Senate Elections 2018: Exploratory Data Analysis
Project Report

1. The pivot_table() is used for converting long dataset format to wide dataset format. Variables 'Year', 'State', 'County', and 'Office' is used for the index. Then we used 'Party' for the columns and 'Votes' for the values. Before this transformation we have used fillna(0) to fill the missing values in the dataset.

```
In [10]: # Question 1
data_b_tidy = pd.pivot_table(fill_with_zero, index=['Year', 'State','County', 'Office'],columns='Party', values="Votes").reset_i
print(data_b_tidy.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1205 entries, 0 to 1204
Data columns (total 6 columns):
Year          1205 non-null int64
State         1205 non-null object
County        1205 non-null object
Office        1205 non-null object
Democratic    1205 non-null float64
Republican    1205 non-null float64
dtypes: float64(2), int64(1), object(3)
memory usage: 56.6+ KB
None
```

2. The inconsistencies of the data was from the 'States' and the 'County' column. The election_train.csv has their States abbreviated while the demographics.csv had the states with their full name. To deal with that, we made a custom changed_value_state variable that basically kept the 'State' column consistent by having it keep its full name. The 'County' column was changed by removing the word County and storing it in our variable data_b_tidy. After that we uppercased all the elements in the 'County' column because some of the counties had an uppercase in Of when the counties in the other csv did not. After dealing with the inconsistencies we finally merged the data with an inner join on the 'State' and 'County' columns.

```
In [14]: # Merge dataset A and dataset B
merged_data_set = pd.merge(data_a,data_b_tidy, how='inner', on=['State', 'County'])
print(merged_data_set.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
```

3. The dataset has 21 variables. The variables in the dataset has types object or int64 or float64. The irrelevant variables are Year and Office because year contains all the observations from the same year (2018). Also, Office contains the same values -> US Senator. These variables did not provide useful information so we dropped the columns from our dataset.

```
#Question 3
# The Dataset has 21 variables
# The variables in this dataset has types object or int64 or float64
# The irrelevant irrelevant are Year because all the observations are from the same year 2018, Office which contains same value
# We decided to drop these variables since they do not provide a any useful information for the current analysis.
merged_data_set = merged_data_set.drop(columns=['Year', 'Office'])
merged_data_set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1200 entries, 0 to 1199
Data columns (total 19 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
Democratic                           1200 non-null float64
Republican                           1200 non-null float64
dtypes: float64(13), int64(4), object(2)
memory usage: 187.5+ KB
```

4. There are missing values in the following variables in the dataset (Citizen Voting-Age Population, Democratic, Republican). Citizen Voting-Age Population has many 0 values in the variable according our inference the zero signifies missing value. So, we decided drop the column since 680 of 1200 observations have 0 in it (more than 50 % of observations). There are five observations in the dataset which have both Republican and Democratic variables zero (missing observation). These observations has been deleted from the dataset since we cannot decide the party of the county based on missing values.

```
#Question 4
#There are missing values in the following variables in the datase.(Citizen Voting-Age Population,Democratic,Republican)
#Citizen Voting-Age Population has many 0 values in the variable according to this column the zero signifies missing value.
#So we decided drop the column since 680 of 1200 observations have 0 in it.(more than 50 % of observations)
#There are five observations in the dataset which have both republican and Democratic variables both zero(filled with zero for m
# These observations will be deleted from the dataset since we cannot decide the party of the county based on missing values.
merged_data_set = merged_data_set.drop(columns=['Citizen Voting-Age Population'])
merged_data_set = merged_data_set.drop(merged_data_set[(merged_data_set.Democratic == 0) & (merged_data_set.Republican == 0)].in
merged_data_set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1195 entries, 0 to 1199
Data columns (total 18 columns):
State                                1195 non-null object
County                               1195 non-null object
FIPS                                 1195 non-null int64
Total Population                     1195 non-null int64
Percent White, not Hispanic or Latino    1195 non-null float64
Percent Black, not Hispanic or Latino    1195 non-null float64
Percent Hispanic or Latino           1195 non-null float64
Percent Foreign Born                 1195 non-null float64
Percent Female                       1195 non-null float64
Percent Age 29 and Under             1195 non-null float64
Percent Age 65 and Older             1195 non-null float64
Median Household Income              1195 non-null int64
Percent Unemployed                   1195 non-null float64
Percent Less than High School Degree     1195 non-null float64
Percent Less than Bachelor's Degree      1195 non-null float64
Percent Rural                        1195 non-null float64
Democratic                           1195 non-null float64
Republican                           1195 non-null float64
dtypes: float64(13), int64(3), object(2)
memory usage: 177.4+ KB
```

