Exploratory Data Analysis

Introduction to Data Science – CS 418

Project 1

Submitted by:

Yukthi Papanna Suresh

Emily Lin

Caglar Kurtkaya

INDEX

Abstract:

Used demographics and election data to do Exploratory data analysis to identify if a county had major Democratic votes or Republican votes. The Party attribute was then used to identify the type and background of people that voted for the two parties.

Tasks:

Task 1:

Reshaping dataset election_train from long format to wide format:

Used pivot table to reshape the dataset to get the dimensions of election_train data as follows:

```
#Inspecting the dimensions of election data
print('election_tidy dimensions:{}'.format(election_tidy.shape))
```

```
election_tidy dimensions:(1205, 6)
```

Task 2:

Merging reshaped dataset election_tidy with dataset demographics_train:

Accounted for the following inconsistencies:

- As in the election_tidy dataset, each county name is followed by the string 'County' which is not the case with demographics_train dataset – ***replaced string County with ' ' in election_tidy dataset***
- As in the election_tidy dataset, each State name is the abbreviation which is not the case with demographics_train dataset – ***changed State values in demographics_train dataset to abbreviations using map function***
- As there are some case-sensitive issues in names of Counties in both datasets, ***converted the values of County variable in both datasets to lower case***

Finally, merged both datasets to get the following dimensions for the merged data:

```
#Inspecting the dimensions of merged data
print('merged_data dimensions:{}'.format(merged_data.shape))
```

```
merged_data dimensions:(1200, 21)
```

Task 3:

Exploring the merged data:

No. of variables = 21

```
#Task 3:
#Exploring the merged dataset
merged_data.shape
```

```
(1200, 21)
```

Types of variables:

```
Year                                    int64
State                                  object
County                                 object
Office                                 object
Democratic                            float64
Republican                            float64
FIPS                                    int64
Total Population                        int64
Citizen Voting-Age Population           int64
Percent White, not Hispanic or Latino  float64
Percent Black, not Hispanic or Latino  float64
Percent Hispanic or Latino             float64
Percent Foreign Born                   float64
Percent Female                         float64
Percent Age 29 and Under               float64
Percent Age 65 and Older               float64
Median Household Income                 int64
Percent Unemployed                     float64
Percent Less than High School Degree   float64
Percent Less than Bachelor's Degree    float64
Percent Rural                          float64
dtype: object
```

Irrelevant variables: As the variables Year and Office have same values for all observations, these are not useful for any kind of analysis.

Dealing with Irrelevant variables: Dropped them

Task 4:

Searching the merged data for missing values:

```
Int64Index: 1200 entries, 0 to 1199
Data columns (total 19 columns):
State                                1200 non-null object
County                               1200 non-null object
Democratic                           1195 non-null float64
Republican                           1195 non-null float64
FIPS                                 1200 non-null int64
Total Population                     1200 non-null int64
Citizen Voting-Age Population        520 non-null float64
Percent White, not Hispanic or Latino    1200 non-null float64
Percent Black, not Hispanic or Latino    1155 non-null float64
Percent Hispanic or Latino           1195 non-null float64
Percent Foreign Born                 1197 non-null float64
Percent Female                       1200 non-null float64
Percent Age 29 and Under             1200 non-null float64
Percent Age 65 and Older             1200 non-null float64
Median Household Income              1200 non-null int64
Percent Unemployed                   1197 non-null float64
Percent Less than High School Degree     1200 non-null float64
Percent Less than Bachelor's Degree      1200 non-null float64
Percent Rural                        1181 non-null float64
dtypes: float64(14), int64(3), object(2)
```

- As seen, the variables Democratic and Republican have 5 missing values – ***dropped the 5 observations***.

- The variable Citizen Voting-Age Population has 680 missing values – this does account for missing values as we observe the Total population variable for each county and the corresponding Citizen Voting-Age Population variable has 0 value where population is substantially high – ***dropped the variable Citizen Voting-Age Population.***

- The missing values for rest of the variables makes sense as they are percentages with respective to a given category such as age, gender, race and ethnicity and education.

Task 5:

Created a new variable named "Party" that labels each county as Democratic or Republican.

