Shawn John
Furqaan Ali
Bhavana laxmi Radharapu

CS 418 Project 1 Report

For this project we prepare two data files and do some exploratory data analysis. The sets of data we have to merge are election_train.csv and demographics_train.csv. The first one is the result of the 2018 senate election which is based on the county and shows the votes each party received for that county. The other dataset contains demographic data on counties in the United States. The first major task was to merge these two datasets into one. What we wanted to do was merge the datasets on county, that means the key would be the county name and variables for both would show up in the new dataset for each key. So it's an inner join where the key will be county and state. There are some inconsistencies in the data that we had to deal with before we could merge. The first was that states were abbreviated in elections_data while demographics were full names. So we mapped the values in the state column of demographics to the associated abbreviations. Next was to convert the names of the counties to lowercase so identical counties with different capitalization would not be treated as entirely different counties. Then we realized election data has the word county in the values for county so we removed those to match the naming convention of the demographics dataset. Now that the datasets have the same types of state and county columns we can merge on those columns.

Percent Age 65 and Older	Median Household Income	Percent Unemployed	Percent Less than High School Degree	Percent Less than Bachelor's Degree	Percent Rural	Year	Office	Democratic	Republican
13.322091	32460	15.807433	21.758252	88.941063	74.061076	2018	US Senator	16298.0	7810.0
19.756275	45383	8.567108	13.409171	76.837055	36.301067	2018	US Senator	17383.0	26929.0
10.873943	51106	8.238305	11.085381	65.791439	31.466066	2018	US Senator	34240.0	19249.0
26.397638	40593	12.129932	15.729958	82.262624	41.062000	2018	US Senator	7643.0	12180.0
12.315809	47422	14.424104	14.580797	86.675944	46.437399	2018	US Senator	3368.0	6870.0

Figure 1.0

Figure 1.0 shows a portion of the new merged dataset with columns from both demographic_data and elections_data.

Task 3

This resulting merged dataset has a total of 21 variables. These variables are of type object, int64, and float64. A variable that seemed somewhat redundant was Citizen Voting Age Population, since a lot of this data is gathered by the other age-related variables. We also noticed that this variable held over 600 zero-values and therefore we decided to simply remove this feature from the dataset. We do not believe it would have been a great predictor of county party, especially since we already have other features regarding age. After doing so, every variable seems to be unique to all other variables in the dataset. However, there are some variables such as Year and Office that arguably can be seen as unnecessary because they hold the same value for all observations, but we have decided to keep them so as to not lose any important information regarding our data. These variables could be removed if we included year and office information elsewhere such as in a title.

Variable	Datatype

State County Office	object
FIPS Total Population Citizen Voting Age Population Median Household Income Year	int64
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 Percent Unemployed Percent Less than High School Degree Percent Less than Bachelor's Degree Percent Rural Democratic Republican	float64

Figure 2

Figure 2 demonstrates the variables in the merged dataset as well as their data types.

Task 4

The merged dataset has a total of 5 missing values, 3 of which are in the 'Democratic' column and 2 in the 'Republican' column. Since these are vital features of the observation and there are only 5 total missing values, we believe that it is best to simply drop these observations from the dataset. After doing so, our dataset still has 1195 observations which is sufficient. There were also missing/zero values in the 'Citizen Voting Age Pop.' feature, but as mentioned in task 3, that feature was removed.

State	**County**	**Missing data for:**
TX	Bee	Republican
WI	Lafayette	Republican
NE	Lancaster	Democratic
TN	Meigs	Democratic
TX	Menard	Democratic

Figure 3

Figure 3 displays the observations with missing values.

Task 6

The mean median household income for Democratic counties in our dataset was $53798.73, whereas the mean median household income for the Republican counties was $48746.82. To test whether these results are statistically significant, we performed a hypothesis test, where the null hypothesis is that the median household incomes for both democratic and republican counties are equivalent and the alternate hypothesis is that democratic counties' median household income is greater. After receiving a p-value of nearly zero (statistic=5.479141589767387, pvalue=7.149437363182572e-08) (well below the significance level), we conclude by rejecting the null hypothesis and noting that our results are statistically significant. Based on our results, The median household income of democratic counties in the US are likely higher than that of republican counties.

