Final Project, Mina Mohammadi: Is there an association between voting for Donald Trump and hate crimes?

-Pre-step

I have a deep interest in political leanings and how they affect certain communities. I have read articles about how since 2016, hate crimes in the United States have increased. I would like to see whether there is a clear association between voting for Donald Trump and hate crimes in a certain state.

-Data

I found this hate crimes dataset after browsing through the FiveThirtyEight github. I found a similar dataset on Buzzfeed News' github, but that dataset was a more generic piece regarding hate crimes and did not provide Trump-related data.

The two variables that I will be concentrating on for my analysis are the Share of 2016 U.S. presidential voters who voted for Donald Trump (share_voters_voted_trump) and the Hate crimes per 100,000 population, by Southern Poverty Law Center, Nov. 9-18, 2016 (hate_crimes_per_100k_splc). The unit of observation are U.S states.

While I recognize that this dataset was used in class, the variables that I will be specifically working to conduct my analysis are different than those used in class. I also recieved an OK to use this dataset from my TA, Angela Lai.

-Data

One part of this dataset that I would change is the limited time that the hate_crimes_per_100k_splc records. This data only records the hate crimes per 100,000 population from the dates of November 9-18th 2016. If I wanted to see the larger scope of hate crime activity beyond the immediate post-election/election activity, this data does not necessarily provide that. If it was extended to record hate crimes per 100,000 population for a full year, it would probably provide a better respresntation to analyze the association between hate crimes and the share of Donald Trump voters.

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
```

In [2]:

```
info = pd.read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/hate-crimes/hate_crimes.csv")
info.head(5)
```

Out[2]:

	state	median_household_income	share_unemployed_seasonal	share_population_in_metro_ar
0	Alabama	42278	0.060	(
1	Alaska	67629	0.064	(
2	Arizona	49254	0.063	(
3	Arkansas	44922	0.052	(
4	California	60487	0.059	(

-Data

No pre-work was required to get the data into an uploadable format.

In [3]:

```
info.describe()
```

Out[3]:

	median_household_income	share_unemployed_seasonal	share_population_in_metro_areas	s
count	51.000000	51.000000	51.000000	
mean	55223.607843	0.049569	0.750196	
std	9208.478170	0.010698	0.181587	
min	35521.000000	0.028000	0.310000	
25%	48657.000000	0.042000	0.630000	
50%	54916.000000	0.051000	0.790000	
75%	60719.000000	0.057500	0.895000	
max	76165.000000	0.073000	1.000000	

In [4]:

```
n_by_state = info.groupby("share_voters_voted_trump")["hate_crimes_per_100k_splc"].count()
n_by_state.head()
```

Out[4]:

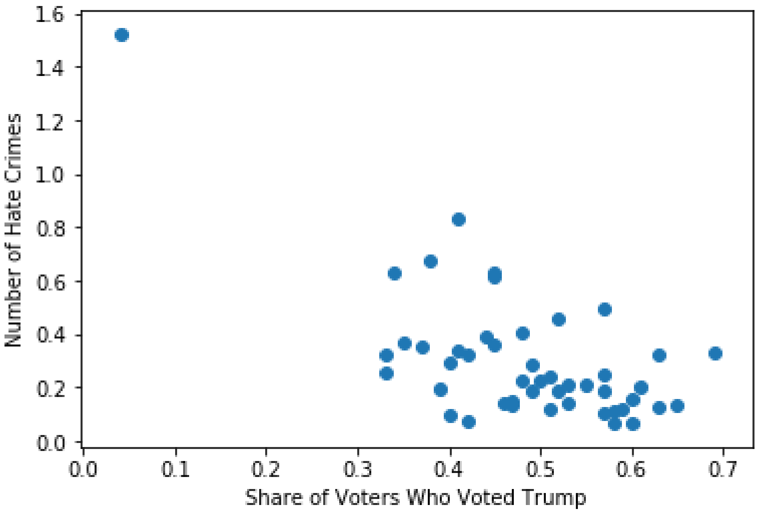
```
share_voters_voted_trump
0.04    1
0.30    0
0.33    2
0.34    1
0.35    1
Name: hate_crimes_per_100k_splc, dtype: int64
```

-Initial analysis

I chose info.describe() because descriptive statistics help to understand the basic features of the data, and can be helpful when drawing conclusions later. The second manipulation I did was to group the 2 varaibales by their counts. I did this so I would be able to concentrate on these two specific variables.

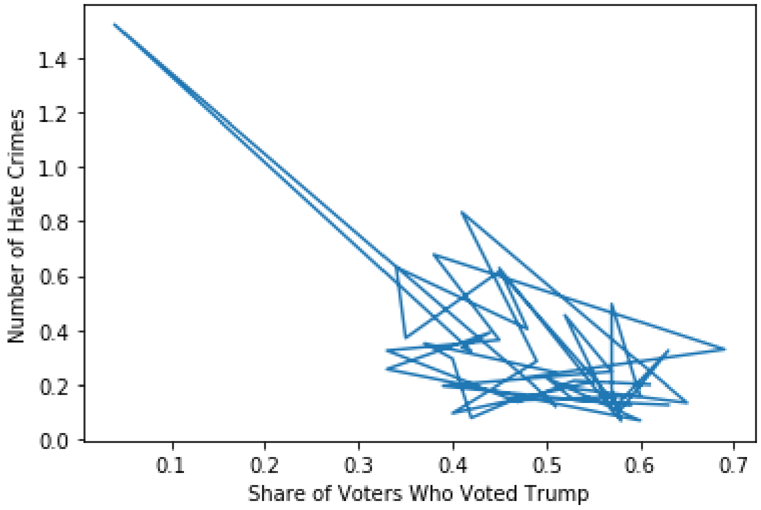
In [5]:
```
x = info.share_voters_voted_trump
y = info.hate_crimes_per_100k_splc
plt.scatter(x, y)
plt.xlabel("Share of Voters Who Voted Trump")
plt.ylabel("Number of Hate Crimes")
```

Out[5]:
```
Text(0, 0.5, 'Number of Hate Crimes')
```



