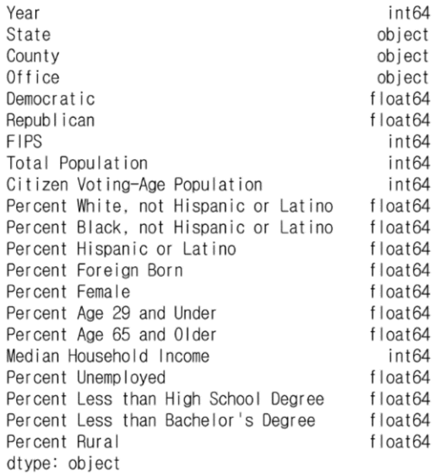
Project 01: Exploratory Data Analysis

Fall 2020

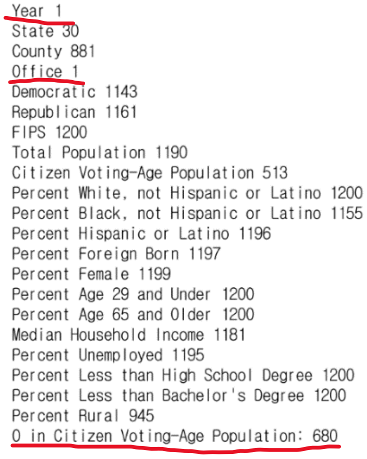
1. **(5 pts.) Reshape datase election_train from long format to wide format. Hint: the reshaped dataset should contain 1205 rows and 6 columns.**
   We use pd.pivot_table() function to transfer long format data to wide format data.

2. **(20 pts.) Merge reshaped dataset election_train with dataset demographics_train. Make sure that you address all inconsistencies in the names of the states and the counties before merging. Hint: the merged dataset should contain 1200 rows.**
   we use str.replace('County','') function to remove "County" on election_train data and replace states' name with its abbreviations on demographics_train data. Then use .str.lower().str.strip() on both data to make them consistent.

3. **(5 pts.) Explore the merged dataset. How many variables does the dataset have? What is the type of these variables? Are there any irrelevant or redundant variables? If so, how will you deal with these variables?**
   21 variables, float64(13), int64(5), object(3), 'Year' and 'Office' are irrelevant they only have one obeservation,'Citizen Voting-Age Population' has too many 0 we can drop them.

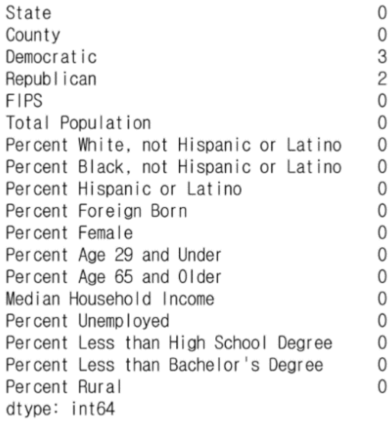


```
Year                                        int64
State                                      object
County                                     object
Office                                     object
Democratic                                float64
Republican                                float64
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
```



```
Year 1
State 30
County 881
Office 1
Democratic 1143
Republican 1161
FIPS 1200
Total Population 1190
Citizen Voting-Age Population 513
Percent White, not Hispanic or Latino 1200
Percent Black, not Hispanic or Latino 1155
Percent Hispanic or Latino 1196
Percent Foreign Born 1197
Percent Female 1199
Percent Age 29 and Under 1200
Percent Age 65 and Older 1200
Median Household Income 1181
Percent Unemployed 1195
Percent Less than High School Degree 1200
Percent Less than Bachelor's Degree 1200
Percent Rural 945
0 in Citizen Voting-Age Population: 680
```

4. **(10 pts.) Search the merged dataset for missing values. Are there any missing values? If so, how will you deal with these values?**



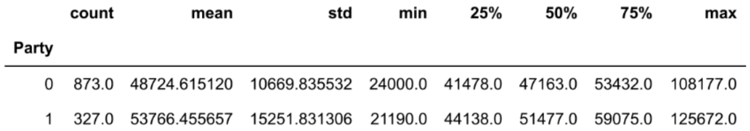
```
State                                   0
County                                  0
Democratic                              3
Republican                              2
FIPS                                    0
Total Population                        0
Percent White, not Hispanic or Latino   0
Percent 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
```

There are three missing values in Democratic and two missing values in Republican, since there are not that much, we set them to zero, it is fair for each other.

