Kalyan Kumar Paladugula- Michael Ybarra - Zoheb Mohammed
CS418 Data Science
Project 1 - Fall 2019

Project Description:

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

	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
5		2018	AZ	La Paz County	US Senator	1609.0	3265.0
6		2018	AZ	Maricopa County	US Senator	732671.0	672505.0
7		2018	AZ	Mohave County	US Senator	19214.0	50209.0
8		2018	AZ	Navajo County	US Senator	16624.0	18767.0
9		2018	AZ	Pima County	US Senator	221242.0	160550.0
10		2018	AZ	Santa Cruz County	US Senator	9241.0	3828.0
11		2018	AZ	Yavapai County	US Senator	40160.0	65308.0
12		2018	CT	Fairfield County	US Senator	210899.0	131321.0
13		2018	CT	Hartford County	US Senator	203591.0	123864.0
14		2018	CT	Middlesex County	US Senator	42383.0	32836.0
15		2018	CT	New Haven County	US Senator	179714.0	126004.0
16		2018	CT	Tolland County	US Senator	34732.0	28046.0
17		2018	CT	Windham County	US Senator	20490.0	19032.0
18		2018	DE	Sussex County	US Senator	40675.0	50391.0
19		2018	FL	Alachua County	US Senator	74493.0	40599.0
20		2018	FL	Baker County	US Senator	1945.0	8579.0
21		2018	FL	Bay County	US Senator	16723.0	46681.0
22		2018	FL	Bradford County	US Senator	2879.0	7576.0
23		2018	FL	Brevard County	US Senator	121112.0	160305.0
24		2018	FL	Broward County	US Senator	472239.0	211397.0
25		2018	FL	Charlotte County	US Senator	33525.0	52916.0
26		2018	FL	Citrus County	US Senator	22660.0	48008.0
27		2018	FL	Collier County	US Senator	54390.0	101266.0
28		2018	FL	Desoto County	US Senator	3328.0	5503.0
29		2018	FL	Dixie County	US Senator	1322.0	4442.0
...		...	...	...	...	...	...
1180		2018	WV	Roane County	US Senator	2165.0	1899.0
1181		2018	WV	Summers County	US Senator	2069.0	1868.0
1182		2018	WV	Taylor County	US Senator	2376.0	2642.0
1183		2018	WV	Tucker County	US Senator	1469.0	1502.0
1184		2018	WV	Tyler County	US Senator	1065.0	1603.0
1185		2018	WV	Upshur County	US Senator	3102.0	4010.0
1186		2018	WV	Wayne County	US Senator	6395.0	5954.0
1187		2018	WV	Wetzel County	US Senator	2518.0	2135.0
1188		2018	WV	Wood County	US Senator	14189.0	13696.0
1189		2018	WV	Wyoming County	US Senator	2607.0	3096.0
1190		2018	WY	Albany County	US Senator	7576.0	6366.0
1191		2018	WY	Campbell County	US Senator	1628.0	11020.0
1192		2018	WY	Carbon County	US Senator	1359.0	3673.0
1193		2018	WY	Converse County	US Senator	834.0	3959.0
1194		2018	WY	Fremont County	US Senator	4734.0	9262.0
1195		2018	WY	Goshen County	US Senator	1020.0	3658.0
1196		2018	WY	Johnson County	US Senator	722.0	3085.0
1197		2018	WY	Lincoln County	US Senator	1152.0	5846.0
1198		2018	WY	Natrona County	US Senator	7285.0	16359.0
1199		2018	WY	Niobrara County	US Senator	144.0	980.0
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. **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.**

	Year	State	County	Office	Democratic	Republican	FIPS
0	2018	AZ	apache	US Senator	16298.0	7810.0	4001
1	2018	AZ	cochise	US Senator	17383.0	26929.0	4003
2	2018	AZ	coconino	US Senator	34240.0	19249.0	4005
3	2018	AZ	gila	US Senator	7643.0	12180.0	4007
4	2018	AZ	graham	US Senator	3368.0	6870.0	4009