```
In [6]: x = info.share_voters_voted_trump
        y = info.hate_crimes_per_100k_splc
        plt.plot(x, y)
        plt.xlabel("Share of Voters Who Voted Trump")
        plt.ylabel("Number of Hate Crimes")

Out[6]: Text(0, 0.5, 'Number of Hate Crimes')
```



Initial analysis:

I chose to use a scatterplot because they are best for visualizing assocaitions between 2 variables, and considering that these are both continous it would make sense to choose this visualization. The second visualization I chose was a line graph because it can work for both discrete and continous data. Because of the time that passes in the Number of hate crimes data, I thought that a line graph would imply the time in which that variable exists in.

I did learn that there were more data points of hate crimes in states with a larger share of voters who voted Trump. There is also one outlier. I believe that the scatterplot was more useful in my analysis as opposed to the line graph as it is more clear.

Hypothesis formation:

My dependant variable in this analysis is hate_crimes_per_100k_splc, my independent is share_voters_voted_trump. In this study, hate_crimes_per_100k_splc is measured as Hate crimes per 100,000 population, Southern Poverty Law Center, Nov. 9-18, 2016. The share_voters_voted_trump is the Share of 2016 U.S. presidential voters who voted for Donald Trump. The unit of observation is States and the data is meausred with this unit

In [7]:
```
info.corr()
-0.657067
```

Out[7]:
```
-0.657067
```

Hypothesis formation:

```
The correlation coefficient between share_voters_voted_trump and hate_crimes
_per_100k_splc is -.657 which means that there is a negative correlation bet
ween the two. As the share of voters who voted for trump goes up, the hate c
rimes per 100k in this period of time from November 9-18 2016, goes down.
```

Hypothesis formation:

$y=mx+b$
$y=hate\_crimes$
$x=\texttt{share\_voters\_voted\_trump}$
$hate\_crimei=\alpha+\beta*\texttt{share\_voters\_voted\_trump}i+ei$

Hypothesis formation:

```
Null:A larger share of voters voting for Trump in a state has no association
with higher rates of hate crimes during November 9-18th, 2016

Alternative: A larger share of voters voting for Trump in a state is associa
ted with higher rates of hate crimes during November 9-18th, 2016
```

In [8]:
```
x = info.share_voters_voted_trump
y = info.hate_crimes_per_100k_splc

X = sm.add_constant(x) #adding an intercept to the indepedent variable
model = sm.OLS(y,X, missing = 'drop') # Constructing a model
results = model.fit() # fitting the model
print(results.params)
```

```
const                        1.017347
share_voters_voted_trump    -1.474833
dtype: float64

