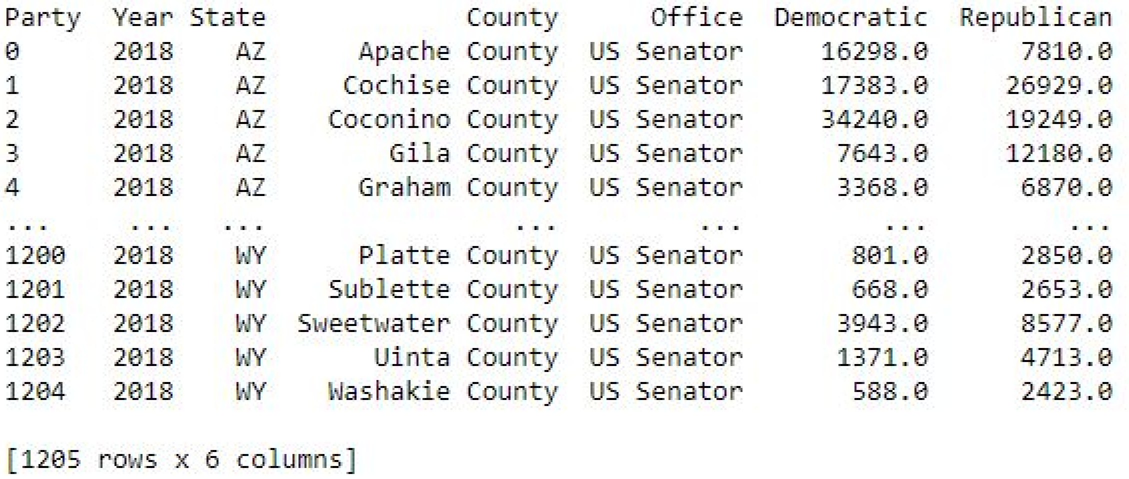
CS 418: Introduction to Data Science
Project 01: Exploratory Data Analysis

Aakash Kotak, Nancy Tacuri Malo, Daisy Sandoval

1. (5 pts.) Reshape dataset election_train from long format to wide format. Hint: the reshaped dataset should contain 1205 rows and 6 columns.

We reshaped the election_train dataset on our Jupyter Notebook file, and this is our result after reshaping the dataset from long format to wide format:

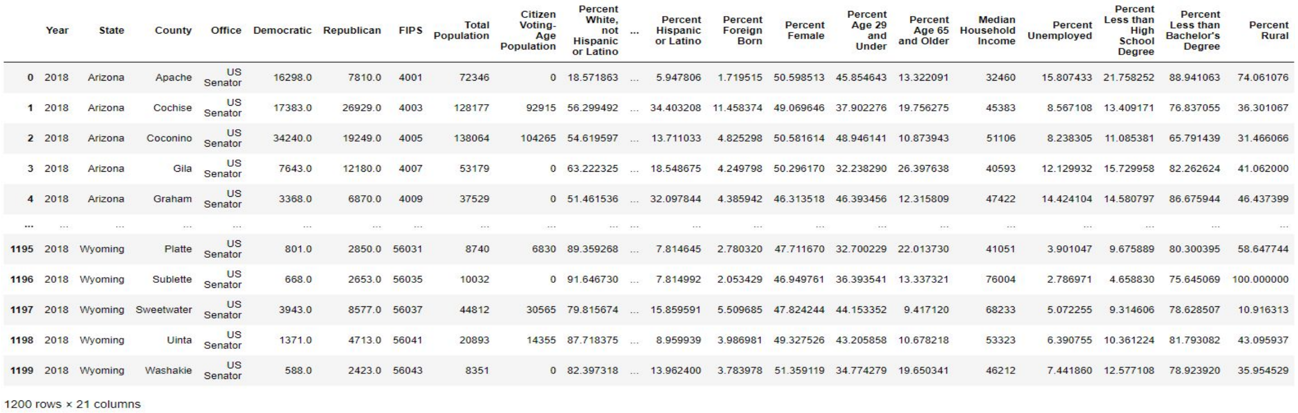


```
Party  Year State             County      Office  Democratic  Republican
0      2018    AZ     Apache County  US Senator     16298.0      7810.0
1      2018    AZ    Cochise County  US Senator     17383.0     26929.0
2      2018    AZ   Coconino County  US Senator     34240.0     19249.0
3      2018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

[1205 rows x 6 columns]
```

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.