Task 7

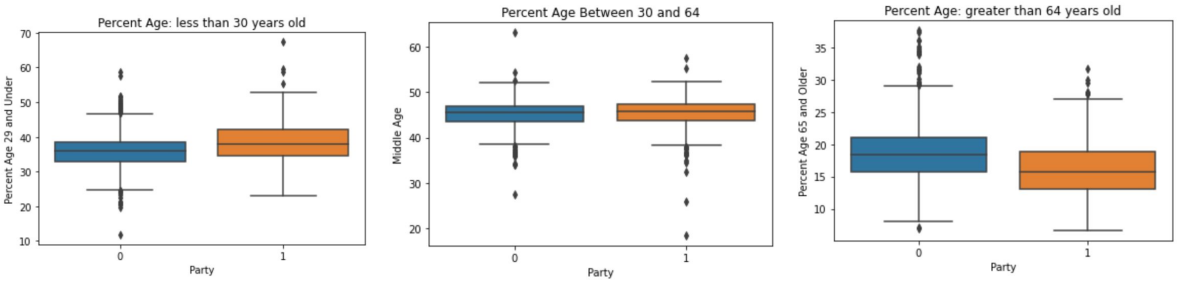
The mean population for Democratic counties in our dataset was 553,600, whereas the mean population for the Republican counties was 94,192. To test whether these results are statistically significant, we performed a hypothesis test, where the null hypothesis is that the populations for both democratic and republican counties are equivalent and the alternate hypothesis is that democratic counties' population is greater. After receiving a p-value of nearly zero (statistic=8.004638577960957, pvalue=2.0478717602973023e-14) (well below the significance level), we conclude by rejecting the null hypothesis and noting that our results are statistically significant. Based on our results, The population of democratic counties in the US are likely higher than that of republican counties.

Task 8

Comparing features for democratic and republican counties

Party 1 = Democratic, Party 0 = Republican

Age:

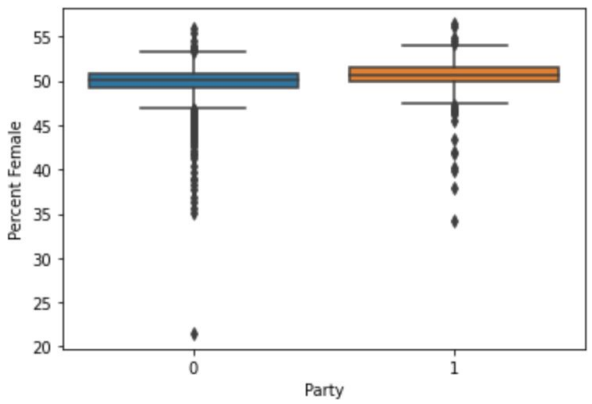


	Middle Age	Percent Age 29 and Under	Percent Age 65 and Older	Party
count	870.000000	870.000000	870.000000	870.0
mean	45.166015	36.005719	18.828267	0.0
std	2.910264	5.181522	4.733155	0.0
min	27.421759	11.842105	6.954387	0.0
25%	43.522522	32.983652	15.784982	0.0
50%	45.553295	35.846532	18.377896	0.0
75%	46.975771	38.539787	21.112847	0.0
max	63.157895	58.749116	37.622759	0.0

	Middle Age	Percent Age 29 and Under	Percent Age 65 and Older	Party
count	325.000000	325.000000	325.000000	325.0
mean	45.078214	38.726959	16.194826	1.0
std	3.907598	6.252786	4.282422	0.0
min	18.433769	23.156452	6.653188	1.0
25%	43.741937	34.488444	13.106233	1.0
50%	45.817819	38.074151	15.698087	1.0
75%	47.448269	42.161162	18.806426	1.0
max	57.478906	67.367823	31.642106	1.0

The figures above visualize the differences in population age percentages between democratic and republican parties. It seems that democratic counties tend to have a greater percentage of younger individuals. Republican counties have a larger percentage of seniors. The percentage of middle-aged individuals, however, seem to be roughly the same for both groups. This variable is significant.

Gender:

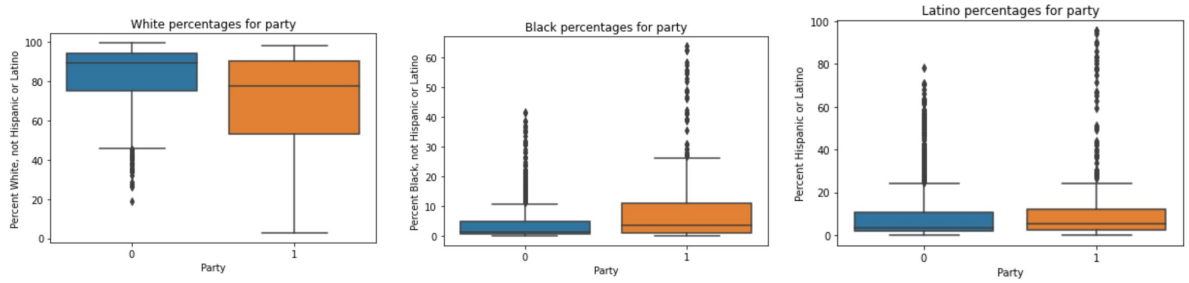


	Percent Female
count	870.000000
mean	49.630898
std	2.429013
min	21.513413
25%	49.222905
50%	50.176792
75%	50.829770
max	55.885023

	Percent Female
count	325.000000
mean	50.385433
std	2.149359
min	34.245291
25%	49.854280
50%	50.653830
75%	51.492075
max	56.418468