/opt/conda/envs/dsua-111/lib/python3.7/site-packages/numpy/core/fromnum
eric.py:2542: FutureWarning: Method .ptp is deprecated and will be remo
ved in a future version. Use numpy.ptp instead.
  return ptp(axis=axis, out=out, **kwargs)
```

In [9]:
```
print(results.params[0])
```

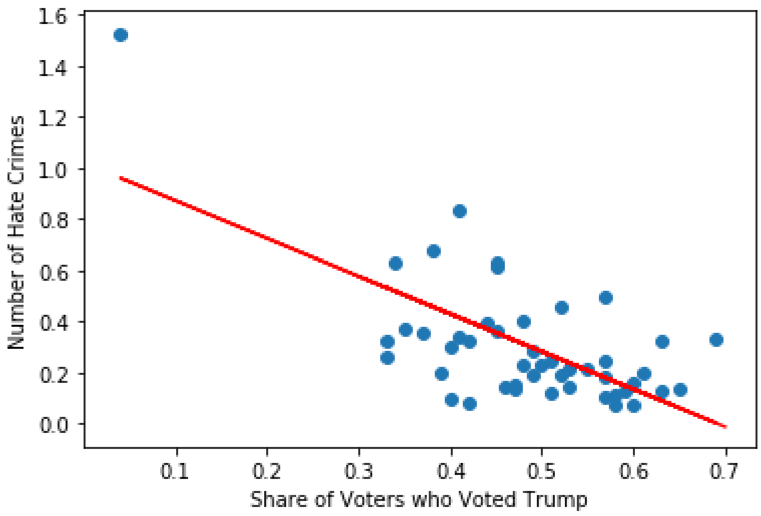
```
1.0173472768742708
```

```
In [10]: print(results.params[1])
```

```
-1.4748329267527818
```

```
In [11]: plt.plot(x,results.params[0] + x*results.params[1], color = "red")
         plt.scatter(x, y)
         plt.xlabel("Share of Voters who Voted Trump")
         plt.ylabel("Number of Hate Crimes")
```

```
Out[11]: Text(0, 0.5, 'Number of Hate Crimes')
```



```
In [12]: results.summary()
```

```
Out[12]:
OLS Regression Results

        Dep. Variable:  hate_crimes_per_100k_splc          R-squared:     0.432
               Model:                        OLS      Adj. R-squared:     0.419
              Method:              Least Squares         F-statistic:     34.19
                Date:           Wed, 06 May 2020  Prob (F-statistic): 5.26e-07
                Time:                   22:09:03      Log-Likelihood:    11.746
    No. Observations:                         47                 AIC:    -19.49
        Df Residuals:                         45                 BIC:    -15.79
            Df Model:                          1
     Covariance Type:                  nonrobust

                             coef   std err       t   P>|t|   [0.025   0.975]
                  const    1.0173     0.125   8.127   0.000    0.765    1.269
 share_voters_voted_trump -1.4748     0.252  -5.847   0.000   -1.983   -0.967

        Omnibus:  6.589     Durbin-Watson:  1.862
  Prob(Omnibus):  0.037  Jarque-Bera (JB):  5.498
           Skew:  0.788          Prob(JB): 0.0640
       Kurtosis:  3.566          Cond. No.   11.1


Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Regression analysis:

What do the results in the regression output tell you? Interpret the coefficient, p-value, and confidence interval for your independent variable (you don’t have to do the intercept) and the R2?

The results of the regression output tell us a lot about the project. As coefficent in this case was -1.4748, and a caluclated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement.The standard error is pretty large, at .252, so I definitely think there may have been some issues within this analysis.

The P-value indicates the level of statistical significance between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis. My p-value is O, so due to this I believe I should reject my null hypothesis.

In this case the confidence interval is [-1.983, -0.967]. The null 0, is not in this interval, which means I can reject my null.

The R-squared value tells me what percent of the variation x and y explain in each other. When the value is closer to zero, the less variation they explain in each other. The closer to 1, the more they explain in each other. My R-squared value is 0.432, which is pretty high. This suggests that 43% of the variation in hate crimes is explained by the share of voters who voted trump. 57% of the variation must be explained by other variables.

