Identification Algorithm****:* The core component of our project work is to design an algorithm that could identify couples from the given voter registration data with high accuracy. We improved our initial algorithm from last sprint and used it for our final analysis. Below is a high-level pseudo-code describing various steps involved in our algorithm.

Algorithm 1 Couple Identification Algorithm

```
countyList ← List(counties)
coupleList ← List()
for each county in countyList do
    preprocessData(county)
    families ← identifyHouseholds(county)
    for each family in families do
        couples ← createMatching(family)
        Append coupleList ← couples
    end for
    Save countyList
end for
```

The algorithm starts by fixing all the addresses that have small typos and are within 2 edit-distance of one another. We use Levinstein distance module for performing this task. Next we call the module for identifying all the households in the given county. This module once again uses edit-distance algorithm to group people that have the same addresses. Once completed, we have a list of families in the counties and each object stores all the members that make up that family. Next, our couple matching module tries to find best matching between the members by eliminating pairs that are more likely to be siblings as well as pairs that have wide age gap as they are more likely to be children-parent relationships. We identify maximum possible matchings that satisfy the rules and save them in our couple's dataset.
In terms of matching couples, we identify all the couples even when multiple of them are staying in a single household. Our matching algorithm detects and forms pairs weighing the utility of each feature appropriately.

2) ***Separated Couples Identification Algorithm****:* Once the couples are identified, the next important task is to identify subset of couples

that have separated post the elections of 2016 due to their political affinity towards different parties. Similar to couples identification algorithm, we

Algorithm 2 Separated Couples Identification Algorithm

$countyList \leftarrow List(counties - 2016 - data)$
$seperatedCouplesList \leftarrow List()$
for each $county$ in $countyList$ **do**
  $preprocessData(county)$
  $couples \leftarrow retrieveCouples(county)$
  for each $couple$ in $couples$ **do**
    if $couple.male$ in $county$ and $couple.female$ in $county$ and $isAddressDifferent(address, address)$ **then**
      $seperatedCouplesList \leftarrow couple$
    end if
  end for
  Save $seperatedCouplesList$
end for

have improved our couple separation identification algorithm shown in Algorithm-2 since the last sprint. The algorithm uses combination of address fields to build tuples that show separation due to political affinity of the couples. We keep fields like original address of the couples along with new addresses of the male and female partner in order to find insights about which partner in the couple changed address. We also retain the age and race information of the couples to analyze age group and race based distribution of separated couples. Our algorithm was able to filter almost 0.15 million couples from 3 million couples who separated over the period of 2015 and 2016.

IV. Analysis and Results

This section presents the results from the analysis of data from Florida and New York. We present the analysis for each state separately along with further granularity based on three main feature : Age, Race and Demographics.

A. *Florida Data*

In this section we explore interesting trends in Florida voter registration data over the years from 2015 to 2017. We use age, race and demographics as our key features in describing trends across all couples vs divorced couples in Florida.

*1) **Feature Analysis - Age**:* We begin by analyzing the age-group distribution of identified couples in the dataset to get an understanding of overall married population in Florida. We choose the younger partner in the relationship to determine the age-group for the couple. Fig 1 shows a distribution of all the couples in Florida and the age group that they fall in.

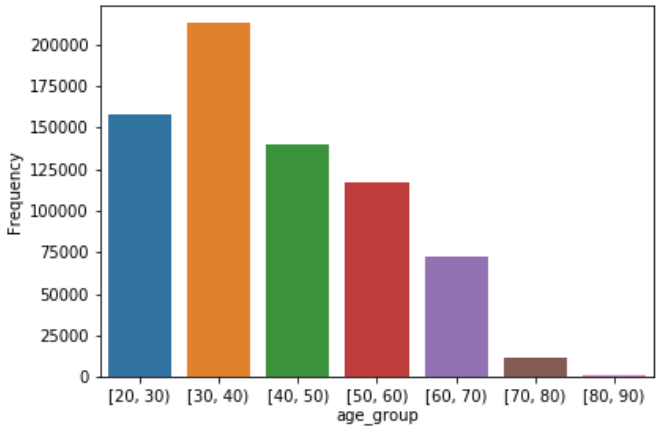

Fig. 1: Distribution of couples in different age groups

We see that highest number of couples fall in the age group of 30-40 years. Followed by age-group of 20-30.
Next, we study the distribution of each age-group

