Who Elected Trump in the 2016 Elections - the Role of Income, Race and Education

By Janhavi Agarwal

In [2]:
```
# importing libraries

import geopandas as gpd
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.iolib.summary2 import summary_col

from shapely.geometry import Point

%matplotlib inline
import qeds
qeds.themes.mpl_style();
import warnings
warnings.filterwarnings('ignore')

from bokeh.io import output_notebook
from bokeh.plotting import figure, ColumnDataSource
from bokeh.io import output_notebook, show, output_file
from bokeh.plotting import figure
from bokeh.models import GeoJSONDataSource, LinearColorMapper, ColorBar, HoverTool
from bokeh.palettes import brewer
output_notebook()
import json
```

BokehJS 2.1.1 successfully loaded.

Introduction

The US presidential are always tumultuous and influenced by a variety of factors. In 2016, the world saw Donald Trump's unexpected win. In this paper, I will analyze the relation between several factors such as education levels (the percentage of population that has attained a Bachelor's degree), share of population that is White and median household income, and their share of Republican voters at a county level. I think that this is an important research topic since the world's affairs are dictated by who is in the White House and it is important to get to root of the people that determine that. I think this adds to the existing research since it will be a stepping stone to understanding and analyzing the 2020 US elections which recently took place.

Throughout the paper, I use scatterplots and maps to observe these relationships. For the income data, I use web-scraping from a WikiPedia page containing data on median household income. In the end, I will use regressions to reach specific numbers which we can use to describe the relationship.

Research Question: Is there a correlation between education, race, income, and the share of Republican voters at a county or state level, and if so, what is the relation?

I do think that there will be a positive relation between share of White population and share of Republican voters since Republican ideologies on issues relating to gun possession, abortion laws, taxation policies and much more align with the majority of the population, and most of the US counties are inhabited by white people (as we will see in a histogram). In addition, I think that there will be a negative relation between the percentage of county population that has attained a Bachelor's degree and the share of Republican voters. This is because, people that have attained a university degree tend to be more liberal in their thought, and liberalism is a Democratic symbol. When it comes to income, I believe that there will be a negative relationship between median household income in each state and that state's share of Rpeublican voters.

The X variables we will be working with are "Share of White population", "Share of College-educated population" and "Median Household Income". The Y variable is the level of Republican voters.

These are some of the previous research papers I have looked at while writing this paper:

- Gelman, A., Kenworthy, L., & Su, Y. (2010). Income Inequality and Partisan Voting in the United States*. Social Science Quarterly, 91(5), 1203-1219. doi:10.1111/j.1540-6237.2010.00728.x
- Gonino, L. (fall 2017). Blacks' and Whites' Attitudes toward Race-Based Policies: Is there an Obama Effect? Michigan Sociological Review, 31, 173-188.
- Sides, J., Tesler, M., & Vavreck, L. (September 2016). The Electoral Landscape of 2016. The Annals of the American Academy of Political and Social Science, 667, elections in america, 50-71.

Project 1

Data Cleaning Process

The first step to analyzing the data is cleaning the data so that we have a workable DataFrame ready for analysis. On cleaning, the DataFrame we will be using for this part of the project is `county_char`.

