1. **Reshape dataset election_train from long format to wide format.**
   We simply made use of pandas pivot_table() function to do this step.
2. **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.**
   First, we removed the substring 'County' from every row in the 'County' feature from the election data. This was done so we could match the 'County' feature from the demographic data. Second, we replaced the state abbreviations with their full abbreviation in the 'State' feature from the election data. This was done for similar reasons. Third, we had to lowercase all rows in the 'County' feature from the election data and the demographic data. This was done to solve any lowercase or uppercase discrepancies among both sets of data. Finally, we merged the two datasets using pandas merge() function.
3. **Explore the merged dataset. How many variables does the dataset have?**
   21 variables (including the 2 recently added 'Democratic' and 'Republican' features).
   What is the type of these variables?
   The types of data include float64, int64, and object.
   Are there any irrelevant or redundant variables?
   The 'Year' and 'Office' features are irrelevant or redundant variables because across all rows they are all the same values. 'Year' is just 2018 and 'Office' is just 'US Senator'.
   If so, how will you deal with these variables?
   We got rid of the 'Office' and 'Year' columns all together using the drop() function.
4. **Search the merged dataset for missing values. Are there any missing values?**
   There exists missing values for 3 Democratic observations and 2 Republican observations.
   If so, how will you deal with these values?
   Since the rows in which these missing values exist only lack that value, it's best to replace/estimate the missing values. We decided to fill forward on these missing values using the fillna() function.
5. **Create a new variable named “Party” that labels each county as Democratic or Republican.**
   To perform this step we simply made use of the loc() function.
6. **Compute the mean median household income for Democratic counties and Republican counties. Which one is higher?**
   The Democratic counties have a higher mean.
   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?
   The test statistic is 5.419 and the p-value is 0.0000000979.
   What conclusion do you make from this result?
   We reject the null hypothesis. The result is statistically significant. There is sufficient evidence to conclude that Democratic counties may have a higher median household income than Republican counties.
7. **Compute the mean population for Democratic counties and Republican counties. Which one is higher?**

The Democratic counties have a higher mean.

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?

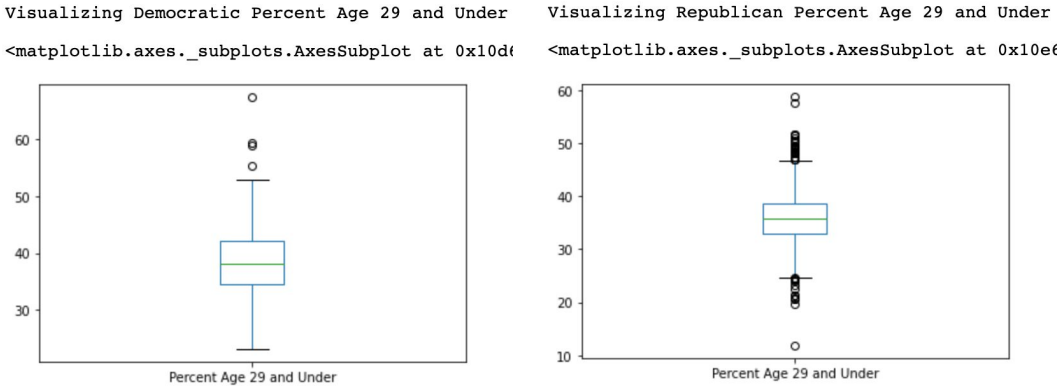
The test statistic is 5.419 and the p-value is 0.0000000000000232

What conclusion do you make from this result?

We reject the null hypothesis. The result is statistically significant. There is sufficient evidence to conclude that Democratic counties may have a higher population than Republican counties.

8. **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?**

Democratic v Republican (Age 29 and Under)



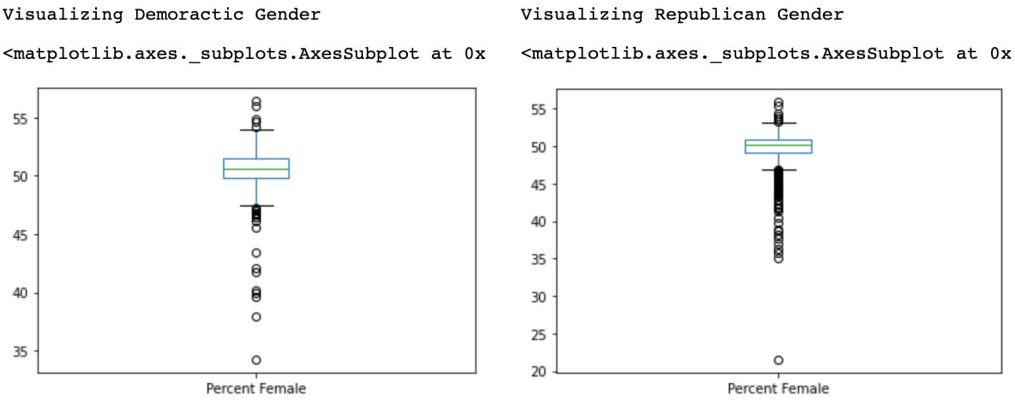
Conclusion: Since the Republican data is infected with outliers, we're going to use the median (35.85%) to compare it to the mean (38.73%) of the Democratic data. The difference is pretty significant and suggests Democratic counties may have a slightly higher population of people aged 29 and under.

