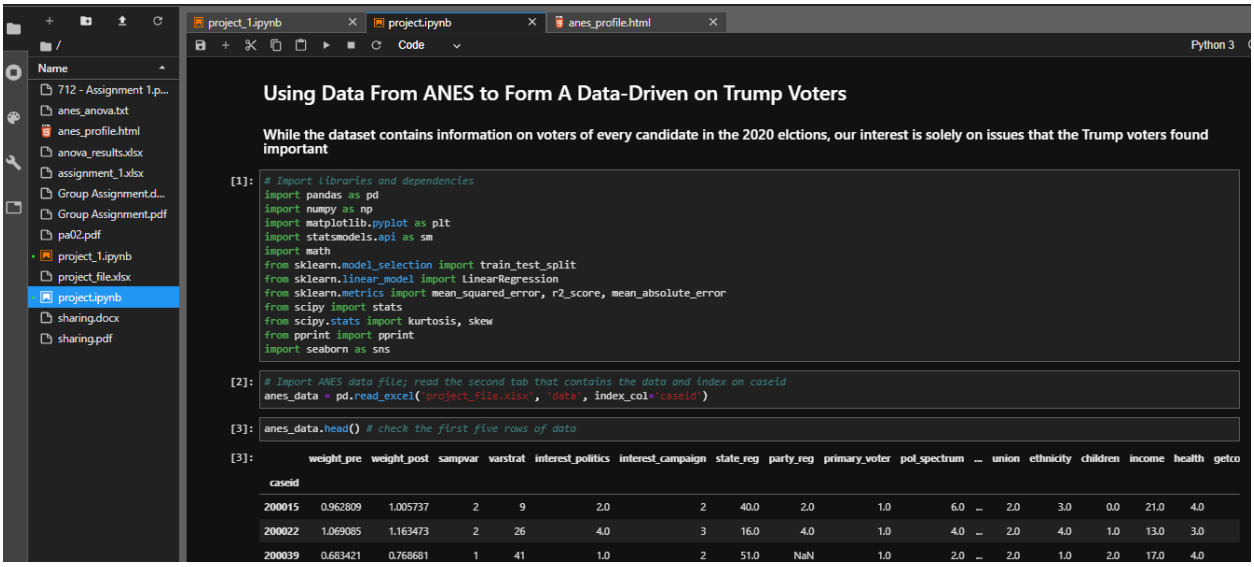
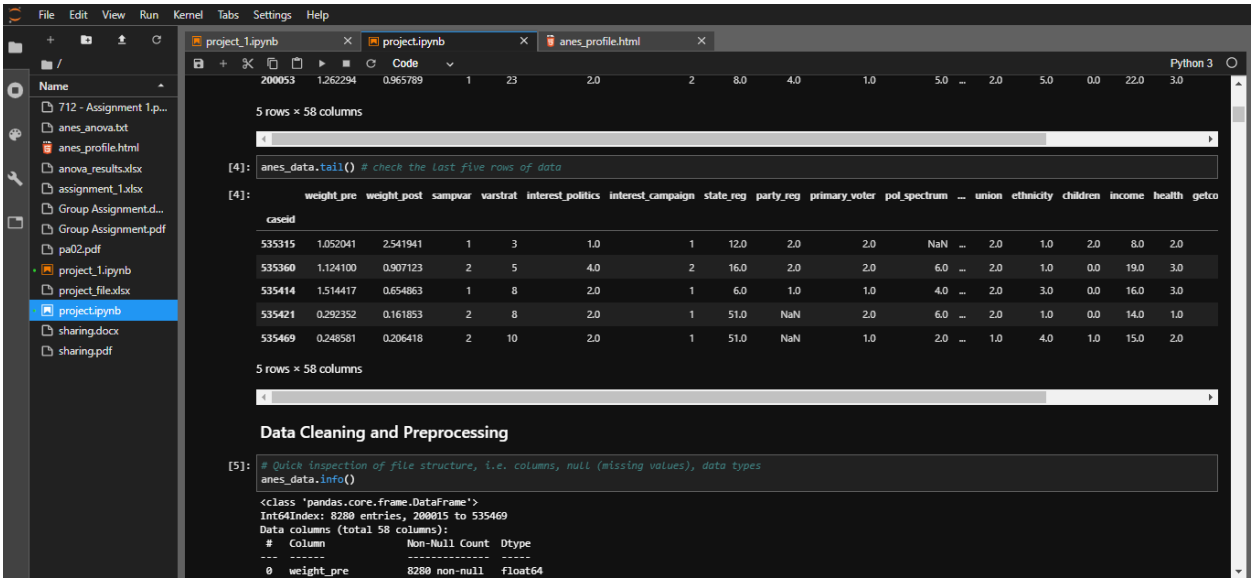
