Data Science

Analysis of 2018 US Senate Office election result

Name	Net Id
Sneha Tikare	stikar2
Cesar Hernandez	cherna83
Ubemio Romero	uromer2

Task 1:

Election_train consists of data in wide format. We convert it to long format to reduce redundancy and increase readability.

Before conversion

Democratic and Republican votes are represented as an observation in each row due to which the data is Year, State, County and office is repeated twice.Our goal is to remove this redundancy.

	Year	State	County	Office	Party	Votes
0	2018	AZ	Apache County	US Senator	Democratic	16298.0
1	2018	AZ	Apache County	US Senator	Republican	7810.0
2	2018	AZ	Cochise County	US Senator	Democratic	17383.0
3	2018	AZ	Cochise County	US Senator	Republican	26929.0
4	2018	AZ	Coconino County	US Senator	Democratic	34240.0

After conversion

Now, the Democratic and Republican votes are represented as two columns which reduced the redundancy and has increased the interpretability.

Party	Year	State	County	Office	Democratic	Republican
0	2018	AZ	Apache County	US Senator	16298.0	7810.0
1	2018	AZ	Cochise County	US Senator	17383.0	26929.0
2	2018	AZ	Coconino County	US Senator	34240.0	19249.0
3	2018	AZ	Gila County	US Senator	7643.0	12180.0
4	2018	AZ	Graham County	US Senator	3368.0	6870.0

Task 2:

Here we merge the Election and demographic data to get the count of demographic and republican votes for different demographics features.
Before merging we handle all inconsistencies like lowering the cases, formatting the states names so both the datasets have same format.

The merged data represents different statistics for each county.

	Year	State	County	Office	Democratic	Republican	FIPS	Total Population	Citizen Voting-Age Population	Percent White, not Hispanic or Latino	...	Percent Hispanic or Latino	Percent Foreign Born	Percent Female	Percent Age 29 and Under	P... and
0	2018	AZ	apache	US Senator	16298.0	7810.0	4001	72346	0	18.571863	...	5.947806	1.719515	50.598513	45.854643	13.3
1	2018	AZ	cochise	US Senator	17383.0	26929.0	4003	128177	92915	56.299492	...	34.403208	11.458374	49.069646	37.902276	19.7
2	2018	AZ	coconino	US Senator	34240.0	19249.0	4005	138064	104265	54.619597	...	13.711033	4.825298	50.581614	48.946141	10.8
3	2018	AZ	gila	US Senator	7643.0	12180.0	4007	53179	0	63.222325	...	18.548675	4.249798	50.296170	32.238290	26.3
4	2018	AZ	graham	US Senator	3368.0	6870.0	4009	37529	0	51.461536	...	32.097844	4.385942	46.313518	46.393456	12.3
5	2018	AZ	la paz	US Senator	1609.0	3265.0	4012	20304	15245	58.884949	...	26.182033	11.372143	48.946020	28.073286	36.0

Task 3:

a) There are 21 variables in the data set. Type of these variables are int64, Object, float64.The column 'Year' and 'Office' are irrelevant variable, since they don't provide any important information. The only information it gives is that data is for 2018 US Senator office. We can drop the irrelevant variables (Year and Office).
b) They are no redundant variables.

Task 4:

Yes, the data has missing values in columns ‘Citizen Voting-Age Population’, ‘Percent Black, not Hispanic or Latino’, ‘Percent Hispanic or Latino’, ‘Percent Foreign Born’,’ Percent Female’, ’Percent Unemployed’, ’Democratic’, ‘Republican’.
We can drop observations where Democratic and Republican votes are zeroes since these are less in numbers and keeping the null values will further impact the analysis in future.

We have dropped column ‘Citizen Voting-Age Population’ because the amount of missing data is higher than the available data.Also, the data seems incorrect.
Example :
For county Apache in state Arizona, the total population is 72346 but the Citizen Voting Age population is 0.There are a number of ways we can interpret this.
1)No one voted in that county.
2)There were no eligible voting population
3)For year 2018, the county did not vote.
It’s a indeterminant data.Hence, we drop the column.

	Year	State	County	Office	Democratic	Republican	FIPS	Total Population	Citizen Voting-Age Population
0	2018	AZ	apache	US Senator	16298.0	7810.0	4001	72346	0

For the other columns we can ignore them, since their impact is low on the analysis.

