CS 418: Project 1 Report

Authors: Anusha Sagi, Fatima Kahack, Lydia Tse

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

Note that the following is only a preview of the data. For a full table, please view the P1_Exploratory_Data_Analysis.pdf file.

	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

……..

[1205 rows x 6 columns]

Task 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.

Note that the following is only a preview of the data. For a full table, please view the P1_Exploratory_Data_Analysis.pdf file.

	Year	State	County	Office	Democratic	Republican
0	2018	arizona	apache	US Senator	16298.0	7810.0
1	2018	arizona	cochise	US Senator	17383.0	26929.0
2	2018	arizona	coconino	US Senator	34240.0	19249.0
3	2018	arizona	gila	US Senator	7643.0	12180.0
4	2018	arizona	graham	US Senator	3368.0	6870.0
5	2018	arizona	la paz	US Senator	1609.0	3265.0
6	2018	arizona	maricopa	US Senator	732671.0	672505.0
7	2018	arizona	mohave	US Senator	19214.0	50209.0
8	2018	arizona	navajo	US Senator	16624.0	18767.0
9	2018	arizona	pima	US Senator	221242.0	160550.0
10	2018	arizona	santa cruz	US Senator	9241.0	3828.0
11	2018	arizona	yavapai	US Senator	40160.0	65308.0
12	2018	connecticut	fairfield	US Senator	210899.0	131321.0
13	2018	connecticut	hartford	US Senator	203591.0	123864.0
14	2018	connecticut	middlesex	US Senator	42383.0	32836.0
15	2018	connecticut	new haven	US Senator	179714.0	126004.0
16	2018	connecticut	tolland	US Senator	34732.0	28046.0
17	2018	connecticut	windham	US Senator	20490.0	19032.0
18	2018	delaware	sussex	US Senator	40675.0	50391.0
19	2018	florida	alachua	US Senator	74493.0	40599.0
20	2018	florida	baker	US Senator	1945.0	8579.0
21	2018	florida	bay	US Senator	16723.0	46681.0
22	2018	florida	bradford	US Senator	2879.0	7576.0
23	2018	florida	brevard	US Senator	121112.0	160305.0
24	2018	florida	broward	US Senator	472239.0	211397.0
25	2018	florida	charlotte	US Senator	33525.0	52916.0
26	2018	florida	citrus	US Senator	22660.0	48008.0
27	2018	florida	collier	US Senator	54390.0	101266.0

…….

[1200 rows x 21 columns]

Task 3: 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?

Number of Variables
There were 1200 rows and 21 columns. So, there were 21 variables

Type of the Variables
float64, int64, object

Irrelevant & Redundant Attributes
Because the 'Year' and 'Office' attribute hold the same values across all of the observations, these attributes do not provide any meaningful insight into the data set. Thus, both of the columns can be dropped.

Task 4: Search the merged dataset for missing values. Are there any missing values? If so, how will you deal with these values?

Of the 1200 observations, 677 observations for the 'Citizen Voting-Age Population' have the value of 0. This means that over 50% of the observations for this attribute are missing. Thus, the 'Citizen Voting-Age Population' column was dropped.

Task 5: 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.
Note that the following is only a preview of the data. For a full table, please view the P1_Exploratory_Data_Analysis.pdf file.

	Percent Less than Bachelor's Degree	Percent Rural	Party
0	88.941063	74.061076	1
1	76.837055	36.301067	0
2	65.791439	31.466066	1
3	82.262624	41.062000	0
4	86.675944	46.437399	0
5	89.563407	56.327786	0
6	69.031137	2.363800	1
7	88.121178	22.963644	0
8	85.507970	54.138242	0
9	69.199391	7.523491	1
10	77.506775	26.883172	1

…….

[1200 rows x 19 columns]

Task 6: 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?

Computation of Means

Mean Population for	**Result**
Democratic Counties	300998.3169230769
Republican Counties	53864.6724137931

Therefore, the mean population for Democratic counties is greater than the mean population for Republican counties.

Hypothesis Testing

Null Hypothesis (H_0)	$\mu_{Democratic\ Population} = \mu_{Republican\ Population}$
Alternative Hypothesis (H_a)	$\mu_{Democratic\ Population} \neq \mu_{Republican\ Population}$
t-Test Statistic	8.004638577960957
p-Value	2.0478717602973023e-14
Conclusion	Since p-value < α, then there is statistical significance. Therefore, we reject the null hypothesis and there is sufficient evidence to conclude that the mean population for Democratic counties is different from the mean population for Republican counties.

Task 7: 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?

Computation of Means

Mean Median Household Income for	**Result**
Democratic Counties	53798.732307692306
Republican Counties	48746.81954022989

Therefore, the mean median household income for Democratic counties is greater than the mean population for Republican counties.