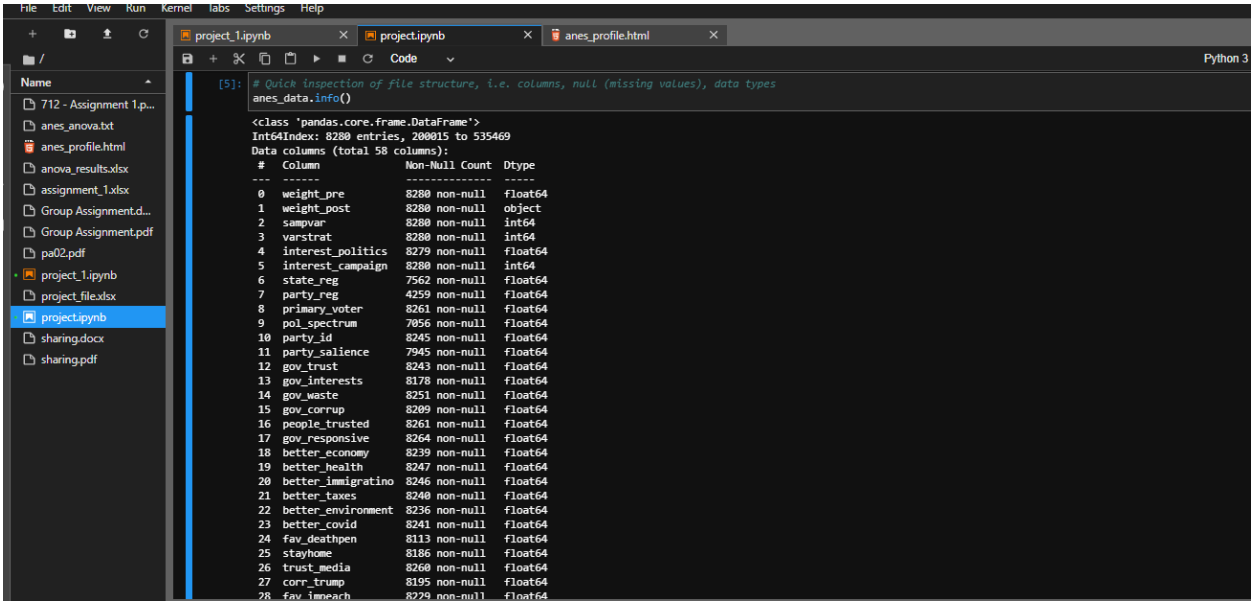
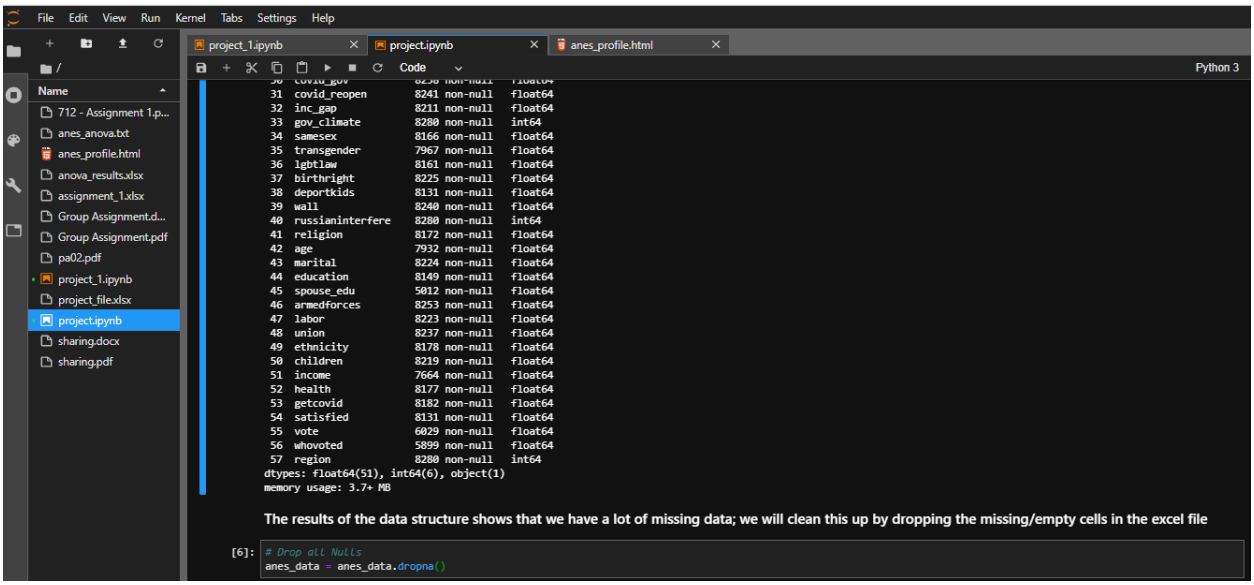