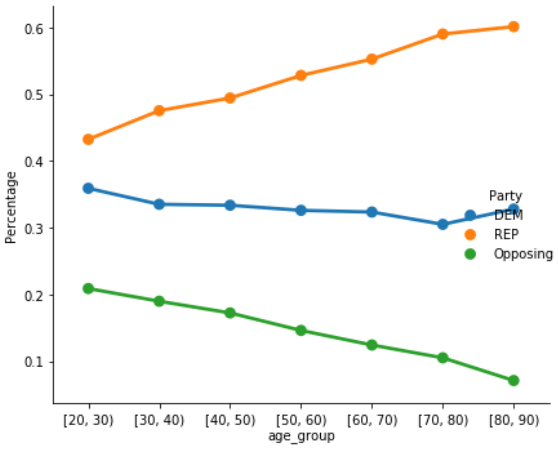

Fig. 2: Distribution of couples in different age groups with their party affiliations

in Florida based on their party affiliation. Fig 2 shows a distribution of percentage of couples in each age-group based on weather they support

DEM, REP or conflicting parties.

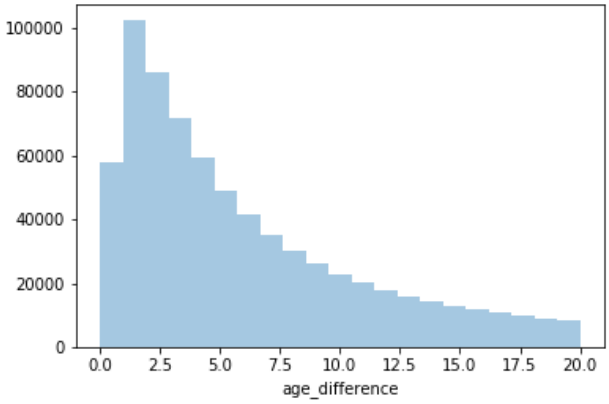

Fig. 3: Distribution of couples in different age difference range

We can observe from the distribution that REP dominates in Florida majorly. We also notice the trends over increasing age-group such that older couples support REP. We also see that couples supporting conflicting parties reduce as the age-group increases i.e. older couples have more harmony in terms of supporting same party. Next, we analyze

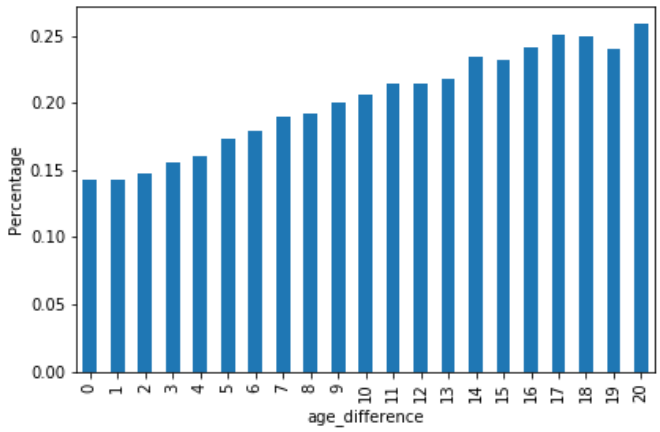

Fig. 4: Couples with opposite party affiliations in different age difference range

trends based on age-difference between partners in a relationship. We start by plotting a distribution for all the couples in the dataset showing their age-differences. Fig 3 shows a distribution for number of couples in each age-difference, from 1 to 20.