After merging all the columns of election_train with demographics_train, we should end up with 21 columns. After fixing all the inconsistencies and getting rid of duplicates, we end up with 1200 rows.



	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	Percent Age 65 and Older	Median Household Income	Percent Unemployed	Percent Less than High School Degree	Percent Less than Bachelor's Degree	Percent Rural
0	2018	Arizona	Apache	US Senator	16298.0	7810.0	4001	72346	0	18.571863	...	5.947806	1.719515	50.598513	45.854643	13.322091	32460	15.807433	21.758252	88.941063	74.061076
1	2018	Arizona	Cochise	US Senator	17383.0	26929.0	4003	128177	92915	56.299492	...	34.403208	11.458374	49.069646	37.902276	19.756275	45383	8.567108	13.409171	76.837055	36.301067
2	2018	Arizona	Coconino	US Senator	34240.0	19249.0	4005	138064	104265	54.619597	...	13.711033	4.825298	50.581614	48.946141	10.873943	51106	8.238305	11.085381	65.791439	31.466066
3	2018	Arizona	Gila	US Senator	7643.0	12180.0	4007	53179	0	63.222325	...	18.548675	4.249798	50.296170	32.238290	26.397638	40593	12.129932	15.729958	82.262624	41.062000
4	2018	Arizona	Graham	US Senator	3368.0	6870.0	4009	37529	0	51.461536	...	32.097844	4.385942	46.313518	46.393456	12.315809	47422	14.424104	14.580797	86.675944	46.437399
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1195	2018	Wyoming	Platte	US Senator	801.0	2850.0	56031	8740	6830	89.359268	...	7.814645	2.780320	47.711670	32.700229	22.013730	41051	3.901047	9.675889	80.300395	58.647744
1196	2018	Wyoming	Sublette	US Senator	668.0	2653.0	56035	10032	0	91.646730	...	7.814992	2.053429	46.949761	36.393541	13.337321	76004	2.786971	4.658830	75.645069	100.000000
1197	2018	Wyoming	Sweetwater	US Senator	3943.0	8577.0	56037	44812	30565	79.815674	...	15.859591	5.509685	47.824244	44.153352	9.417120	68233	5.072255	9.314606	78.628507	10.916313
1198	2018	Wyoming	Uinta	US Senator	1371.0	4713.0	56041	20893	14355	87.718375	...	8.959939	3.986981	49.327526	43.205858	10.678218	53323	6.390755	10.361224	81.793082	43.095937
1199	2018	Wyoming	Washakie	US Senator	588.0	2423.0	56043	8351	0	82.397318	...	13.962400	3.783978	51.359119	34.774279	19.650341	46212	7.441860	12.577108	78.923920	35.954529

1200 rows × 21 columns

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?