After dropping the irrelevant columns, the data now as 1200 observation and 18 attributes

	State	County	Democratic	Republican	FIPS	Total Population	Percent White, not Hispanic or Latino	Percent Black, not Hispanic or Latino	Percent Hispanic or Latino	Percent Foreign Born	Percent Female	Percent Age 29 and Under	Percent Age 65 and Older	Median Household Income	Per Unemplo
0	AZ	apache	16298.0	7810.0	4001	72346	18.571863	0.486551	5.947806	1.719515	50.598513	45.854643	13.322091	32460	15.807
1	AZ	cochise	17383.0	26929.0	4003	128177	56.299492	3.714395	34.403208	11.458374	49.069646	37.902276	19.756275	45383	8.567
2	AZ	coconino	34240.0	19249.0	4005	138064	54.619597	1.342855	13.711033	4.825298	50.581614	48.946141	10.873943	51106	8.238
3	AZ	gila	7643.0	12180.0	4007	53179	63.222325	0.552850	18.548675	4.249798	50.296170	32.238290	26.397638	40593	12.129

Task 5:

A new column named 'Party' is created to label each county as Democratic or Republic based on majority vote casted in that county.

a) Label '1' shows that the county voted Democratic as majority.
b) Label '0' shows that the county voted Republican as majority.

FIPS	Total Population	Percent White, not Hispanic or Latino	Percent Black, not Hispanic or Latino	Percent Hispanic or Latino	Percent Foreign Born	Percent Female	Percent Age 29 and Under	Percent Age 65 and Older	Median Household Income	Percent Unemployed	Percent Less than High School Degree	Percent Less than Bachelor's Degree	Percent Rural	Party
4001	72346	18.571863	0.486551	5.947806	1.719515	50.598513	45.854643	13.322091	32460	15.807433	21.758252	88.941063	74.061076	1
4003	128177	56.299492	3.714395	34.403208	11.458374	49.069646	37.902276	19.756275	45383	8.567108	13.409171	76.837055	36.301067	0
4005	138064	54.619597	1.342855	13.711033	4.825298	50.581614	48.946141	10.873943	51106	8.238305	11.085381	65.791439	31.466066	1

Task 6:

Mean population for Democratic counties (m1) = 300998.317
Mean population for Republican counties (m2) = 53864.673
Mean population of Democratic counties is higher than Republican counties

Hypothesis analysis

Null hypothesis: m1=m2
Alternative Hypothesis: m1 != m2
t-Test value = 8.0046
P-value = 2.0478 e-14
At alpha =0.05 level of significance, p-value is less than alpha, we reject the null hypothesis
And we see that Mean population of Democratic counties is higher than Republican counties.

Task 7:

Mean Median household income for Democratic counties (m1) = 53798.732
Mean Median household income for Republican counties (m2) = 48746.819
Mean Median household income of Democratic counties is higher than Republican counties

Hypothesis analysis

Null hypothesis: m1=m2
Alternative Hypothesis: m1 != m2
t-Test value = 5.4791
P-value = 7.149 xe-08
At alpha =0.05 level of significance, p-value is less than alpha, we reject the null hypothesis
And we see that Mean Median household income of Democratic counties is higher than Republican counties.

Task 8:

To visualize the data, we have used box plot to display the descriptive statistics.

Figure 1 : Statistics for different age group vs party (1 - Democratic, 0-Republican)

