Report for Project1

Task1

- After Task1 is completed, 1205 rows and 6 columns remained in the dataset.

	State	County	Year	Office	Votes	
Party					Democratic	Republican
0	AZ	Apache County	2018	US Senator	16298.0	7810.0
1	AZ	Cochise County	2018	US Senator	17383.0	26929.0
2	AZ	Coconino County	2018	US Senator	34240.0	19249.0
3	AZ	Gila County	2018	US Senator	7643.0	12180.0
4	AZ	Graham County	2018	US Senator	3368.0	6870.0
...	...	...	...	...	...	...
1200	WY	Platte County	2018	US Senator	801.0	2850.0
1201	WY	Sublette County	2018	US Senator	668.0	2653.0
1202	WY	Sweetwater County	2018	US Senator	3943.0	8577.0
1203	WY	Uinta County	2018	US Senator	1371.0	4713.0
1204	WY	Washakie County	2018	US Senator	588.0	2423.0

[1205 rows x 6 columns]

Task2

- After Task2 is completed, 1200 rows remained in the dataset.

	State	County	(State,)	(County,)	(Year,)	(Office,)	\
0	AZ	apache	AZ	apache	2018	US Senator	
1	AZ	cochise	AZ	cochise	2018	US Senator	
2	AZ	coconino	AZ	coconino	2018	US Senator	
3	AZ	gila	AZ	gila	2018	US Senator	
4	AZ	graham	AZ	graham	2018	US Senator	
...	...	...	...	...	...	...	
1195	WY	platte	WY	platte	2018	US Senator	
1196	WY	sublette	WY	sublette	2018	US Senator	
1197	WY	sweetwater	WY	sweetwater	2018	US Senator	
1198	WY	uinta	WY	uinta	2018	US Senator	
1199	WY	washakie	WY	washakie	2018	US Senator	

… (omitted: For entire result, please find the Jupyter source code or pdf file) ...

```
      Percent Less than Bachelor's Degree  Percent Rural
0                                88.941063      74.061076
1                                76.837055      36.301067
2                                65.791439      31.466066
3                                82.262624      41.062000
4                                86.675944      46.437399
...                                    ...            ...
1195                             80.300395      58.647744
1196                             75.645069     100.000000
1197                             78.628507      10.916313
1198                             81.793082      43.095937
1199                             78.923920      35.954529

[1200 rows x 23 columns]
```

Task3

- The merged data set has 23 variables
- The data types of each column are follows:

```
State                                      object
County                                     object
Votes_Democratic                           float64
Votes_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
Party                                      int64
```

· Irrelevant or redundant variables:

1. ('State', '')
2. ('County', '')
3. ('Year', '')
4. ('Office', '')

- First two State and Country are just redundant columns they are appearing twice in the data set, and the last two column Year and Office contain same data through the dataset hence not adding much information to our analysis. We removed these columns from our data set.
- These variables were irrelevant or redundant hence we remove them from the data set.

Task4

- We checked whether there are any explicit missing values like null, but there were no null values.
- After that, we started to define what are the missing values and valid values among the 0's.
- Valid variables
    - race variables: There are some 0's in the three (white / black / hispanic) variables. However, when we calculated the sum of the three variables, they mostly reached to 100% although there are 0's. Hence, we concluded that the 0's in the race variables (white / black / hispanic) are valid.
    - Percent Rural: There are some 0's, but at the same time, there are some 100's in the variable. We concluded Percent Rural is valid variable.
    - Unemployment Rate: There are 0's, however, there are also the unemployment rate reaching almost to 0. In addition, when we look at the Total population variable where the Unemployment Rate is 0, the number of population was very small number. So, it is possible for the unemployment rate to be 0.
- Missing variables
    - Citizen Voting-Age Population: This variable includes a lot of 0's, and when we look at the next tasks of the project, there is no relevant issue on this variable. We dropped this variable.
    - Votes democratic / Votes republican: There are 5 observations where the values are 0's. Although they are important variable by nature, there was not any related task regarding the number of votes itself. We could replace or estimate the values by interpolation, the five counties with 0's will anyway follow the trend of the same state. Later, we set the Party variable as -1 for the five rows to distinguish the 1 or 0 observations. Ultimately, the five rows are removed from the dataset.