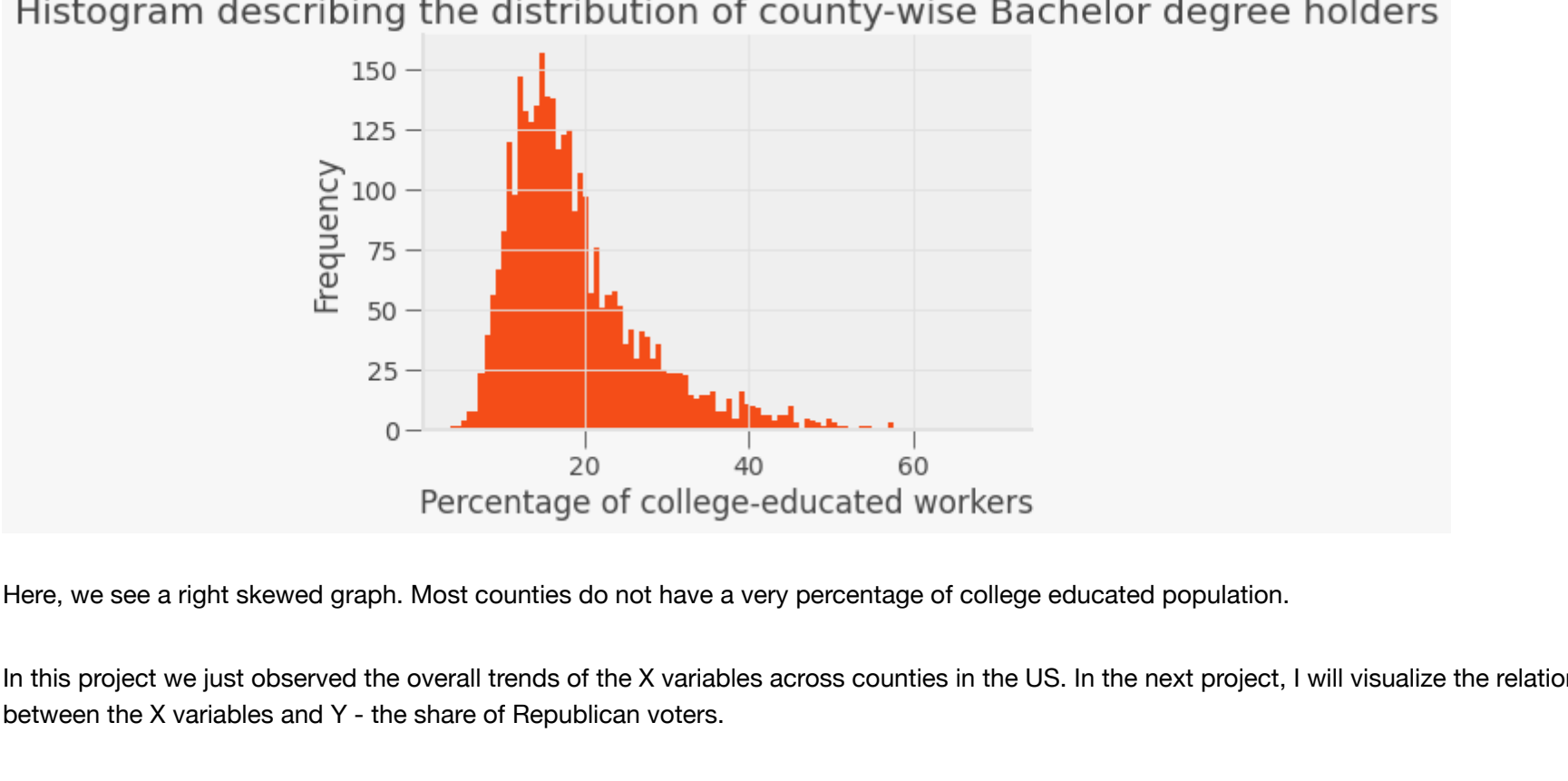
In [3]:
```
county_characteristics = pd.read_csv("/Users/janhavi/Desktop/eco225/project1/usa-2016-presidential-elec
tion-by-county.csv",
                                      sep=';')
county_char = county_characteristics[["State", "County", "Votes", "Republicans 2016", "Democrats 2016",

                                      "White (Not Latino) Population", "African American Population",
                                      "Native American Population", "Asian American Population",
                                      "Latino Population", "At Least Bachelors's Degree"]]
county_char["Other races"] = 100 - county_char["White (Not Latino) Population"]
county_char = county_char[["State", "County", "Votes", "Republicans 2016", "Democrats 2016",
                          "White (Not Latino) Population", "Other races", "At Least Bachelors's Degre
e"]]
county_char.rename(columns = {"White (Not Latino) Population": "White",
                              "At Least Bachelors's Degree":"Bachelor's Degree",
                             "Republicans 2016":"Republicans share",
                             "Democrats 2016": "Democrats share"}, inplace=True)
county_char = county_char.dropna()
county_char.head()
```