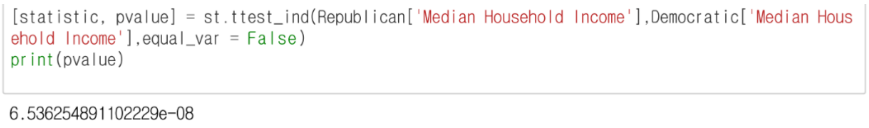
5. **(5 pts.) Create a new variable named "Party" that labels each county as Democratic or Republican. This new variable should be equal to 1 if there were more votes cast for the Democratic party than the Republican party in that county and it should be equal to 0 otherwise.**

We use .apply(lambda row:1 if row.Democratic > row.Republican else 0,axis=1) function to achieve that.

6. **(10 pts.) Compute the mean median household income for Democratic counties and Republican counties. Which one is higher? Perform a hypothesis test to determine whether this difference is statistically significant at the $\alpha = 0.05$ significance level. What is the result of the test? What conclusion do you make from this result?**



Party	count	mean	std	min	25%	50%	75%	max
0	873.0	48724.615120	10669.835532	24000.0	41478.0	47163.0	53432.0	108177.0
1	327.0	53766.455657	15251.831306	21190.0	44138.0	51477.0	59075.0	125672.0

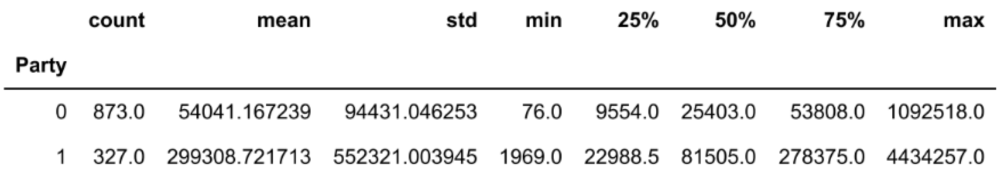


```
[statistic, pvalue] = st.ttest_ind(Republican['Median Household Income'],Democratic['Median Household Income'],equal_var = False)
print(pvalue)
```

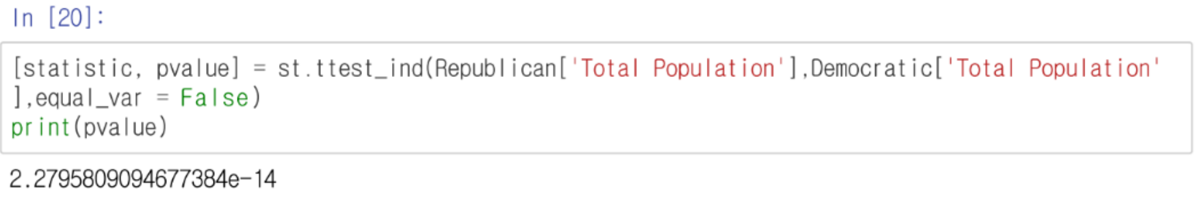
```
6.536254891102229e-08
```

Democratic mean median household is higher. p-value is less than 0.05 which means there is sufficient data to reject the null hypothesis that mean median household income of Republican countries and Republican counties are equal.

7. **(10 pts.) Compute the mean population for Democratic counties and Republican counties. Which one is higher? Perform a hypothesis test to determine whether this difference is statistically significant at the $\alpha = 0.05$ significance level. What is the result of the test? What conclusion do you make from this result?**



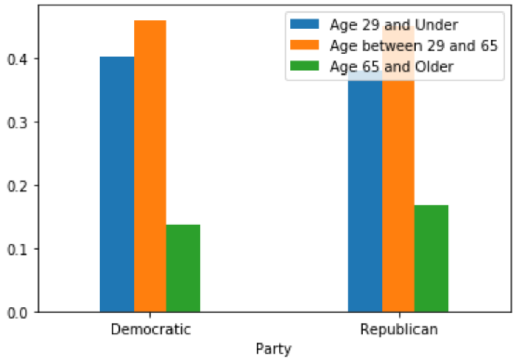
	count	mean	std	min	25%	50%	75%	max
Party								
0	873.0	54041.167239	94431.046253	76.0	9554.0	25403.0	53808.0	1092518.0
1	327.0	299308.721713	552321.003945	1969.0	22988.5	81505.0	278375.0	4434257.0