Task5

- We created a Party variable indicating which party the county is supporting. The value of the Party is 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.
- In addition, we set the value as -1 if the number of votes does not exist in the county. We found five rows regarding this, and later in the code, it was used to delete the five rows from the dataset.

Task6

- Democratic mean Total Population is: 300,998
- Republic mean Total Population is: 53,864

- **We observe that the mean population of Democratic Counties is greater in the dataset.**
- **Null Hypothesis:** The mean of population between Democratic and Republican countries are the same.
- **Alternative Hypothesis:** The mean of population between Democratic and Republican countries are different. (two-tailed test)
- We get the p-value which is approaching to 0 that indicates strong evidence against the null hypothesis, so we reject the null hypothesis. And we conclude that this difference in the mean of population of Rebublican and democratic countries is statistically significant at the $\alpha = 0.05$ significance level.
- The interpretation of the statistic finds that the means are different, with a significance of at least 5%, and this difference in the means is not due to some chance there a strong evidence behind it.
- **Conclusion**: The mean population of Democratic and Repubican counties are not the same.

Task7

- Democratic mean Median Household Income is: $53,798
- Republican mean Median Household Income is: $48,746

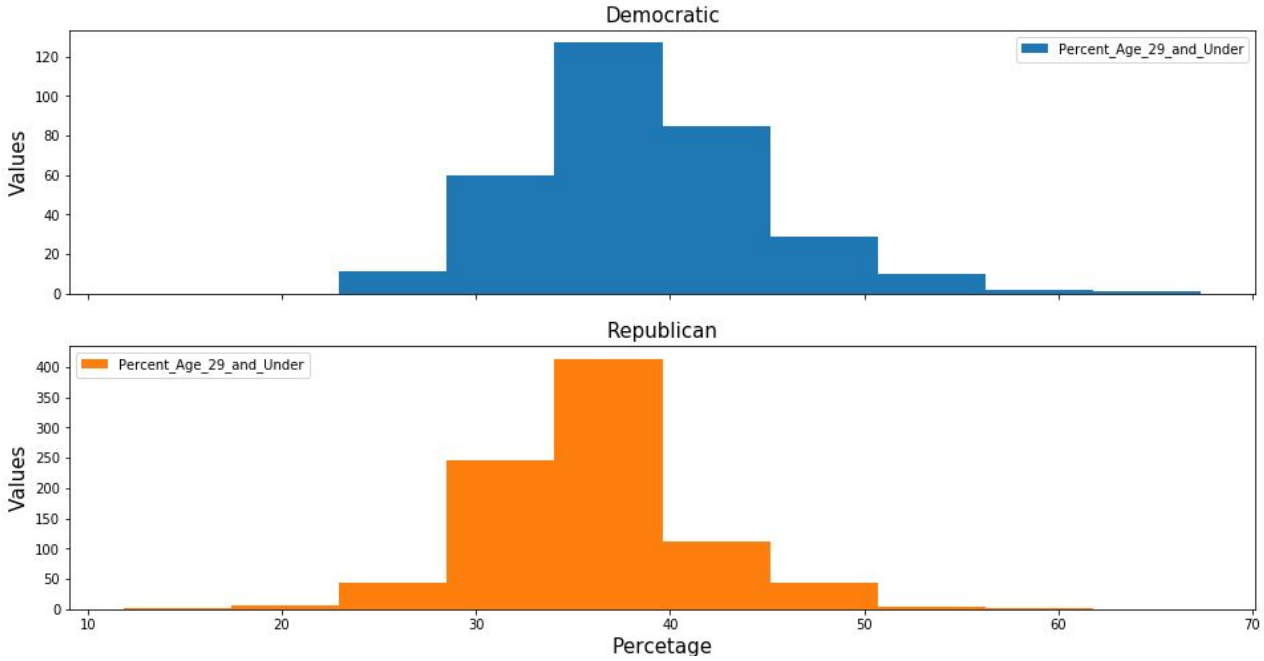
- **We observe that the Democratic mean Median Household Income is greater.**
- **Null Hypothesis:** The mean of median household income between Democratic and Republican counties are the same.
- **Alternative Hypothesis:** The mean of population between Democratic and Republic countries are different. (two-tailed test)
- We get the p-value which is approaching to 0 that indicates strong evidence against the null hypothesis, so we reject the null hypothesis. And we conclude that this difference in the mean of median household income of Rebublican and democratic countries is statistically significant at the $\alpha = 0.05$ significance level.The interpretation of the statistic finds that the means are different, with a significance of at least 5%, and this difference in the means is not due to some chance there a strong evidence behind it.
- **Conclusion**: The mean of median household income between Democratic and Repubican counties are not the same.

