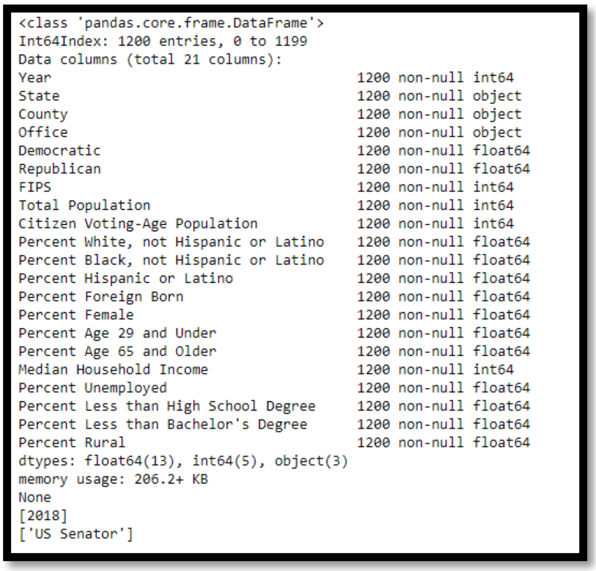
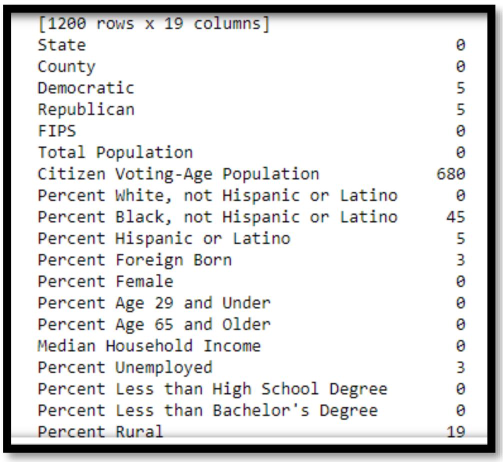
IDS Project 01 REPORT

1. We use the pivot_table() function to reshape the dataset election_train from long to wide format.
2. We remove the word ‘County’ from the reshaped dataset and convert all the county names to lower case. We convert all the county names to lower case in the demographics_train dataset too.
3. There is a total of 21 variables in the dataset. The types of these variables are integers, floats, and strings. There are some irrelevant and redundant variables which include ‘Year’ and ‘Office’. To deal with these variables, we will remove (or drop) them from the dataset.

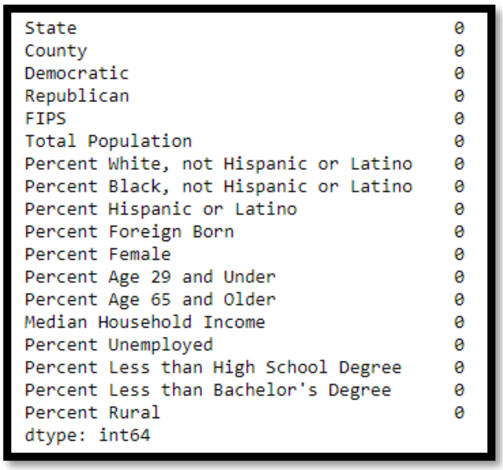


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1200 entries, 0 to 1199
Data columns (total 21 columns):
Year                                         1200 non-null int64
State                                        1200 non-null object
County                                       1200 non-null object
Office                                       1200 non-null object
Democratic                                   1200 non-null float64
Republican                                   1200 non-null float64
FIPS                                         1200 non-null int64
Total Population                             1200 non-null int64
Citizen Voting-Age Population                1200 non-null int64
Percent White, not Hispanic or Latino        1200 non-null float64
Percent Black, not Hispanic or Latino        1200 non-null float64
Percent Hispanic or Latino                   1200 non-null float64
Percent Foreign Born                         1200 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                           1200 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                                1200 non-null float64
dtypes: float64(13), int64(5), object(3)
memory usage: 206.2+ KB
None
[2018]
['US Senator']
```

4. There are missing values in the dataset. The missing values are in the column of “Citizen Voting-Age Population”. To deal with these missing values, we will remove (or drop) them from the dataset.