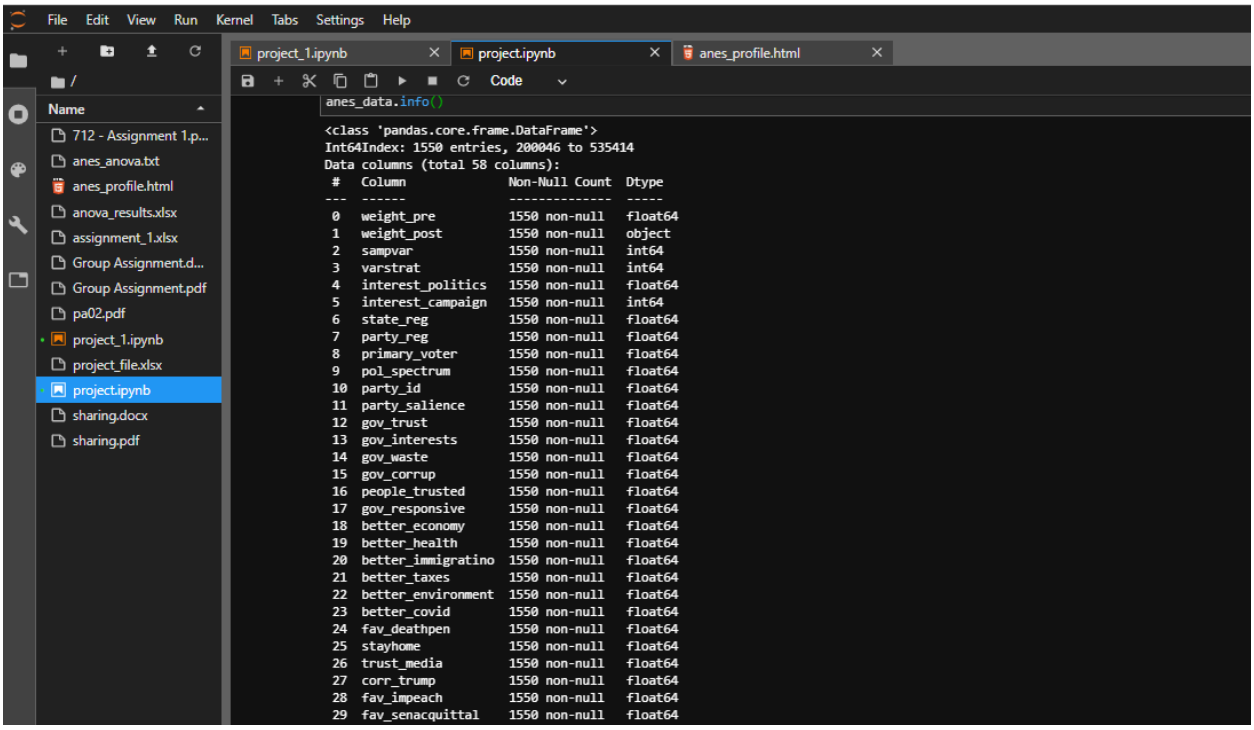
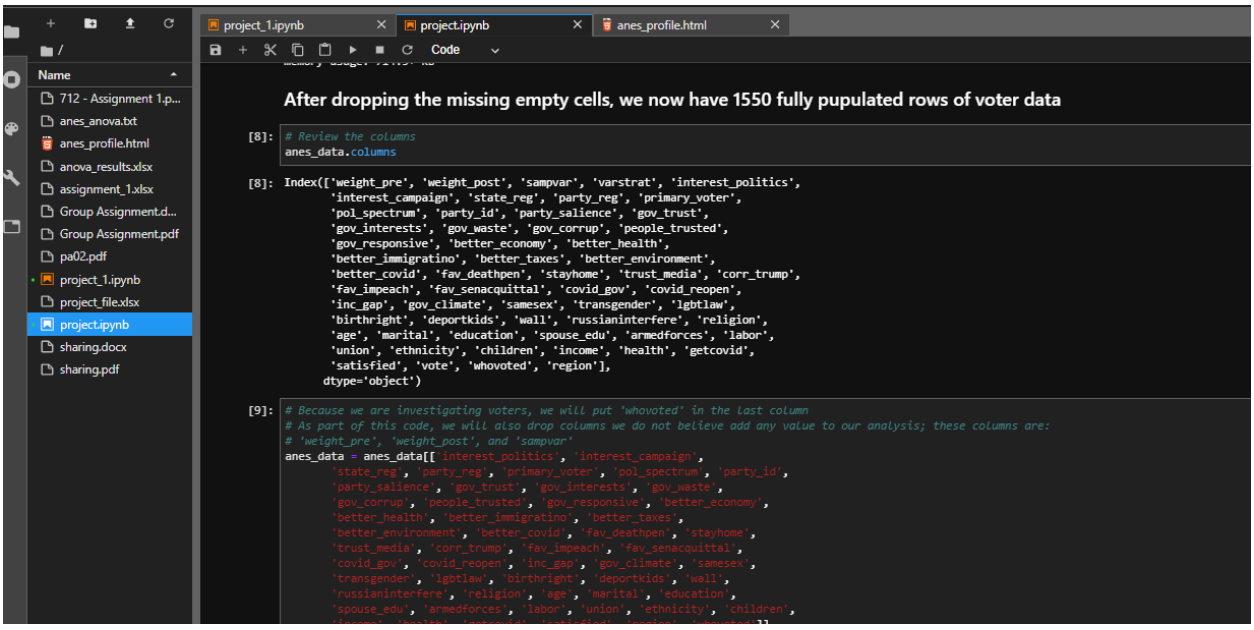