Task8

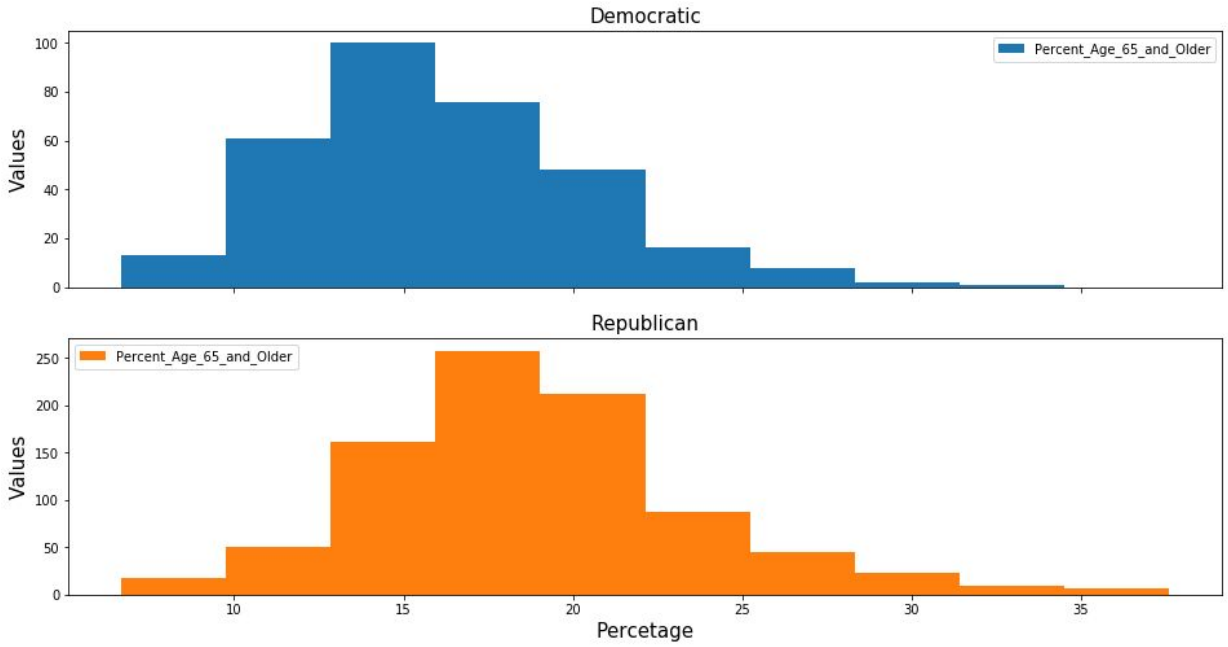
- We divided the data set into democratic and republican tables and found the descriptive statistics of each table and compared the certain variables of each table.
- We compared the following variables between democratic and republican tables.