Hypothesis Testing

Null Hypothesis (H_0)	$\mu_{Democratic\ Median\ Household\ Income} = \mu_{Republican\ Median\ Household\ Income}$
Alternative Hypothesis (H_a)	$\mu_{Democratic\ Median\ Household\ Income} \neq \mu_{Republican\ Median\ Household\ Income}$
t-Test Statistic	5.479141589767387
p-Value	7.149437363182598e-08
Conclusion	Since p-value < α, then there is statistical significance. Therefore, we reject the null hypothesis and there is sufficient evidence to conclude that the mean median household income of Democratic counties is different from that of Republican counties.

Task 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?

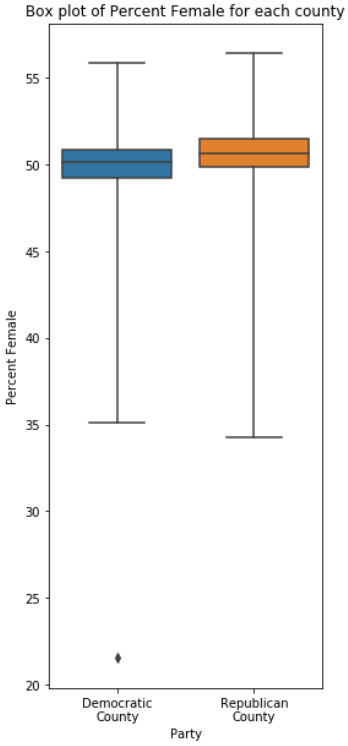
Descriptive Statistics for Democratic Counties

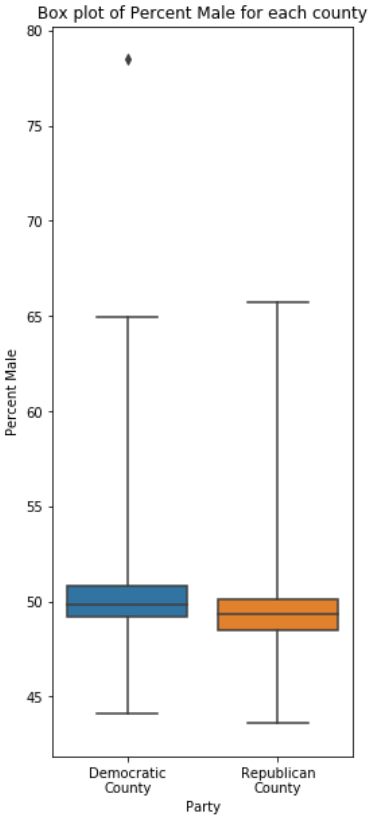
	count	mean	std	min	25%	50%	75%	max
Percent White, not Hispanic or Latino	325.0	69.683766	24.981502	2.776702	53.271579	77.786090	90.300749	98.063495
Percent Black, not Hispanic or Latino	325.0	9.242649	13.351340	0.000000	0.839103	3.485992	11.058843	63.953279
Percent Hispanic or Latino	325.0	12.587391	19.575030	0.193349	2.531017	5.039747	11.857116	95.479801
Percent Foreign Born	325.0	7.986330	8.330740	0.179769	2.470508	5.105490	10.144555	52.229868
Percent Female	325.0	50.385433	2.149359	34.245291	49.854280	50.653830	51.492075	56.418468
Percent Male	325.0	49.614567	2.149359	43.581532	48.507925	49.346170	50.145720	65.754709
Percent Age 29 and Under	325.0	38.726959	6.252786	23.156452	34.488444	38.074151	42.161162	67.367823
Percent Age 65 and Older	325.0	16.194826	4.282422	6.653188	13.106233	15.698087	18.806426	31.642106
Percent Less than High School Degree	325.0	11.883760	6.505613	3.215803	7.893714	10.370080	13.637059	49.673777
Percent Less than Bachelor's Degree	325.0	71.968225	11.192404	26.335440	65.711800	72.736143	79.903653	94.849957