Percent Age 65 and Older	Median Household Income	Percent Unemployed	Percent Less than High School Degree	Percent Less than Bachelor's Degree	Percent Rural	Party
13.322091	32460	15.807433	21.758252	88.941063	74.061076	1
19.756275	45383	8.567108	13.409171	76.837055	36.301067	0
10.873943	51106	8.238305	11.085381	65.791439	31.466066	1
26.397638	40593	12.129932	15.729958	82.262624	41.062000	0
12.315809	47422	14.424104	14.580797	86.675944	46.437399	0

Task 6:

Compute the mean population for Democratic counties and Republican counties. Which one is higher? - ***Mean population of Democratic counties is higher than that of Republican counties***

Party	Total Population
0	53864.672414
1	300998.316923

2-sample hypothesis test:

t-test statistic = 8.004638577960957
p-value= 2.0478717602973023e-14

Since p-value < significance level -> we reject the null hypothesis.

This means that there is a substantial difference in the mean population for Democratic counties and Republican counties.

Task 7:

Compute the mean median household income for Democratic counties and Republican counties. Which one is higher? - ***Mean median household income of Democratic counties is higher than that of Republican counties***

Party	Median Household Income
0	48746.819540
1	53798.732308

2-sample hypothesis test:

t-test statistic = 5.479141589767387
p-value= 7.149437363182598e-08

Since p-value < significance level -> we reject the null hypothesis.

This means that there is a substantial difference in the mean median household income for Democratic counties and Republican counties.

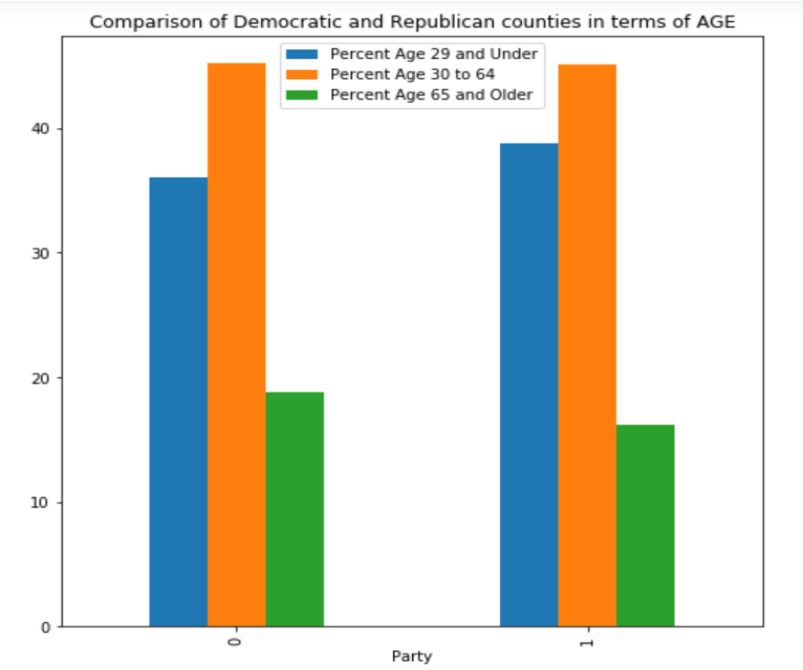
Task 8:

Comparing Democratic counties and Republican counties in terms of age:

Descriptive statistics(mean):

	Party	Percent Age 29 and Under	Percent Age 30 to 64	Percent Age 65 and Older
0	0	36.005719	45.166015	18.828267
1	1	38.726959	45.078214	16.194826

Plot:



As seen from the statistics and plot,
For Democratic counties - the mean values of the percentages for young population(Percent Age 29 and under) are slightly high and the older population(Percent Age 65 and Older) are slightly low compared to Republican counties.

The middle age population(Percent Age 30 to 64) mean values of percentages are almost same for both, hence not useful for analysis.