Age Comparison: Democratic Vs Republican

Histogram for Percent Age 29 and Under



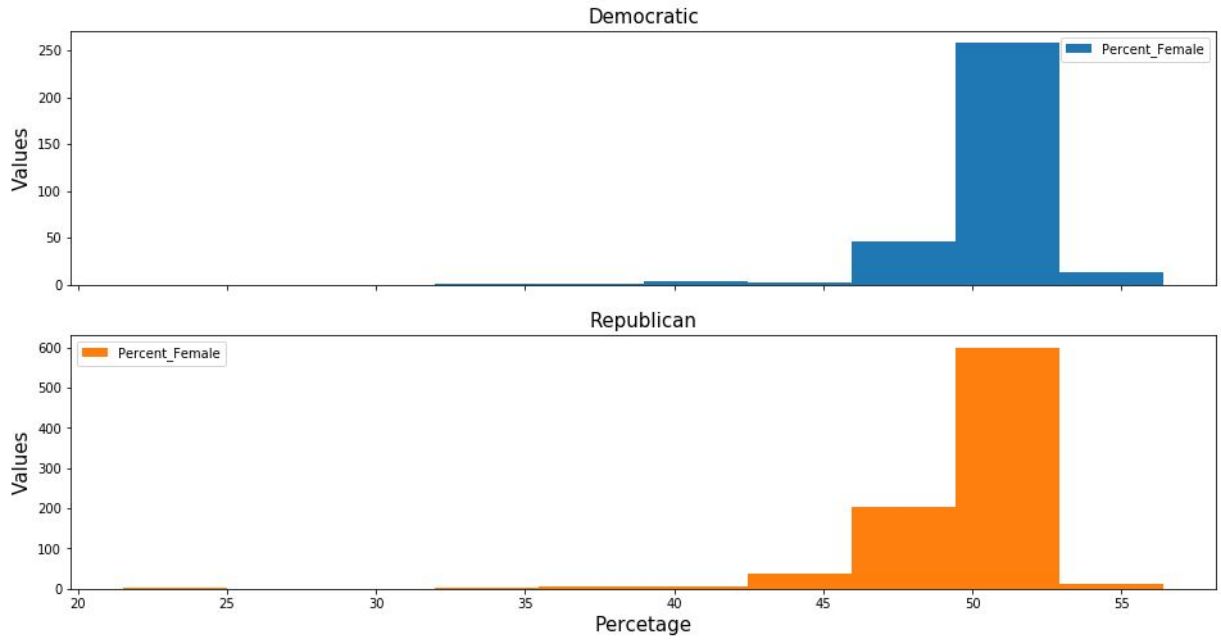
Histogram for Percent Age 65 and Older



Explanation: The first noticeable difference between the variables are their ranges. There seems to exist a wider variety of our data in the age 29 and under histogram. Another observation is that there seems to be on average more young voters than elderly voters. You can find many counties that have almost a third of their voters being people who are age 29 and under. Whereas for the histogram for percent age 65 and older, the max value is about the mean for age 29 and under. These two histograms are very different from each other.

Gender Comparison: Democratic Vs Republican

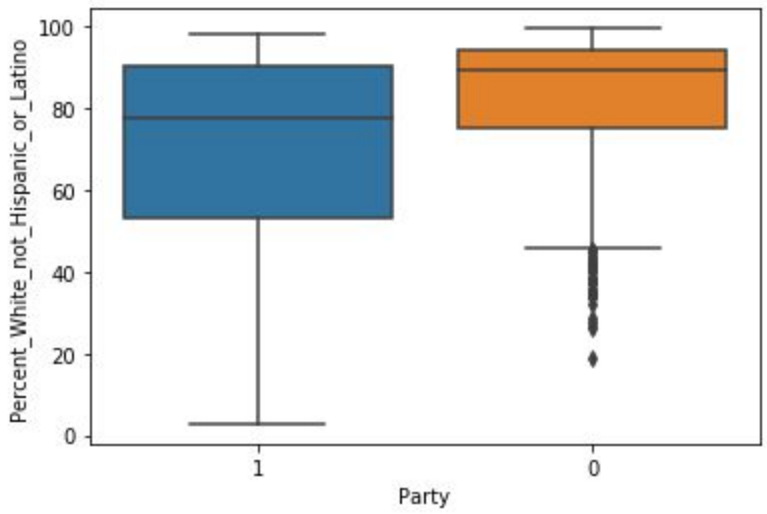
Histogram for Percent Female



Explanation: There isn't a huge difference between these two histograms. Some counties have around 30%-40% of their voters being female and some have around 55%. This also means that there are counties that have over 60% of their voters being male. This might infer to us that females may not be voting as much as males. But any given county will likely have a 1:1 ratio of female/male voters.