project_1.ipynb
project.ipynb
anes_profile.html
Code
Python 3
Name
712 - Assignment 1.p...
anes_anova.txt
anes_profile.html
anova_results.xlsx
assignment_1.xlsx
Group Assignment.d...
Group Assignment.pdf
pa02.pdf
project_1.ipynb
project_file.xlsx
project.ipynb
sharing.docx
sharing.pdf
Using Data From ANES to Form A Data-Driven on Trump Voters
While the dataset contains information on voters of every candidate in the 2020 elctions, our interest is solely on issues that the Trump voters found important
[1]:
# Import libraries and dependencies
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
import math
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
from scipy import stats
from scipy.stats import kurtosis, skew
from pprint import pprint
import seaborn as sns
[2]:
# Import ANES data file; read the second tab that contains the data and index on caseid
anes_data = pd.read_excel('project_file.xlsx', 'data', index_col='caseid')
[3]:
anes_data.head() # check the first five rows of data
[3]:
weight_pre weight_post sampvar varstrat interest_politics interest_campaign state_reg party_reg primary_voter pol_spectrum ... union ethnicity children income health getco
caseid
200015 0.962809 1.005737 2 9 2.0 2 40.0 2.0 1.0 6.0 ... 2.0 3.0 0.0 21.0 4.0
200022 1.069085 1.163473 2 26 4.0 3 16.0 4.0 1.0 4.0 ... 2.0 4.0 1.0 13.0 3.0
200039 0.683421 0.768681 1 41 1.0 2 51.0 NaN 1.0 2.0 ... 2.0 1.0 2.0 17.0 4.0