Democratic v Republican (Age 65 and Older)

Conclusion: Both graphs seem to be skewed by the amount of outliers. As such, we're going to compare them using their medians (D: 15.70% vs R: 18.38%). The difference is pretty

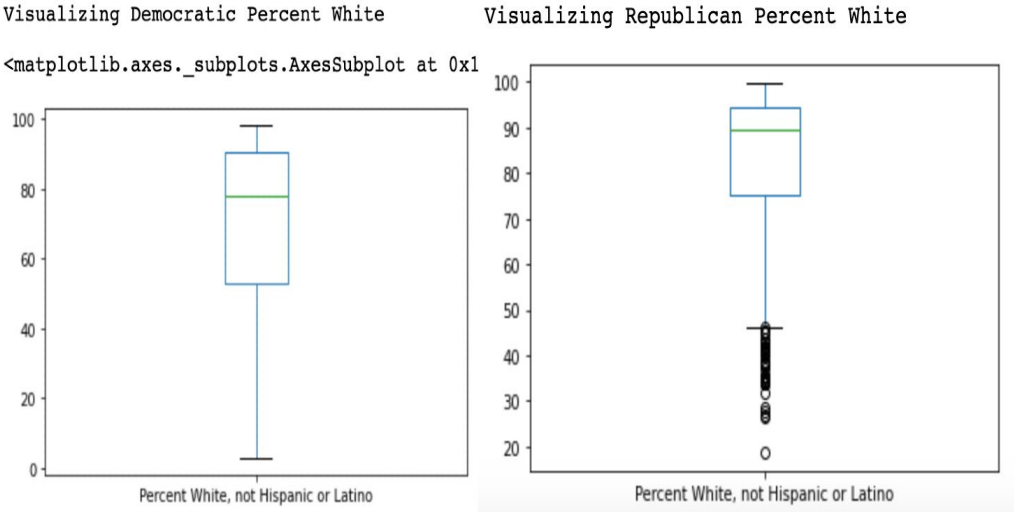
significant and suggests Democratic counties may have a slightly higher population of people aged 65 and older.

Democratic v Republican (Female)



Conclusion: Both graphs seem to be skewed by the amount of outliers. As such, we're going to compare them using their medians (D: 50.65% vs R: 50.17%). The difference is barely noticeable and as such, the female feature wouldn't be a good predictor variable.

Democratic v Republican (White)

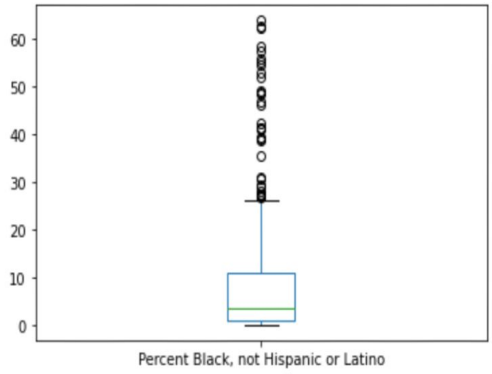


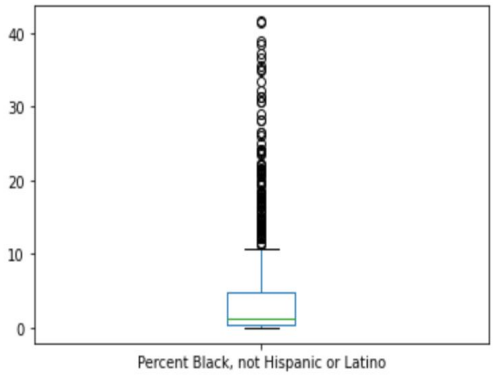
Conclusion: Since the Republican data is infected with outliers, we're going to use the median (89.45%) to compare it to the mean (69.53%) of the Democratic data. The difference is pretty significant and suggests Republican counties may have a slightly higher percent of white people than Democratic counties.

Democratic v Republican (Black)

```
Visualizing Democratic Percent Black
<matplotlib.axes._subplots.AxesSubplot at 0x
```

```
Visualizing Republican Percent Black
<matplotlib.axes._subplots.AxesSubplot at 0x
```



