CS 418: Introduction to Data Science
Project 01: Exploratory Data Analysis
Fall 2019

Project Report

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.

Answer : To reshape the dataset from long to wide, we use pandas.pivot_table. Since the pivot_table resets the missing values (NaN) to 0, we reset them to NaN.

```
# convert from long format to wide format using pivot_table
data_election_tidy = pd.pivot_table(data_election, values='Votes', index=['Year', 'State', 'County', 'Office'],
                                    columns='Party', aggfunc = np.sum).reset_index()

# reset the missing values which were set to 0 by the above pivot table
data_election_tidy.loc[data_election_tidy['Democratic'] == 0,'Democratic'] = np.nan
data_election_tidy.loc[data_election_tidy['Republican'] == 0, 'Republican'] = np.nan
print(data_election_tidy.head(10))
data_election_tidy.shape
```

```
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

(1205, 6)
```

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.

Answer :

To address the inconsistencies in state names in election_train dataset, we replace the short forms of the states with full names by using dataframe.replace.

To address the inconsistencies in county names in election_train dataset, we remove the 'county' word from all the values in County column using str.replace.

To address the inconsistencies with lower and uppercase letters in counties and states in both election_train and demographics_train dataset, we are using str.title() to follow the same format "Harrisonburg City" (first letter of each word is a capital case and remaining letters are small case).

```
# address the inconsistencies in the states and counties
data_election_tidy = data_election_tidy.replace({'AZ': 'Arizona', 'CT': 'Connecticut', 'DE': 'Delaware', 'FL' : 'Floric
'HI': 'Hawaii', 'IN':'Indiana', 'ME': 'Maine', 'MD': 'Maryland', 'MA':'Massachusetts', 'MI': 'Michigan', 'MN':'Minnesot
'MT':'Montana','NE':'Nebraska', 'NV':'Nevada', 'NJ':'New Jersey', 'NM':'New Mexico', 'NY':'New York', 'ND':'North Dakot
'OH':'Ohio','PA':'Pennsylvania', 'RI':'Rhode Island', 'TN':'Tennessee','TX':'Texas', 'UT':'Utah', 'VT':'Vermont',
'VA':'Virginia','WA':'Washington', 'WV':'West Virginia','WI':'Wisconsin', 'WY':'Wyoming'})
data_election_tidy['County'] = data_election_tidy['County'].str.replace("County", "").str.strip()
data_election_tidy['State'] = data_election_tidy['State'].str.title()
data_election_tidy['County'] = data_election_tidy['County'].str.title()
data_demographics['State'] = data_demographics['State'].str.title()
data_demographics['County'] = data_demographics['County'].str.title()
data_election_tidy.head(10)

# merge data_demographics and data_election_wide
data_merged = pd.merge(data_demographics, data_election_tidy, how = 'inner', on = ['State','County'], sort = True)
print(data_merged.shape)
data_merged.head(3)
```

```
(1200, 21)
```

	State	County	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 65 and Older	Median Household Income	Percent Unemployed	Percent Less than High School Degree	I
0	Arizona	Apache	4001	72346	0	18.571863	0.486551	5.947806	1.719515	50.598513	...	13.322091	32460	15.807433	21.758252	
1	Arizona	Cochise	4003	128177	92915	56.299492	3.714395	34.403208	11.458374	49.069646	...	19.756275	45383	8.567108	13.409171	
2	Arizona	Coconino	4005	138064	104265	54.619597	1.342855	13.711033	4.825298	50.581614	...	10.873943	51106	8.238305	11.085381	