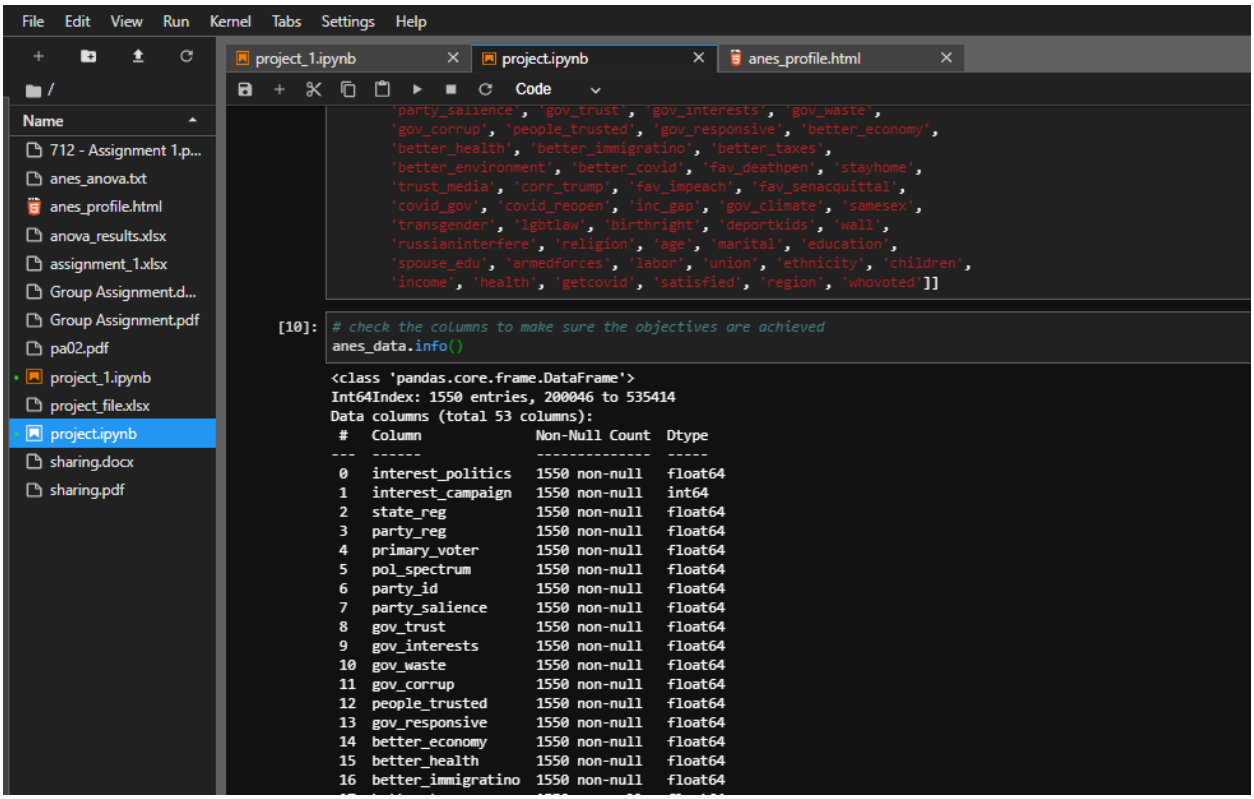

File Edit View Run Kernel Tabs Settings Help
project_1.ipynb
project.ipynb
anes_profile.html
Code
Python 3
Name
712 - Assignment 1.p...
anes_anova.txt
anes_profile.html
anova_results.xlsx
assignment_1.xlsx
Group Assignment.d...
Group Assignment.pdf
pa02.pdf
project_1.ipynb
project_file.xlsx
project.ipynb
sharing.docx
sharing.pdf
200053 1.262294 0.965789 1 23 2.0 2 8.0 4.0 1.0 5.0 ... 2.0 5.0 0.0 22.0 3.0
5 rows × 58 columns
[4]:
anes_data.tail() # check the last five rows of data
[4]:
weight_pre weight_post sampvar varstrat interest_politics interest_campaign state_reg party_reg primary_voter pol_spectrum ... union ethnicity children income health getco
caseid
535315 1.052041 2.541941 1 3 1.0 1 12.0 2.0 2.0 NaN ... 2.0 1.0 2.0 8.0 2.0
535360 1.124100 0.907123 2 5 4.0 2 16.0 2.0 2.0 6.0 ... 2.0 1.0 0.0 19.0 3.0
535414 1.514417 0.654863 1 8 2.0 1 6.0 1.0 1.0 4.0 ... 2.0 3.0 0.0 16.0 3.0
535421 0.292352 0.161853 2 8 2.0 1 51.0 NaN 2.0 6.0 ... 2.0 1.0 0.0 14.0 1.0
535469 0.248581 0.206418 2 10 2.0 1 51.0 NaN 1.0 2.0 ... 1.0 4.0 1.0 15.0 2.0
5 rows × 58 columns
Data Cleaning and Preprocessing
[5]:
# Quick inspection of file structure, i.e. columns, null (missing values), data types
anes_data.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8280 entries, 200015 to 535469
Data columns (total 58 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
0 weight_pre 8280 non-null float64

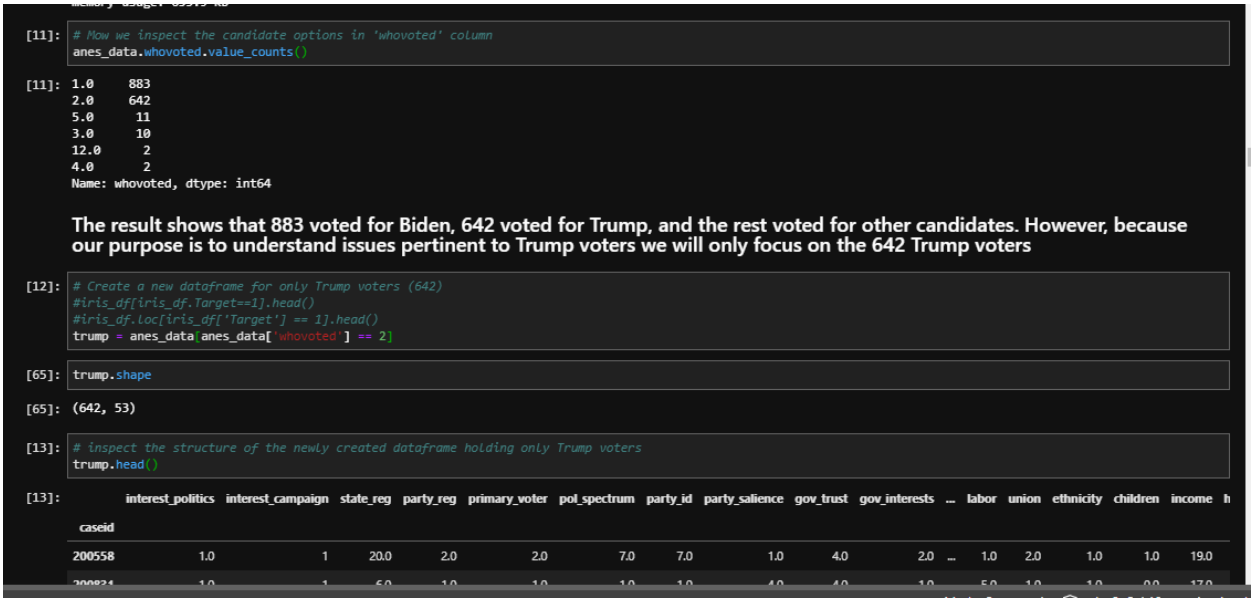

File Edit View Run Kernel Tabs Settings Help
project_1.ipynb
project.ipynb
anes_profile.html
Code
Python 3
Name
712 - Assignment 1.p...
anes_anova.txt
anes_profile.html
anova_results.xlsx
assignment_1.xlsx
Group Assignment.d...
Group Assignment.pdf
pa02.pdf
project_1.ipynb
project_file.xlsx
project.ipynb
sharing.docx
sharing.pdf
[5]:
# Quick inspection of file structure, i.e. columns, null (missing values), data types
anes_data.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8280 entries, 200015 to 535469
Data columns (total 58 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
0 weight_pre 8280 non-null float64
1 weight_post 8280 non-null object
2 sampvar 8280 non-null int64
3 varstrat 8280 non-null int64
4 interest_politics 8279 non-null float64
5 interest_campaign 8280 non-null int64
6 state_reg 7562 non-null float64
7 party_reg 4259 non-null float64
8 primary_voter 8261 non-null float64
9 pol_spectrum 7056 non-null float64
10 party_id 8245 non-null float64
11 party_salience 7945 non-null float64
12 gov_trust 8243 non-null float64
13 gov_interests 8178 non-null float64
14 gov_waste 8251 non-null float64
15 gov_corrup 8209 non-null float64
16 people_trusted 8261 non-null float64
17 gov_responsive 8264 non-null float64
18 better_economy 8239 non-null float64
19 better_health 8247 non-null float64
20 better_immigratino 8246 non-null float64
21 better_taxes 8240 non-null float64
22 better_environment 8236 non-null float64
23 better_covid 8241 non-null float64
24 fav_deathpen 8113 non-null float64
25 stayhome 8186 non-null float64
26 trust_media 8260 non-null float64
27 corr_trump 8195 non-null float64
28 fav_impeach 8229 non-null float64