Descriptive Statistics for Republican Counties

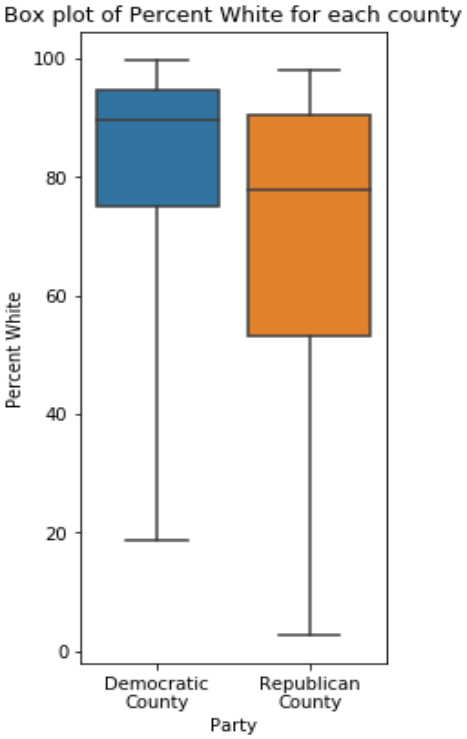
	count	mean	std	min	25%	50%	75%	max
Percent White, not Hispanic or Latino	870.0	82.656646	16.056122	18.758977	75.016397	89.434849	94.466596	99.627329
Percent Black, not Hispanic or Latino	870.0	4.189241	6.721695	0.000000	0.460419	1.318311	4.753831	41.563041
Percent Hispanic or Latino	870.0	9.733094	14.049576	0.000000	1.704539	3.427435	10.709696	78.397012
Percent Foreign Born	870.0	3.990096	4.507786	0.000000	1.320101	2.326317	5.149429	37.058317
Percent Female	870.0	49.630898	2.429013	21.513413	49.222905	50.176792	50.829770	55.885023
Percent Male	870.0	50.369102	2.429013	44.114977	49.170230	49.823208	50.777095	78.486587
Percent Age 29 and Under	870.0	36.005719	5.181522	11.842105	32.983652	35.846532	38.539787	58.749116
Percent Age 65 and Older	870.0	18.828267	4.733155	6.954387	15.784982	18.377896	21.112847	37.622759
Percent Less than High School Degree	870.0	14.009112	6.303126	2.134454	9.662491	12.572435	17.447168	47.812773
Percent Less than Bachelor's Degree	870.0	81.095427	6.815537	43.419470	78.108424	82.406700	85.546272	97.014925