Task 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?*

Answer : The dataset has 21 variables. Year and Office have the same value for all the rows. '2018' for Year and 'US Senator' for Office. So, they are not useful for the data analysis. So, we can drop them using dataframe.drop().

```
data_merged.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1200 entries, 0 to 1199
Data columns (total 21 columns):
State                                1200 non-null object
County                               1200 non-null object
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
Year                                 1200 non-null int64
Office                               1200 non-null object
Democratic                           1195 non-null float64
Republican                           1195 non-null float64
dtypes: float64(13), int64(5), object(3)
memory usage: 206.2+ KB
```

```
data_merged.Year.unique()
```

```
array([2018])
```

```
data_merged.Office.unique()
```

```
array(['US Senator'], dtype=object)
```

```
data_merged = data_merged.drop(columns=['Year', 'Office'])
data_merged.head()
```

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	Democratic	Republican
18.571863	0.486551	5.947806	1.719515	50.598513	45.854643	13.322091	32460	15.807433	21.758252	88.941063	74.061076	16298.0	7810.0
56.299492	3.714395	34.403208	11.458374	49.069646	37.902276	19.756275	45383	8.567108	13.409171	76.837055	36.301067	17383.0	26929.0
54.619597	1.342855	13.711033	4.825298	50.581614	48.946141	10.873943	51106	8.238305	11.085381	65.791439	31.466066	34240.0	19249.0
63.222325	0.552850	18.548675	4.249798	50.296170	32.238290	26.397638	40593	12.129932	15.729958	82.262624	41.062000	7643.0	12180.0
51.461536	1.811932	32.097844	4.385942	46.313518	46.393456	12.315809	47422	14.424104	14.580797	86.675944	46.437399	3368.0	6870.0

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

Answer : We search the merged dataset for missing values using dataframe.isnull().

Yes, there are missing values in Democratic and Republic columns.We drop the corresponding rows using dataframe.dropna(), to avoid any impact on the summary statistics(like mean). If we fill the missing values with 0, the summary statistics(like mean) would be deviated.

Though the columns like Citizen Voting-Age Population, Percent Black, not Hispanic or Latino, Percent Hispanic or Latino, Percent Foreign Born, Percent Unemployed and Percent Rural have zero values, since they are the percentages, 0 is considered to be a valid value.

```
data_merged[data_merged.isnull().any(axis=1)]
```

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	Democratic	Republican
82.659667	3.783472	6.531834	7.523856	49.891782	45.457016	12.175057	53730	4.372984	6.636272	62.697684	8.253126	NaN	NaN
94.713656	1.330058	1.465605	1.067435	50.626906	33.361572	19.662826	35209	12.544170	21.141176	91.176471	100.000000	NaN	NaN
32.660674	7.989360	57.909864	4.344769	39.579894	40.448236	11.514707	44005	6.065048	28.602944	91.094503	42.977308	NaN	NaN
56.310680	1.248266	39.389736	4.299584	46.833102	36.430883	22.468793	37917	8.360836	21.113990	84.909326	100.000000	NaN	NaN
94.771631	0.339427	3.507414	2.286667	49.199071	37.688323	16.709343	53038	2.998885	10.123457	82.336861	100.000000	NaN	NaN

```
data_merged = data_merged.dropna(how = 'any')
data_merged[data_merged.isnull().any(axis=1)]
```

State	County	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

Task 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.

Answer : A new column 'Party' is created whose value is 1 if Democratic column has more votes than Republican column.

```
data_merged['Party'] = np.where(data_merged['Democratic'] > data_merged['Republican'], 1, 0)
data_merged.head(10)
```