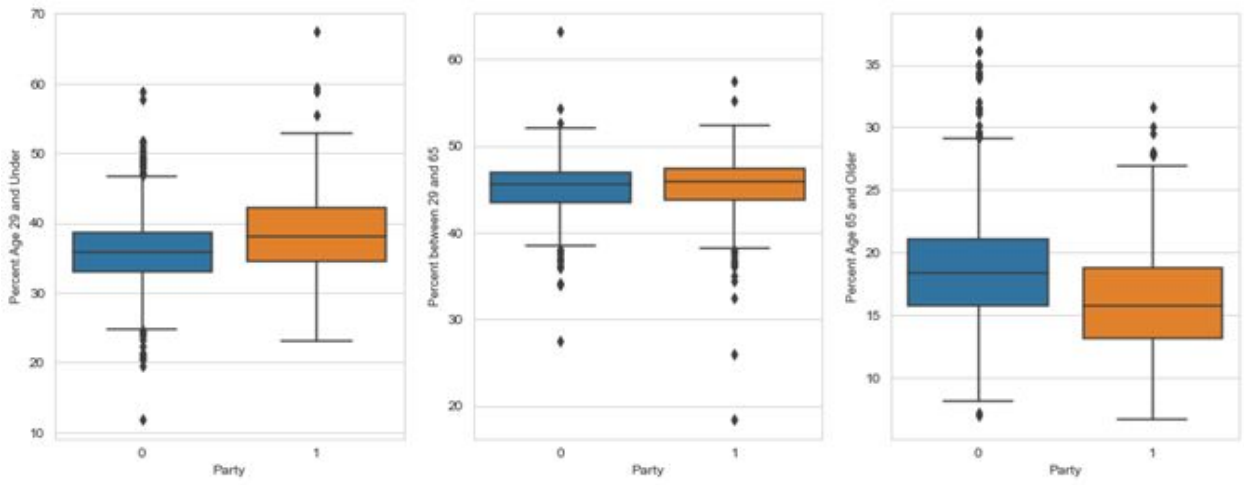


Figure 2 : Statistics for gender vs party (1 - Democratic,0-Republican)

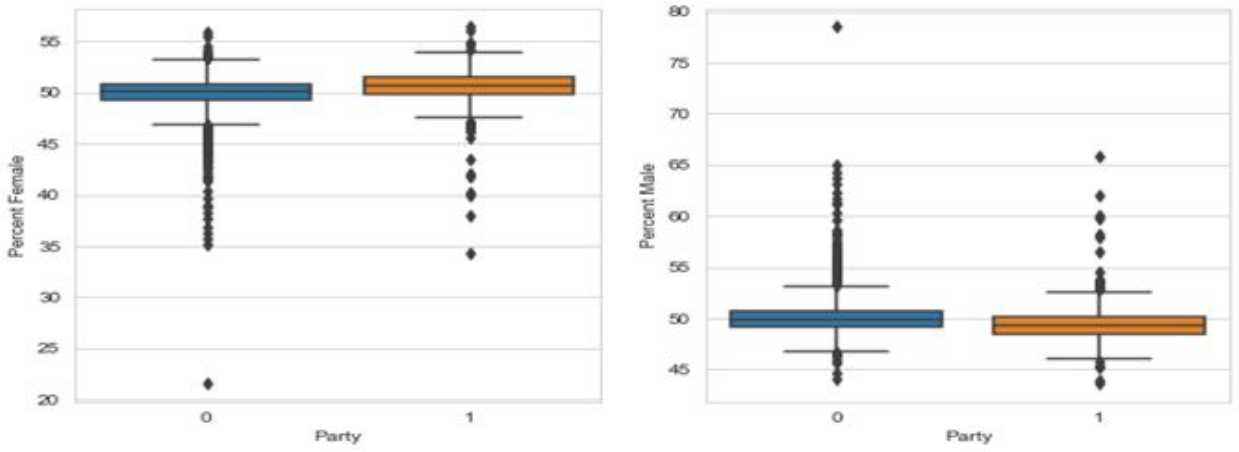


Figure 3 : Statistics for different race and ethnicity vs party (1 - Democratic, 0-Republican)

