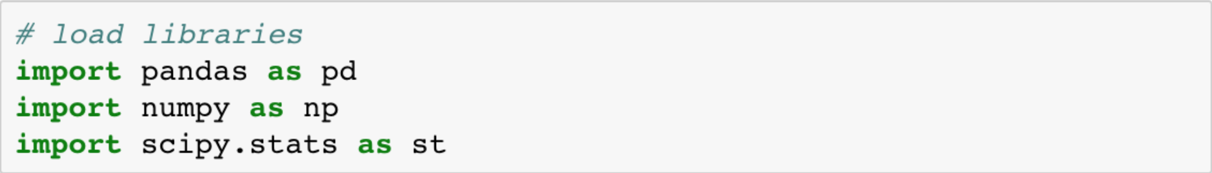
CS 418 Data Science

Project 01: Exploratory Data Analysis

Group Members: Vani Jaiswal, Mudit Kumar, Ahmed Khan

The libraries we used to accomplish this project are:



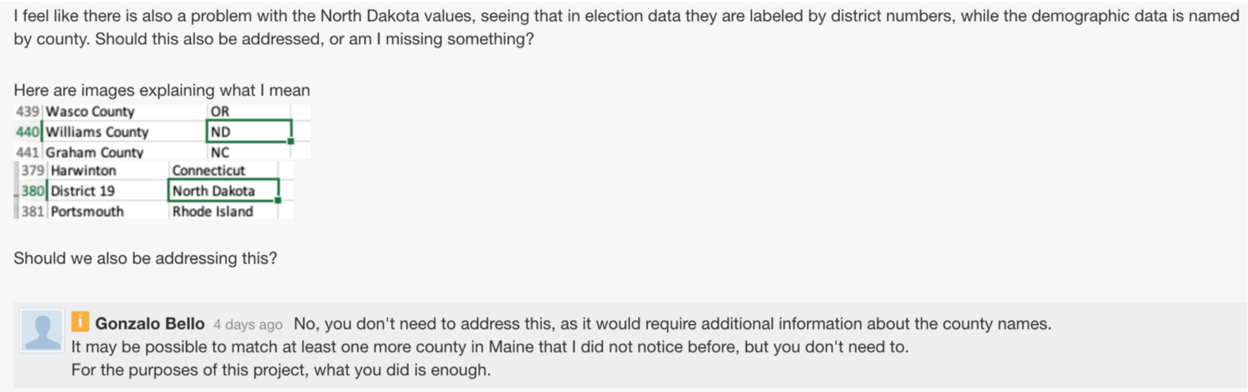
```
# load libraries
import pandas as pd
import numpy as np
import scipy.stats as st
```

1) There were multiple inconsistencies within the data that we addressed.

a. Counties do not have the same naming conventions

i. The demographic data contained the suffix ‘County’ and ‘Parish’ at the end of each county, but the election data did not. Therefore, we stripped all inconsistent suffixes.

ii. We found an inconsistency with the naming convention of North Dakota within the datasets, but Professor Bello had not caught this inconsistency and therefore told us here to ignore it.



I feel like there is also a problem with the North Dakota values, seeing that in election data they are labeled by district numbers, while the demographic data is named by county. Should this also be addressed, or am I missing something?

Here are images explaining what I mean

439	Wasco County	OR
440	Williams County	ND
441	Graham County	NC

379	Harwinton	Connecticut
380	District 19	North Dakota
381	Portsmouth	Rhode Island

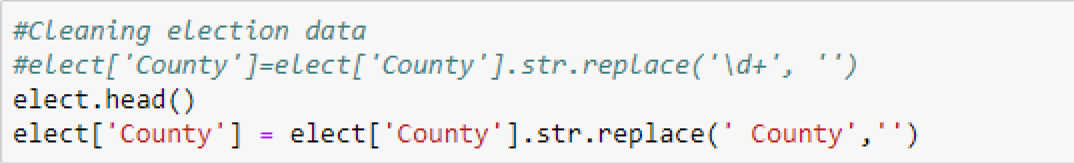
Should we also be addressing this?

Gonzalo Bello 4 days ago No, you don't need to address this, as it would require additional information about the county names.
It may be possible to match at least one more county in Maine that I did not notice before, but you don't need to.
For the purposes of this project, what you did is enough.

b. States do not have the same naming conventions

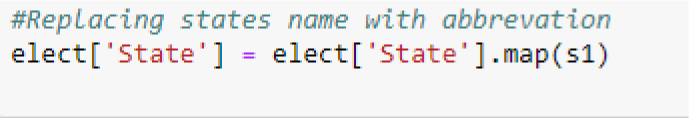
i. The demographic data named U.S. states using their appropriate abbreviations, but the election data spelled the full state name out. Therefore, we used a dictionary to swap all of the election data state attributes to their respective abbreviation. The dictionary we used can be found here.

Here are the commands we used to clean the data:



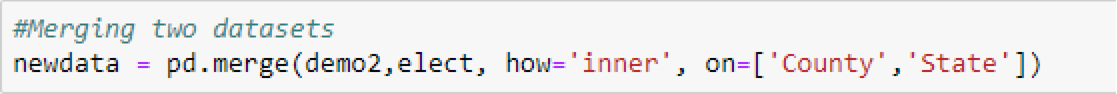
```
#Cleaning election data
#elect['County']=elect['County'].str.replace('\d+', '')
elect.head()
elect['County'] = elect['County'].str.replace(' County','')
```

For replacing state names with abbreviation, we created a dictionary for all the states