	Total Population	Citizen Voting-Age Population
0	72346	0
1	128177	92915
2	138064	104265
3	53179	0
4	37529	0

	Percent White, not Hispanic or Latino	...	Percent Hispanic or Latino
0	18.571863	...	5.947806
1	56.299492	...	34.403208
2	54.619597	...	13.711033
3	63.222325	...	18.548675
4	51.461536	...	32.097844

	Percent Foreign Born	Percent Female	Percent Age 29 and Under
0	1.719515	50.598513	45.854643
1	11.458374	49.069646	37.902276
2	4.825298	50.581614	48.946141
3	4.249798	50.296170	32.238290
4			46.393456

[1200 rows x 21 columns]

	Percent Age 65 and Older	Median Household Income	Percent Unemployed
0	13.322091	32460	15.807433
1	19.756275	45383	8.567108
2	10.873943	51106	8.238305
3	26.397638	40593	12.129932
4	12.315809	47422	14.424104

	Percent Less than High School Degree	Percent Less than Bachelor's Degree
0	21.758252	88.941063
1	13.409171	76.837055
2	11.085381	65.791439
3	15.729958	82.262624
4	14.580797	86.675944

	Percent Rural
0	74.061076
1	36.301067
2	31.466066
3	41.062000
4	46.437399

[5 rows x 21 columns]

3. **Search the merged dataset for missing values. Are there any missing values? If so, how will you deal with these values?**

   Yes. There are many zeroes (missing values) in the Citizen Voting-Age Population. We are replacing them with NaN.

4. **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.**

	State	County	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
0	AZ	apache	16298.0	7810.0	4001	72346	NaN	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	AZ	cochise	17383.0	26929.0	4003	128177	92915.0	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	AZ	coconino	34240.0	19249.0	4005	138064	104265.0	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	AZ	gila	7643.0	12180.0	4007	53179	NaN	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	AZ	graham	3368.0	6870.0	4009	37529	NaN	51.461536	1.811932	32.097844	4.385942	46.313518	46.393456	12.315809	47422	14.424104	14.580797	86.675944	46.437399	0

5. **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?**

   There are 21 attributes in the merged dataset.

```
Data columns (total 21 columns):
Year                                    1200 non-null int64
State                                   1200 non-null object
County                                  1200 non-null object
Office                                  1200 non-null object
Democratic                              1200 non-null float64
Republican                              1200 non-null float64
FIPS                                    1200 non-null int64
Total Population                        1200 non-null int64
Citizen Voting-Age Population           1200 non-null int64
Percent White, not Hispanic or Latino   1200 non-null float64
Percent Black, not Hispanic or Latino   1200 non-null float64
Percent Hispanic or Latino              1200 non-null float64
Percent Foreign Born                    1200 non-null float64
Percent Female                          1200 non-null float64
Percent Age 29 and Under                1200 non-null float64
Percent Age 65 and Older                1200 non-null float64
Median Household Income                 1200 non-null int64
Percent Unemployed                      1200 non-null float64
Percent Less than High School Degree    1200 non-null float64
Percent Less than Bachelor's Degree     1200 non-null float64
Percent Rural                           1200 non-null float64
dtypes: float64(13), int64(5), object(3)
```

The attributes "Year", "Office" are the irrelevant attributes. However, there are no redundant attributes. We removed the irrelevant variables from the dataset.