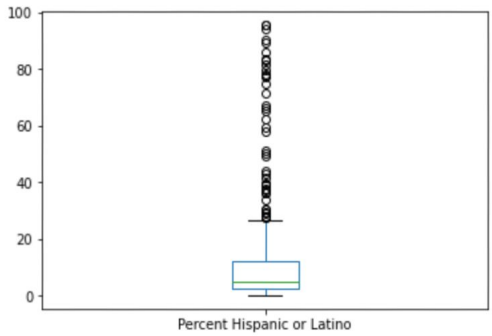


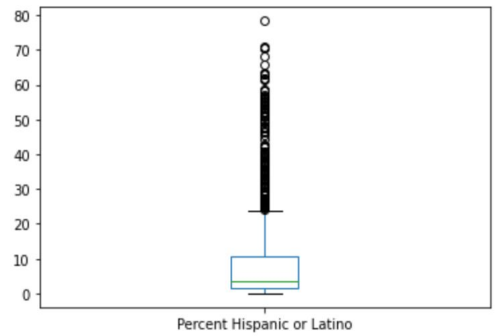
Conclusion: Both graphs seem to be skewed by the amount of outliers. As such, we're going to compare them using their medians (D: 3.49% vs R: 1.32%). The difference is somewhat noticeable and as such, the percent black feature could be a good predictor variable.

Democratic v Republican (Hispanic or Latino)

```
Visualizing Democratic Percent Hispanic
<matplotlib.axes._subplots.AxesSubplot at 0x1
```

```
Visualizing Republican Percent Hispanic
<matplotlib.axes._subplots.AxesSubplot at 0x
```



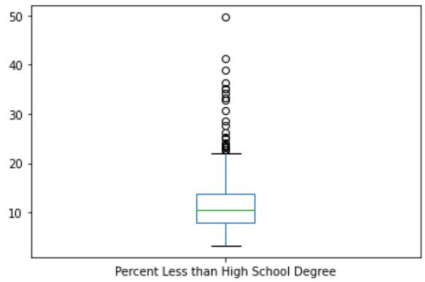


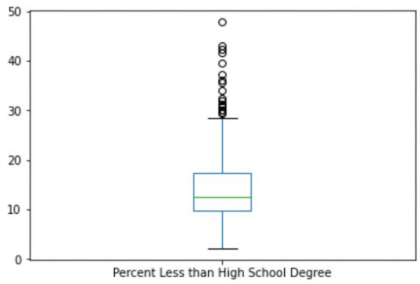
Conclusion: Both graphs seem to be skewed by the amount of outliers. As such, we're going to compare them using their medians (D: 5.07% vs R: 3.44%). The difference is somewhat noticeable and as such, the percent hispanic feature could be a good predictor variable.

Democratic v Republican (Less than High School Degree)

```
Visualizing Democratic Percent Less than High School Degree
<matplotlib.axes._subplots.AxesSubplot at 0x11d1f40d0>
```

```
Visualizing Republican Percent Less than High School Degree
<matplotlib.axes._subplots.AxesSubplot at 0x11d610d90>
```



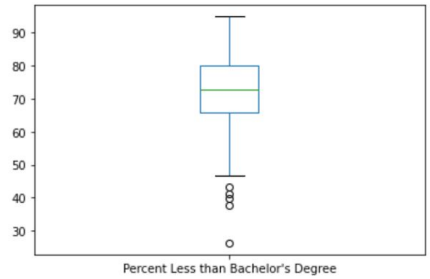


Conclusion: The Democratic counties have most of their high school degree percentages in the range 7.91-13.77%. The Republican counties have most of their high school degree

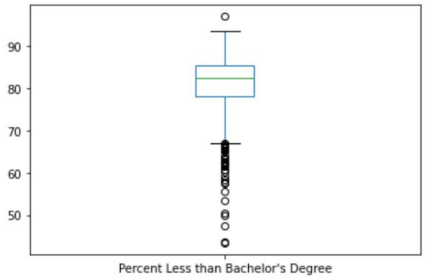
percentages in the range 9.66-17.45%. The outliers are very similar and the percentages are very low for parties. Therefore, this feature will not be a good potential predictor.

Democratic v Republican (Less than Bachelor's Degree)

```
Visualizing Democratic Percent Less than Bachelor's Degree
<matplotlib.axes._subplots.AxesSubplot at 0x11d2c58b0>
```




```
Visualizing Republican Percent Less than Bachelor's Degree
<matplotlib.axes._subplots.AxesSubplot at 0x11d6d8b50>
```



Conclusion: Since the Republican data is infected with outliers, we're going to use the median (82.40%) to compare it to the mean (72.07%) of the Democratic data. The difference is pretty significant and suggests Republican counties may have a higher percentage of people that have less than a bachelor's degree.

9. **Based on your 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.**
   The most important variables to determine whether a county is labeled as Democratic or Republican are, Percent Less than Bachelor's Degree, Total Population, Percent Hispanic, Percent White, not Hispanic or Latino, and Percent Age 65 and Older, because these variables help to give a more accurate representation of whether or not a county is Democratic or Republican and could be used as good predictors, for example, percentages show that counties that have less than a bachelor's degree tend to be Republican. Also there is a slight similarity between hispanic democrats and republicans, both are high for each party, but the hispanic democratic percentage is higher. Republicans that are white have a more concise percentage while democrats that are white are more spread out. Republican counties also tend to have a higher percentage of people that are age 65 or older than the democratic counties.
10. **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.**