5. Created a new variable “Party” which signifies if the county is Democratic or Republican. The value of new variable value is 1 (if Votes cast Democratic > Votes Cast Republican) else 0.

```
# Question 5
# Created a new column called 'Party' dependending on which party has more votes
merged_data_set['Party'] = np.where(merged_data_set['Democratic'] > merged_data_set['Republican'],1,0)
```

```
merged_data_set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1195 entries, 0 to 1199
Data columns (total 19 columns):
State                                1195 non-null object
County                               1195 non-null object
FIPS                                 1195 non-null int64
Total Population                     1195 non-null int64
Percent White, not Hispanic or Latino    1195 non-null float64
Percent Black, not Hispanic or Latino    1195 non-null float64
Percent Hispanic or Latino           1195 non-null float64
Percent Foreign Born                 1195 non-null float64
Percent Female                       1195 non-null float64
Percent Age 29 and Under             1195 non-null float64
Percent Age 65 and Older             1195 non-null float64
Median Household Income              1195 non-null int64
Percent Unemployed                   1195 non-null float64
Percent Less than High School Degree     1195 non-null float64
Percent Less than Bachelor's Degree      1195 non-null float64
Percent Rural                        1195 non-null float64
Democratic                           1195 non-null float64
Republican                           1195 non-null float64
Party                                1195 non-null int32
dtypes: float64(13), int32(1), int64(3), object(2)
memory usage: 182.1+ KB
```

6. The mean population of Democratic counties is 300998.3169230769. The mean population of Republican counties is 53864.6724137931. The mean population of Democratic counties is higher.

   $\boldsymbol{\mu 1}=$ mean population of Democratic counties

   $\boldsymbol{\mu 2}=$ mean population of Republican counties

   $\boldsymbol{H0}:\boldsymbol{\mu 1}=\boldsymbol{\mu 2}$

   $\boldsymbol{Ha}:\boldsymbol{\mu 1}\neq\boldsymbol{\mu 2}$ (so it's a two tailed test)

   t-test statistic 8.004638577960957

   pvalue 2.0478717602973023e-14

   Since pvalue 2.0478717602973023e-14 less than $\boldsymbol{\alpha}$=0.05

   We reject the null hypothesis.

```
#Question 6
#data_population_mean=merged_data_set.groupby('Party')['Total Population'].mean()
republican_population_mean=merged_data_set[merged_data_set.Party == 0]['Total Population'].mean()
democratic_population_mean=merged_data_set[merged_data_set.Party == 1]['Total Population'].mean()
print("Mean population for Democratic counties "+str(democratic_population_mean))
print("Mean population for Republican counties "+str(republican_population_mean))
print("The mean population of Democratic counties is higher")
```

```
Mean population for Democratic counties 300998.3169230769
Mean population for Republican counties 53864.6724137931
The mean population of Democratic counties is higher
```

```
[statistic, pvalue] = st.ttest_ind(merged_data_set[merged_data_set.Party == 1]['Total Population'], merged_data_set[merged_data_
print("t-test statistic "+str(statistic))
print("pvalue "+str(pvalue))
print("Since pvalue "+str(pvalue)+" less than "+"α=0.05")
print("We reject the null hypothesis")
```

```
t-test statistic 8.004638577960957
pvalue 2.0478717602973023e-14
Since pvalue 2.0478717602973023e-14 less than α=0.05
We reject the null hypothesis
```

7. The Mean Median Household Income for Democratic counties 53798.732307692306. The Mean Median Household Income for Republican counties 48746.81954022989. The Mean Median Household Income of Democratic counties is higher

   $\boldsymbol{\mu 1}=$ Mean Median Household Income of Democratic counties

   $\boldsymbol{\mu 2}=$ Mean Median Household Income of Republican counties

   $\boldsymbol{H0}:\boldsymbol{\mu 1}=\boldsymbol{\mu 2}$

   $\boldsymbol{Ha}:\boldsymbol{\mu 1}\neq\boldsymbol{\mu 2}$ (so it's a two tailed test)