Out[3]:

	State	County	Votes	Republicans share	Democrats share	White	Other races	Bachelor's Degree
1	Nebraska	Red Willow County, Nebraska	5061.0	83.224659	12.606204	94.15	5.85	21.6
2	California	Del Norte County, California	9361.0	54.844568	37.228929	66.05	33.95	14.3
3	Florida	Duval County, Florida	428734.0	48.995648	47.494950	58.00	42.00	24.9
4	Kansas	Ellsworth County, Kansas	2626.0	73.610053	19.611577	88.90	11.10	19.8
5	Kentucky	Cumberland County, Kentucky	3066.0	81.604697	14.970646	94.70	5.30	7.9

Plotting a Histogram for the share of White population

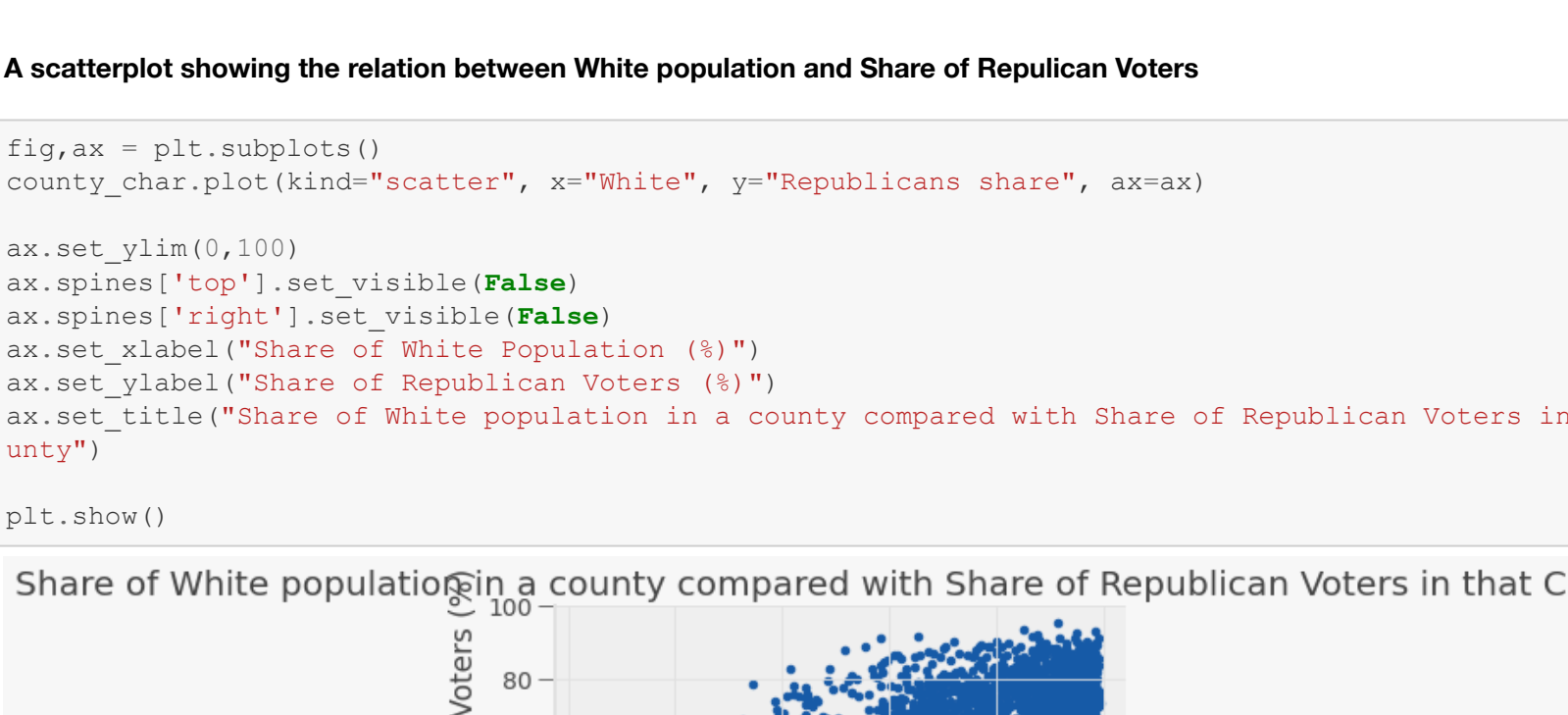
In [4]:
```
fig,ax = plt.subplots()

county_char.plot(kind="hist", y="White", color=(244/255, 77/255, 24/255),
                 bins=100, legend=False, density=False, ax=ax)

ax.set_xlabel("Percentage of Republican voters")
ax.set_title("Histogram describing the distribution of county-wise Republican voting percentages")
ax.spines['right'].set_visible(False)
ax.spines['top'].set_visible(False)
```