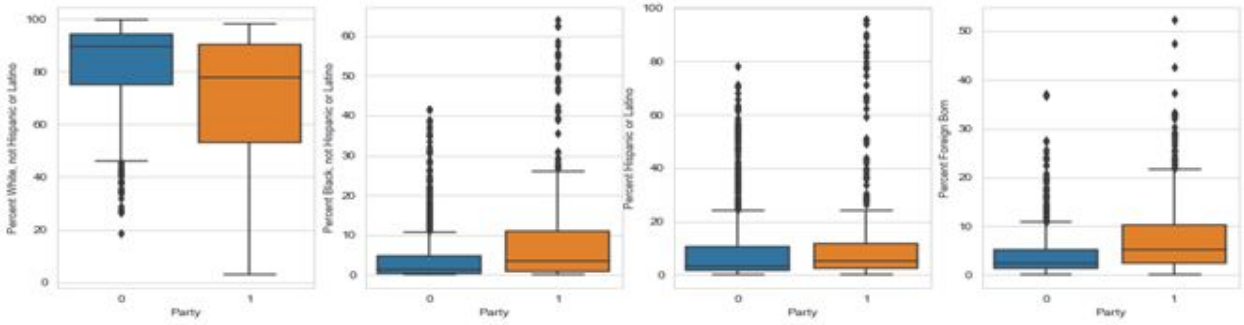
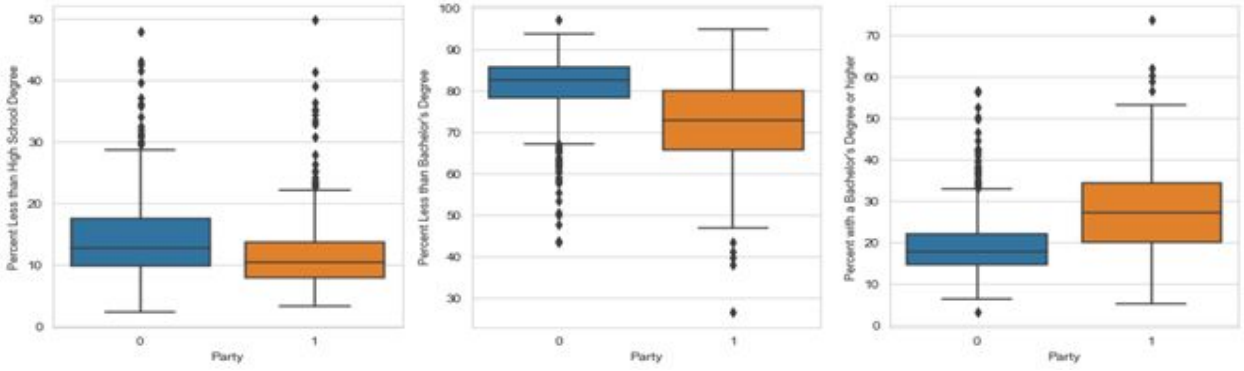


Figure 4 : Statistics for different literacy group vs party (1 - Democratic, 0-Republican)



Race and Ethnicity:

The conclusions we make from the plots is that the Republicans win counties where there is a high White Demographic. Although both relatively win counties where the white population is high.But the Republicans seem to win in the white county. Whereas, the Democrats win Counties with higher

black and foreign-born populated counties compared to the Republicans. In the case of the Hispanic population it is too close to say whether a higher Hispanic population can help either party win a county.

Gender:

In the case of gender, the data is very tight along party lines but there is a small difference seen in the boxplots and that Democrats primarily win where there is a greater female population and Republicans win where there is a greater male population.

Age:

Now in case of age we can conclude that people aged 29 and under tend to vote Democrats which is why Democrats win counties with a younger population that votes.And people aged 65 and over vote Republican more and that is why Republicans win counties with older populations.But as far as the ages in between it is too close to say whether they lean more towards Democrat or Republican as the statistics are indistinguishable towards one party.

Literacy:

Lastly, in the case of education we can conclude based on our data and plots that counties with higher populations with people with less than a high school degree and a Bachelor's degree will vote Republican whereas in counties with a high population with more than a Bachelor's degree will vote Democrat which makes sense as seen in other data reports that republicans do well in the deep south and the deep south has a higher than average high school dropout rates than in other parts of the country.

So, with these box plots it helps us visualize how certain demographics can influence an election result even before the election happens based on how certain counties with a high population of a particular demographic attribute vote.

Task 9:.

We believe that Education and Age are the most important as seen in the difference in the plots and data is more marginal than other variables like race and gender where the data isn't strong enough to support whether a county will be Democrat or Republican.Although with age and education we can see it and also agree based on how election turnouts go in the case of education as seen in today's political landscape the less educated tend to be more Republican as seen in the south and rural areas like Nebraska compared to more educated areas like California and New York that tend to vote Democrat.But this is mostly on the basis that education plays a big role on an individual to decide how to vote. In the case of Age, it makes sense that Democrats out win the Republicans based on the party's pillars. Young voters tend to be more for free education as this is the age, they

are going to school so what will be better than getting to go to school without worrying about the finances of it or student loans.Democrats are also keen on more progressive ideas like same-sex marriage, legalization of recreational cannabis, and women's reproductive rights which attract young crowd.