   t-test statistic= 5.479141589767388 pvalue=7.149437363182572e-08

   Since pvalue 7.149437363182572e-08 less than $\boldsymbol{\alpha}$=0.05

   We reject the null hypothesis

```
#Question 7 Median Household Income(mean)
republican_mhi_mean=merged_data_set[merged_data_set.Party == 0]['Median Household Income'].mean()
democratic_mhi_mean=merged_data_set[merged_data_set.Party == 1]['Median Household Income'].mean()
print("Mean Median Household Income for Democratic counties "+str(democratic_mhi_mean))
print("Mean Median Household Income for Republican counties "+str(republican_mhi_mean))
print("The Mean Median Household Income of Democratic counties is higher")
```

```
Mean Median Household Income for Democratic counties 53798.732307692306
Mean Median Household Income for Republican counties 48746.81954022989
The Mean Median Household Income of Democratic counties is higher
```

```
[statistic_mhi, pvalue_mhi] = st.ttest_ind(merged_data_set[merged_data_set.Party == 1]['Median Household Income'], merged_data_s
print("t-test statistic "+str(statistic_mhi))
print("pvalue "+str(pvalue_mhi))
print("Since pvalue "+str(pvalue_mhi)+" less than "+"α=0.05")
print("We reject the null hypothesis")
```

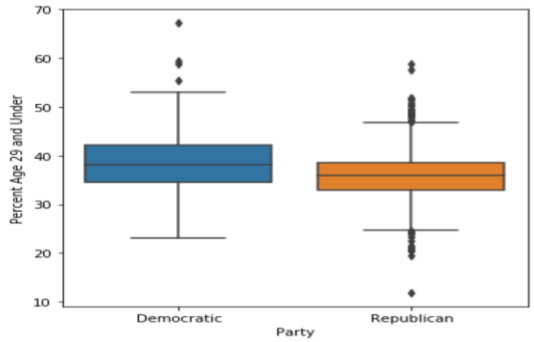
```
t-test statistic 5.479141589767388
pvalue 7.149437363182572e-08
Since pvalue 7.149437363182572e-08 less than α=0.05
We reject the null hypothesis
```

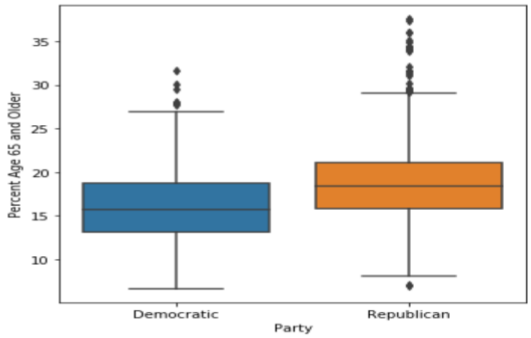
8. The Age category has two variables in the dataset ("Percent Age 29 and Under" and "Percent Age 65 and Older"). According to the statistics and the box plot we can determine that "Percent Age 29 and Under" more percentage prefers Democratic Party little more than the Republican Party in their counties. But more percentage prefers the Republican than Democratic in category "Percent Age 65 and Older" in their counties.

Descriptive Statistics:

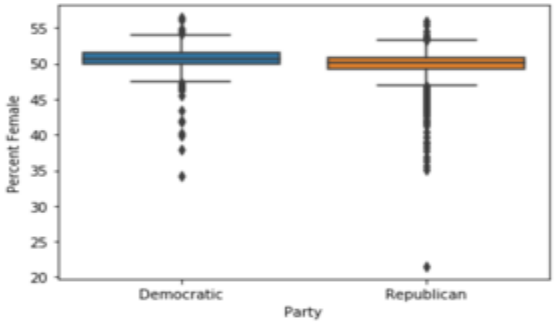
Percent Age 29 and Under	**count**	870.000000	325.000000
	mean	36.005719	38.726959
	std	5.181522	6.252786
	min	11.842105	23.156452
	25%	32.983652	34.488444
	50%	35.846532	38.074151
	75%	38.539787	42.161162
	max	58.749116	67.367823

Percent Age 65 and Older	**count**	870.000000	325.000000
	mean	18.828267	16.194826
	std	4.733155	4.282422
	min	6.954387	6.653188
	25%	15.784982	13.106233
	50%	18.377896	15.698087
	75%	21.112847	18.806426
	max	37.622759	31.642106





Gender: "Percent Female" variable dataset have same percentage of population voting for Democratic and Republican party in their won counties.



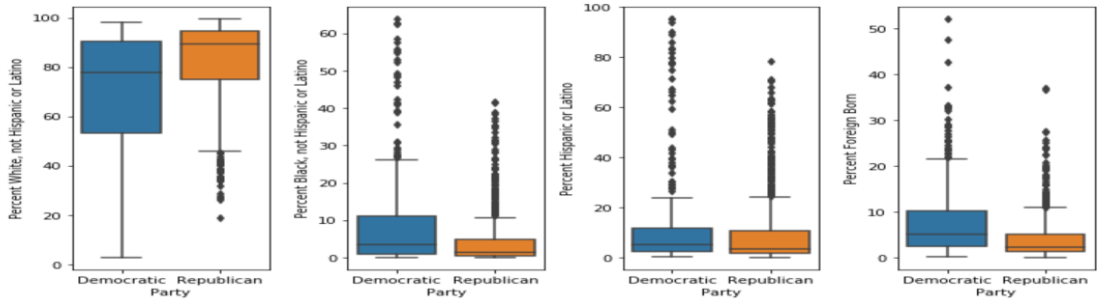
Percent Female	**count**	870.000000	325.000000
	mean	49.630898	50.385433
	std	2.429013	2.149359
	min	21.513413	34.245291
	25%	49.222905	49.854280
	50%	50.176792	50.653830
	75%	50.829770	51.492075
	max	55.885023	56.418468