File Edit View Run Kernel Tabs Settings Help
project_1.ipynb
project.ipynb
anes_profile.html
Code
Python 3
Name
712 - Assignment 1.p...
anes_anova.txt
anes_profile.html
anova_results.xlsx
assignment_1.xlsx
Group Assignment.d...
Group Assignment.pdf
pa02.pdf
project_1.ipynb
project_file.xlsx
project.ipynb
sharing.docx
sharing.pdf
31 covid_reopen 8241 non-null float64
32 inc_gap 8211 non-null float64
33 gov_climate 8280 non-null int64
34 samesex 8166 non-null float64
35 transgender 7967 non-null float64
36 lgbtlaw 8161 non-null float64
37 birthright 8225 non-null float64
38 deportkids 8131 non-null float64
39 wall 8240 non-null float64
40 russianinterfere 8280 non-null int64
41 religion 8172 non-null float64
42 age 7932 non-null float64
43 marital 8224 non-null float64
44 education 8149 non-null float64
45 spouse_edu 5012 non-null float64
46 armedforces 8253 non-null float64
47 labor 8223 non-null float64
48 union 8237 non-null float64
49 ethnicity 8178 non-null float64
50 children 8219 non-null float64
51 income 7664 non-null float64
52 health 8177 non-null float64
53 getcovid 8182 non-null float64
54 satisfied 8131 non-null float64
55 vote 6029 non-null float64
56 whovoted 5899 non-null float64
57 region 8280 non-null int64
dtypes: float64(51), int64(6), object(1)
memory usage: 3.7+ MB
The results of the data structure shows that we have a lot of missing data; we will clean this up by dropping the missing/empty cells in the excel file
[6]:
# Drop all Nulls
anes_data = anes_data.dropna()


File Edit View Run Kernel Tabs Settings Help
project_1.ipynb
project.ipynb
anes_profile.html
Code
Name
712 - Assignment 1.p...
anes_anova.txt
anes_profile.html
anova_results.xlsx
assignment_1.xlsx
Group Assignment.d...
Group Assignment.pdf
pa02.pdf
project_1.ipynb
project_file.xlsx
project.ipynb
sharing.docx
sharing.pdf
anes_data.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1550 entries, 200046 to 535414
Data columns (total 58 columns):
Column Non-Null Count Dtype
0 weight_pre 1550 non-null float64
1 weight_post 1550 non-null object
2 sampvar 1550 non-null int64
3 varstrat 1550 non-null int64
4 interest_politics 1550 non-null float64
5 interest_campaign 1550 non-null int64
6 state_reg 1550 non-null float64
7 party_reg 1550 non-null float64
8 primary_voter 1550 non-null float64
9 pol_spectrum 1550 non-null float64
10 party_id 1550 non-null float64
11 party_salience 1550 non-null float64
12 gov_trust 1550 non-null float64
13 gov_interests 1550 non-null float64
14 gov_waste 1550 non-null float64
15 gov_corrup 1550 non-null float64
16 people_trusted 1550 non-null float64
17 gov_responsive 1550 non-null float64
18 better_economy 1550 non-null float64
19 better_health 1550 non-null float64
20 better_immigratino 1550 non-null float64
21 better_taxes 1550 non-null float64
22 better_environment 1550 non-null float64
23 better_covid 1550 non-null float64
24 fav_deathpen 1550 non-null float64
25 stayhome 1550 non-null float64
26 trust_media 1550 non-null float64
27 corr_trump 1550 non-null float64
28 fav_impeach 1550 non-null float64
29 fav_senacquittal 1550 non-null float64