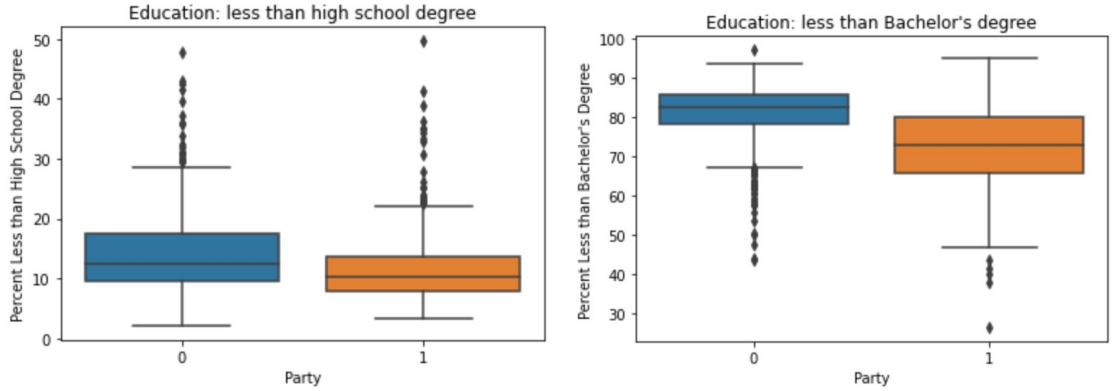
The figures above visualize gender statistics for democratic and republican counties. Gender does not seem to play a significant role, as the statistics do not vary significantly between the parties.

Race/Ethnicity:



The figures above visualize the race and ethnicity statistics between the democratic and republican counties. We can see that there is quite a significant difference in the statistics between the parties for these features. Republican counties seem to have higher percentages of White individuals and democratic counties seem to have much higher percentages of Black individuals and slightly higher percentages of Latino individuals. Based on these results, race and ethnicity seem to play a strong role in a county being republican or democratic, where more diversity increases the likelihood of a county being republican.

Education:



The figures above visualize the differences in educational levels for democratic and republican counties. We observe that democratic counties seem to have higher percentages of individuals with more education (both in terms of a high school degree and a bachelor's degree). The differences seem to be quite significant, especially in terms of a bachelor's degree, indicating that education is a significant feature in determining the party of a county.

Task 9

	Percent White, not Hispanic or Latino	Party
count	870.000000	870.0
mean	82.656646	0.0
std	16.056122	0.0
min	18.758977	0.0
25%	75.016397	0.0
50%	89.434849	0.0
75%	94.466596	0.0
max	99.627329	0.0

	Percent White, not Hispanic or Latino	Party
count	325.000000	325.0
mean	69.683766	1.0
std	24.981502	0.0
min	2.776702	1.0
25%	53.271579	1.0
50%	77.786090	1.0
75%	90.300749	1.0
max	98.063495	1.0

	Percent Black, not Hispanic or Latino	Party
count	870.000000	870.0
mean	4.189241	0.0
std	6.721695	0.0
min	0.000000	0.0
25%	0.460419	0.0
50%	1.318311	0.0
75%	4.753831	0.0
max	41.563041	0.0

	Percent Black, not Hispanic or Latino	Party
count	325.000000	325.0
mean	9.242649	1.0
std	13.351340	0.0
min	0.000000	1.0
25%	0.839103	1.0
50%	3.485992	1.0
75%	11.058843	1.0
max	63.953279	1.0

	Percent Less than Bachelor's Degree
count	870.000000
mean	81.095427
std	6.815537
min	43.419470
25%	78.108424
50%	82.406700
75%	85.546272
max	97.014925

	Percent Less than Bachelor's Degree
count	325.000000
mean	71.968225
std	11.192404
min	26.335440
25%	65.711800
50%	72.736143
75%	79.903653
max	94.849957

The variables in the dataset that seem to play significant roles in determining the party of a county are the following: Total Population, Median Household Income, Percent Age 29 and Under, Percent Age 65 and Older, White Black and Latino Percentage, and Education percentages of less than high school and bachelor's degrees. From these, we would argue that the most important variables are White Percentage, Black Percentage, Percent Less than Bachelor's Degree, and Total Population. The mean of the White population percentage in republican counties is over 12% higher than the mean White percentage in democratic counties, clearly indicating that large percentages of White individuals increases the likelihood of a county being Republican. The mean Black Percentage for democratic counties is over twice that of republican counties, indicating that higher percentages of Black individuals increases the likelihood of a county being Democratic. The mean Percentage Less than Bachelor's degree for democratic counties is almost 10% lower than that of republican counties, indicating that having a higher percentage of individuals with a bachelor's degree increases the likelihood of a county being democratic. Lastly, the mean total population of democratic counties is almost 6 times higher than that of republicans, so a county having a larger population makes it more likely to be democratic.