We observe that most of the couples have an age-difference between 0 to 7 years with the counts decreasing as age difference increases. Next, we try to analyze correlation between the age-difference in couples and their party affiliations. Figure 4 shows a distribution for party-affiliation between couples as the age-difference increases.

We observe that there is a correlation between increasing age-difference and party affinity of the partners. Couples with higher age-difference tend to have higher chances of conflicting party-association in the marriage then compared to couples that have smaller age-difference.

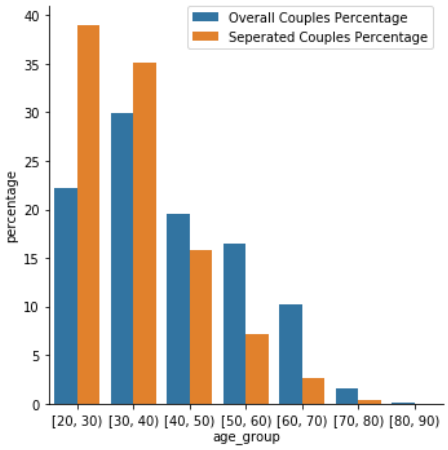

Fig. 5: Age group of separated vs all couples

Next, we try to analyze separated couples in Florida and contrast the result with entire couple distribution. We start by comparison between the distribution of age-groups for separated couples vs age-group distribution of all the couples in Florida. Fig 5 shows this comparison. We see that even though highest number of the couples lie in the age-group of 30-40, maximum number of separated couples lie in the age-group of 20-30 i.e. most of the couples separated during the early phase of the marriage and divorces decreased substantially after age of 40.

*2) **Feature Analysis - Race**:* In this section, we analyze couples based on their race and study their party-affiliations. First we start with analysis on overall party affiliation of couples belonging to different race groups. Figure 6 shows distribution of couples grouped by their race and intensity of their affinity towards different parties.

We observe that there are two race groups that stand out in terms of their party affiliation. Black

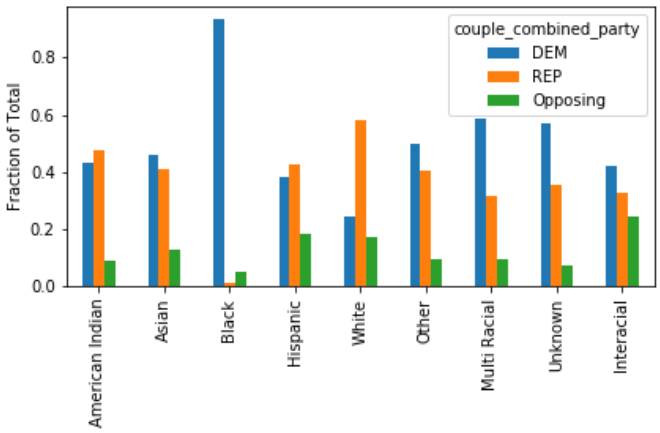

Fig. 6: Distribution of couples from different races based on their party affiliations

community seems to favour DEM by a large percentage on the other hand, white community seems to favour REP in large numbers.

Next, we try to analyze races that have highest number of couples with conflicting party associations.

Figure 7 shows a bar plot for each race and the percentage of couples that have conflicting party affinity between the partners.

We observe that Interracial couples have the highest number in terms of opposite party association followed by White and Hispanic people.

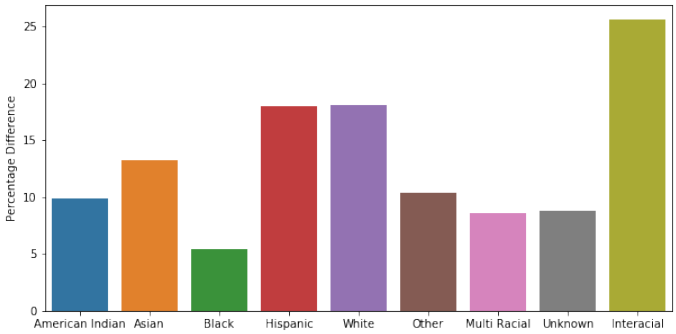

Fig. 7: Distribution of age difference for separated vs all couples

*3) **Feature Analysis - Demographics**:* In this section we analyze the identified couples based on demographics of the state.