We see a heavily left-skewed histogram which means that most counties do tend to have a very high percentage of White population.

Plotting a Histogram for the share of population that holds a Bachelor's degree

In [5]:
```
fig,ax = plt.subplots()

county_char.plot(kind="hist", y="Bachelor's Degree", color=(244/255, 77/255, 24/255),
                 bins=100, legend=False, density=False, ax=ax)

ax.set_xlabel("Percentage of college-educated workers")
ax.set_title("Histogram describing the distribution of county-wise Bachelor degree holders")
ax.spines['right'].set_visible(False)
ax.spines['top'].set_visible(False)
```



Here, we see a right skewed graph. Most counties do not have a very percentage of college educated population.

In this project we just observed the overall trends of the X variables across counties in the US. In the next project, I will visualize the relation between the X variables and Y - the share of Republican voters.

Project 2

In this project I want to see what kind of relationship there exists between the X variables (Share of White population, share of college-educated population) and the Y-variable (Share of Republican voters) through visualizations.

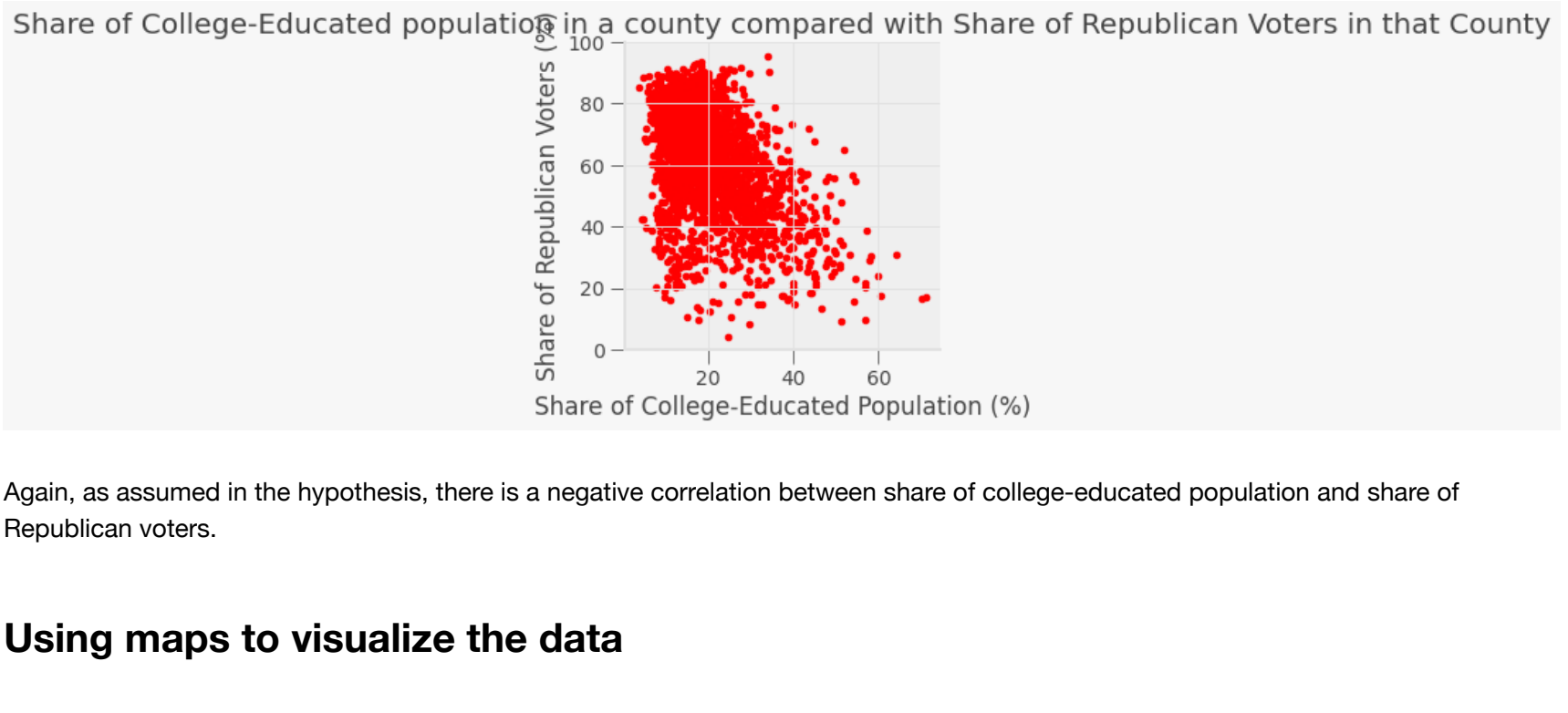
The Message

There should be a relationship between the share of White population, college educated population and the share of Republican voters in the respective counties.

Visualizing the Data through Scatterplots

A scatterplot showing the relation between White population and Share of Repulican Voters

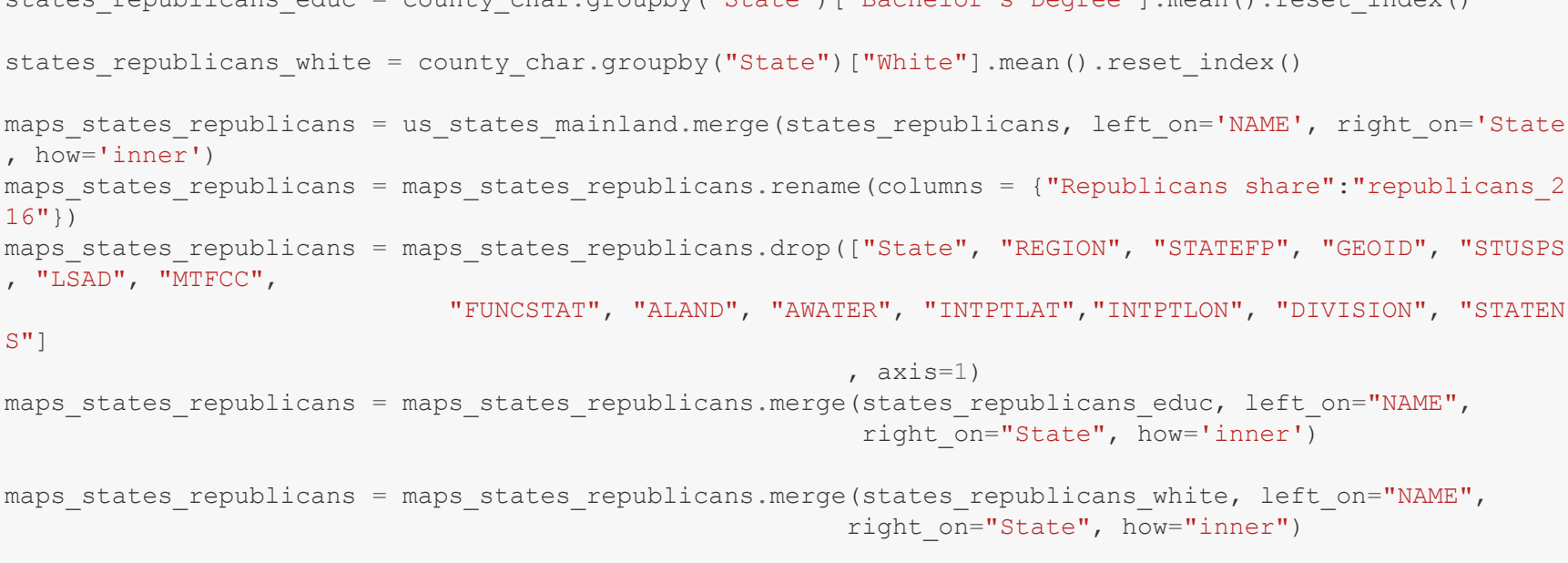
In [6]:
```
fig,ax = plt.subplots()
county_char.plot(kind="scatter", x="White", y="Republicans share", ax=ax)

ax.set_ylim(0,100)
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)
ax.set_xlabel("Share of White Population (%)")
ax.set_ylabel("Share of Republican Voters (%)")
ax.set_title("Share of White population in a county compared with Share of Republican Voters in that Co
unty")

plt.show()
```



