For this project, we were tasked with performing multiple different types of data analyses on multiple datasets regarding population and voting statistics in over 1000 counties across the United States. Here are our results:

1) See our result in the Jupyter Notebook. To reshape the data set, first fill in the NaN values with zeros and then use “pd.pivot_table” function to change the shape from long to wide format. Pivot on the ‘Party’ column and on the value of ‘Votes’. The resulting shape is (1205, 6).

Year	State	County	Office	Democratic	Republican
2018	AZ	Apache County	US Senator	16298.0	7810.0
2018	AZ	Cochise County	US Senator	17383.0	26929.0
2018	AZ	Coconino County	US Senator	34240.0	19249.0

2) To merge our two datasets, we use the Pandas merge function. Before we can do this, we reconciled differences between the column names of the two dataframes, and addressed irregularities and inconsistencies. The most obvious example is the differences of the State column where one set used the full name and the other used the abbreviation. Another example is that some values were capitalized in one dataset, but not in the other set. After theses differences are reconciled, merge using how=’inner’. The resulting shape is (1200, 21)

```
merged_data = pd.merge(election_data_wide,demographic_data,how='inner')
merged_data.shape
display(merged_data)
```

Year	State	County	Office	Democratic	Republican	FIPS	Total Population	Voting-Age Population	White, not Hispanic or Latino	...
2018	AZ	apache county	US Senator	16298.0	7810.0	4001	72346	0	18.571863	...
2018	AZ	cochise county	US Senator	17383.0	26929.0	4003	128177	92915	56.299492	...
2018	AZ	coconino county	US Senator	34240.0	19249.0	4005	138064	104265	54.619597	...

3) To analyze the dataset variables apart from their names, we used the Pandas info function which tells us about the quantities and qualities of our dataset variables. From this we can find that we have 21 variables with the types: float64(15), int64(3), object(3). Upon taking a closer look at the variables, there seems to be a few redundant variables. “Year” seems to be redundant because the data all refers to 2018--at least for the election. “Office” seems redundant because each election is for US Senator. FIPS seems irrelevant -- it is only used to describe government-used document processing which does not seem useful for our purposes. To deal with this irrelevant data, we could simply remove those columns.

4) For this question, upon exploration of the data, there were many cases of missing values, and they were dealt with in a couple of different ways. The most serious case of missing data was in the “Citizen Voting-Age Population” variable. This variable was missing approximately 50% of its values, which would imply a flawed variable. As such, according to what was discussed in lecture, when a variable is missing a substantial number of values, it is fitting to remove the variable altogether.

   The other missing values were found in multiple variables. These variables are found in the “Percent Black, not Hispanic or Latino,” “Percent White, not Hispanic or Latino,” “Percent Hispanic or Latino,” and Percent Foreign Born.” Each of these variables had an insignificant number of missing values among their variables, and therefore it was most fitting to remove the observations. Estimating the values was an option, but due to the range of possible values among the observations, it did not seem fit to potentially completely mis-classify a county due to an estimation.

5) In order to answer this question, I started with creating a new variable called “Party.” The approach I took was an on-pass approach that assigned values to the “Party” variable on creation based on the condition of whether the value for the “Democratic” variable was greater or less than

"Republican" variable. The resulting variable can be seen in the dataframe below:

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

6) The approach we used for this question was to use the "stats.ttest_ind" function, which takes in two series to perform the hypothesis test. In order to do this, we performed a groupby on "Party" and "Total Population." This operation allowed us to create two separate dataframes called "republican" and "democrat" for which we could easily feed in their "Total Population" variables as series into the t-test function.

   The resulting mean populations obtained from Democratic and Republican counties revealed that Democratic counties tend to have higher populations than their republican counterparts. Then, the t-test was performed. As seen below, the mean population for Democratic counties came out to approximately 300,000, where the mean population of Republican counties came out to approximately 50,000.

   Below is the code and output from these operations:

```
Republican mean:  53974.214857142855
Democratic mean:  300998.3169230769

Ttest_indResult(statistic=-8.001207114045041, pvalue=2.0965719353509958e-14)
```

   Due to the resulting p-value being lower (and significantly at that) than .05, we conclude that the difference is statistically significant and reject the null hypothesis.

7) The same approach as question 6 was taken in question 7, where variables were created to hold the two different series holding Median_Household_Income divided based on "Party." Next, the mean was calculated on those two series, and the t-test was set up. The null hypothesis is that the mean median household incomes for Democratic and Republican counties are equal, and the alternative hypothesis is that the mean median household income for Democrats is higher than that of Republicans.

   Below are the results from the operations performed:

```
The mean median household income for Democratic counties:
53798.732307692306
The mean median household income for Republican counties:
48724.15085714286
```

The mean median household income for Democratic counties is higher

Let mu1 be the mean median household income for Democartic counties and mu2 the the mean median household income for Republican counties.

H0: mu1 = mu2
Ha: mu1 > mu2

Perform an unpaired t-test:

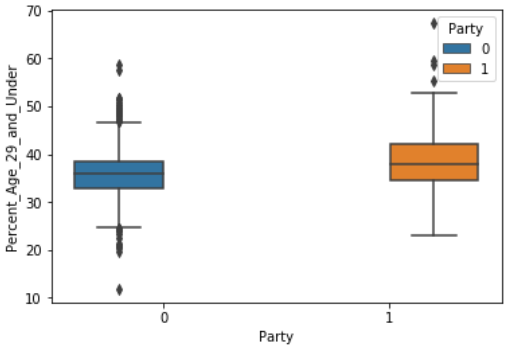
```
5.507012409466501
3.0866199456151866e-08
```

Since pvalue/2 < 0.05, reject null hypothesis, i.e., the difference is statistically significant at the alpha = 0.05 significance level.

   As can be seen from the t-test, a p-value/2 of less than the significance level is produced, therefore the difference is statistically significant.

8) For these analyses, we start each with a groupby on "Party" and the requested variables for testing. Next, we obtain box plots for each. Below are the resulting plots:

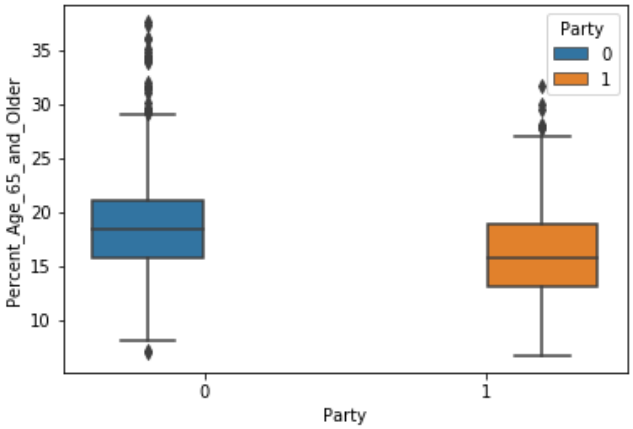
Percent_Age_29_and_Under



Conclusion: We observe that the median 'Percent_Age_29_and_Under' is roughly the same for Democratic counties and Republican counties.

Here we see that the median of the “Percent_Age_29_and_Under” variable is roughly the same across both types of counties, meaning that this would likely not be a good metric to use to determine the party of a given county.

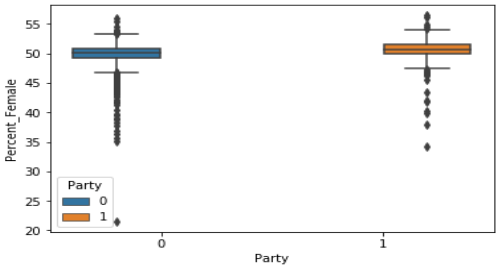
Percent_Age_65_and_Older



Conclusion: We observe that the median 'Percent_Age_65_and_Older' is higher for Republican counties.

Here we see that the median of the “Percent_Age_65_and_Older” variable is higher among Republican counties than in Democratic counties, meaning that this might be a good metric to use to determine the party of a given county.

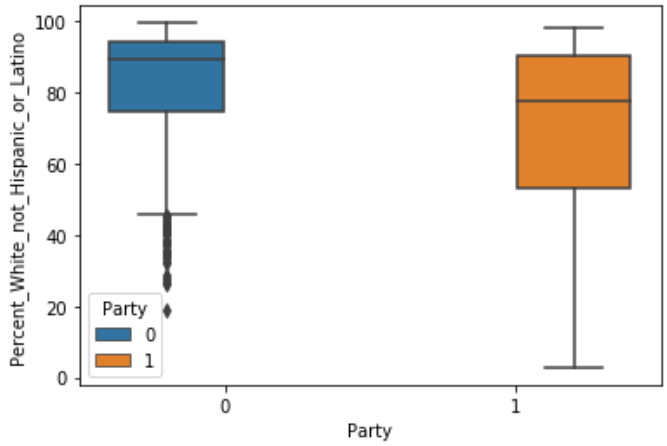
Percent_Female



Conclusion: We observe that the 'Percent_Female' is almost the same for Democratic counties and Republican counties.

Here we see that the median of the “Percent_Female” variable is almost the same across both types of counties, meaning that this would likely not be a good metric to use to determine the party of a given county.

Percent_White_not_Hispanic_or_Latino

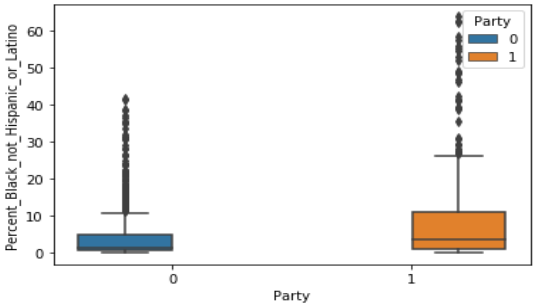


Conclusion: We observe that the median 'Percent_White_not_Hispanic_or_Latino' is higer for Republican counties.

Here we see that the median of the “Percent_White_not_Hispanic_or_Latino” variable is higher among Republican counties than in Democratic counties, meaning that this might be a good metric to use to determine the party of a given county.

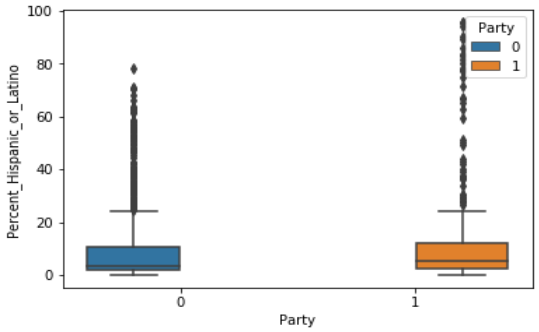
Percent_Black_not_Hispanic_or_Latino

Here we see that the median



Conclusion: We observe that the median 'Percent_Black_not_Hispanic_or_Latino' is higher for Democratic counties.

of the “Percent_Black_not_Hispanic_or_Latino” variable is higher among Democratic counties than in Republican counties, meaning that this might be a good metric to use to determine the party of a given county.

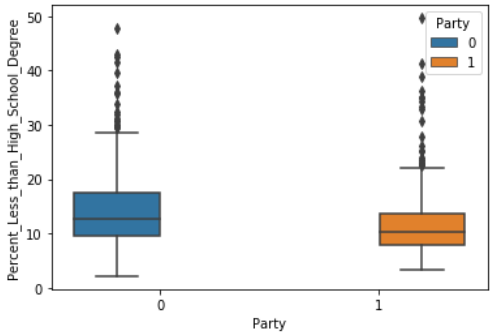


Conclusion: We observe that the 'Percent_Hispanic_or_Latino' is roughly the same for Democratic counties and Republican counties.

Percent_Hispanic_or_Latino

Here we see that the median of the “Percent_Hispanic_or_Latino” variable is almost the same across both types of counties, meaning that this would likely not be a good metric to use to determine the party of a given county.

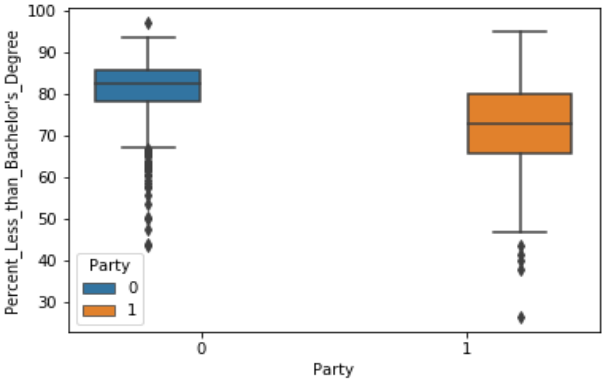
Percent_Less_than_High_School_Degree



Conclusion: We observe that the median 'Percent_Less_than_High_School_Degree' is roughly the same for Democratic counties and Republican counties.

Here we see that the median of the “Percent_Less_than_High_School_Degree” variable is almost the same across both types of counties, meaning that this would likely not be a good metric to use to determine the party of a given county.

Percent_Less_than_Bachelor’s_Degree



Conclusion: We observe that the median 'Percent_Less_than_Bachelor's_Degree' is higher for Republican counties.

Here we see that the median of the “Percent_Less_than_Bachelor’s_Degree” variable is higher among Republican counties, meaning that this might be a good metric to use to determine the party of a given county.

9) Features 'Percent_Age_65_and_Older', 'Percent_White_not_Hispanic_or_Latino', 'Percent_Black_not_Hispanic_or_Latino', 'Percent_Less_than_High_School_Degree' and 'Percent_Less_than_Bachelor's_Degree' seem to be more important to determine whether a county is labeled as Democratic or Republican.
10) Below is the map created from the plotly library of the United States according to county