Race and ethnicity: “Percent White, not Hispanic or Latino” this variable has higher percentage value for mean, median and higher percentage values in quartile 3 and 4 for republican counties than democratic counties. This says in this category more people prefer Republican Party in their county. “Percent Black, not Hispanic or Latino”, “Percent Hispanic or Latino” and “Percent Foreign Born” higher percentage values prefer Democratic Party in their county where the party has come.

Percent White, not Hispanic or Latino		
count	870.000000	325.000000
mean	82.656646	69.683766
std	16.056122	24.981502
min	18.758977	2.776702
25%	75.016397	53.271579
50%	89.434849	77.786090
...	...	...

Percent Hispanic or Latino		
std	14.049576	19.575030
min	0.000000	0.193349
25%	1.704539	2.531017
50%	3.427435	5.039747
75%	10.709696	11.857116
max	78.397012	95.479801

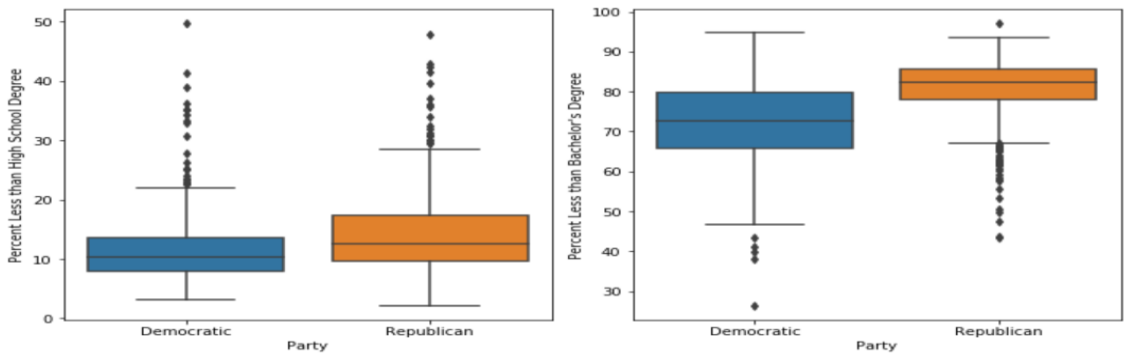
Percent Foreign Born		
count	870.000000	325.000000
mean	3.990096	7.986330
std	4.507786	8.330740
min	0.000000	0.179769
25%	1.320101	2.470508
50%	2.326317	5.105490
75%	5.149429	10.144555
max	37.058317	52.229868



Education: The variable “Percent Less than High School Degree” has little higher percentage values in their statistics for Republican Party than Democratic Party counties and for “Percent Less than Bachelor's Degree” has significant difference in higher values for Republican Party counties than Democratic Party Counties.

Percent Less than High School Degree		
count	870.000000	325.000000
mean	14.009112	11.883760
std	6.303126	6.505613
min	2.134454	3.215803
25%	9.662491	7.893714
50%	12.572435	10.370080
75%	17.447168	13.637059
max	47.812773	49.673777

Percent Less than Bachelor's Degree		
count	870.000000	325.000000
mean	81.095427	71.968225
std	6.815537	11.192404
min	43.419470	26.335440
25%	78.108424	65.711800
50%	82.406700	72.736143
75%	85.546272	79.903653
max	97.014925	94.849957



9. Based on the previous statistics we can see that the number of counties where Republican Party have more votes (870) is higher than counties where Democratic Party have more votes (325).The important variables to determine that the county is labeled as Democratic or Republican are “Percent White, not Hispanic or Latino”, “Percent Less than Bachelor's Degree” and “Percent Less than High School Degree” because we can see significant variation in the values of the descriptive statistics (like median, mean and higher percent values in higher quartiles). The Republican Counties have higher values (Percentage values) when compared to Democratic Counties.

10. The map contains the counties in which the Democratic Party (Blue color) have majority votes and Republican Party (Red color) have majority votes. From the map we can see that majority of the counties have red color associated with them.