```
[1200 rows x 19 columns]
State                                      0
County                                     0
Democratic                                 5
Republican                                 5
FIPS                                       0
Total Population                           0
Citizen Voting-Age Population            680
Percent White, not Hispanic or Latino      0
Percent Black, not Hispanic or Latino     45
Percent Hispanic or Latino                 5
Percent Foreign Born                       3
Percent Female                             0
Percent Age 29 and Under                   0
Percent Age 65 and Older                   0
Median Household Income                    0
Percent Unemployed                         3
Percent Less than High School Degree       0
Percent Less than Bachelor's Degree        0
Percent Rural                             19
```



```
State                                    0
County                                   0
Democratic                               0
Republican                               0
FIPS                                     0
Total Population                         0
Percent White, not Hispanic or Latino    0
Percent Black, not Hispanic or Latino    0
Percent Hispanic or Latino               0
Percent Foreign Born                     0
Percent Female                           0
Percent Age 29 and Under                 0
Percent Age 65 and Older                 0
Median Household Income                  0
Percent Unemployed                       0
Percent Less than High School Degree     0
Percent Less than Bachelor's Degree      0
Percent Rural                            0
dtype: int64
```

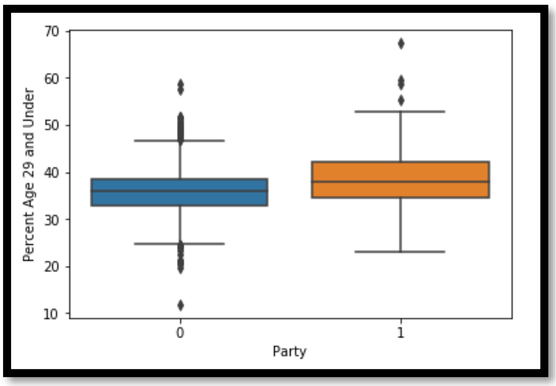
5. We make a function for setting the value to 0 or 1 as required for ‘Democratic’ or ‘Republican’ respectively.

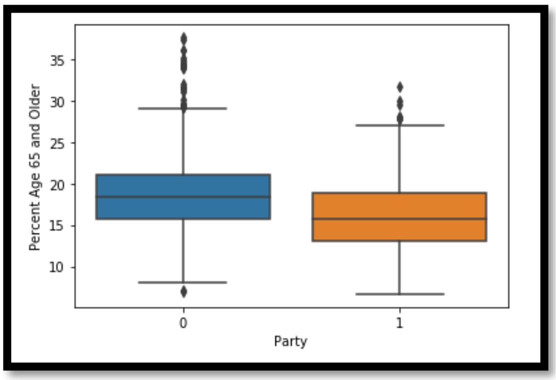
6. The mean population that is higher between the Democratic and Republican counties is in the Democratic counties. The result of the hypothesis test is that the p-value is greater than

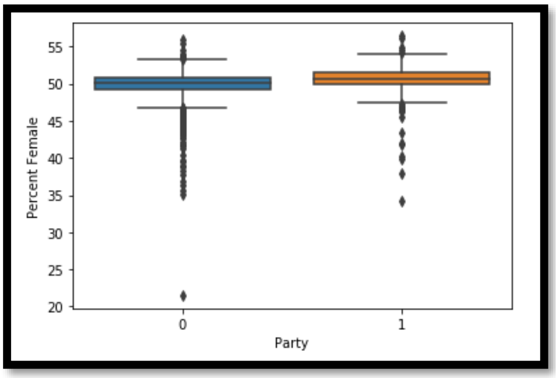
the significance value. Therefore, the result of the hypothesis test is that we reject the null hypothesis.