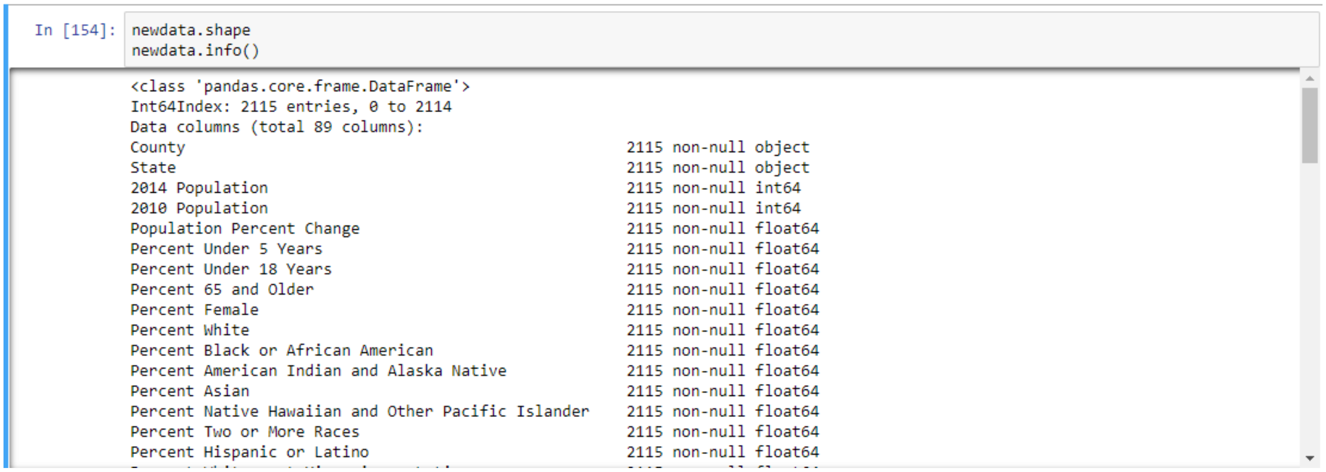
```
#Replacing states name with abbrevation
elect['State'] = elect['State'].map(s1)
```

After cleaning the data, we performed an inner join of the two datasets



```
#Merging two datasets
newdata = pd.merge(demo2,elect, how='inner', on=['County','State'])
```

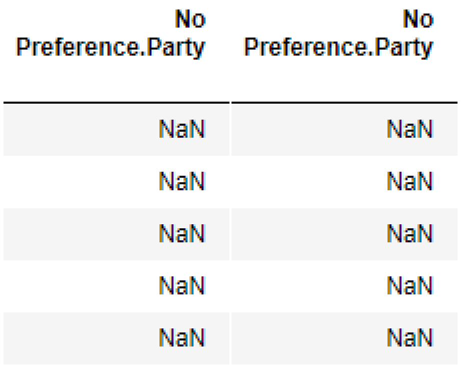
2) The different datatypes of these variables are float64, int64, and object. Below is the command we used to get the list of the types of variables.



```
In [154]: newdata.shape
          newdata.info()
```

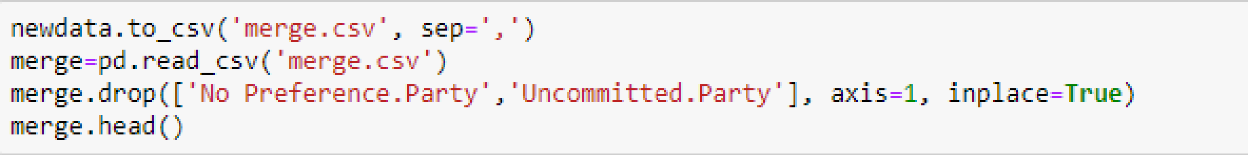
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2115 entries, 0 to 2114
Data columns (total 89 columns):
County                                               2115 non-null object
State                                                2115 non-null object
2014 Population                                      2115 non-null int64
2010 Population                                      2115 non-null int64
Population Percent Change                            2115 non-null float64
Percent Under 5 Years                                2115 non-null float64
Percent Under 18 Years                               2115 non-null float64
Percent 65 and Older                                 2115 non-null float64
Percent Female                                       2115 non-null float64
Percent White                                        2115 non-null float64
Percent Black or African American                    2115 non-null float64
Percent American Indian and Alaska Native            2115 non-null float64
Percent Asian                                        2115 non-null float64
Percent Native Hawaiian and Other Pacific Islander   2115 non-null float64
Percent Two or More Races                            2115 non-null float64
Percent Hispanic or Latino                           2115 non-null float64
```

These columns contain missing values:



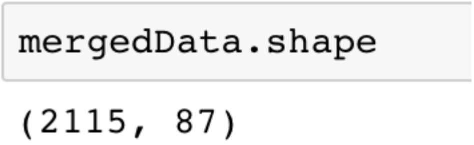
No Preference.Party	No Preference.Party
NaN	NaN
NaN	NaN
NaN	NaN
NaN	NaN
NaN	NaN

We chose to drop these columns since we had no basis to estimate from.



```
newdata.to_csv('merge.csv', sep=',')
merge=pd.read_csv('merge.csv')
merge.drop(['No Preference.Party','Uncommitted.Party'], axis=1, inplace=True)
merge.head()
```

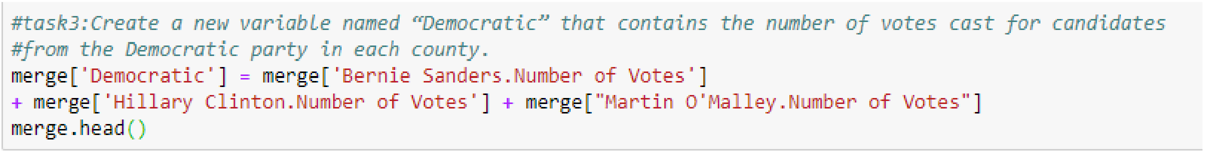
After cleaning, the merged dataset has 2115 observations, and 87 attributes. Here is how we found the dimensions of the merged dataset:



```
mergedData.shape
```

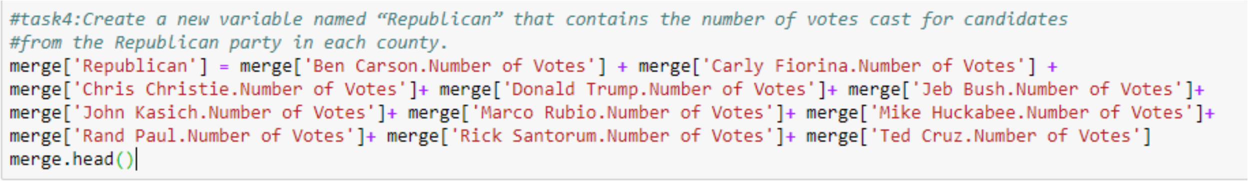
```
(2115, 87)
```

3)



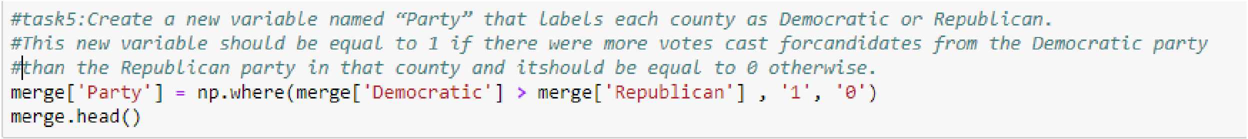
```
#task3:Create a new variable named “Democratic” that contains the number of votes cast for candidates
#from the Democratic party in each county.
merge['Democratic'] = merge['Bernie Sanders.Number of Votes']
+ merge['Hillary Clinton.Number of Votes'] + merge["Martin O'Malley.Number of Votes"]
merge.head()
```

4)



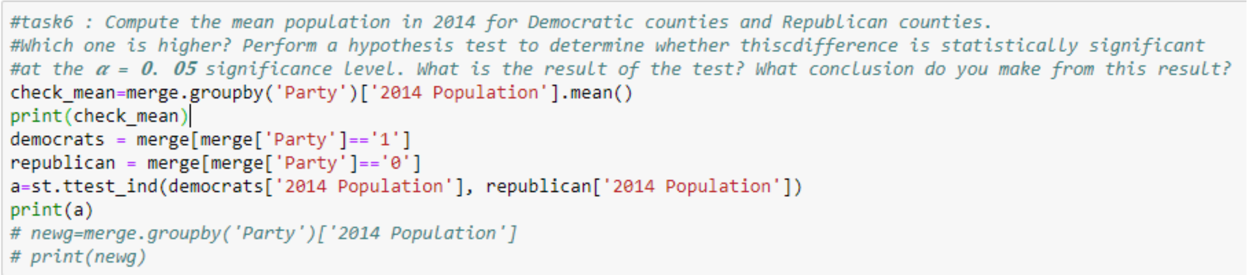
```
#task4:Create a new variable named “Republican” that contains the number of votes cast for candidates
#from the Republican party in each county.
merge['Republican'] = merge['Ben Carson.Number of Votes'] + merge['Carly Fiorina.Number of Votes'] +
merge['Chris Christie.Number of Votes']+ merge['Donald Trump.Number of Votes']+ merge['Jeb Bush.Number of Votes']+
merge['John Kasich.Number of Votes']+ merge['Marco Rubio.Number of Votes']+ merge['Mike Huckabee.Number of Votes']+
merge['Rand Paul.Number of Votes']+ merge['Rick Santorum.Number of Votes']+ merge['Ted Cruz.Number of Votes']
merge.head()
```

5)



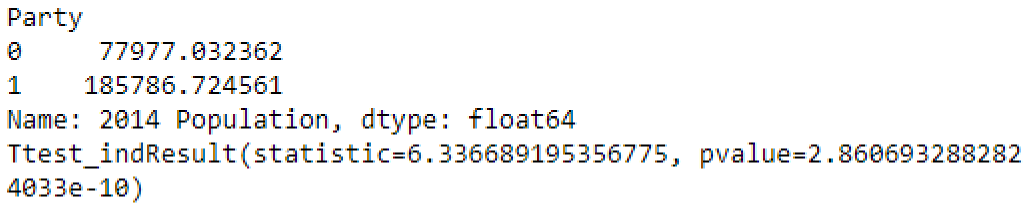
```
#task5:Create a new variable named “Party” that labels each county as Democratic or Republican.
#This new variable should be equal to 1 if there were more votes cast forcandidates from the Democratic party
#than the Republican party in that county and itshould be equal to 0 otherwise.
merge['Party'] = np.where(merge['Democratic'] > merge['Republican'] , '1', '0')
merge.head()
```

6)



```
#task6 : Compute the mean population in 2014 for Democratic counties and Republican counties.
#Which one is higher? Perform a hypothesis test to determine whether thiscdifference is statistically significant
#at the α = 0. 05 significance level. What is the result of the test? What conclusion do you make from this result?
check_mean=merge.groupby('Party')['2014 Population'].mean()
print(check_mean)
democrats = merge[merge['Party']=='1']
republican = merge[merge['Party']=='0']
a=st.ttest_ind(democrats['2014 Population'], republican['2014 Population'])
print(a)
# newg=merge.groupby('Party')['2014 Population']
# print(newg)
```

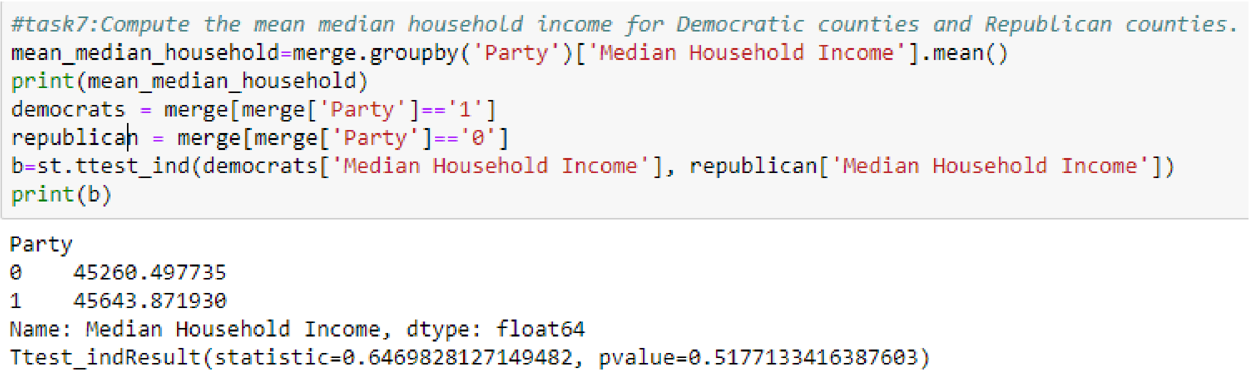
Result of the test:



```
Party
0     77977.032362
1    185786.724561
Name: 2014 Population, dtype: float64
Ttest_indResult(statistic=6.336689195356775, pvalue=2.860693288282
4033e-10)
```

The mean population in 2014 for democratic counties is higher than republican counties. After performing the t-test, we get the p-value of 2.86 e-10 which is significantly lower than our significance level. So we conclude this difference is statistically significant at the 0.05 significance level.

7)



```
#task7:Compute the mean median household income for Democratic counties and Republican counties.
mean_median_household=merge.groupby('Party')['Median Household Income'].mean()
print(mean_median_household)
democrats = merge[merge['Party']=='1']
republican = merge[merge['Party']=='0']
b=st.ttest_ind(democrats['Median Household Income'], republican['Median Household Income'])
print(b)
```

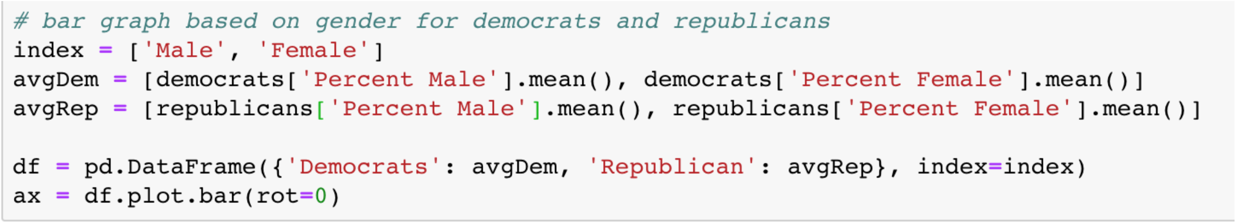
```
Party
0    45260.497735
1    45643.871930
Name: Median Household Income, dtype: float64
Ttest_indResult(statistic=0.6469828127149482, pvalue=0.5177133416387603)
```

The mean median household income is higher for democratic counties than republican counties. After performing the t-test, we get the p-value of 0.518 which is greater than the significance level of .05 so we can conclude that we don't have sufficient evidence say that this observation is statistically significant.

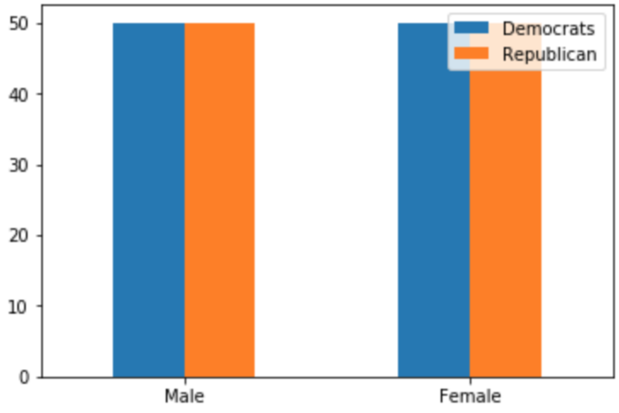
8)

We then attempted to visualize the data in order to analyze it better and arrive at certain conclusions.

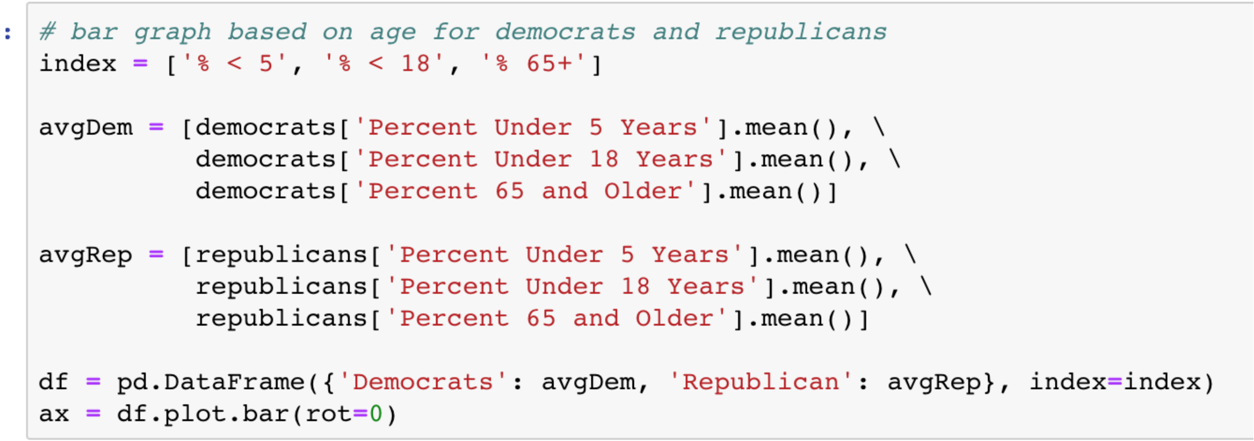
We first plotted average gender according to party to see if gender plays a role in determining one's political party. As you can see, the male and female percentages are almost identical for each party, therefore we do not believe that gender is a good determining factor of political party.



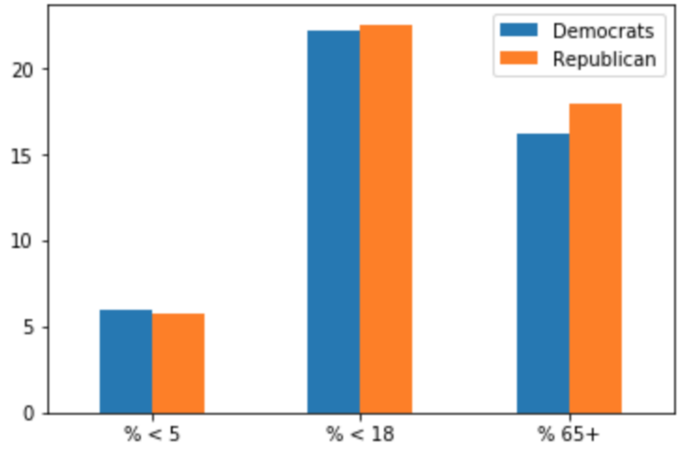
```
# bar graph based on gender for democrats and republicans
index = ['Male', 'Female']
avgDem = [democrats['Percent Male'].mean(), democrats['Percent Female'].mean()]
avgRep = [republicans['Percent Male'].mean(), republicans['Percent Female'].mean()]

df = pd.DataFrame({'Democrats': avgDem, 'Republican': avgRep}, index=index)
ax = df.plot.bar(rot=0)
```



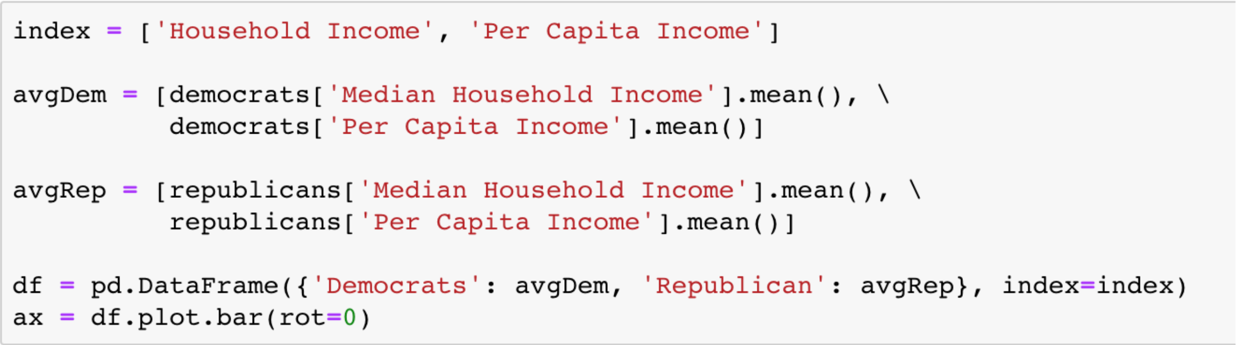
We then analyzed a graph of age based on political party. As you can see, the data could allow us to draw a conclusion, but seeing that U.S. citizens can only vote if they are at least eighteen years of age and we do not have any data for people between the ages of 18 and 64, we cannot recommend age to be a determining factor of political party based on this dataset.



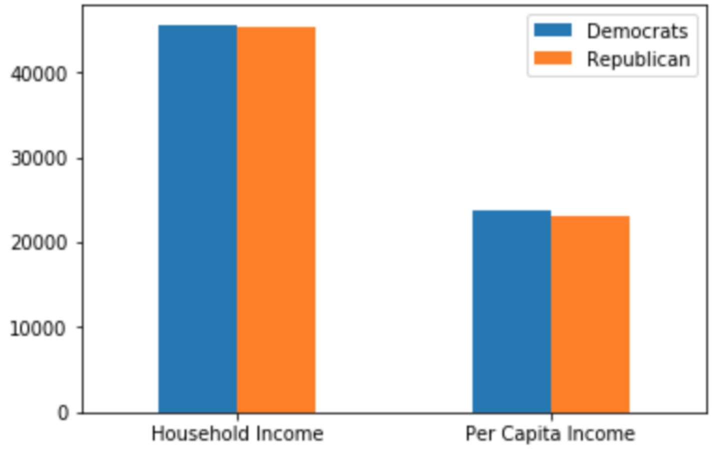
```
# bar graph based on age for democrats and republicans
index = ['% < 5', '% < 18', '% 65+']

avgDem = [democrats['Percent Under 5 Years'].mean(), \
          democrats['Percent Under 18 Years'].mean(), \
          democrats['Percent 65 and Older'].mean()]

avgRep = [republicans['Percent Under 5 Years'].mean(), \
          republicans['Percent Under 18 Years'].mean(), \
          republicans['Percent 65 and Older'].mean()]

df = pd.DataFrame({'Democrats': avgDem, 'Republican': avgRep}, index=index)
ax = df.plot.bar(rot=0)
```

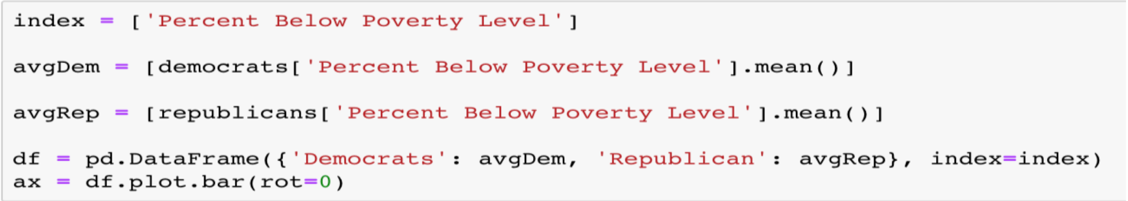


We then compared average median income and poverty percentages. As you can see, income does not seem to play a vital role in political affiliation, but poverty levels has interesting results. Republicans have a lower average poverty percentage than democrats do.

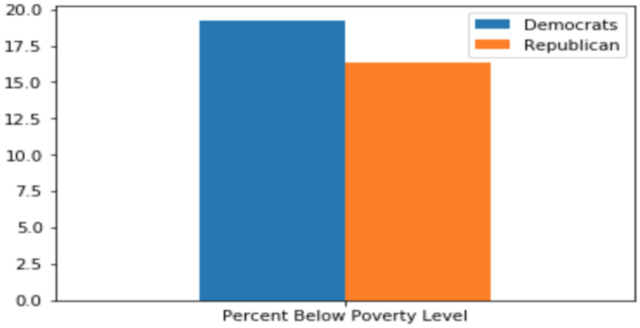