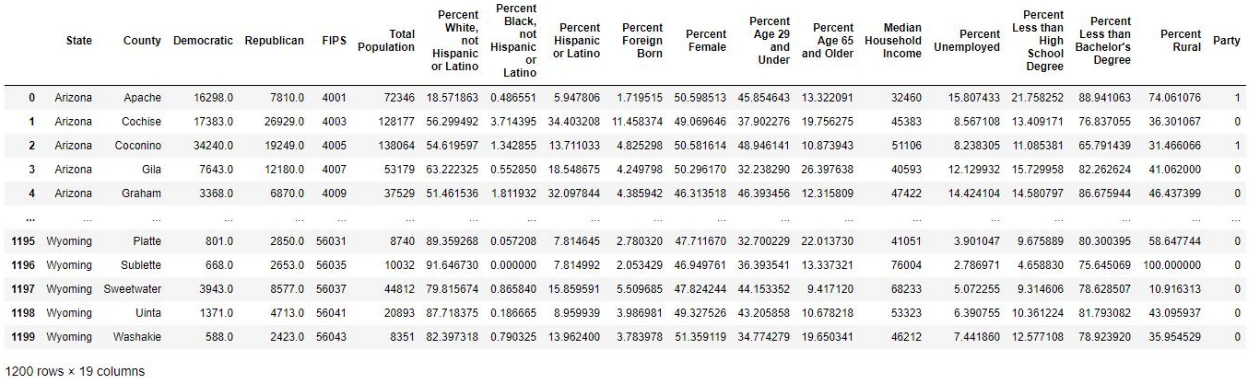
Since there are 21 columns in our merged dataset, this means that these are our variables, so we have 21 variables in our merged dataset. To check the types of these variables, we used the info() function to determine the types. From this, we figured that there are 13 float64 variables, 5 int64 variables, and 3 object variables. Our Year, FIPS, Total Population, Citizen Voting-Age Population, and Median Household Income columns are all int64 type variables. Our State, County and Office columns are all object type variables. The rest of the columns that weren't mentioned are all float64 type variables. There are a few variables that we feel are irrelevant or redundant variables since we do not end up using them in Tasks 1-10 such as "Year", "Citizen Voting-Age Population", and "Office". Since we ended up merging our 2018 election data with the 2012-2016 demographics data, our year is unnecessary and it's redundant to have every row say "2018" in the Year column. This is basically the same case for the "Office" column. We don't use this column anywhere, and it's being redundant by having every row say "US Senator". For our "Citizen Voting-Age Population", we don't use this information anywhere for Tasks 1-10, so it's also not needed. Therefore, to deal with these variables, we will remove these columns from our dataset. So our merged dataset will now be 1200 rows x 18 columns since we removed 3 of these variables.

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

After searching the merged dataset for missing values, we noticed that there were a few missing values in our Democratic and Republican columns. To deal with these values, we decided that we will replace the missing values with 0. This is to indicate that the county doesn't have any Democrats or Republicans. So basically, if there is a row that has a value in Democratic column, but it has a 0 in the Republican column, this means that the county is full Democratic. This will be the same case if it was the other way around. So then if a county has a 0 in Democratic, then the county is full Republican.

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 added a new variable named "Party" and it created a new column in our dataframe, so now we have 19 columns. Party will be equal to 1 if there were more votes for the Democratic column than the Republican column. Otherwise, if there were more votes for the Republican column than Democratic column, then Party will be equal to 0. For example, in our index 0, we see that there are more Democratic votes than Republican votes, so therefore, our Party is equal to 1. However, if we go to index 1, we see that there are more Republican votes now than Democratic votes, so therefore, our Party is equal to 0, and so on for every other row.



	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	Percent Unemployed	Percent Less than High School Degree	Percent Less than Bachelor's Degree	Percent Rural	Party
0	Arizona	Apache	16298.0	7810.0	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
1	Arizona	Cochise	17383.0	26929.0	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
2	Arizona	Coconino	34240.0	19249.0	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
3	Arizona	Gila	7643.0	12180.0	4007	53179	63.222325	0.552850	18.548675	4.249798	50.296170	32.238290	26.397638	40593	12.129932	15.729958	82.262624	41.062000	0
4	Arizona	Graham	3368.0	6870.0	4009	37529	51.461536	1.811932	32.097844	4.385942	46.313518	46.393456	12.315809	47422	14.424104	14.580797	86.675944	46.437399	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1195	Wyoming	Platte	801.0	2850.0	56031	8740	89.359268	0.057208	7.814645	2.780320	47.711670	32.700229	22.013730	41051	3.901047	9.675889	80.300395	58.647744	0
1196	Wyoming	Sublette	668.0	2653.0	56035	10032	91.646730	0.000000	7.814992	2.053429	46.949761	36.393541	13.337321	76004	2.786971	4.658830	75.645069	100.000000	0
1197	Wyoming	Sweetwater	3943.0	8577.0	56037	44812	79.815674	0.865840	15.859591	5.509685	47.824244	44.153352	9.417120	68233	5.072255	9.314606	78.628507	10.916313	0
1198	Wyoming	Uinta	1371.0	4713.0	56041	20893	87.718375	0.186665	8.959939	3.986981	49.327526	43.205858	10.678218	53323	6.390755	10.361224	81.793082	43.095937	0
1199	Wyoming	Washakie	588.0	2423.0	56043	8351	82.397318	0.790325	13.962400	3.783978	51.359119	34.774279	19.650341	46212	7.441860	12.577108	78.923920	35.954529	0