6. **Compute the mean population for the Democratic counties and Republican counties. Which one is higher? What is the result of the test? What conclusion do you make from this result?**

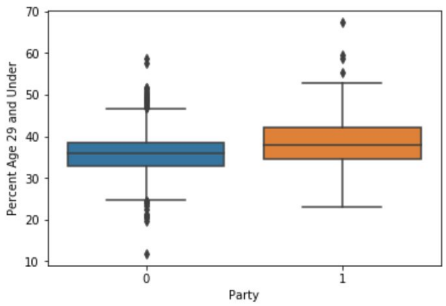
   The mean population for the Democratic counties (300998) is higher than that of Republican counties (53974). The result of the test produces a p-value (`2.097e-14`) that is less than the given significance level (0.05). Therefore, we can reject the null hypothesis that the mean of the Democratic and Republican populations are equal.

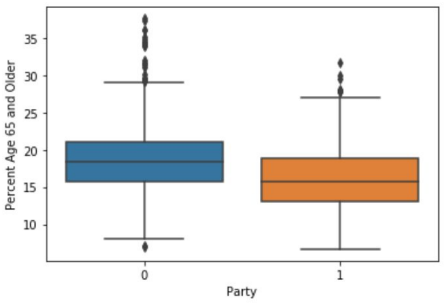
7. **Compute the mean median household income for Democratic counties and Republican parties. Which one is higher? Perform a hypothesis test to determine whether the difference is statistically significant at the 0.05 significance level. What is the result of the test? What conclusion did you make from this result?**

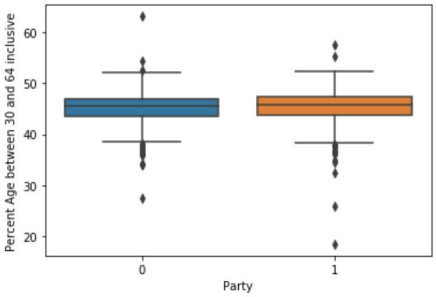
   The mean median household income for Democratic counties (53798) is higher than that of Republican parties (48724). The result of the test produces a p-value (`6.173e-08`) that is less than the given significance level (0.05). Therefore, we can reject the null hypothesis that the mean of the Democratic and Republican household incomes are equal.

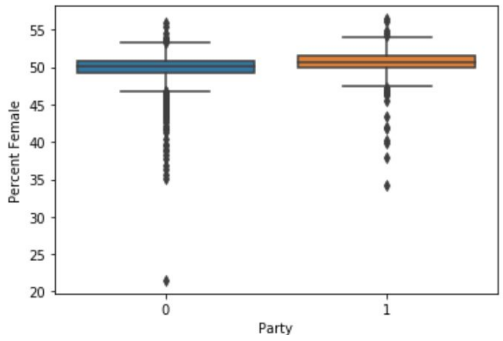
8. **Compare Democratic counties and Republican counties in terms of age, gender, race and ethnicity, and education by computing descriptive statistics and creating plots to visualize the results. What conclusions do you make for each variable from the descriptive statistics and the plots?**

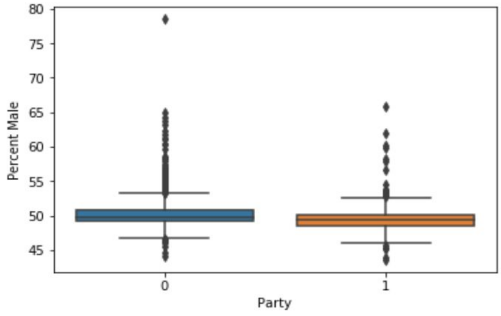
   The age group 29 and under has not much effect on the voting. The difference between the medians for both parties is not significant. The age group 65 and over has not much effect on the voting. The difference between the medians for both parties is not significant. The age group b etween 30 and 64 inclusive has almost no effect on the voting. The difference between the medians for both parties is almost zero.