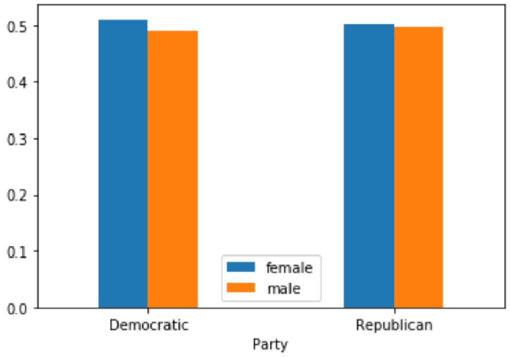


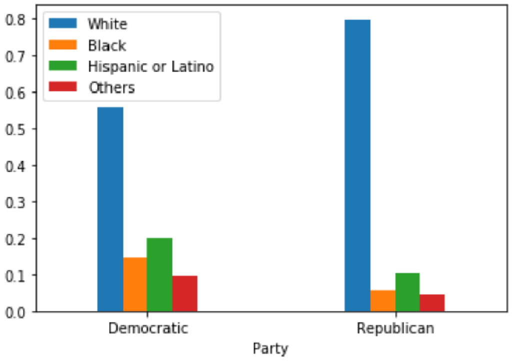
```
In [20]:
[statistic, pvalue] = st.ttest_ind(Republican['Total Population'],Democratic['Total Population'
],equal_var = False)
print(pvalue)
2.2795809094677384e-14
```

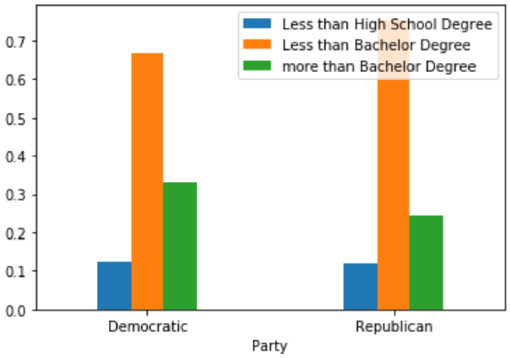
Democratic counties' mean population is higher. p-value is less than 0.05 which means there is sufficient data to reject the null hypothesis that population of Republican countries and Republican counties are equal.

8. **(20 pts.) Compare Democratic counties and Republican counties in terms of age, gender, race and ethnicity, and education by computing descriptive statistics and creating plots to visualize the results. What conclusions do you make for each variable from the descriptive statistics and the plots?**









we use the formula $P\left(\frac{X}{party}\right) = \frac{\Sigma_1^n\left(p_n\left(\frac{x}{county}\right)*population_n\right)}{total\ population\ of\ this\ party}$ to reorganize the data, then plot out.

from the age, Democratic counties have more percent of people whose age is 29 and under, and have less people whose age is 65 and above than that in Republican counties.

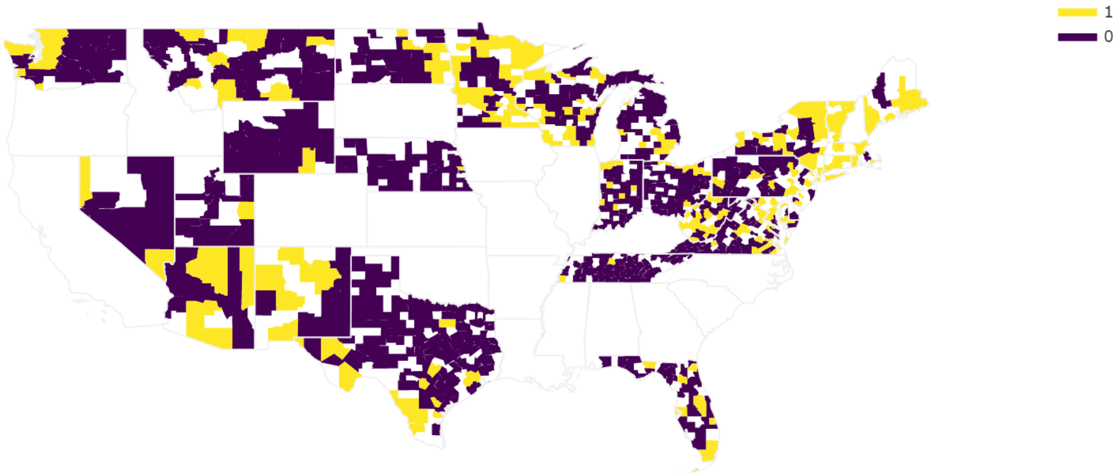
from the gender, the percent of female in Democratic counties is a litter higher.

from the race and ethnicity, the percent of White people in Democratic counties is less than that in Republican counties

from the education, the percent of people in Democratic counties who has a degree more than Bachelor is more than that in Republican counties

9. **(5 pts.) Based on y our results for tasks 6-8, which variables in the dataset do you think are more important to determine whether a county is labeled as Democratic or Republican? Justify your answer.**
'Median Household Income', 'Total population' and 'race and ethnicity' are important to determine a county is labeled as Democratic or Republican. they have significant difference from different parties. Education, gender and age are almost same between different parties which means they are not affected by the parties but maybe be affected by different counties.

10. **(10 pts.) Create a map of Democratic counties and Republican counties using the counties' FIPS codes and Python's Plotly library (plot.ly/python/county-choropleth/). Note that this dataset does not include all United States counties.**


1
0