1200 rows × 19 columns

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

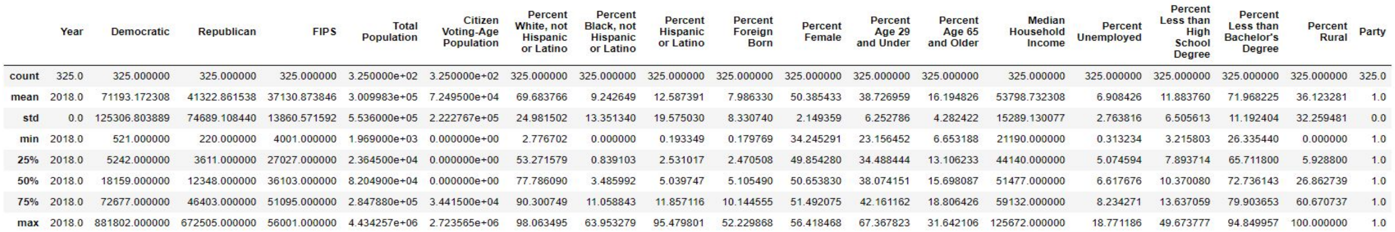
statistically significant at the $\alpha = 0.05$ significance level. What is the result of the test? What conclusion do you make from this result?

We computed our mean median household income for Democratic counties and Republican counties to be 53798.732307692306 and 48724.15085714286, respectively. We can clearly see from these means that the Democratic mean median household income is higher than the Republican mean median household income. For performing our hypothesis test, we decided to do a 2-sample t-test and our alternative hypothesis is µ1 > µ2. We determined our t-test statistic to be 5.507012409466501 and our p-value is 3.0866199456151866e-08. Since our p-value result is less than 0.05, this means for our conclusion that we are rejecting the null hypothesis.

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?

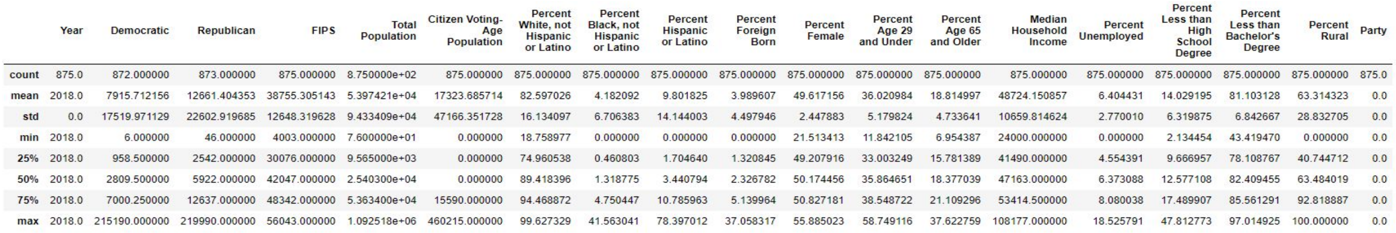
We computed our mean median household income for Democratic counties and Republican counties to be 300998.3169230769 and 53974.214857142855, respectively. We can clearly see from these means that the Democratic mean total population is higher than the Republican mean total population. For performing our hypothesis test, we decided to do a 2-sample t-test and our alternative hypothesis is µ1 > µ2. We determined our t-test statistic to be 8.001207114045041 and our p-value is 1.0482859676754979e-14. Since our p-value result is less than 0.05, this means for our conclusion that we are rejecting the null hypothesis.

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?