project_1.ipynb
project.ipynb
anes_profile.html
Code
Name
712 - Assignment 1.p...
anes_anova.txt
anes_profile.html
anova_results.xlsx
assignment_1.xlsx
Group Assignment.d...
Group Assignment.pdf
pa02.pdf
project_1.ipynb
project_file.xlsx
project.ipynb
sharing.docx
sharing.pdf
After dropping the missing empty cells, we now have 1550 fully pupulated rows of voter data
[8]:
Review the columns
anes_data.columns
[8]: Index(['weight_pre', 'weight_post', 'sampvar', 'varstrat', 'interest_politics',
'interest_campaign', 'state_reg', 'party_reg', 'primary_voter',
'pol_spectrum', 'party_id', 'party_salience', 'gov_trust',
'gov_interests', 'gov_waste', 'gov_corrup', 'people_trusted',
'gov_responsive', 'better_economy', 'better_health',
'better_immigratino', 'better_taxes', 'better_environment',
'better_covid', 'fav_deathpen', 'stayhome', 'trust_media', 'corr_trump',
'fav_impeach', 'fav_senacquittal', 'covid_gov', 'covid_reopen',
'inc_gap', 'gov_climate', 'samesex', 'transgender', 'lgbtlaw',
'birthright', 'deportkids', 'wall', 'russianinterfere', 'religion',
'age', 'marital', 'education', 'spouse_edu', 'armedforces', 'labor',
'union', 'ethnicity', 'children', 'income', 'health', 'getcovid',
'satisfied', 'vote', 'whovoted', 'region'],
dtype='object')
[9]:
# Because we are investigating voters, we will put 'whovoted' in the last column
# As part of this code, we will also drop columns we do not believe add any value to our analysis; these columns are:
# 'weight_pre', 'weight_post', and 'sampvar'
anes_data = anes_data[['interest_politics', 'interest_campaign',
'state_reg', 'party_reg', 'primary_voter', 'pol_spectrum', 'party_id',
'party_salience', 'gov_trust', 'gov_interests', 'gov_waste',
'gov_corrup', 'people_trusted', 'gov_responsive', 'better_economy',
'better_health', 'better_immigratino', 'better_taxes',
'better_environment', 'better_covid', 'fav_deathpen', 'stayhome',
'trust_media', 'corr_trump', 'fav_impeach', 'fav_senacquittal',
'covid_gov', 'covid_reopen', 'inc_gap', 'gov_climate', 'samesex',
'transgender', 'lgbtlaw', 'birthright', 'deportkids', 'wall',
'russianinterfere', 'religion', 'age', 'marital', 'education',
'spouse_edu', 'armedforces', 'labor', 'union', 'ethnicity', 'children',


File Edit View Run Kernel Tabs Settings Help
project_1.ipynb
project.ipynb
anes_profile.html
Code
Name
712 - Assignment 1.p...
anes_anova.txt
anes_profile.html
anova_results.xlsx
assignment_1.xlsx
Group Assignment.d...
Group Assignment.pdf
pa02.pdf
project_1.ipynb
project_file.xlsx
project.ipynb
sharing.docx
sharing.pdf
'party_salience', 'gov_trust', 'gov_interests', 'gov_waste',
'gov_corrup', 'people_trusted', 'gov_responsive', 'better_economy',
'better_health', 'better_immigratino', 'better_taxes',
'better_environment', 'better_covid', 'fav_deathpen', 'stayhome',
'trust_media', 'corr_trump', 'fav_impeach', 'fav_senacquittal',
'covid_gov', 'covid_reopen', 'inc_gap', 'gov_climate', 'samesex',
'transgender', 'lgbtlaw', 'birthright', 'deportkids', 'wall',
'russianinterfere', 'religion', 'age', 'marital', 'education',
'spouse_edu', 'armedforces', 'labor', 'union', 'ethnicity', 'children',
'income', 'health', 'getcovid', 'satisfied', 'region', 'whovoted']]
[10]:
# check the columns to make sure the objectives are achieved
anes_data.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1550 entries, 200046 to 535414
Data columns (total 53 columns):
Column Non-Null Count Dtype
0 interest_politics 1550 non-null float64
1 interest_campaign 1550 non-null int64
2 state_reg 1550 non-null float64
3 party_reg 1550 non-null float64
4 primary_voter 1550 non-null float64
5 pol_spectrum 1550 non-null float64
6 party_id 1550 non-null float64
7 party_salience 1550 non-null float64
8 gov_trust 1550 non-null float64
9 gov_interests 1550 non-null float64
10 gov_waste 1550 non-null float64
11 gov_corrup 1550 non-null float64
12 people_trusted 1550 non-null float64
13 gov_responsive 1550 non-null float64
14 better_economy 1550 non-null float64
15 better_health 1550 non-null float64
16 better_immigratino 1550 non-null float64


[11]:
# Now we inspect the candidate options in 'whovoted' column
anes_data.whovoted.value_counts()
[11]:
1.0 883
2.0 642
5.0 11
3.0 10
12.0 2
4.0 2
Name: whovoted, dtype: int64
The result shows that 883 voted for Biden, 642 voted for Trump, and the rest voted for other candidates. However, because our purpose is to understand issues pertinent to Trump voters we will only focus on the 642 Trump voters
[12]:
Create a new dataframe for only Trump voters (642)
#iris_df[iris_df.Target==1].head()
#iris_df.loc[iris_df['Target'] == 1].head()
trump = anes_data[anes_data['whovoted'] == 2]
[65]:
trump.shape
[65]:
(642, 53)
[13]:
# inspect the structure of the newly created dataframe holding only Trump voters
trump.head()
[13]:
interest_politics interest_campaign state_reg party_reg primary_voter pol_spectrum party_id party_salience gov_trust gov_interests ... labor union ethnicity children income h
caseid
200558 1.0 1 20.0 2.0 2.0 7.0 7.0 1.0 4.0 2.0 ... 1.0 2.0 1.0 1.0 19.0

File Edit View Run Kernel Tabs Settings Help

project_1.ipynb | project.ipynb | anes_profile.html

Code

Name

712 - Assignment 1.p...
anes_anova.txt
anes_profile.html
anova_results.xlsx
assignment_1.xlsx
Group Assignment.d...
Group Assignment.pdf
pa02.pdf
project_1.ipynb
project_file.xlsx
project.ipynb
sharing.docx
sharing.pdf