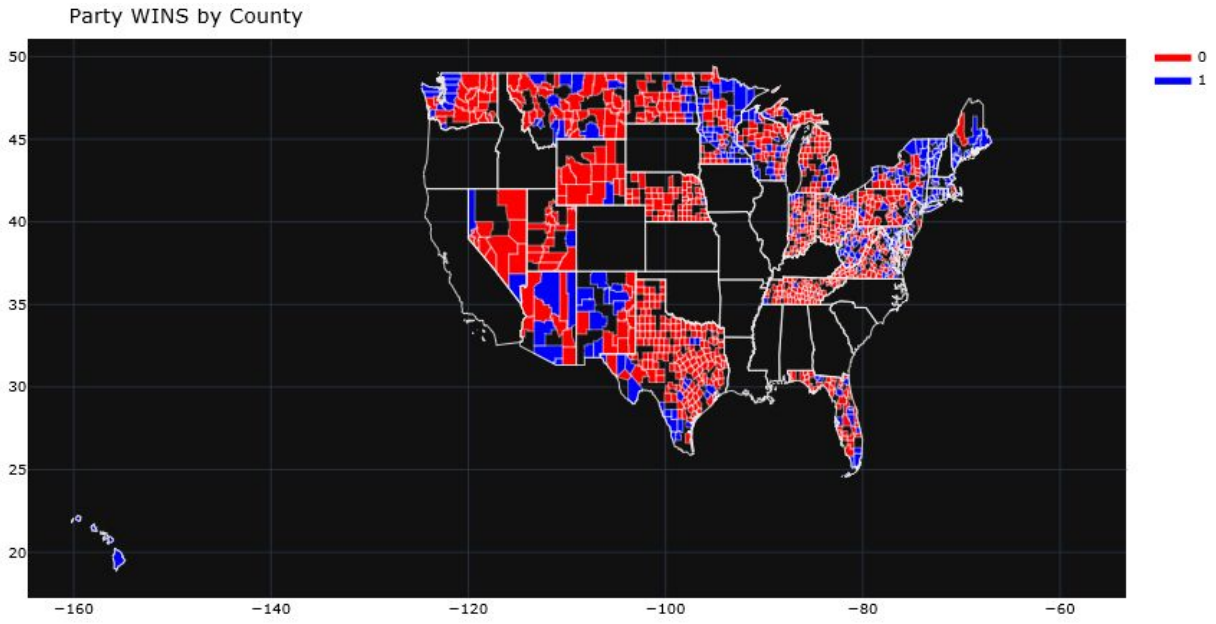
Whereas the senior populations vote Republican majorly as the party is more of conservative and traditional values that older people cherish more.Not to mention this is an age where they are retired and are more economically vulnerable because changes in their society can affect their livelihood and even threaten their current way of life.

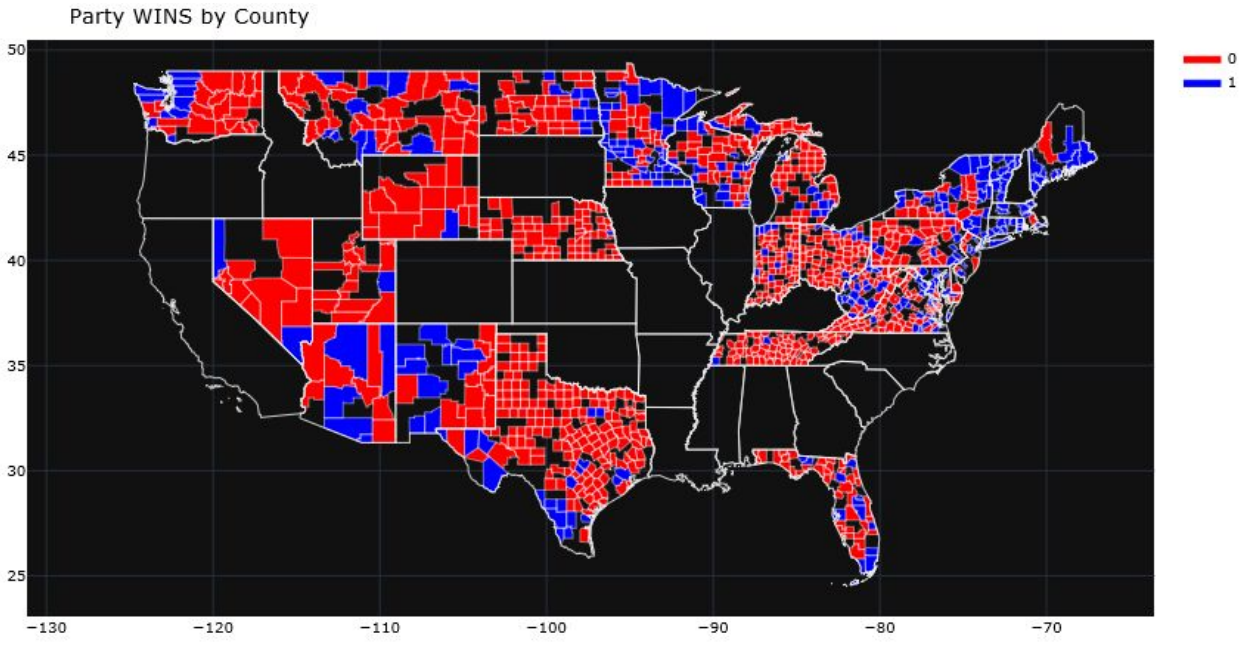
But the in between age of these two groups is what really decides whether a county will vote a certain way and perhaps with more thorough information we can also investigate whether high younger aged population will influence the low population of the middle age group to vote for democrats or vice versa.