Race and Ethnicity Comparison: Democratic Vs Republican

Descriptive Statistics/Box Plot for Percent White, not Hispanic or Latino



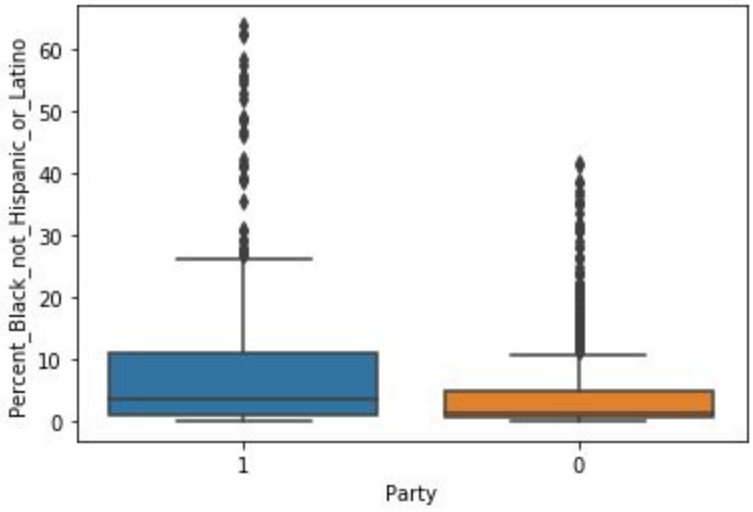

```
count    325.000000
mean      69.683766
std       24.981502
min        2.776702
25%       53.271579
50%       77.786090
75%       90.300749
max       98.063495
Name: Percent_White_not_Hispanic_or_Latino, dtype: float64
```

```
count    870.000000
mean      82.656646
std       16.056122
min       18.758977
25%       75.016397
50%       89.434849
75%       94.466596
max       99.627329
Name: Percent_White_not_Hispanic_or_Latino, dtype: float64
```

Explanation: The first thing that pops out at my eye is how white folks are usually the majority of the voters in any given county. There are a bit more white folks that are republican voters

than not. There are also many outliers for the republican party, whereas there aren't any outliers for the democratic party. Also, in some counties, 99% of the voters are white folks.

Descriptive Statistics/Box Plot for Percent Black, not Hispanic or Latino

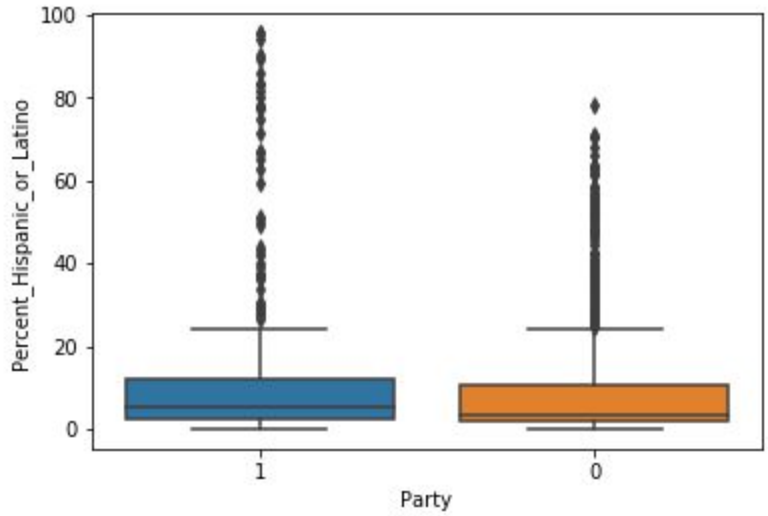



```
count    325.000000
mean       9.242649
std       13.351340
min        0.000000
25%        0.839103
50%        3.485992
75%       11.058843
max       63.953279
Name: Percent_Black_not_Hispanic_or_Latino, dtype: float64

count    870.000000
mean       4.189241
std        6.721695
min        0.000000
25%        0.460419
50%        1.318311
75%        4.753831
max       41.563041
Name: Percent_Black_not_Hispanic_or_Latino, dtype: float64
```

Explanation: There aren't as many black folks who are voting as there are white folks. But this plot does shower many outliers for both parties. Some counties can have up to 63% of their voters be black folks. What is surprising is that there exists counties that don't have any black folks voting. Maybe those votes weren't accounted for? The minimum for white folks in their respective parties is not anywhere close to this.