First, we look at top 5 counties that have have highest opposing polarity in terms of party affiliation in marriages. Figure 8 shows a distribution of counties that have opposing party affiliation in couples. We also highlight top 5 counties among all.

Next we try to analyze top 5 counties that have highest number of couples separated due to conflicting party-affiliations. Figure 9 shows a bar graph with top 5 counties that have highest number of seperated couples with conflicting party-affiliations. From the graph we can observe that counties Miami-Dade, Osceola, Hillsborough, Monroe and Volusia have highest percentage of separation among married couples due to their political affiliations.

Lastly, we analyze separated couples to understand which partner in the relationship left the house when the couple separated. Fig 10 shows the distribution of separated couples based on which gender left the home post divorce.

We clearly see that female partners in the relationship left the house more as compared to male partners post separation.

B. New York Data

This section explores interesting trends in New York voter registration data between year 2012 and 2017. We use age and demographics as key features in identifying trends across party affiliations of couples and separated couples.

*1) **Feature Analysis - Age**:* We start by analyzing the distribution of couples based on their age group. We classify couples into a particular age-group based on the age of younger partner. Fig 11 shows a distribution between couples in different age-groups.

The plot shows that more than 60 percent of the total couples identified lie in the age-group of 20-40. Next, we analyze the party affiliation of couples in different age-groups. Fig 12 shows a distribution about the number of couples that support DEM, REP and Conflicting party-affiliations.

From the above plot we can make the following observations. First, New York is predominantly a Liberal state and hence we see more than 50 percent of the couples supporting Democrats. In terms of age-groups, couples in younger age groups are supporting democrats but the percentage shows a fall as the age group is higher. The opposite is true for REP i.e. couples in older age-groups tend to support REP more as compared to Democrats. The

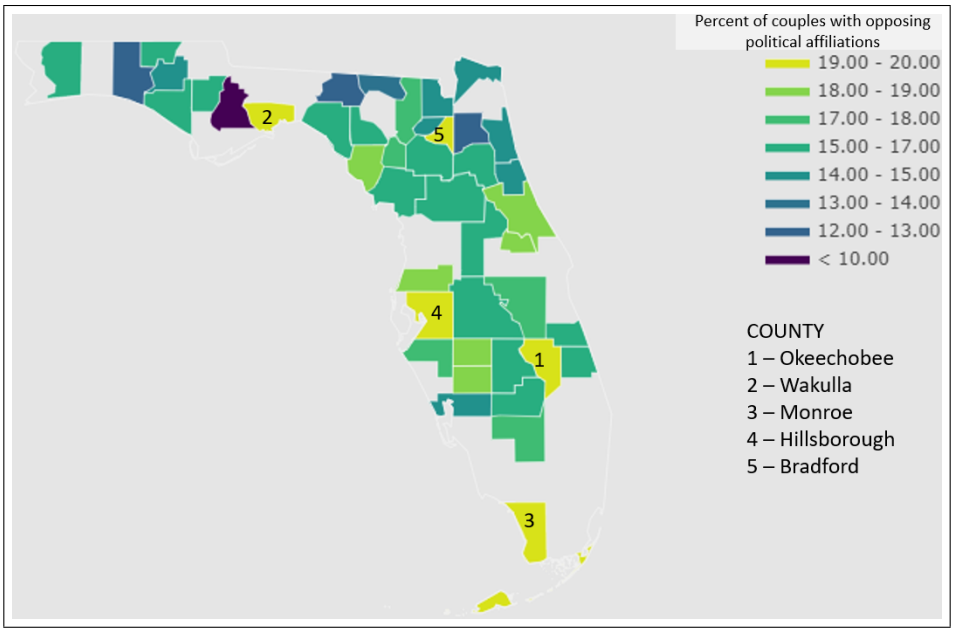

Fig. 8: Percent of couples with opposing political affiliations over different counties in Florida