ınt te, ıot ıic no	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	Democratic	Republican	Party
63	0.486551	5.947806	1.719515	50.598513	45.854643	13.322091	32460	15.807433	21.758252	88.941063	74.061076	16298.0	7810.0	1
92	3.714395	34.403208	11.458374	49.069646	37.902276	19.756275	45383	8.567108	13.409171	76.837055	36.301067	17383.0	26929.0	0
97	1.342855	13.711033	4.825298	50.581614	48.946141	10.873943	51106	8.238305	11.085381	65.791439	31.466066	34240.0	19249.0	1
25	0.552850	18.548675	4.249798	50.296170	32.238290	26.397638	40593	12.129932	15.729958	82.262624	41.062000	7643.0	12180.0	0
36	1.811932	32.097844	4.385942	46.313518	46.393456	12.315809	47422	14.424104	14.580797	86.675944	46.437399	3368.0	6870.0	0
49	0.379236	26.182033	11.372143	48.946020	28.073286	36.056935	36321	10.599013	24.842215	89.563407	56.327786	1609.0	3265.0	0
14	5.013612	30.286833	14.729333	50.549278	41.886620	13.837843	55676	6.808454	13.051927	69.031137	2.363800	732671.0	672505.0	1
06	0.951731	15.708470	6.969047	49.676618	30.485835	26.858650	39856	11.680953	16.145850	88.121178	22.963644	19214.0	50209.0	0
96	0.672772	11.049913	2.914730	49.846131	43.243168	15.745456	36868	18.525791	18.494087	85.507970	54.138242	16624.0	18767.0	0
79	3.199719	36.105978	12.903428	50.807405	40.087388	17.801778	46764	9.214114	12.252238	69.199391	7.523491	221242.0	160550.0	1

Task 6:(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?

```
#get the mean based on the party and then use loc to get the row and column that correspond to the democratic and repul
repMean = data_merged.groupby(['Party']).mean().loc[0:1,'Total Population'][0]
demMean = data_merged.groupby(['Party']).mean().loc[0:1,'Total Population'][1]

#print('Republican Total Population Mean',repMean)
#print('Demoractic Total Population Mean',demMean)

#isolating data for democratic and republican total population
dem = data_merged[data_merged['Party']==1].loc[:,'Total Population']
rep = data_merged[data_merged['Party']==0].loc[:,'Total Population']

print('Democratic Total Population Mean:',demMean)
print('Republican Total Population Mean:',repMean)

#compute unpaired t test
import scipy.stats as st
[statistic, pvalue] = st.ttest_ind(dem , rep, equal_var = False)
print('T test statistic:',statistic)
print('Pvalue:',pvalue)
```

```
Democratic Total Population Mean: 300998.3169230769
Republican Total Population Mean: 53864.6724137931
T test statistic: 8.004638577960957
Pvalue: 2.0478717602973023e-14
```

Answer: The mean for democrat total population is **300988.32** and the mean for the republican total population is **53864.67**. The democratic mean total population is higher than the republican. To perform a hypothesis test an unpaired t test was performed in which the resulting statistic is **8.00**. The p value for that statistic is **2.047e-14.** Given that the significance level is .05 and p value < .05 we reject the null

hypothesis which means it is not statistically significant. The null hypothesis is that the total population mean for democrats and republicans is equal (M1 = M2) and the alternative hypothesis is that the total population mean for democrats and republicans is not equal (M1 != M2).

Task 7:(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?

```
repIncome = data_merged.groupby(['Party']).mean().loc[0:1,'Median Household Income'][0]
demIncome = data_merged.groupby(['Party']).mean().loc[0:1,'Median Household Income'][1]
print('Democrats Mean Household Income:' ,demIncome)
print('Republican Mean Household Income:',repIncome)

#isolating data for democratic and republican median household income
demIncome = data_merged[data_merged['Party']==1].loc[:,'Median Household Income']

repIncome = data_merged[data_merged['Party']==0].loc[:,'Median Household Income']

#compute unpaired t test
import scipy.stats as st
[statistic, pvalue] = st.ttest_ind(demIncome , repIncome, equal_var = False)

print('Test stastic:',statistic)
print('Pvalue:',pvalue)
```