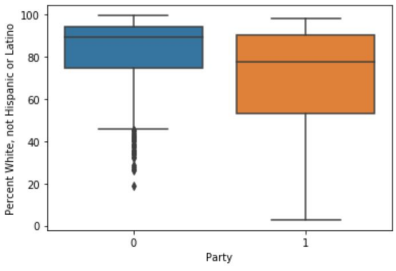


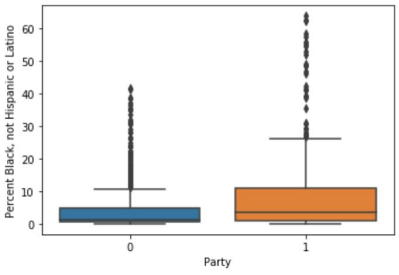


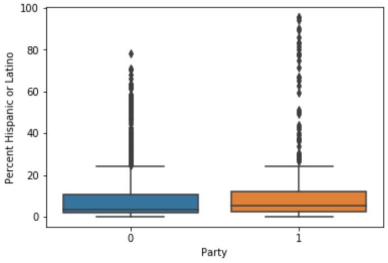


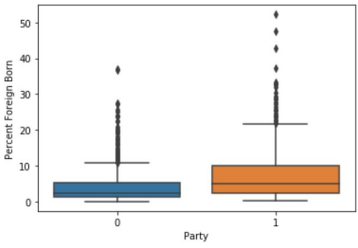


The gender has almost no effect on the voting. The difference between the medians for both parties is almost zero.

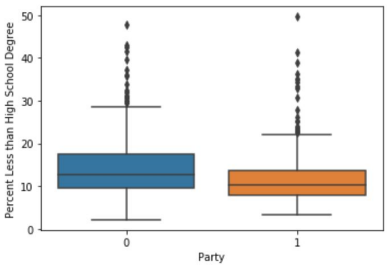


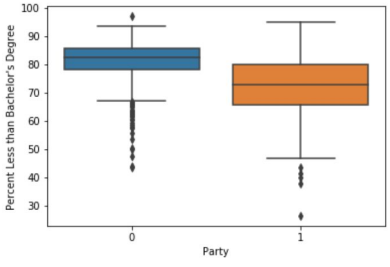


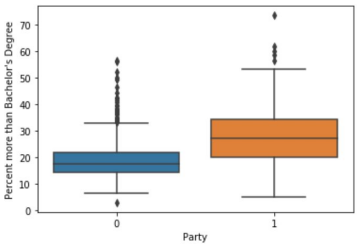




The attribute Percent White, not Hispanic or Latino has significant effect on the voting. The difference between the medians for both parties is significant. The attribute Percent White, not Hispanic or Latino has significant effect on the voting. The difference between the medians for both parties is significant. The attribute Percent Hispanic or Latino has very little effect on the voting. The difference between the medians for both parties is not that much. The attribute Percent Foreign Born has very little effect on the voting. The difference between the medians for both parties is not that much.







The attribute **Percent Less than High School Degree** has very little effect on the voting. The difference between the medians for both parties is not that much. The attribute **Percent Less than Bachelor's Degree** has significant effect on the voting. The difference between the medians for both parties is significant. The

attribute **Percent more than Bachelor's Degree** has significant effect on the voting. The difference between the medians for both parties is significant.

- **Based on your previous analysis, 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 variables "Percent White, not Hispanic or Latino", "Percent Less than Bachelor's Degree" and "Percent more than Bachelor's Degree" are more important to determine whether a county is labeled as Democratic or Republican because of the following reasons:
  1) In the plot of "Percent White, not Hispanic or Latino", nearly 656 democratic counties have White (not Hispanic or Latino population) more than 75 percent . And the median of the White (not Hispanic or Latino population) percent for the democratic counties is greater by 12 percent.
  2) In the plot of "Percent Less than Bachelor's Degree", nearly 656 democratic counties have more than 78 percent of people with less than a bachelor's degree. And the median of the Percent Less than Bachelor's Degree for the democratic counties is greater by 10 percent.
  3) In the plot of "Percent More than Bachelor's Degree", nearly 244 republican counties have more than 27 percent of people with more than bachelor's degrees. And the median of the Percent More than Bachelor's Degree for the republican counties is greater by 10 percent.

- **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.**