In this scatterplot, we see that our hypothesis about share of White population and share of Republican voters having a positive relation is true. I will now compare the share of College educated share of population and share of Republican voters.

A scatterplot showing the relation between College-educated population and Share of Repulican Voters

In [7]:
```
fig,ax = plt.subplots(figsize=(4,4))
county_char.plot(kind="scatter", x="Bachelor's Degree", y="Republicans share", ax=ax, color='red')


ax.set_ylim(0,100)
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)
ax.set_xlabel("Share of College-Educated Population (%)")
ax.set_ylabel("Share of Republican Voters (%)")
ax.set_title("Share of College-Educated population in a county compared with Share of Republican Voters
 in that County")

plt.show()
```



Again, as assumed in the hypothesis, there is a negative correlation between share of college-educated population and share of Republican voters.

Using maps to visualize the data

We look at the correlation between our variables at a state level to make the visualizations easier.

The first step is to make a GeoDataFrame that consists of the geometries of each state. I read a shapefile and then merge the DataFrame with the respective columns from `county_char`. I take the mean of the percentage of people with Bachelor's degrees and the percentage of White population across the counties in each state. The GeoDataFrame is then cleaned and saved under the name `maps_states_republicans`.

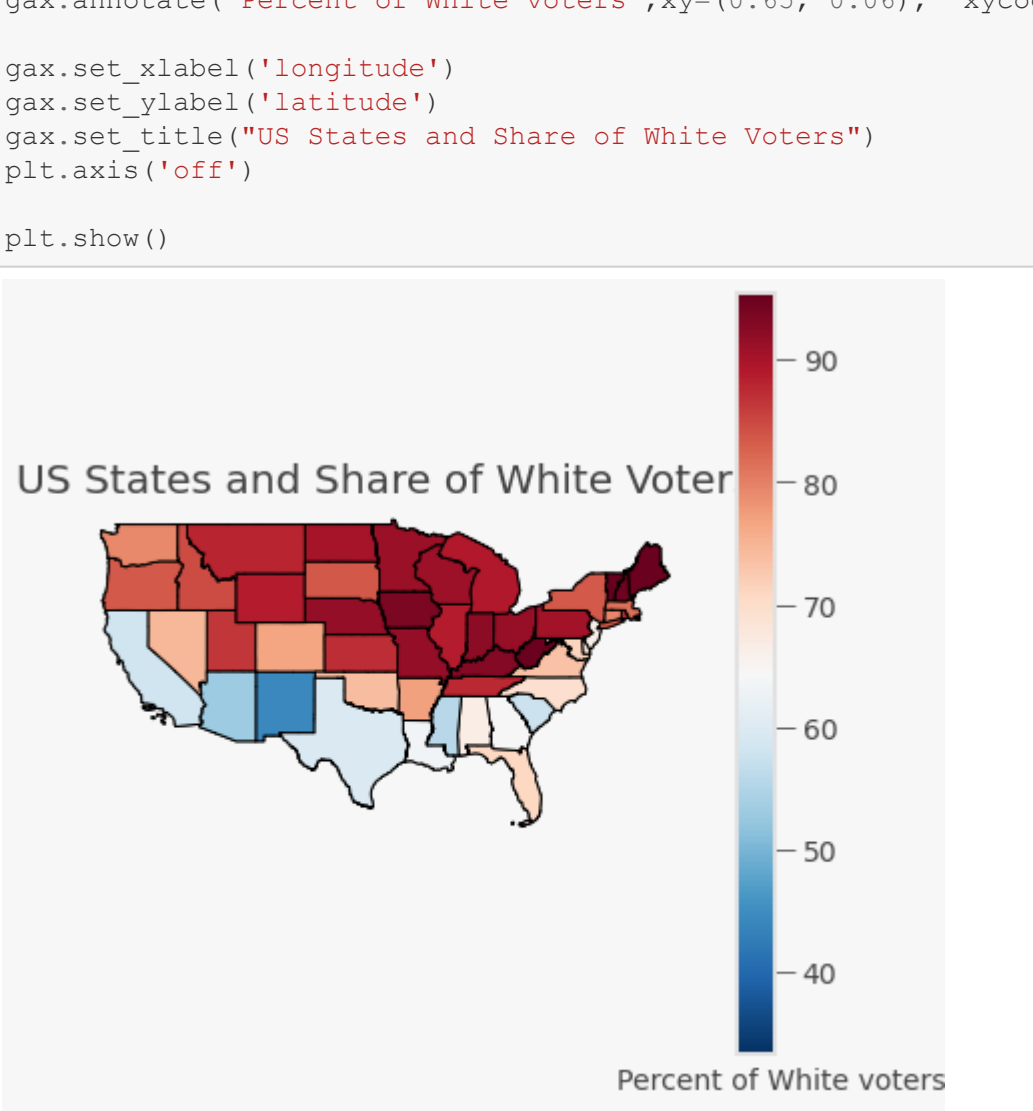
In [8]:
```
us_states = gpd.read_file('/Users/janhavi/Downloads/tl_2017_us_state/tl_2017_us_state.shp')
us_states_mainland = us_states.drop([31, 34, 35, 36, 41, 49, 40])
us_states_mainland.head()

states_republicans = county_char.groupby("State")["Republicans share"].mean().reset_index()
states_republicans = states_republicans.sort_values("Republicans share", ascending=True).reset_index().
drop(columns="index")

states_republicans_educ = county_char.groupby("State")["Bachelor's Degree"].mean().reset_index()

states_republicans_white = county_char.groupby("State")["White"].mean().reset_index()

maps_states_republicans = us_states_mainland.merge(states_republicans, left_on='NAME', right_on='State'
, how='inner')
maps_states_republicans = maps_states_republicans.rename(columns = {"Republicans share":"republicans_20
16"})
maps_states_republicans = maps_states_republicans.drop(["State", "REGION", "STATEFP", "GEOID", "STUSPS"
, "LSAD", "MTFCC",
                          "FUNCSTAT", "ALAND", "AWATER", "INTPTLAT","INTPTLON", "DIVISION", "STATEN
S"]
                                                       , axis=1)
maps_states_republicans = maps_states_republicans.merge(states_republicans_educ, left_on="NAME",
                                                        right_on="State", how='inner')

maps_states_republicans = maps_states_republicans.merge(states_republicans_white, left_on="NAME",
                                                       right_on="State", how="inner")

maps_states_republicans = maps_states_republicans.drop(["State_x", "State_y"], axis=1)

maps_states_republicans = maps_states_republicans.rename(columns = {"Bachelor's Degree":"bachelors",
                                                                   "White":"white", "NAME":"State"})

maps_states_republicans.head()
```

Out[8]:

	State	geometry	republicans_2016	bachelors	white
0	West Virginia	POLYGON ((-81.74725 39.09538, -81.74635 39.096...	73.095301	13.794545	95.169091
1	Florida	MULTIPOLYGON (((-82.98748 24.62538, -82.98748 ...	62.025003	19.407463	70.825373
2	Illinois	POLYGON ((-91.18529 40.63780, -91.17510 40.643...	63.425538	18.777451	88.791667
3	Minnesota	POLYGON ((-96.78438 46.63050, -96.78434 46.630...	58.359900	20.748276	91.134483
4	Maryland	POLYGON ((-77.45881 39.22027, -77.45866 39.220...	51.185476	28.454167	71.277083