```
index = ['Household Income', 'Per Capita Income']

avgDem = [democrats['Median Household Income'].mean(), \
          democrats['Per Capita Income'].mean()]

avgRep = [republicans['Median Household Income'].mean(), \
          republicans['Per Capita Income'].mean()]

df = pd.DataFrame({'Democrats': avgDem, 'Republican': avgRep}, index=index)
ax = df.plot.bar(rot=0)
```

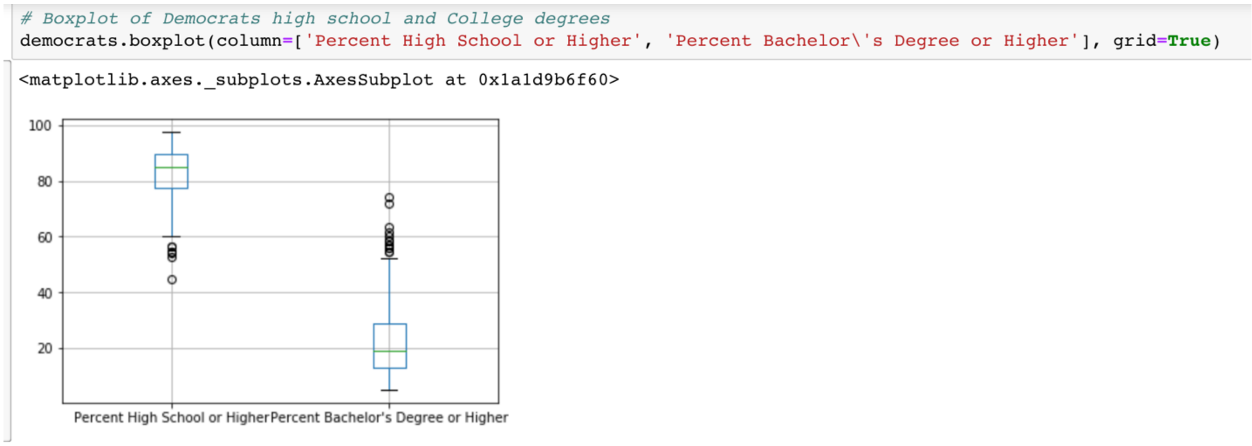




```
index = ['Percent Below Poverty Level']

avgDem = [democrats['Percent Below Poverty Level'].mean()]

avgRep = [republicans['Percent Below Poverty Level'].mean()]

df = pd.DataFrame({'Democrats': avgDem, 'Republican': avgRep}, index=index)
ax = df.plot.bar(rot=0)
```



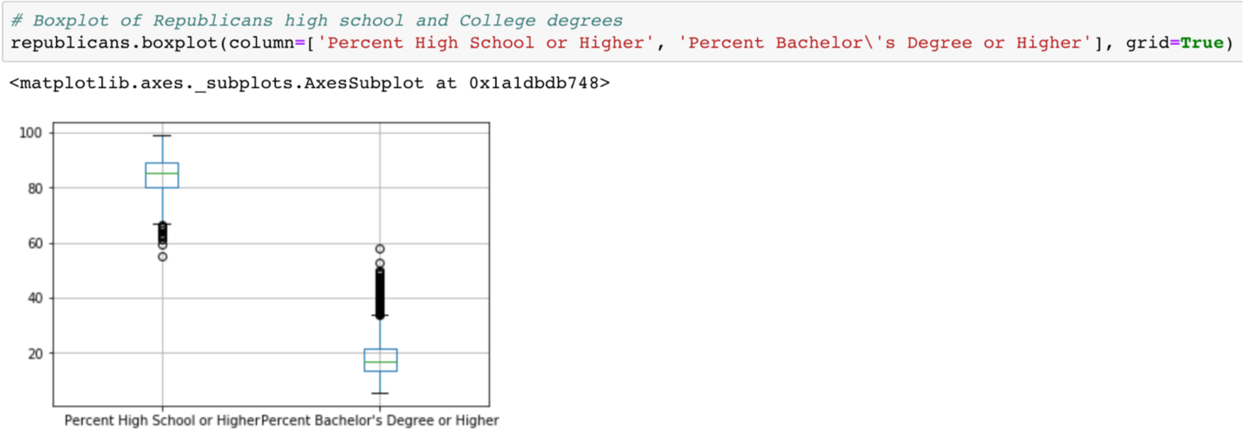
After that, we analyzed education levels based on party affiliation.