Descriptive Statistics/Box Plot for Percent Hispanic or Latino



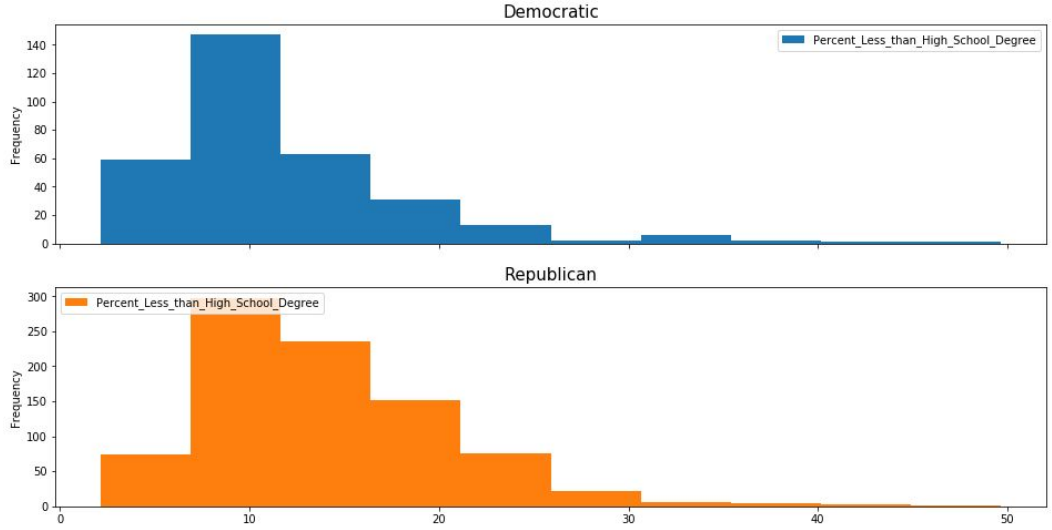

```
count    325.000000
mean      12.587391
std       19.575030
min        0.193349
25%        2.531017
50%        5.039747
75%       11.857116
max       95.479801
Name: Percent_Hispanic_or_Latino, dtype: float64
```

```
count    870.000000
mean       9.733094
std       14.049576
min        0.000000
25%        1.704539
50%        3.427435
75%       10.709696
max       78.397012
Name: Percent_Hispanic_or_Latino, dtype: float64
```

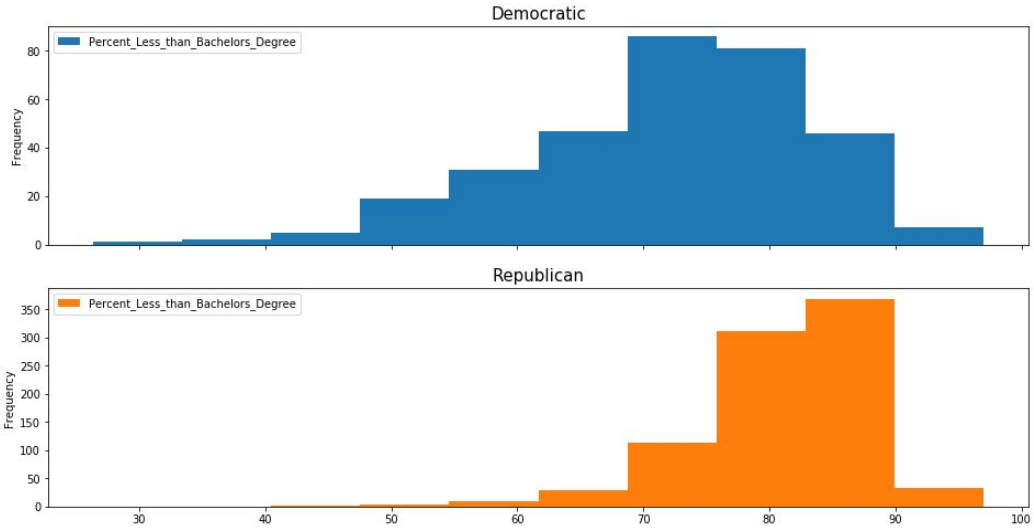
Explanation: Similar story here to the black folks in that hispanic voters are not usually the majority of the voters in any given county. There does exist a few counties that are extreme outliers. Some counties have over 95% of their voters as hispanic folks. This was not the same for the box plots for black folks.

Comparison of Education: Democratic Vs Republican

Histogram Distribution for Percent Less than High School Degree



Histogram Distribution for Percent Less than bachelor's degree



Explanation: One observation to be made here is that there aren't many voters who don't have a high school diploma. Whereas there is a wide variety of percentages in bachelor's degree histogram. The range for bachelor's degree is much larger than high school degree. It is also worth noting that many voters do not have a college education.

Task9

- We can observe from the graph analysis from Task 8 that white, older, American native, less educated, and male people voted for Republican party, and black / hispanic, younger, Foreign born, more educated and female people voted for Democratic party.
- Although most of variables are import factors, the race variable showed the most significant difference between white and non-white in terms of supporting parties.
- Age, foreign-born, and education are also important factors to predict which party the county will support.

Task10

- The map created with plotly is like below. The brown color represents the counties that support Democratic party, and orange color represents the counties that support Republican party.