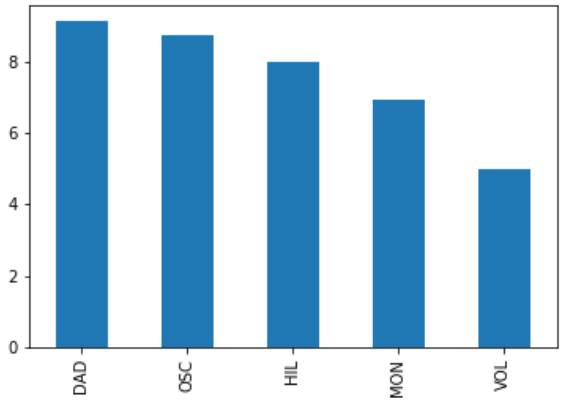

Fig. 9: Top 5 counties with highest number of couples with conflicting party-affiliations

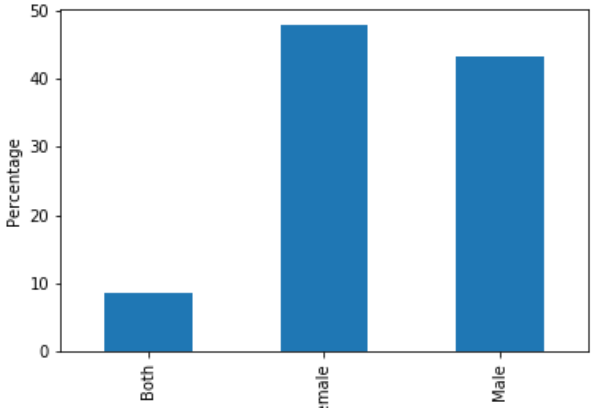

Fig. 10: Percentage of each gender leaving the house on separation

opposite party affiliation couples show a declining trend for higher age-groups.

Next, we try to study age-difference between partners in a marriage in order to measure its effect on their political affinity. First we try to see the distribution of couples based on their age-difference. Fig 13 below shows the overall distribution of couples based on their age-difference. We see that more than 55 percent of the couples lie in the range of 0 to 5 years of age-difference. Next we try to see if the age-difference between couples correlates with their differences in party-affiliations. Figure 14 shows a distribution of couples with conflicting party-affiliations based on their age-difference.

For the figure 14, we can clearly see an increase in percentage of opposite party-affiliation in couples as the age-difference between them increases. Couples of nearly same age tend to have lower conflicting party-affiliations as compared to couples with age-difference of 7 or more.

Next we try to compare distribution of separated couples vs all couples based on same features as analyzed above.