```
# Boxplot of Democrats high school and College degrees
democrats.boxplot(column=['Percent High School or Higher', 'Percent Bachelor\'s Degree or Higher'], grid=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1d9b6f60>
```

100
80
60
40
20
Percent High School or Higher
Percent Bachelor's Degree or Higher



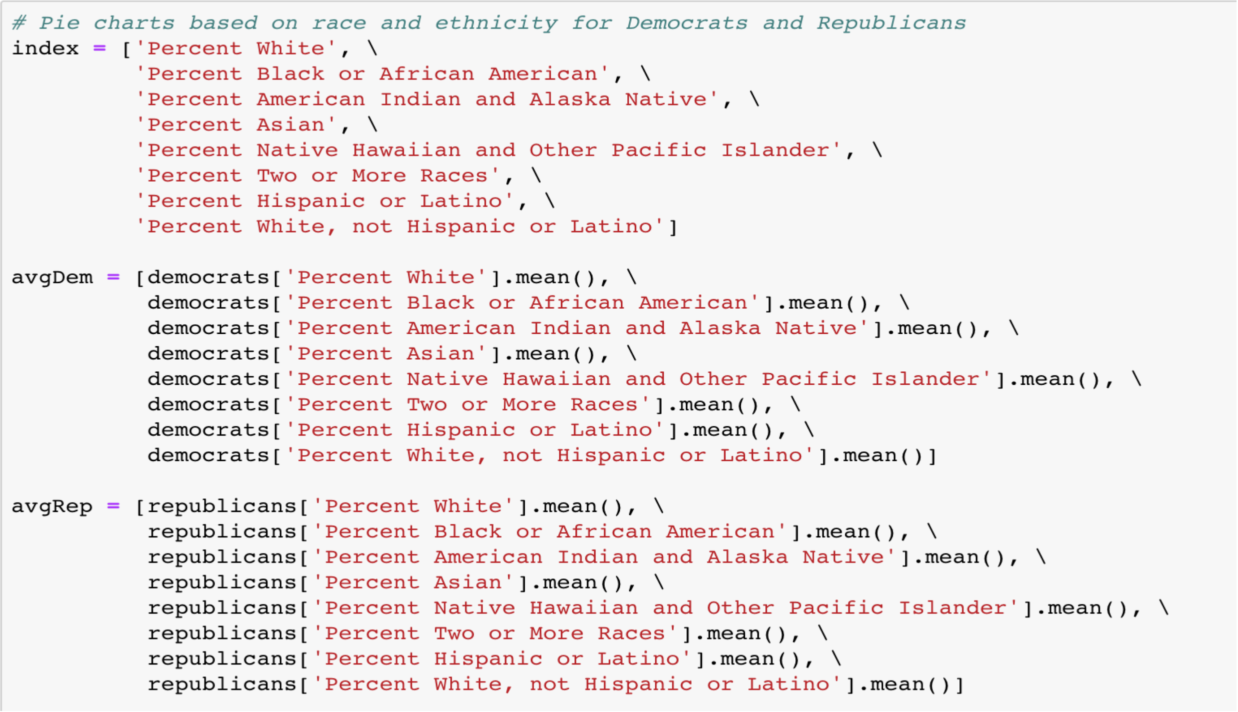
```
# Boxplot of Republicans high school and College degrees
republicans.boxplot(column=['Percent High School or Higher', 'Percent Bachelor\'s Degree or Higher'], grid=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1dbdb748>
```

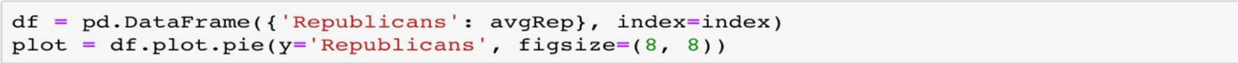
100
80
60
40
20
Percent High School or Higher
Percent Bachelor's Degree or Higher

As you can see, Democrats and Republicans both have high High-School graduation rates. Although Democrats have more outliers that receive college degrees at increased percentages, the median of both Democrat and Republican college degrees are roughly the same, at around 20%. Therefore, we cannot recommend education rates to be a determining factor of political affiliation.

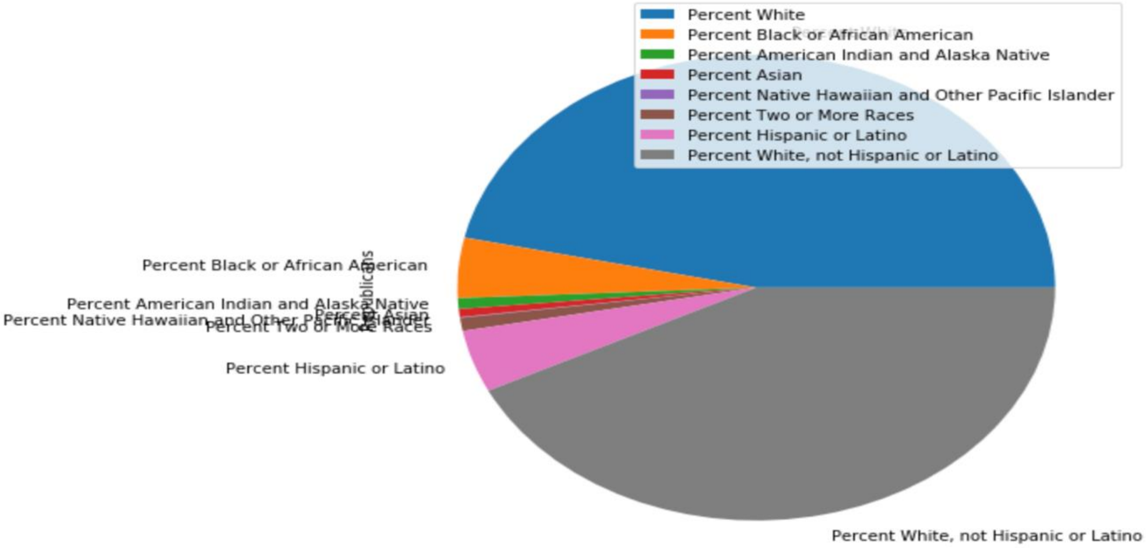
We finally visualized racial and ethnic backgrounds based on political party. Here are our findings:

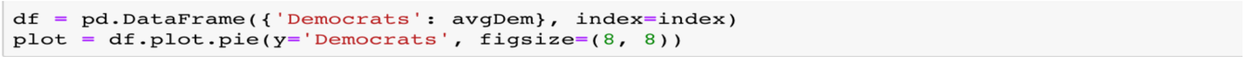