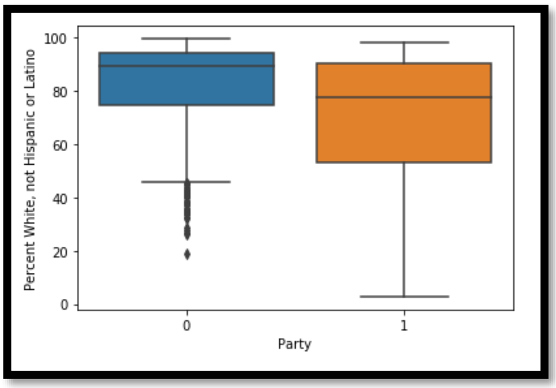
7. The mean median household income that is higher between the Democratic and Republican counties is in the Democratic counties. The result of the hypothesis test is that the p-value is greater than the significance value. Therefore, the result of the hypothesis test is that we reject the null hypothesis.

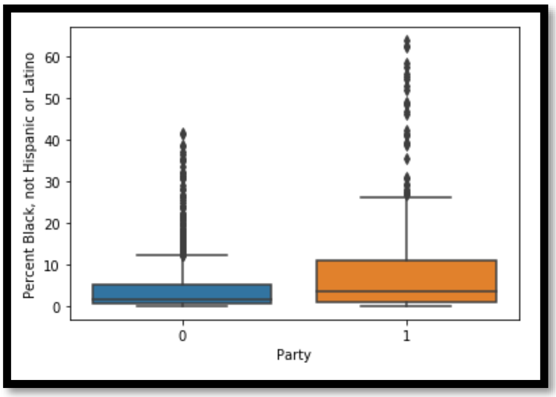
8.

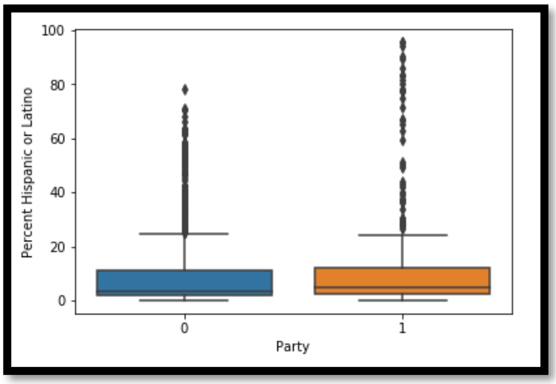


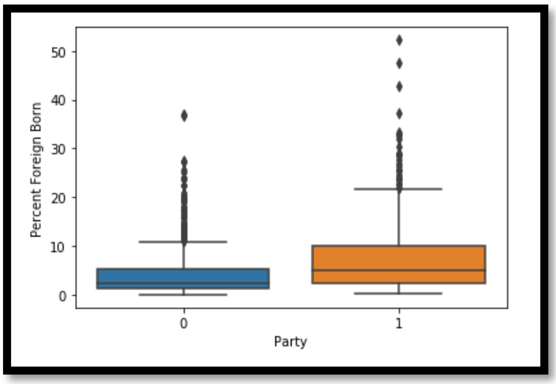


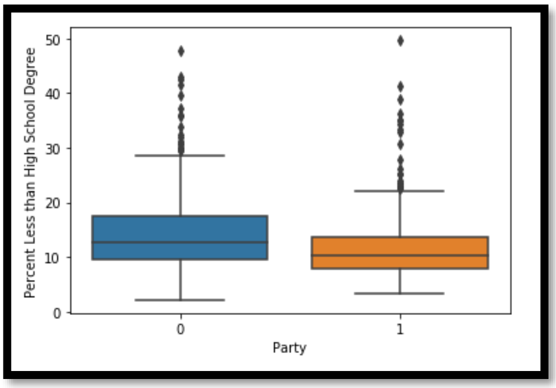


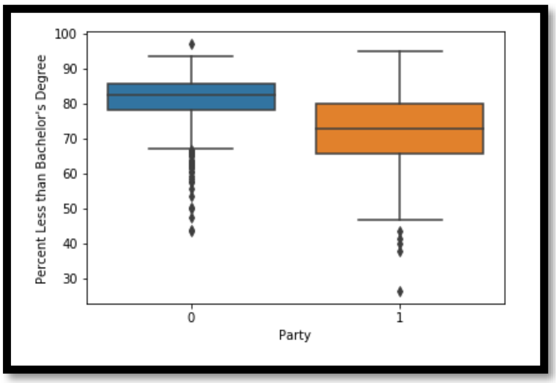








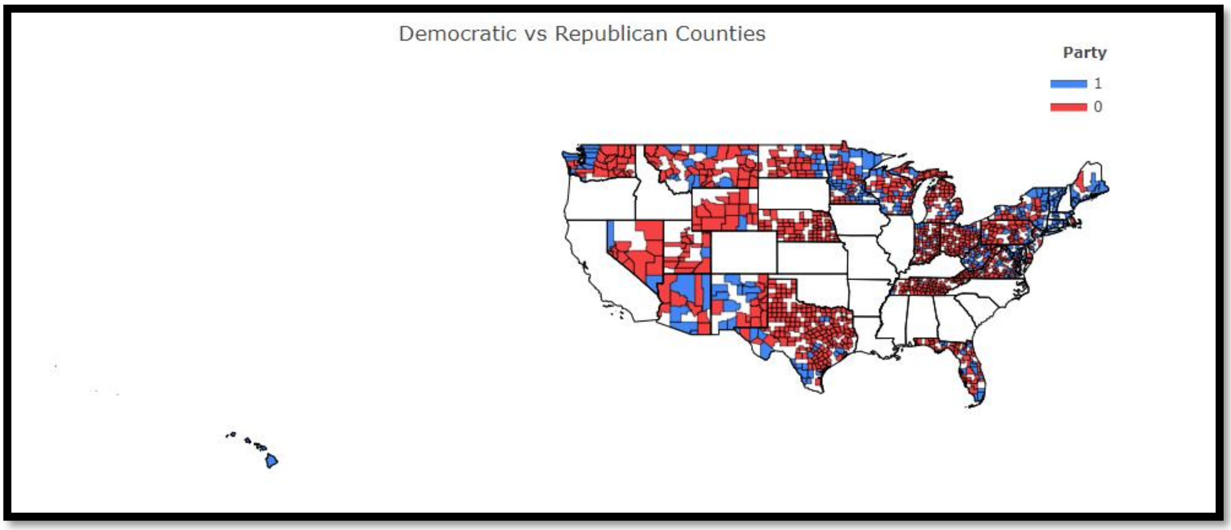




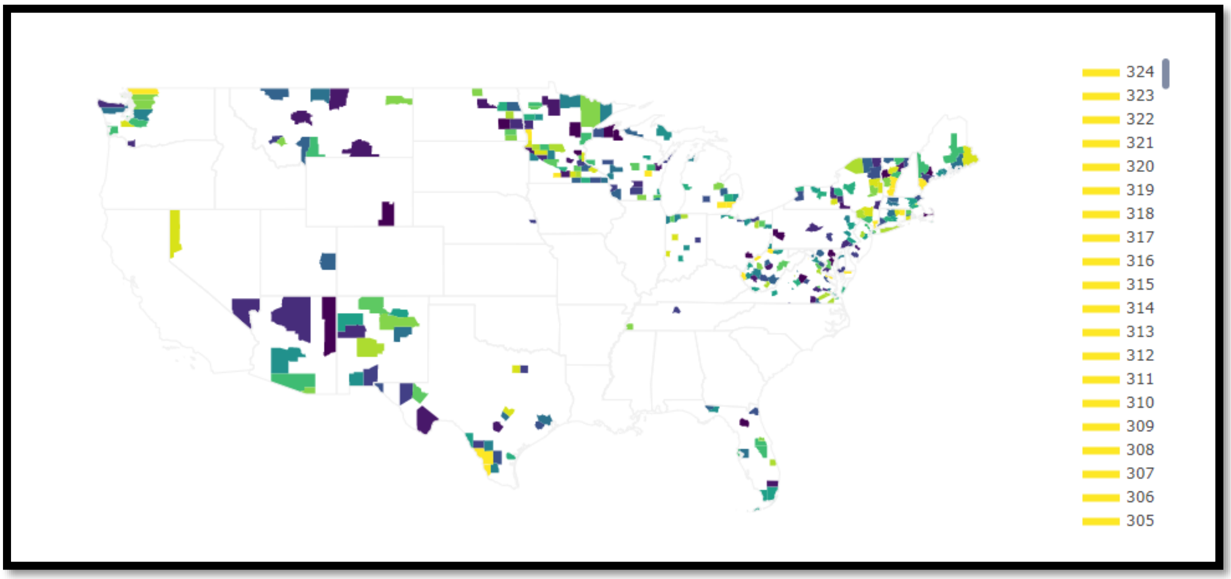