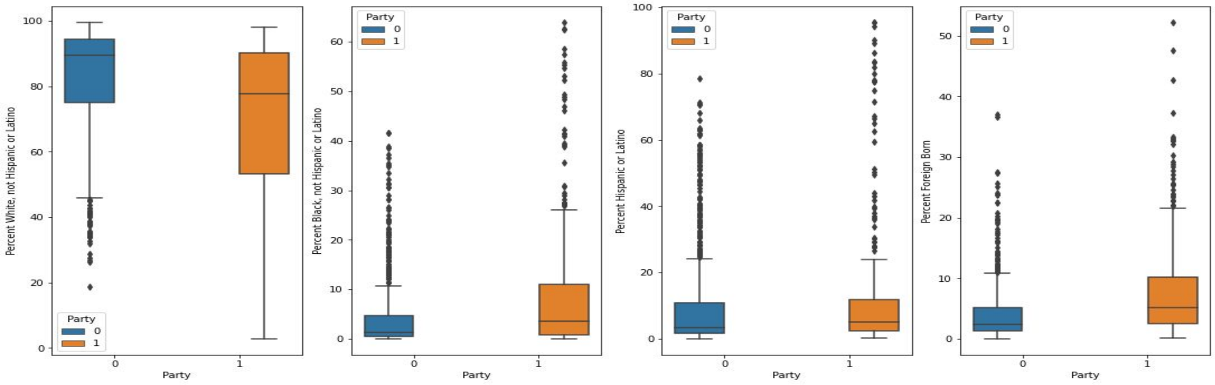
	Year	Democratic	Republican	FIPS	Total Population	Citizen Voting-Age Population	Percent White, not Hispanic or Latino	Percent Black, not Hispanic or Latino	Percent Hispanic or Latino	Percent Foreign Born	Percent Female	Percent Age 29 and Under	Percent Age 65 and Older	Median Household Income	Percent Unemployed	Percent Less than High School Degree	Percent Less than Bachelor's Degree	Percent Rural	Party
count	325.0	325.000000	325.000000	325.000000	3.250000e+02	3.250000e+02	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.0
mean	2018.0	71193.172308	41322.861538	37130.873846	3.009983e+05	7.249500e+04	69.683766	9.242649	12.587391	7.986330	50.385433	38.726959	16.194826	53798.732308	6.908426	11.883760	71.968225	36.123281	1.0
std	0.0	125306.803889	74689.108440	13860.571592	5.536000e+05	2.222767e+05	24.981502	13.351340	19.575030	8.330740	2.149359	6.252786	4.282422	15289.130077	2.763816	6.505613	11.192404	32.259481	0.0
min	2018.0	521.000000	220.000000	4001.000000	1.969000e+03	0.000000e+00	2.776702	0.000000	0.193349	0.179769	34.245291	23.156452	6.653188	21190.000000	0.313234	3.215803	26.335440	0.000000	1.0
25%	2018.0	5242.000000	3611.000000	27027.000000	2.364500e+04	0.000000e+00	53.271579	0.839103	2.531017	2.470508	49.854280	34.488444	13.106233	44140.000000	5.074594	7.893714	65.711800	5.928800	1.0
50%	2018.0	18159.000000	12348.000000	36103.000000	8.204900e+04	0.000000e+00	77.786090	3.485992	5.039747	5.105490	50.653830	38.074151	15.698087	51477.000000	6.617676	10.370080	72.736143	26.862739	1.0
75%	2018.0	72677.000000	46403.000000	51095.000000	2.847880e+05	3.441500e+04	90.300749	11.058843	11.857116	10.144555	51.492075	42.161162	18.806426	59132.000000	8.234271	13.637059	79.903653	60.670737	1.0
max	2018.0	881802.000000	672505.000000	56001.000000	4.434257e+06	2.723565e+06	98.063495	63.953279	95.479801	52.229868	56.418468	67.367823	31.642106	125672.000000	18.771186	49.673777	94.849957	100.000000	1.0