[14]:
```
(642, 53)
```

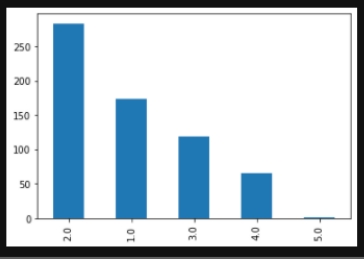
Now that we have prepared our dataset, it is time for Exploratory Data Analysis

[15]:
```
### Interest in Politics
trump.interest_politics.value_counts()
```

[15]:
```
2.0    283
1.0    173
3.0    119
4.0     66
5.0      1
Name: interest_politics, dtype: int64
```

[16]:
```
### Option 2 means 'Most of the time'. We will chart visualize this field
trump.interest_politics.value_counts().plot(kind='bar')
```

[16]:
```
<AxesSubplot:>
```



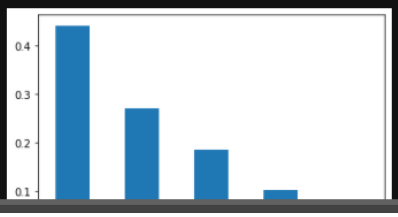
While the visuals are clear, it is better to normalize the numbers as percentages; this sends a much clearer message across the different polling categories

[17]:
```
trump.interest_politics.value_counts(normalize=True)
```

[17]:
```
2.0    0.440810
1.0    0.269470
3.0    0.185358
4.0    0.102804
5.0    0.001558
Name: interest_politics, dtype: float64
```

[18]:
```
trump.interest_politics.value_counts(normalize=True).plot(kind='bar')
```

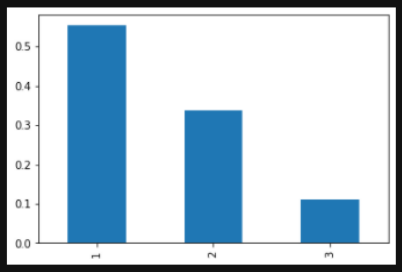
[18]:
```
<AxesSubplot:>
```



```
Name: interest_campaign, dtype: float64
```

```
[20]: trump.interest_campaign.value_counts(normalize=True).plot(kind='bar')
```

```
[20]: <AxesSubplot:>
```



More than 88% of Trump voters are somewhat or very much interested in following campaigns

```
[21]: trump.state_reg.value_counts(normalize=True).sort_values(ascending=False).head(5)
```

```
[21]: 12.0    0.127726
      6.0     0.112150
      42.0    0.077882
      37.0    0.076324
      36.0    0.059190
      Name: state_reg, dtype: float64
```

```
[22]: trump.state_reg.value_counts(normalize=True).sort_values(ascending=False).head(5).plot(kind='bar')
```

```
[22]: <AxesSubplot:>
```

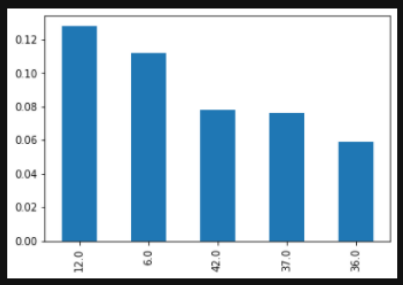
More than 88% of Trump voters are somewhat or very much interested in following campaigns

[21]: trump.state_reg.value_counts(normalize=True).sort_values(ascending=False).head(5)

[21]: 12.0 0.127726
6.0 0.112150
42.0 0.077882
37.0 0.076324
36.0 0.059190
Name: state_reg, dtype: float64

[22]: trump.state_reg.value_counts(normalize=True).sort_values(ascending=False).head(5).plot(kind='bar')

[22]: <AxesSubplot:>



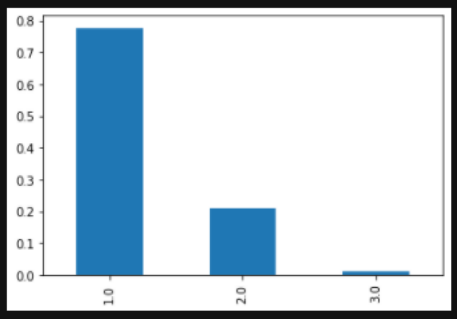
This graph shows the top five states with Trump voters (not as percentage of population). Florida = 12.8%, California = 11.2%, Pennsylvania = 7.8%, N. Carolina = 7.6% and New York = 5.9%

[23]:
```
trump.gov_waste.value_counts(normalize=True).sort_values(ascending=False)
```

[23]:
```
1.0    0.777259
2.0    0.210280
3.0    0.012461
Name: gov_waste, dtype: float64
```

[24]:
```
trump.gov_waste.value_counts(normalize=True).sort_values(ascending=False).plot(kind='bar')
```

[24]: <AxesSubplot:>



Almost all of Republican voters (~99%) of Republican voters believe that the government wastes tax dollars

[25]:
```
trump.party_reg.value_counts(normalize=True).sort_values(ascending=False)
```

[25]:
```
2.0    0.685358
4.0    0.227414
1.0    0.080997
```