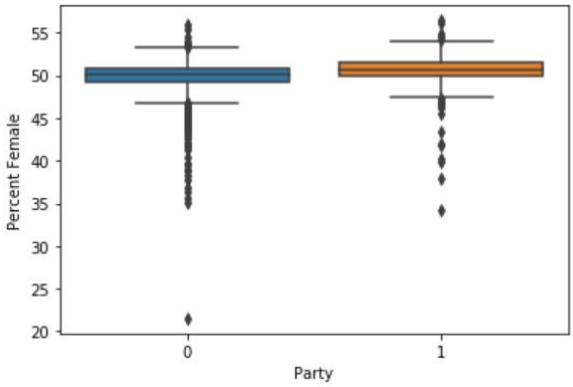
```
Democrats Mean Household Income: 53798.732307692306
Republican Mean Household Income: 48746.81954022989
Test stastic: 5.479141589767387
Pvalue: 7.149437363182572e-08
```

Answer : The mean household income for democrats is **53798.732** and the mean household income for republicans is **48746.81**. Democrats had the higher mean household income. To perform a hypothesis test an unpaired t test was performed which resulted in a test statistic of **5.45**. The p value is **7.14e-08**. The p value < .05 so we reject the null hypothesis which means it is not statistically significant. The null hypothesis is that the mean household income for republicans and democrats is equal (M1 = M2) and the alternative hypothesis is that the mean household income for republicans and democrats is not equal (M1 != M2).

Task 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?

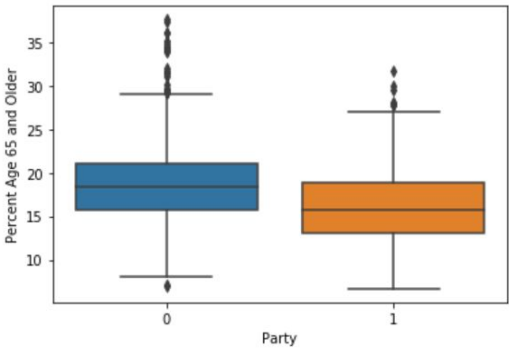
Answer: *descriptive statistics is in an excel file (included in Project1.zip)

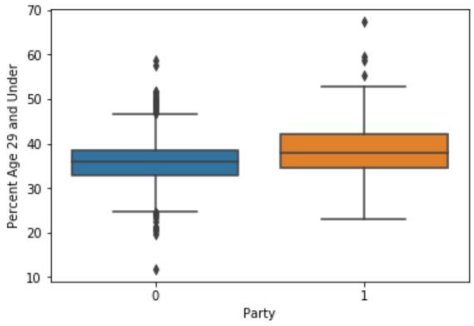
<u>Gender:</u>



To visualize the difference in gender between the democratic and republican party a box plot was used. In the plot above the median for democratic and republican party for gender is very close but the republican party has a smaller percentage of females and has many more outliers than the democratic. It can be seen that the shift to a lower median for the republican party is due to more observation below the minimum

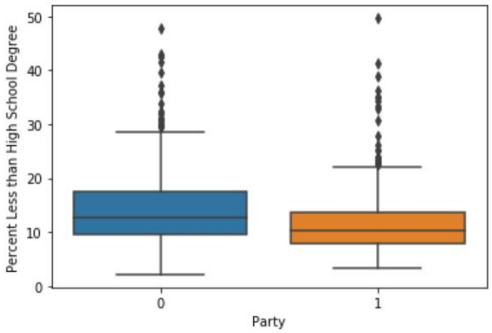
Age:

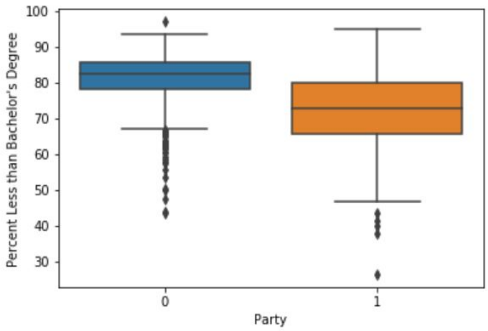




In the above plots the republican party has a higher median for percent of people 65 and older and contains more outliers above the maximum and in the plot for percentage of people 29 years and under is less for republicans. Meanwhile for percentage of people 29 and under the median for democratic party is highers with outliers above the maximum percentage. This illustrates that there is a greater percentage of people 65 and older that are republican than democratic and greater percentage of people 29 and under that are demoractic.