Democratic Counties Descriptive Statistics

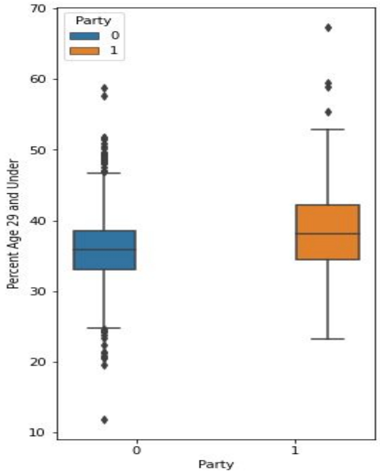


	Year	Democratic	Republican	FIPS	Total Population	Citizen Voting-Age Population	Percent White, not Hispanic or Latino	Percent Black, not Hispanic or Latino	Percent Hispanic or Latino	Percent Foreign Born	Percent Female	Percent Age 29 and Under	Percent Age 65 and Older	Median Household Income	Percent Unemployed	Percent Less than High School Degree	Percent Less than Bachelor's Degree	Percent Rural	Party
count	875.0	872.000000	873.000000	875.000000	8.750000e+02	875.000000	875.000000	875.000000	875.000000	875.000000	875.000000	875.000000	875.000000	875.000000	875.000000	875.000000	875.000000	875.000000	875.0
mean	2018.0	7915.712156	12661.404353	38755.305143	5.397421e+04	17323.685714	82.597026	4.182092	9.801825	3.989607	49.617156	36.020984	18.814997	48724.150857	6.404431	14.029195	81.103128	63.314323	0.0
std	0.0	17519.971129	22602.919685	12648.319628	9.433409e+04	47166.351728	16.134097	6.706383	14.144003	4.497946	2.447883	5.179824	4.733641	10659.814624	2.770010	6.319875	6.842667	28.832705	0.0
min	2018.0	6.000000	46.000000	4003.000000	7.600000e+01	0.000000	18.758977	0.000000	0.000000	0.000000	21.513413	11.842105	6.954387	24000.000000	0.000000	2.134454	43.419470	0.000000	0.0
25%	2018.0	958.500000	2542.000000	30076.000000	9.565000e+03	0.000000	74.960538	0.460803	1.704640	1.320845	49.207916	33.003249	15.781389	41490.000000	4.554391	9.666957	78.108767	40.744712	0.0
50%	2018.0	2809.500000	5922.000000	42047.000000	2.540300e+04	0.000000	89.418396	1.318775	3.440794	2.326782	50.174456	35.864651	18.377039	47163.000000	6.373088	12.577108	82.409455	63.484019	0.0
75%	2018.0	7000.250000	12637.000000	48342.000000	5.363400e+04	15590.000000	94.468872	4.750447	10.785963	5.139964	50.827181	38.548722	21.109296	53414.500000	8.080038	17.489907	85.561291	92.818887	0.0
max	2018.0	215190.000000	219990.000000	56043.000000	1.092518e+06	460215.000000	99.627329	41.563041	78.397012	37.058317	55.885023	58.749116	37.622759	108177.000000	18.525791	47.812773	97.014925	100.000000	0.0

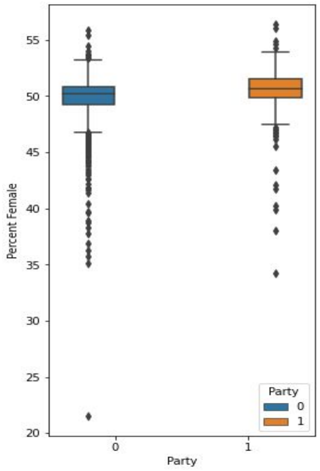
Republican Counties Descriptive Statistics



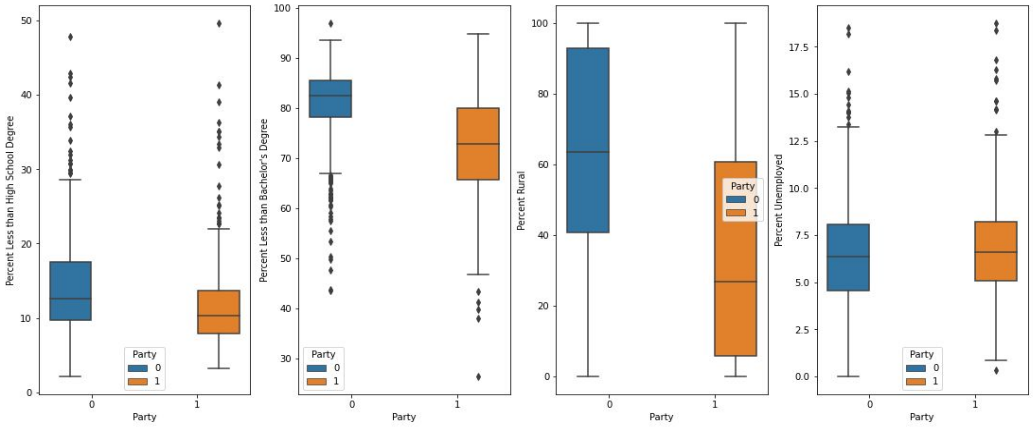
Race and Ethnicity Boxplots



Age Boxplots



Gender Boxplot



Education Boxplots

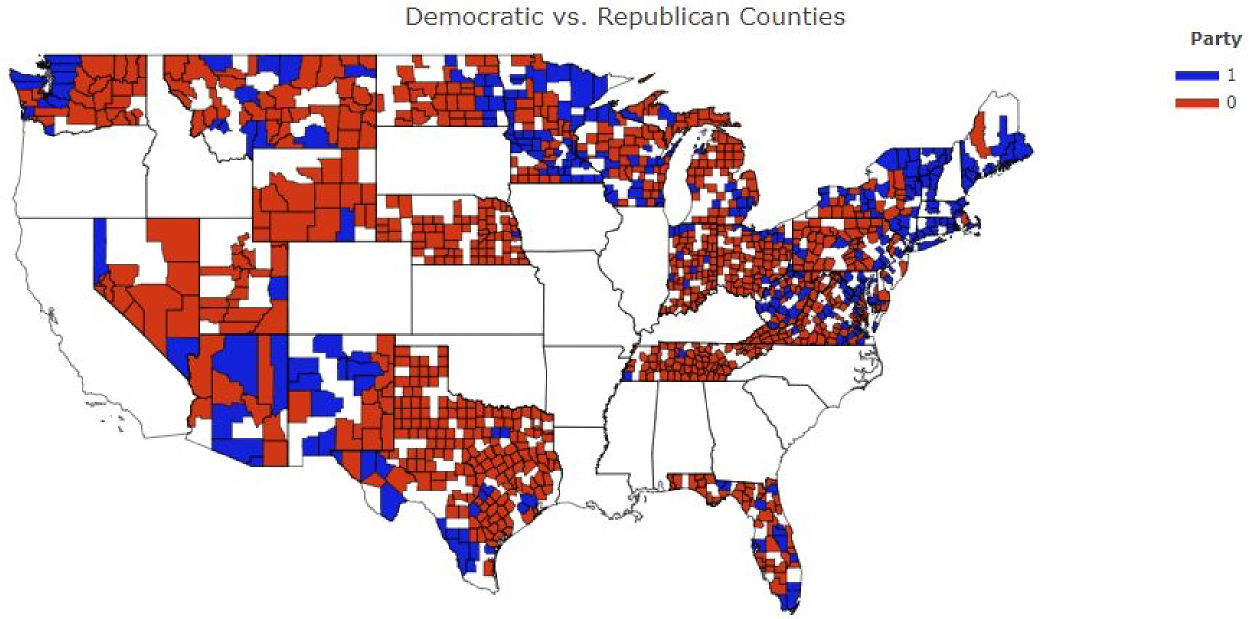
Based on our boxplots, it seems that there are significantly more Democrats than Republicans in terms of Race and Ethnicity, and Age since all of the boxplots for Democrats are larger than Republicans. However, the boxplots for Education and Gender seem to be about the same for Democrats and Republicans, and there are some plots that have Republicans higher than Democrats. From our descriptive statistics though, we see that there are significantly more Republican counties, 875, than Democrat counties, 325. From this data, we can see that although there are more Republican counties than Democrat counties, there are significantly more Democrats since the Population for Democrats is much higher than Republicans.

9. (5 pts.) 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.

Based on our results from tasks 6-8, we just need the Party, Democratic, and Republican variables in our merged_data. The Democratic and Republican variables show how much of the county is Democratic or Republican, while our Party variable determines which counties are mostly Democratic or Republican by checking if the Democratic variable is greater than the Republican variable in each cell. If Democratic is greater than Republican, then it will place a "1" in the Party column in that County row. Otherwise, if Republican is greater than Democratic in the County, then it will place a "0" in the Party column in that County row. This is to distinguish between which county is Democratic or Republican.

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.

The red counties indicate that it is a Republican County while the blue counties indicate it is a Democrat County.


Democratic vs. Republican Counties
Party
1
0