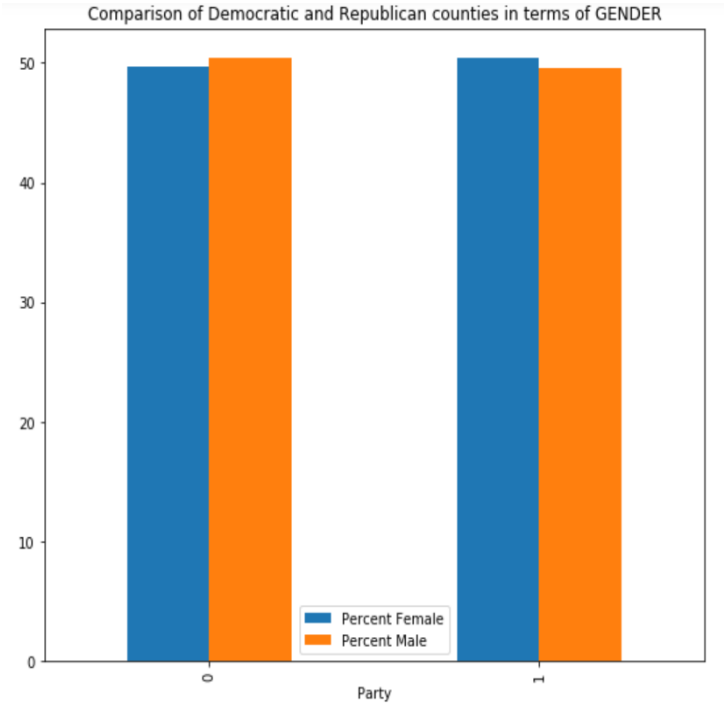
So, the younger population voted more, and older population voted less for Democratic counties than for Republican counties.

Comparing Democratic counties and Republican counties in terms of gender:

Descriptive statistics(mean):

	Party	Percent Female	Percent Male
0	0	49.630898	50.369102
1	1	50.385433	49.614567

Plot:



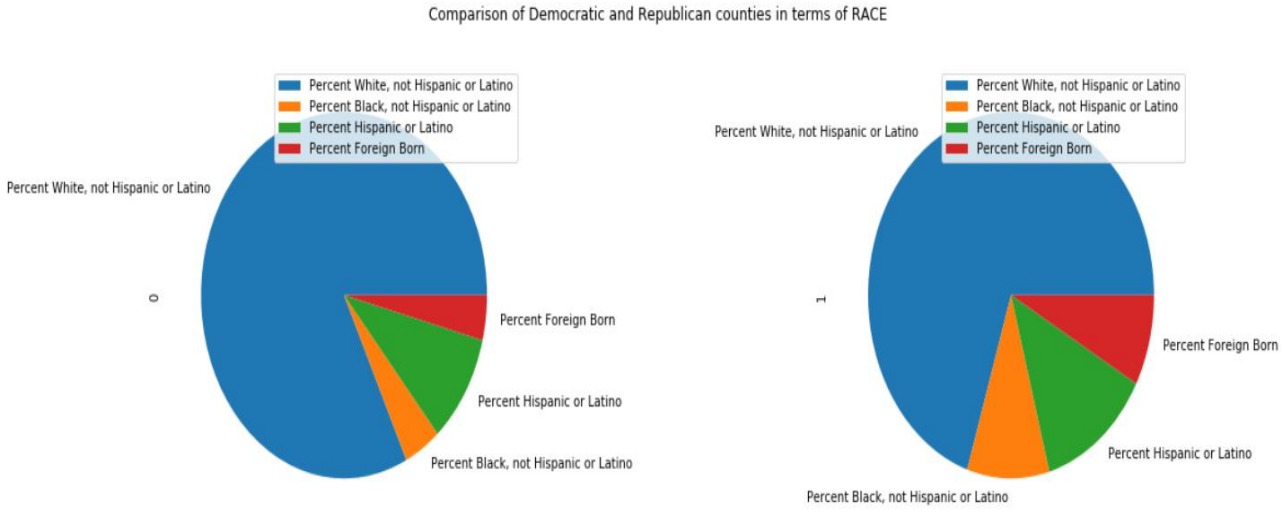
As seen from the statistics and plot,
For Democratic counties - the mean values of the percentages both genders are similar compared to that of Republican counties, hence not very useful for analysis.