Education:

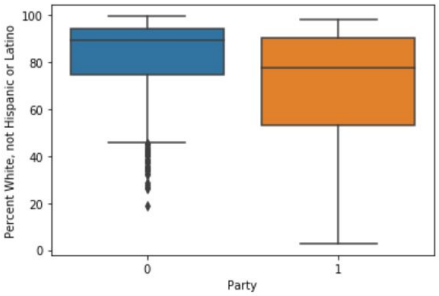


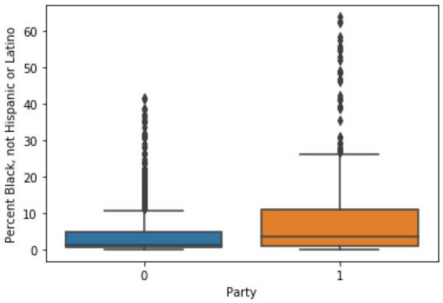


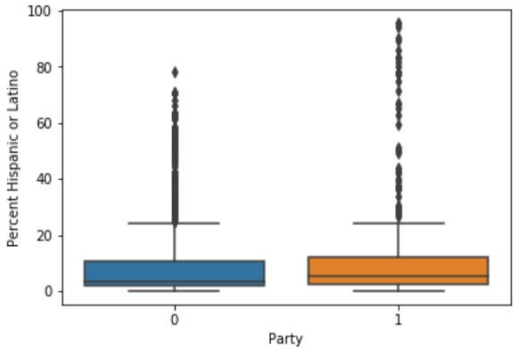
The republican party has a greater percent of people that have less than a bachelor's degree and the lower percentages are outliers. The democratic party has less percentage of people that have less than a bachelor's degree, however, the minimum percentage and outliers are much lower than the republican party.

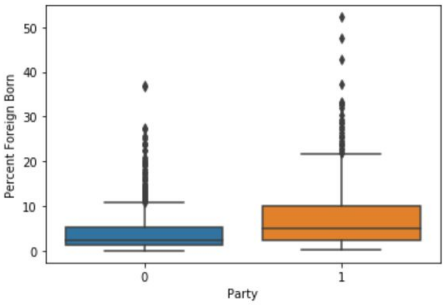
The democratic party has a lower percentage of people with less than a high school degree than republican party as noted by the median but both parties have outliers that lie above the maximum percentage.

Race and Ethnicity:









The percentage of white voters is the highest in the republican party while all the other demographics has a higher median percentage for the democratic party. There are also considerably more outliers for democratic party for non white voters. For white and hispanic or latino voter the maximum values are close, however, there is a big difference in the minimum values for white voters among the two parties. More white voters tend to be republican than the other demographics.

Task 9 : (5 pts.) 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.

Answer : Gender, race/ethnicity, and age are important in determining whether a county is democratic or republican. In the box plot above it can be seen that there is a higher percentage of females that are democratic, higher percentage of white voters tend to be republican while the other demographics have higher percentage for democratic voters. Lastly, there is a higher percentage of younger voters that are democratic and a higher percentage of older voters are republican. The analysis and justification can be found in task 8.

Task 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.

Answer : We get FIPS from FIPS column of the merged dataset and values from the Party column.

We then pass the FIPS and values to plotly.figure_factory.create_choropleth() to create a map with Democratic and Republican counties. The blue ones represent democratic counties(Party = 1) and red ones represent republican counties(Party = 0).

```
fips = data_merged['FIPS'].tolist()
values = data_merged['Party'].tolist()
colorscale = ['rgb(244,109,67)', 'rgb(49,54,149)']
fig = ff.create_choropleth(
    colorscale=colorscale,
    fips=fips, values=values,
    title='Counties by Democratic/Republican',
    legend_title='1 = Democratic Counties, 0 = Republican Counties'
)
fig.layout.template = None
fig.show(sort=True)
```