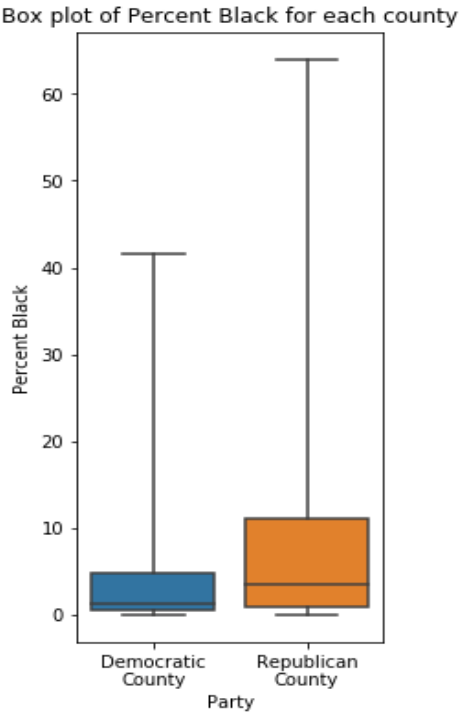
Plots for Gender

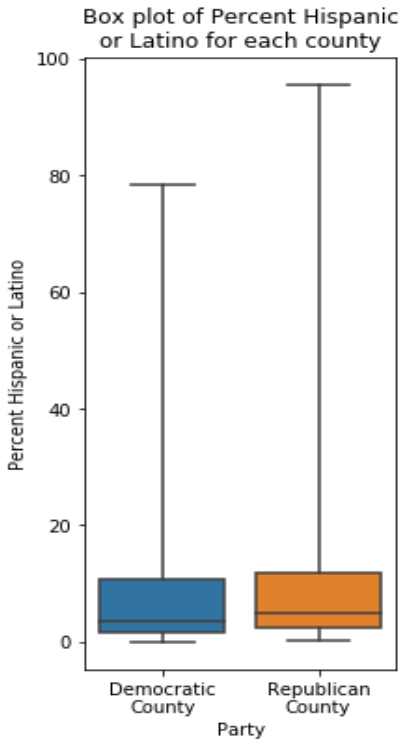


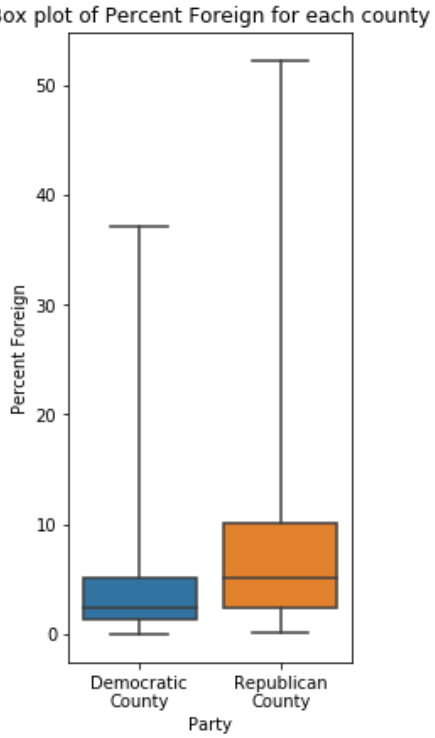


Plots for Ethnicity and Race

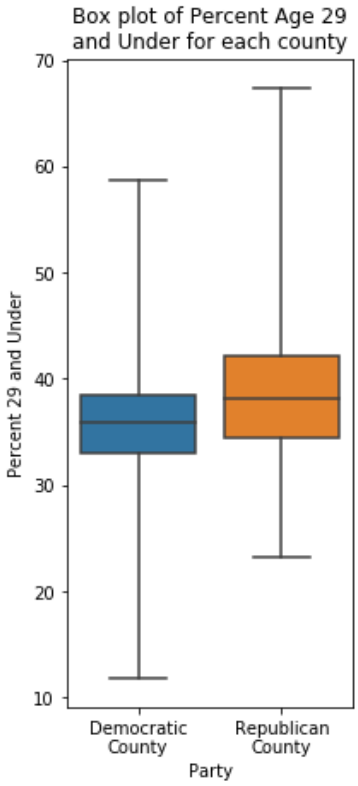


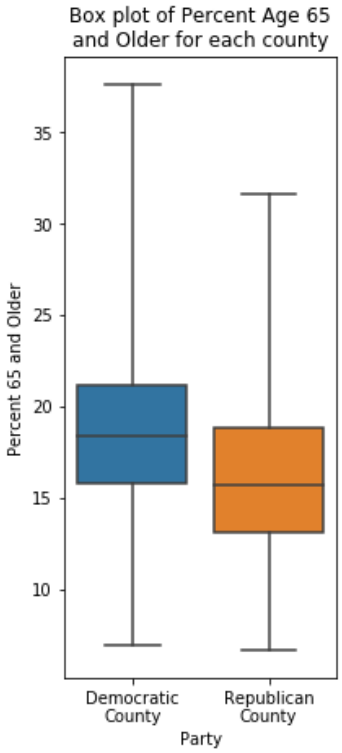




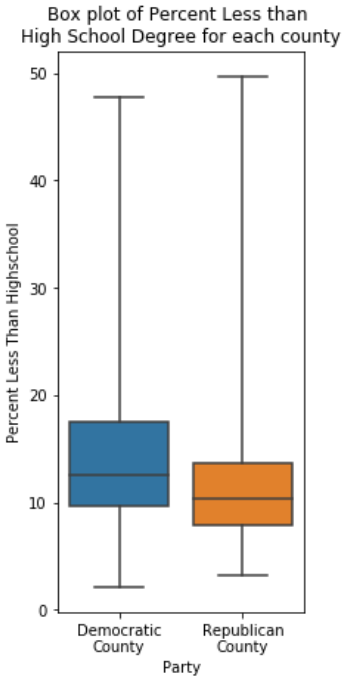


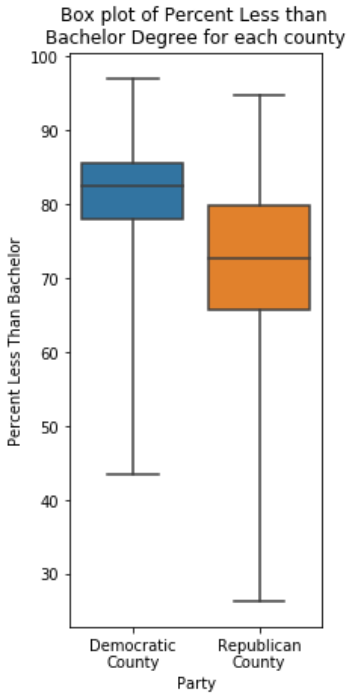
Plots for Age





Plots for Education





Task 8 Conclusions

Race and Ethnicity
The dominant race for both parties is white.

Gender
For both parties the number of males and females are equal.

Age
For both parties, younger people tend to be more politically active than older people.

Education
People with more education tend to be involved with both parties more than people with less education.

Task 9: 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 that help determine whether a county is marked Democratic or Republican are Race and Ethnicity and Education. There is a significant difference in race and ethnicity between the two counties. According to our analysis 30% of Democratic counties are non-white and only 18% of Republican counties are non-white. The white population fills 70% of Democratic counties and 83% of Republican counties. In education, 72% of the population in Democratic counties hold a degree less than a Bachelor's degree and 81% of the population in Republican counties hold a degree less than a Bachelor's degree. As for the other variables, the percentages differ by 1-4%, which is significantly small.

Task 10: 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.