```
# Pie charts based on race and ethnicity for Democrats and Republicans
index = ['Percent White', \
         'Percent Black or African American', \
         'Percent American Indian and Alaska Native', \
         'Percent Asian', \
         'Percent Native Hawaiian and Other Pacific Islander', \
         'Percent Two or More Races', \
         'Percent Hispanic or Latino', \
         'Percent White, not Hispanic or Latino']

avgDem = [democrats['Percent White'].mean(), \
          democrats['Percent Black or African American'].mean(), \
          democrats['Percent American Indian and Alaska Native'].mean(), \
          democrats['Percent Asian'].mean(), \
          democrats['Percent Native Hawaiian and Other Pacific Islander'].mean(), \
          democrats['Percent Two or More Races'].mean(), \
          democrats['Percent Hispanic or Latino'].mean(), \
          democrats['Percent White, not Hispanic or Latino'].mean()]

avgRep = [republicans['Percent White'].mean(), \
          republicans['Percent Black or African American'].mean(), \
          republicans['Percent American Indian and Alaska Native'].mean(), \
          republicans['Percent Asian'].mean(), \
          republicans['Percent Native Hawaiian and Other Pacific Islander'].mean(), \
          republicans['Percent Two or More Races'].mean(), \
          republicans['Percent Hispanic or Latino'].mean(), \
          republicans['Percent White, not Hispanic or Latino'].mean()]
```

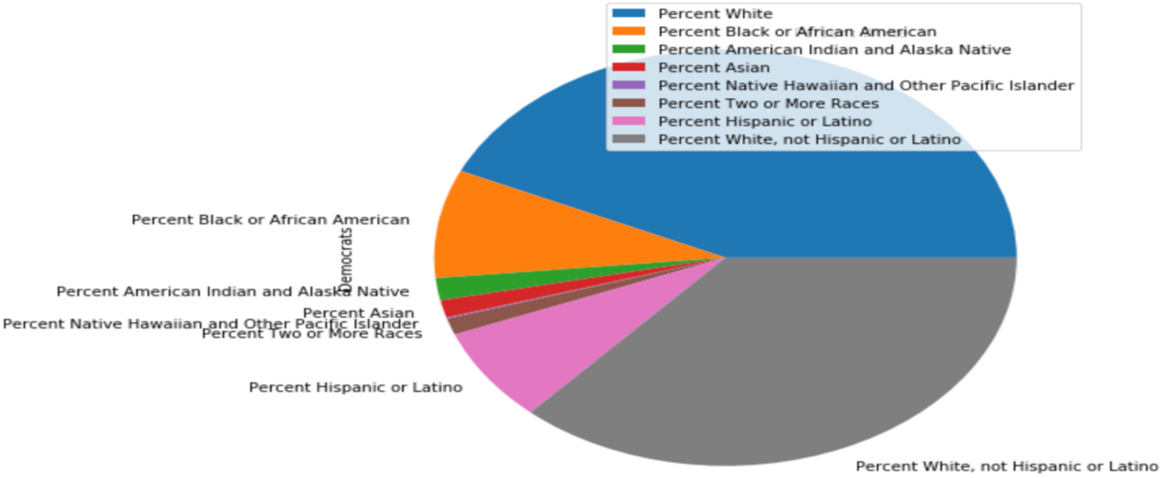


```
df = pd.DataFrame({'Republicans': avgRep}, index=index)
plot = df.plot.pie(y='Republicans', figsize=(8, 8))
```





```
df = pd.DataFrame({'Democrats': avgDem}, index=index)
plot = df.plot.pie(y='Democrats', figsize=(8, 8))
```



As you can see by the given pie charts for Democrats and Republicans, Republicans have far fewer minority percentages than Democrats do.

9)

Male and female percentages are almost identical for each party; therefore, we do not believe that gender is a good determining factor of political party.

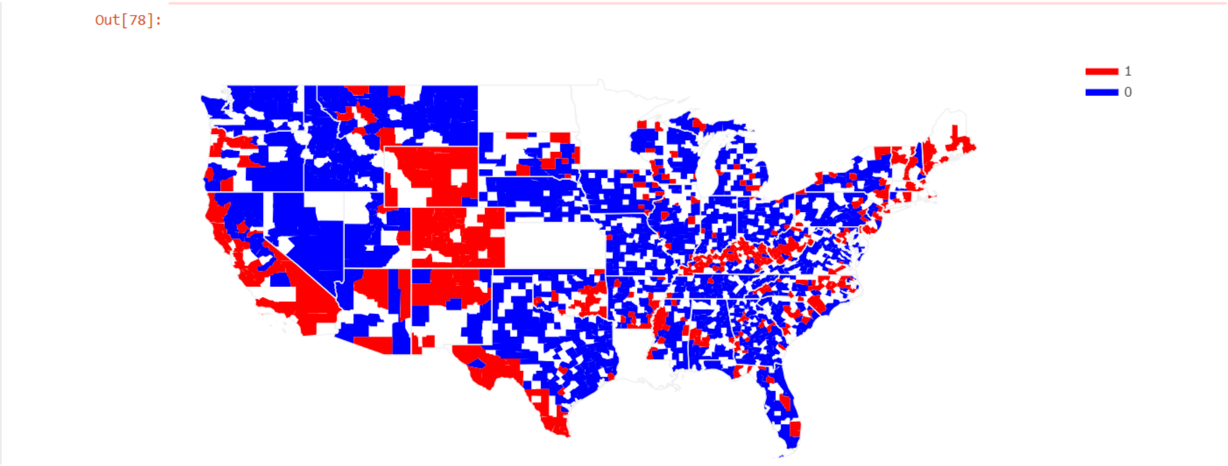
The data could allow us to draw a conclusion based on age but seeing that U.S. citizens can only vote if they are at least eighteen years of age and we do not have any data for people between the ages of 18 and 64, we cannot recommend age to be a determining factor of political party based on this dataset.

Income does not seem to play a vital role in political affiliation, but poverty levels has interesting results. Republicans have a lower average poverty percentage than democrats do.

Although Democrats have more outliers that receive college degrees at increased percentages, the median of both Democrat and Republican college degrees are roughly the same, at around 20%. Therefore, we cannot recommend education rates to be a determining factor of political affiliation.

Altogether, Republicans have far fewer minority percentages than Democrats do. Therefore, we believe that race and ethnicity are the most important variables in the dataset to determine whether a county is Democratic or Republican.

10)



Out[78]:

For task 10, we installed all the necessary libraries. Used credentials to connect to plotly. Used this https://raw.githubusercontent.com/plotly/datasets/master/laucnty16.csv' to merged data by inner join. We required FIPS code to generate map so required the previous data. We used reference for map from following link https://plot.ly/python/county-choropleth/.