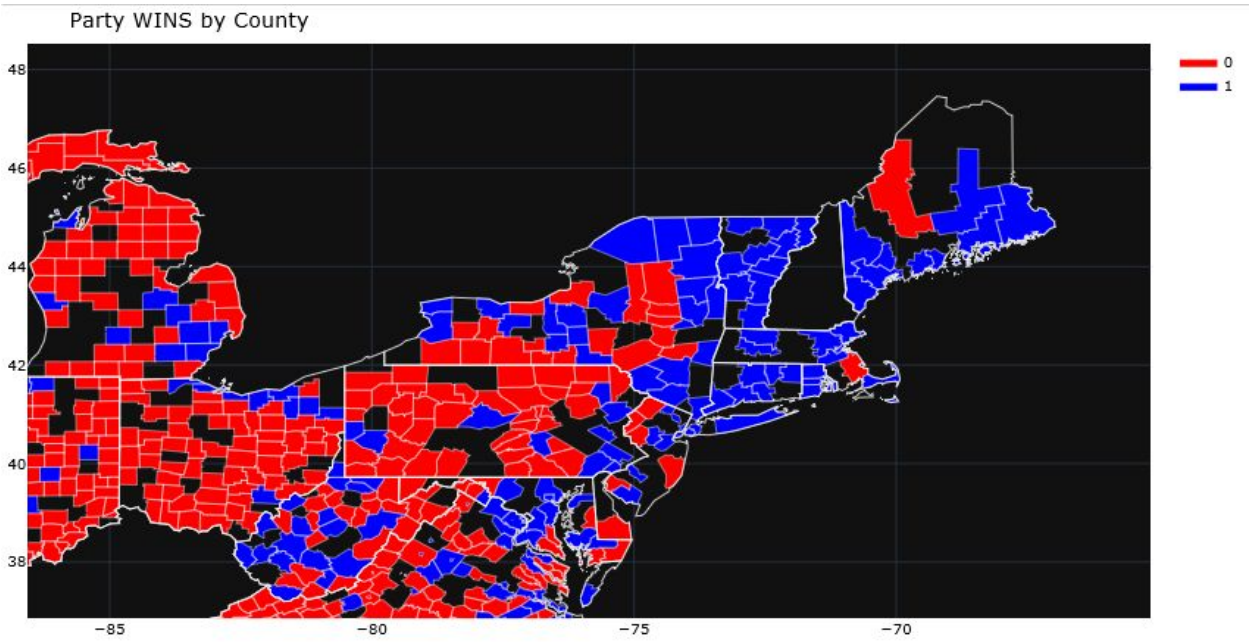
Task 10 :

In order to view Democratic, Republican counties as a map we used "import plotly.figure_factory as ff". To construct a map where we can visualize the given states that voted for a specific party. This is done by using the FIPS column and party values. After gathering the data from the merge_data we simply needed to specify what type of map we want, in this case, "USA" with the addition of "HI" . On the right side of the margin we have a simple legend with a red bar labeled as 0 which identify "Republican" and 1 for "Democratic". This will populate the counties within the states but we don't get a great visualization.

To zoom in and out of the image to get a clearer picture of small states, use the zoom button provided on the top.This is important to see the smallest northeast states or even Hawaii.


Party WINS by County
50
45
40
35
30
25
20
−160
−140
−120
−100
−80
−60
0
1


Party WINS by County
50
45
40
35
30
25
−130
−120
−110
−100
−90
−80
−70
0
1


Party WINS by County
0
1
48
46
44
42
40
38
−85
−80
−75
−70