Fig 15 below shows a comparison between age-

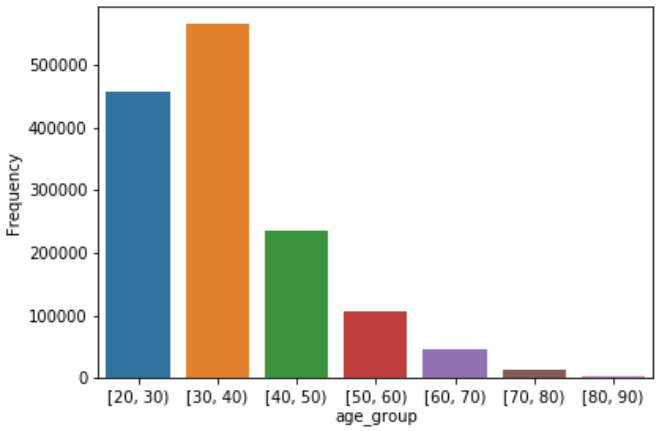

Fig. 11: Distribution of couples in different age-groups

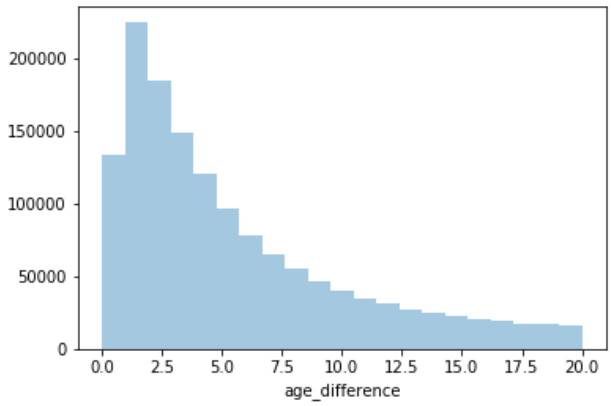

Fig. 13: Distribution of couples based on their age difference

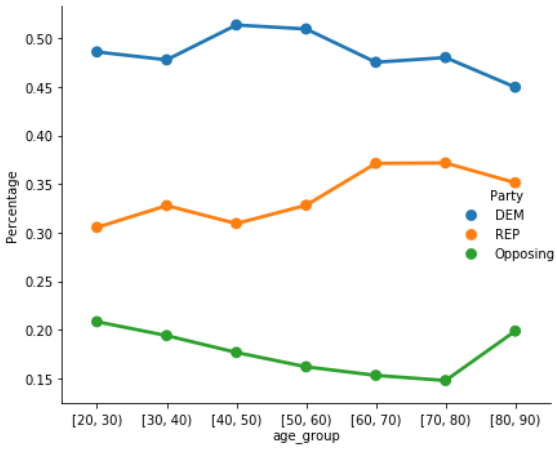

Fig. 12: Distribution of couples with different party affiliations

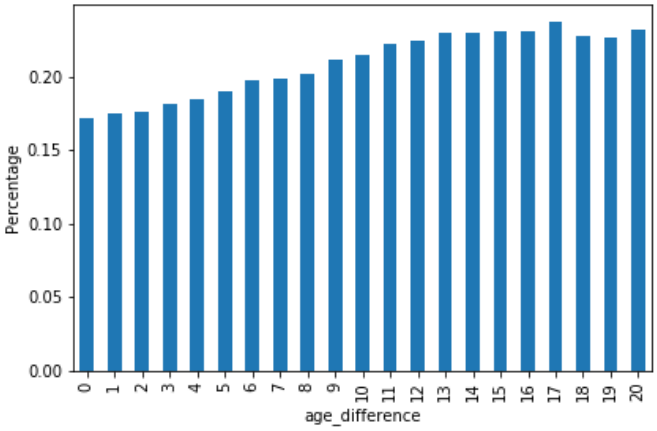

Fig. 14: Distribution of couples with conflicting party-affiliations based on their age-difference

group of all couples and age-groups of separated couples.

The plot shows that even though majority of the couples lie in the 30-40 age group, most of separations happened between couples that are under age-group of 20-30 i.e. young-couples formed majority of the divorces with numbers decreasing rapidly as age-group increases.

Next, we analyze the distribution of divorced couples that share opposing party-affiliations as compared to those that have same party-affiliations. Fig 16 shows this distribution. We observe that percentage of separated couples having opposing party affiliations is much higher as compared to those having same party-affiliation. This reaffirms the fact that divorces have a strong correlation with opposing political affinities of the partners in the marriage.

*2) **Feature Analysis - Demographics**:* In this section we present our analysis on demographics and its effects on political affiliations of couples.

First, we start by identifying top 5 counties that have highest number of couples with opposing party affiliations. Fig 17 shows this distribution.

We can observe that counties Cattaraugus, Cortland, Cayuga, Broome and Montgomery have the highest percentage of couples that have opposing party polarity.

Next we try to find the counties that have maximum percentage of couples that separated due to political affinity. Fig 18 shows this distribution.

From the above figure, we can observe that

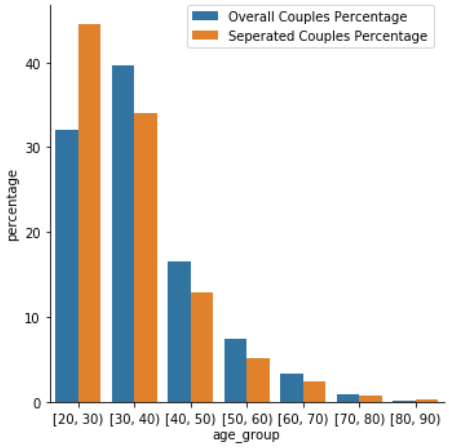

Fig. 15: Comparison of age groups of overall couples and age groups of separated couples

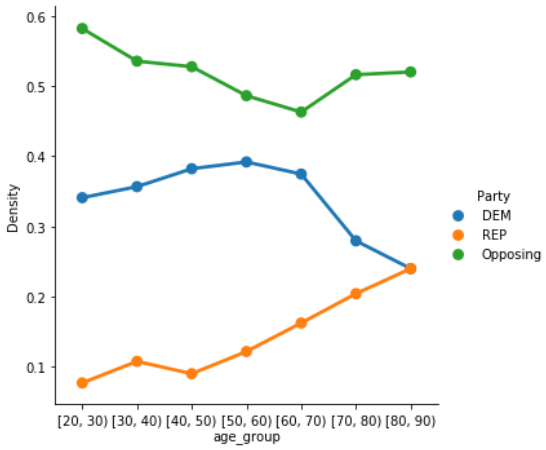

Fig. 16: Distribution of divorced couples with opposing party-affiliations

counties Hamilton, Wyoming, Essex, Orange and Ulster have the maximum percentage for separated couples in New York.

Finally we analyze the separated couples to understand which gender in the relationship left the house when they separated. Fig 19 shows the distribution based on gender, explaining the percentage of each gender in terms of leaving the house after split.

We can observe that in most of the cases, both the partners decided to leave the address that they were staying in, with equal number of males and females deciding to leave the house when the other partner kept staying in the same house.

V. Conclusion

Effects of political polarization on couples have been well documented. Through this project work, we try to study the effects of 2016 elections on the marriages in two politically polarized states in the US, Florida and New York. The two states are interesting in the fact that we have Representative's dominating in Florida and Democrats dominating in state of New York. We designed algorithms to identify all the couples in the two states and a novel algorithm to identify couples that have separated due to conflicting political associations. We performed an extensive analysis on the data from both the states for multiple key features like Age, Race and Demographics. We observed the changes in number of separated couples from year 2012 to 2017 and found that percentage of separated couples post the election of 2016 increased by 0.5 percent than the previous year. As a future work, we will continue to analyze data from other states in order to find more results that could strengthen our hypothesis.

VI. Next Steps

Moving to our next phase of development, below are some steps identified for our project:

- Our analysis could be extended to the other states in the USA.
- We strive to obtain actual marriage data from the state of Florida and New York and validate our results.

References

[1] Hersh, Eitan, and Yair Ghitza. "Mixed partisan households and electoral participation in the United States." PloS one 13, no. 10 (2018): e0203997.

[2] Ansolabehere, Stephen, and Eitan D. Hersh. "ADGN: an algorithm for record linkage using address, date of birth, gender, and name." Statistics and Public Policy 4, no. 1 (2017): 1-10.

[3] https://www.theatlantic.com/family/archive/2019/03/can-families-communicate-across-the-political-divide/585379/

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

[5] https://science.sciencemag.org/content/360/6392/1020

[6] Yujian, Li, and Liu Bo. "A normalized Levenshtein distance metric." IEEE transactions on pattern analysis and machine intelligence 29, no. 6 (2007): 1091-1095.

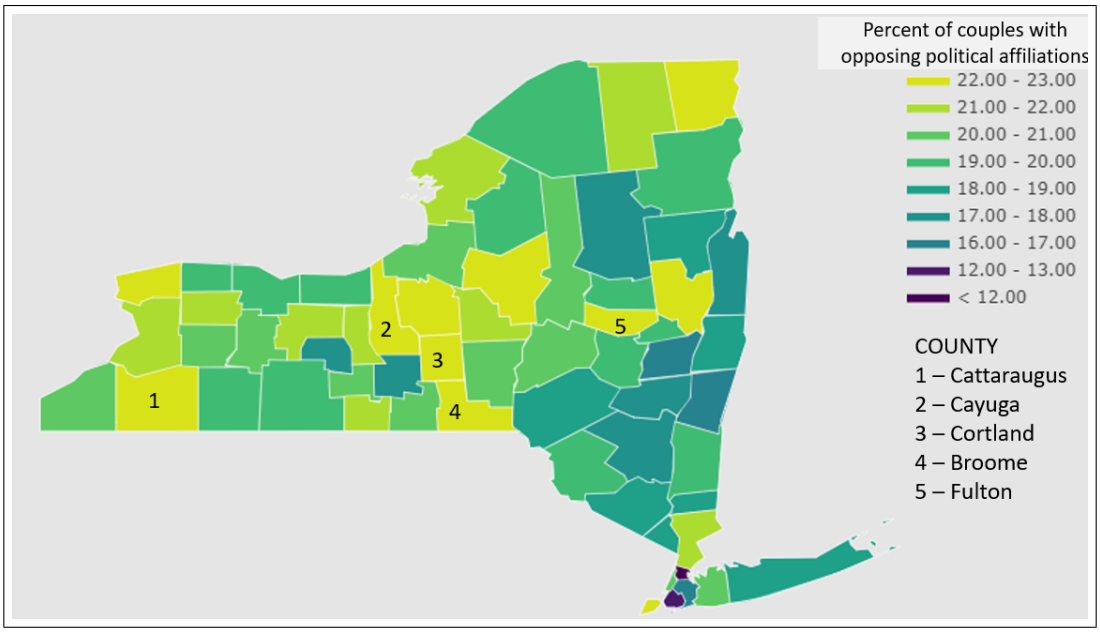

Fig. 17: Percent of couples with opposing political affiliations over different counties in New York

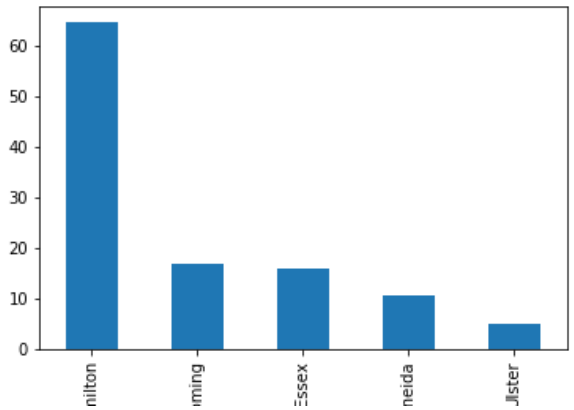

Fig. 18: Top 5 counties with highest number of separated couples with opposing party affiliations

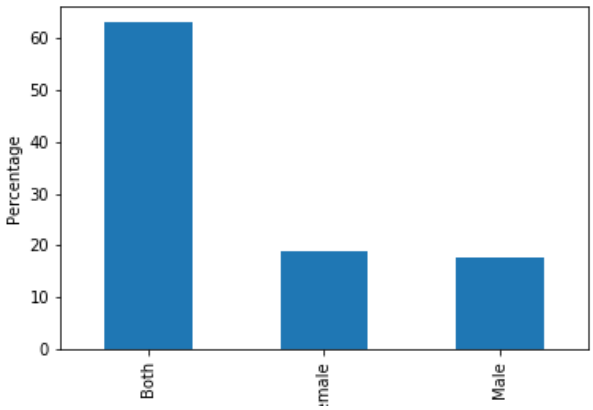

Fig. 19: Percentage of people leaving the house based on gender