For the variable age, the plots show that as the age increases people start to vote for more republicans rather than democrats. For the variable gender, more females vote for democrats rather than republicans. For the variable race, people of not white color tend to go with the democrats rather than going with the republicans while it is the opposite for the people of white color. For the variable ethnicity, people who are foreign-born tend to vote for more democrats rather than voting for republicans. Finally, for the variable education, people with less than a high school degree, tend to go with democrats while people with less than a bachelor's degree tend to go towards the republicans but it is only by a slight margin.

9. The most important variables that are important to the dataset are the age, ethnicity, and education variables because for age, as people grow older, it seems that they are being less involved with politics. Education is important especially for high school graduates and people who don't have a bachelor's degree because they are the ones that are the most involved with politics. Ethnicity plays a huge part because most ethnic groups vote for democrats rather than republicans and this shows that these groups tend to participate more with Democrats.

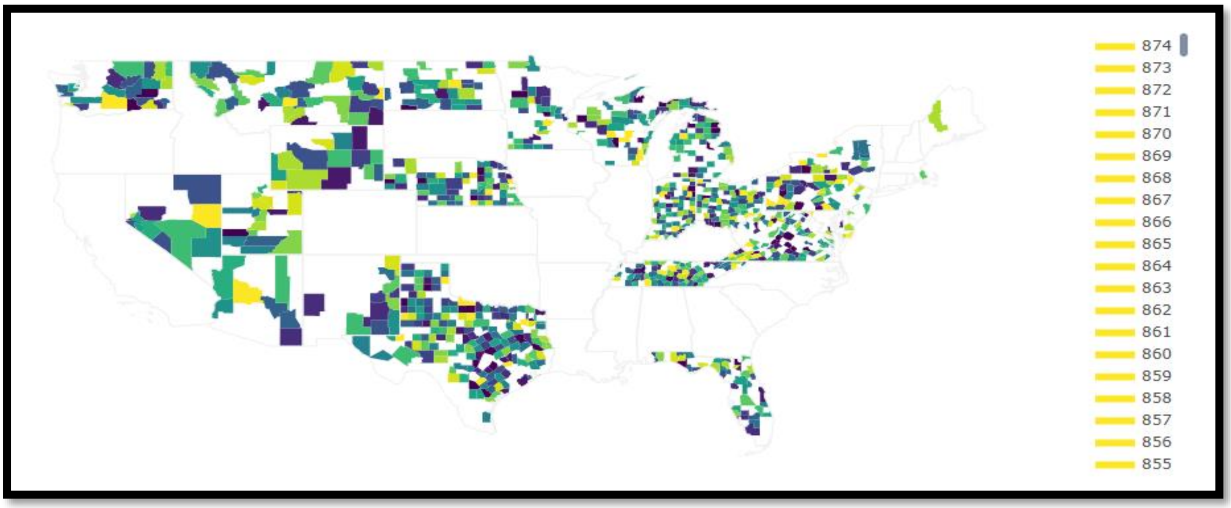
10. Democratic and Republican Counties:



Democratic Counties :



Republican Counties :


874
873
872
871
870
869
868
867
866
865
864
863
862
861
860
859
858
857
856
855