Comparing Democratic counties and Republican counties in terms of race and ethnicity:

Descriptive statistics(mean):

	Party	Percent White, not Hispanic or Latino	Percent Black, not Hispanic or Latino	Percent Hispanic or Latino	Percent Foreign Born
0	0	82.656646	4.189241	9.733094	3.990096
1	1	69.683766	9.242649	12.587391	7.986330

Plot:



As seen from the statistics and plot,
For Democratic counties - the mean values of the percentages for Black, Hispanic and Foreign born are slightly higher compared to that of the Republican counties.
For Democratic counties - the mean values of the percentages for White are slightly lower compared to that of the Republican counties.

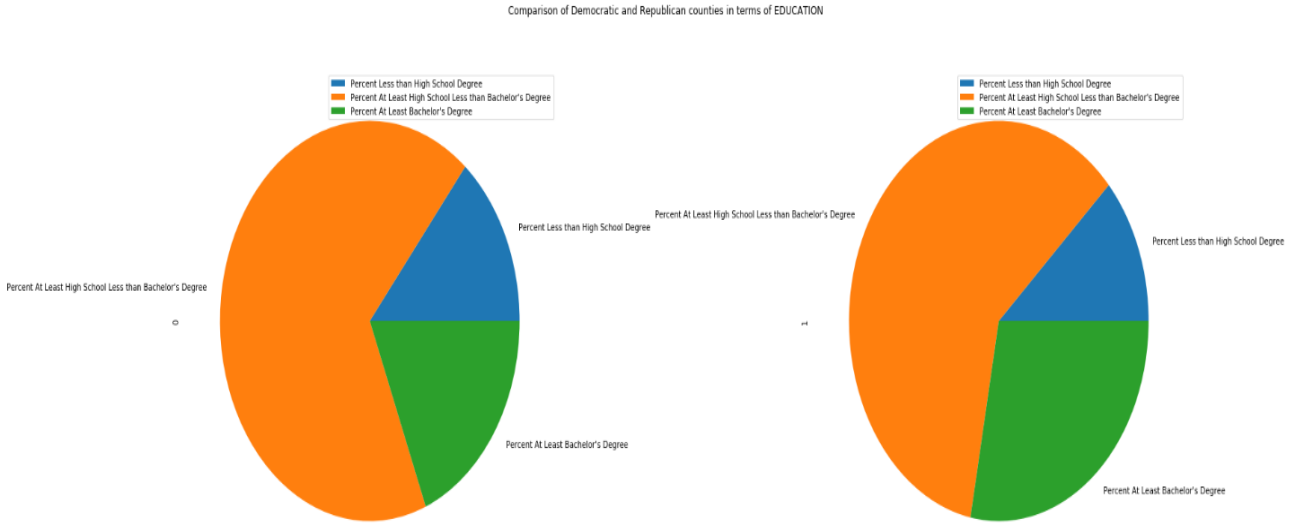
So, the Black, Hispanic or Latino and Foreign-born population voted more for Democratic counties than for Republican counties.

Comparing Democratic counties and Republican counties in terms of education:

Descriptive statistics(mean):

	Party	Percent Less than High School Degree	Percent At Least High School Less than Bachelor's Degree	Percent At Least Bachelor's Degree
0	0	14.009112	67.086315	18.904573
1	1	11.883760	60.084465	28.031775

Plot:



As seen from the statistics and plot,
For Democratic counties - the mean values of the percentages for less than high school degree and for at-least high school degree less than bachelor's are lower compared to that of the Republican counties.
For Democratic counties - the mean values of the percentages for at least bachelor's degree are higher compared to that of the Republican counties.

So, the people who are holding at least bachelor's degree voted more for Democratic counties compared to Republican counties.

Task 9:

The variables 'Percent Black, not Hispanic or Latino' , 'Percent Hispanic or Latino' , 'Percent Foreign Born' ,' Percent Less than bachelor's degree' and 'Percent Less than High School Degree' are more important than others to determine whether a county is labelled as Democratic or Republican.

As observed from the plots previously, these variables give more insights as to identify and classify the type and category of people from each county that either voted for Democratic or Republican.

For a given county to be labelled as either Democratic or Republican, we need to analyze the majority type and category of people that voted for either party. The other variables did not help much in clearly classifying the type and background of people that voted for either parties.

Task 10:

Creating a map of Democratic counties and Republican counties using the counties' FIPS codes