Regression analysis:

Which hypothesis do you reject and fail to reject, and why?

```
As mentioned before, I reject my null hypothesis which states there is no as
sociaiton with higher rates of hate crimes during November 9-18th, 2016. I m
ade this choice due to my p-value and confidence interval.

I fail to reject my alternative hypothesis. There is an association between
 the 2 variables.
```

In [13]:
```
residuals = y - (results.params[0] + x*results.params[1])
```

In [14]:
```
residuals = y - results.predict(X)
```

In [15]:
```
residuals.head()
```

Out[15]:
```
0    0.037636
1   -0.091946
2   -0.054611
3   -0.063387
4   -0.274847
dtype: float64
```

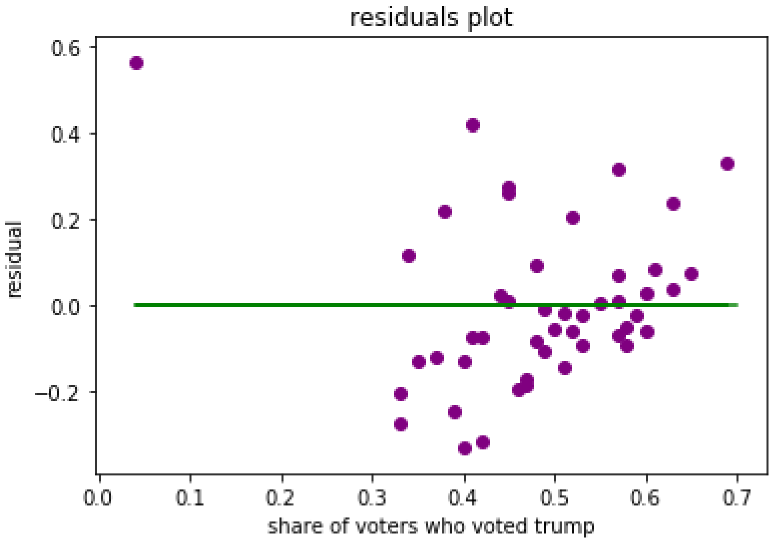
In [16]:

```
plt.scatter(x,residuals, color = "purple")
plt.title("residuals plot")
plt.xlabel("share of voters who voted trump")
plt.ylabel("residual")

plt.plot(x,[0]*len(x), color = "green")
```

Out[16]:

```
[<matplotlib.lines.Line2D at 0x7f2411dfb410>]
```



Regression analysis:

Generate the residual plot and comment on any heteroskedasticity. What does this imply for your inference?

Heteroscedasticity means that the variance of our errors (how spread out the y are) changes over the data. This data does exhibit heteroscedasticity. The variance around .5-.6 of the residuals is much higher than around .2.

Conclusions

-What biases might be present in the sample itself that could be affecting the outcome? Discuss at least two sources of bias.

As mentioned before previously as something I would change about the data is the amount of time the hate crimes data accounted for. The analysis includes only 10 days of post-election data, and most of the time is centered around and immedatley after the 2016 election. This analysis could possibly be mischaracterizing post-election sentiments as completely-Trump related sentiments, as often times violence tends to increase right after polarizing events like presidential elections. In this case, we cannot tell whether there is some obvious increase in hate crimes in the days after an election than is typical. This was also mentioned in the FiveThirtyEight article.

Other biases include what consititus a hate crime. The data originally from the Southern Poverty Law Center records both hate crimes and hate incidents under "hate crimes", which might also exaggerate the extent of hate crimes in this data.

This data is limited by its colleciton, with many of these incidents being self-reported. Much of this data is submitted voluntarily, so there may be more cases than actually seen here.

I also question why the data does not include the share of voters who voted for Hillary. It seems as though the choice to include Trump related data in a hate crime dataset and not to also include voters for Hillary makes it seem quite partisan.

Lastly, this data also includes only the share of voters who voted for Trump. Many people are not registered to vote, so in general the collection of this data may be flawed due to a large majority of people who cannot vote. The sentiment in a state for right wing ideology might be stronger than what the share of voters who voted for Trump shows.

-Considering all the work you've done, including the regression output, the results of your hypothesis tests, and any biases present in the data, what conclusions, however tentative, can you draw from your analysis about the relationship between your two variables of interest?

Because I failed to reject my alternative hypothesis, there can quite possibly be an assocaiton in a larger share of voters voting for Trump in a state and higher rates of hate crimes. I just think that the true conclusion that should be drawn is more analysis of this relationship and the various variables that influence this.

-What is your analysis's greatest weakness? In other words, what are the best reasons to be cautious about what we can learn from it?

```
I generally believe the hypothesis that A larger share of voters voting for
 Trump in a state is associated with higher rates of hate crimes during Nove
mber 9-18th, 2016, is extrememly strong and while we can't fail to reject i
t, we should be cautious about what the suggestion of this may say about a l
arge majority of American voters.
```

In [ ]: