CS418 Project 1

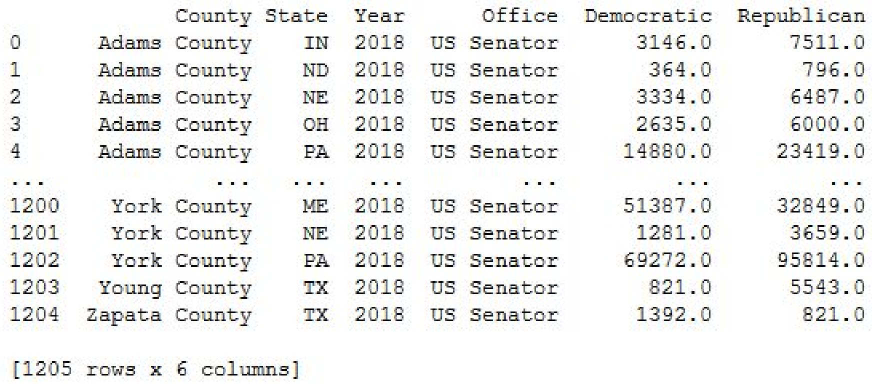
Exploratory Data Analysis

Frank Errichiello, Tomasz Hulka, Ankit Singh

Overview

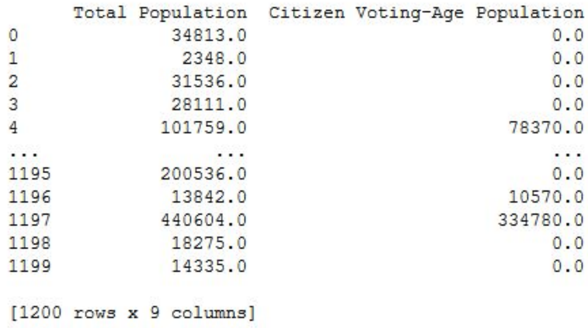
The purpose of this report is to explore datasets of election data and demographics for a large sample of US counties and find if there is any correlation.

Task 1



```
                County State  Year      Office  Democratic  Republican
0      Adams County    IN  2018  US Senator      3146.0      7511.0
1      Adams County    ND  2018  US Senator       364.0       796.0
2      Adams County    NE  2018  US Senator      3334.0      6487.0
3      Adams County    OH  2018  US Senator      2635.0      6000.0
4      Adams County    PA  2018  US Senator     14880.0     23419.0
...             ...   ...   ...         ...         ...         ...
1200    York County    ME  2018  US Senator     51387.0     32849.0
1201    York County    NE  2018  US Senator      1281.0      3659.0
1202    York County    PA  2018  US Senator     69272.0     95814.0
1203   Young County    TX  2018  US Senator       821.0      5543.0
1204  Zapata County    TX  2018  US Senator      1392.0       821.0

[1205 rows x 6 columns]
```

Task 2



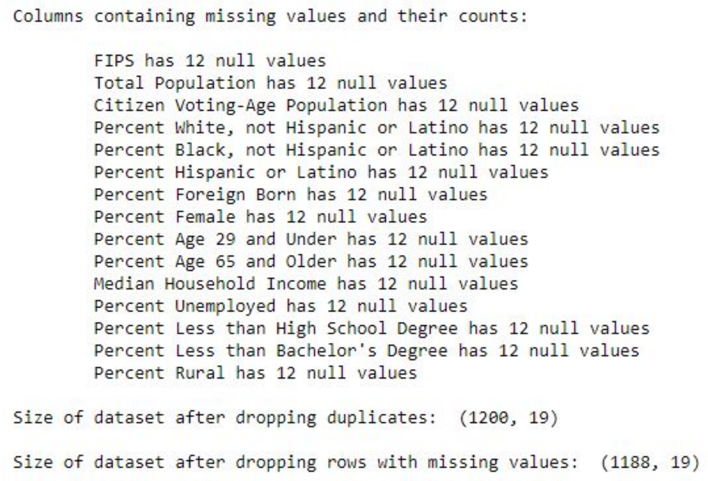
```
      Total Population  Citizen Voting-Age Population
0              34813.0                            0.0
1               2348.0                            0.0
2              31536.0                            0.0
3              28111.0                            0.0
4             101759.0                        78370.0
...                ...                            ...
1195          200536.0                            0.0
1196           13842.0                        10570.0
1197          440604.0                       334780.0
1198           18275.0                            0.0
1199           14335.0                            0.0

[1200 rows x 9 columns]
```

Task 3

The merged dataset has 21 variables, as shown at the top of the output. The County, State, and Office columns are strings, the year is an int64, and all other columns are float64.

It is hard to tell if any of the demographic data is irrelevant at this point in the project, since all of it can be used to make meaningful comparisons based on income, population, age distribution, etc. The join used for Task 2 eliminated the duplicate County and State columns, and the only two redundant columns are the Year and Office ones. After analyzing their values they both only have one unique value (2018 for year, US Senator for office) and we dealt with them by dropping them to reduce the number of columns. This means our dataset now has 19 columns.

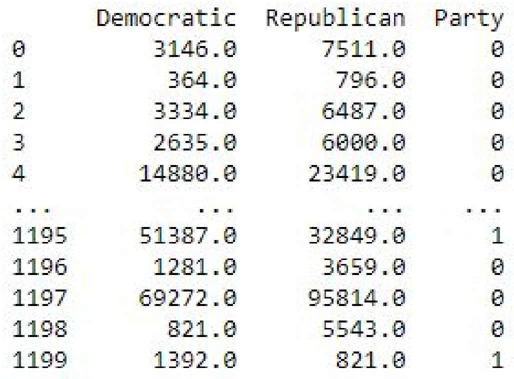
Task 4



```
Columns containing missing values and their counts:

        FIPS has 12 null values
        Total Population has 12 null values
        Citizen Voting-Age Population has 12 null values
        Percent White, not Hispanic or Latino has 12 null values
        Percent Black, not Hispanic or Latino has 12 null values
        Percent Hispanic or Latino has 12 null values
        Percent Foreign Born has 12 null values
        Percent Female has 12 null values
        Percent Age 29 and Under has 12 null values
        Percent Age 65 and Older has 12 null values
        Median Household Income has 12 null values
        Percent Unemployed has 12 null values
        Percent Less than High School Degree has 12 null values
        Percent Less than Bachelor's Degree has 12 null values
        Percent Rural has 12 null values

Size of dataset after dropping duplicates:  (1200, 19)

Size of dataset after dropping rows with missing values:  (1188, 19)
```

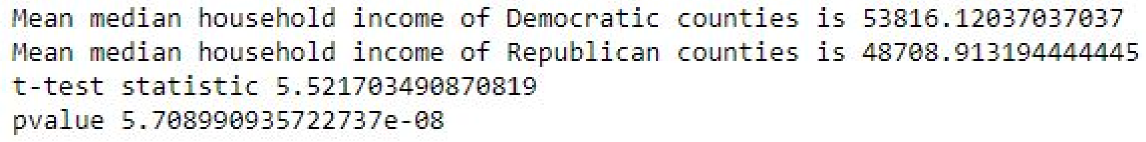
As shown above, the merged data is missing demographic information for 12 counties. Since the merged set has 1200 entries, we can afford to drop the 12 rows containing missing values and still be able to accurately interpret the data.There were no duplicate values.

Task 5



```
        Democratic  Republican  Party
0           3146.0      7511.0      0
1            364.0       796.0      0
2           3334.0      6487.0      0
3           2635.0      6000.0      0
4          14880.0     23419.0      0
...            ...         ...    ...
1195       51387.0     32849.0      1
1196        1281.0      3659.0      0
1197       69272.0     95814.0      0
1198         821.0      5543.0      0
1199        1392.0       821.0      1
```

Task 6



```
Mean median household income of Democratic counties is 53816.12037037037
Mean median household income of Republican counties is 48708.913194444445
t-test statistic 5.521703490870819
pvalue 5.708990935722737e-08
```

For task 6 we created two separate data cells for the Median household income of the Democratic and Republican counties and calculated the mean value of each and hence found that the mean “Median household income” of Democratic counties is greater than Republican.

Hypothesis test:

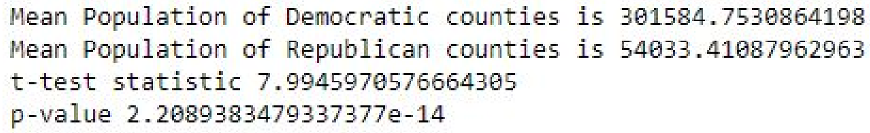
For Hypothesis test we performed t-test on Democrats median household income and Republican median household income and hence made the conclusion as stated below considering significance level α = 0.05.

x¯1 = 53816.120 is the mean median household income of Democratic counties
x¯2 = 48708.913 is the mean median household income of Republican counties

H0: μ1 = μ2 , Hα: μ1 ≠ μ2

Now since, t-test statistic = 5.521and p value = 5.708990935722737e-08. So as p value < α = 0.05, we reject H0: Null hypothesis.

Task 7



```
Mean Population of Democratic counties is 301584.7530864198
Mean Population of Republican counties is 54033.41087962963
t-test statistic 7.9945970576664305
p-value 2.2089383479337377e-14
```

For task 7 we added two extra cells named “Population” to the previously created data frame in task 6 and calculated the mean value of the total population for each Democratic and Republican counties and found that the mean “Total Population” of Democratic counties is greater than Republican.

Hypothesis test:

For Hypothesis test we performed t-test on Democrats county total population and Republican county total population and hence made the conclusion as stated below considering significance level α = 0.05.

x¯1 = 301584.753 is the mean Population of Democratic counties
x¯2 = 54033.410 is the mean Population of Republican counties

H0: μ1 = μ2 , Hα: μ1 ≠ μ2

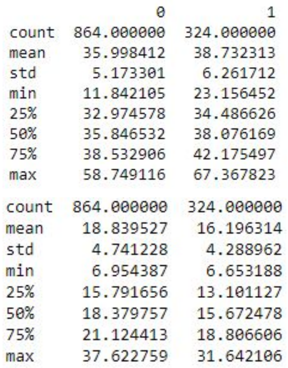
Now since, t-test statistic = 7.994 and p value = 2.2089383479337377e-14. So as p value < α = 0.05, we reject H0: Null hypothesis.

Task 8

For task 8 I split it into two separate cells, that way I could look at the descriptive values of the possible columns that cause a county to be democratic or republican. A lot of these values were either small or large percentages. Images seen below. In the images for the descriptive statistics along with the plots a “0” represents the Republican party and a “1” represents the Democratic party.

Columns that determine age

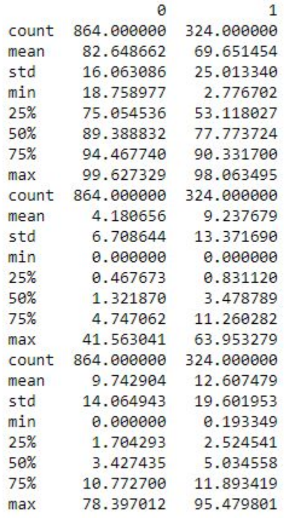
“Percent Age 29 and Under”” And “Percent Age 65 and Over”



```
               0           1
count  864.000000  324.000000
mean    35.998412   38.732313
std      5.173301    6.261712
min     11.842105   23.156452
25%     32.974578   34.486626
50%     35.846532   38.076169
75%     38.532906   42.175497
max     58.749116   67.367823

count  864.000000  324.000000
mean    18.839527   16.196314
std      4.741228    4.288962
min      6.954387    6.653188
25%     15.791656   13.101127
50%     18.379757   15.672478
75%     21.124413   18.806606
max     37.622759   31.642106
```

Columns that determine race

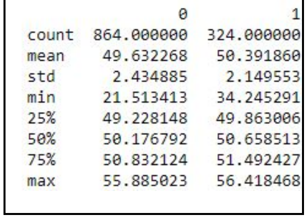
“Percent White, not Hispanic or Latino”, “Percent Black, not Hispanic or Latino”, and “Percent Hispanic or Latino”



```
               0           1
count  864.000000  324.000000
mean    82.648662   69.651454
std     16.063086   25.013340
min     18.758977    2.776702
25%     75.054536   53.118027
50%     89.388832   77.773724
75%     94.467740   90.331700
max     99.627329   98.063495
count  864.000000  324.000000
mean     4.180656    9.237679
std      6.708644   13.371690
min      0.000000    0.000000
25%      0.467673    0.831120
50%      1.321870    3.478789
75%      4.747062   11.260282
max     41.563041   63.953279
count  864.000000  324.000000
mean     9.742904   12.607479
std     14.064943   19.601953
min      0.000000    0.193349
25%      1.704293    2.524541
50%      3.427435    5.034558
75%     10.772700   11.893419
max     78.397012   95.479801
```

Columns that determine gender

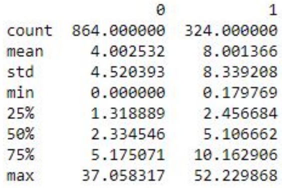
“Percent Female”



```
               0           1
count  864.000000  324.000000
mean    49.632268   50.391860
std      2.434885    2.149553
min     21.513413   34.245291
25%     49.228148   49.863006
50%     50.176792   50.658513
75%     50.832124   51.492427
max     55.885023   56.418468
```

Columns that determine ethnicity

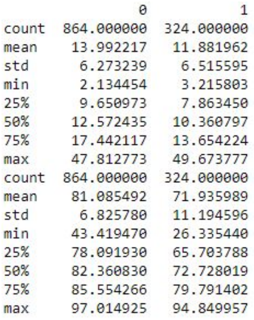
“Percent Foreign Born”



```
               0           1
count  864.000000  324.000000
mean     4.002532    8.001366
std      4.520393    8.339208
min      0.000000    0.179769
25%      1.318889    2.456684
50%      2.334546    5.106662
75%      5.175071   10.162906
max     37.058317   52.229868
```

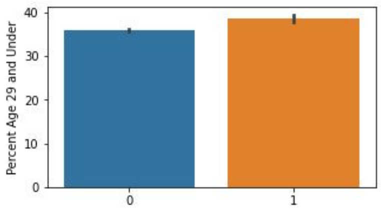
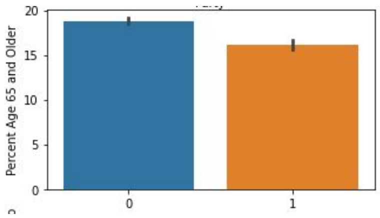
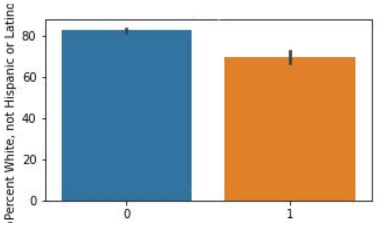
Columns that determine education

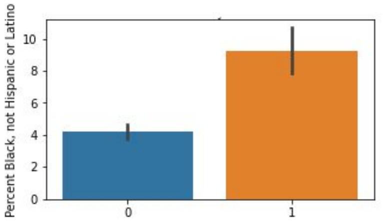
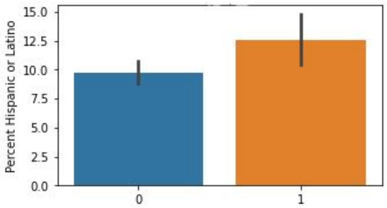
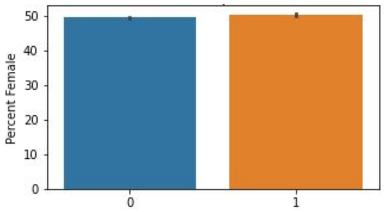
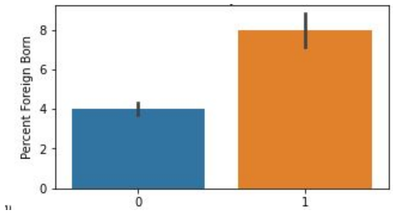
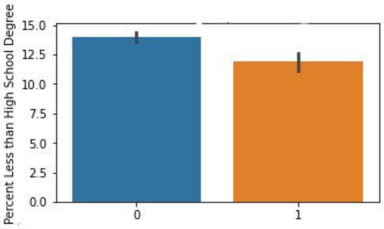
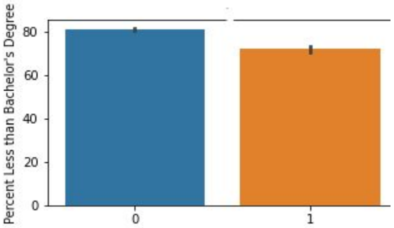
“Percent Less than High School Degree” and “Percent Less than Bachelor's Degree”



```
               0           1
count  864.000000  324.000000
mean    13.992217   11.881962
std      6.273239    6.515595
min      2.134454    3.215803
25%      9.650973    7.863450
50%     12.572435   10.360797
75%     17.442117   13.654224
max     47.812773   49.673777
count  864.000000  324.000000
mean    81.085492   71.935989
std      6.825780   11.194596
min     43.419470   26.335440
25%     78.091930   65.703788
50%     82.360830   72.728019
75%     85.554266   79.791402
max     97.014925   94.849957
```

Plots for task 8 are represented as percentages for the y-axis and the party for the x-axis.

Percent Age 29 and Under	Percent Age 65 and Over	Percent White, not Hispanic or Latino
		
Percent Black, not Hispanic or Latino	Percent Hispanic or Latino	Percent Female

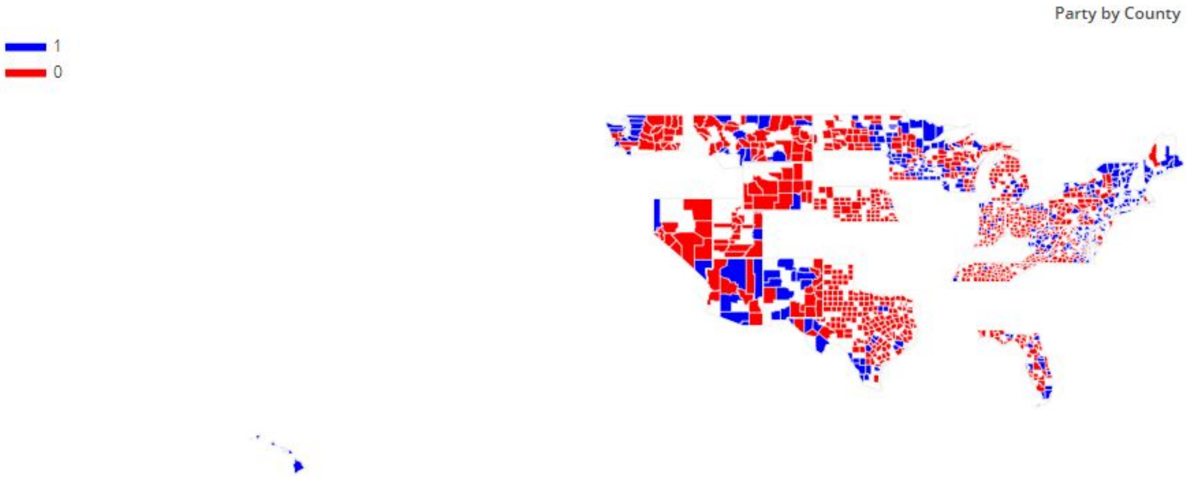
 	 	
Percent Foreign Born	Percent Less than High School Degree	Percent Less than Bachelor's Degree
 	 	

Task 9

For task 9 looking at these descriptive values from, tasks 6-8 the most important variables when determining if a county is Democratic or Republican are Population, Median Household Income, Percent white, not hispanic or latino, Percent Less than Bachelor's degree. These have a much larger percentage and a large difference in their mean. While others like, Hispanic or Latino have a big difference in means, they are at a much lower percent level.

Task 10

For task 10 We compiled a list of the distinct states to be used in the mapping of the counties demographics of political parties. I then used the fips codes and values of the parties from the merged data created. The plot can be seen below with blue correlating to democratic counties and red to republican counties.


Party by County
1
0