Visualizing Share of Republican Voters

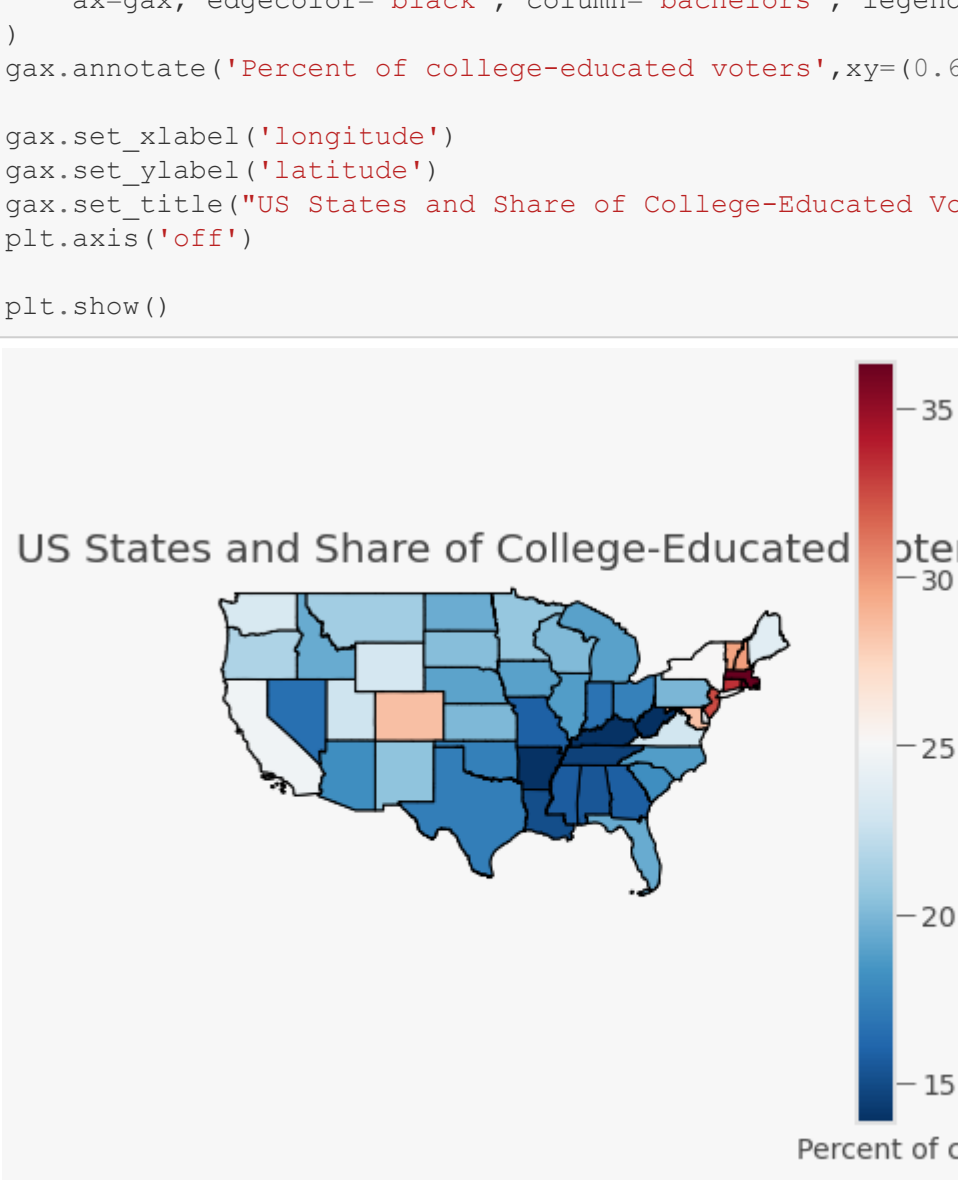
In [9]:
```
fig, gax = plt.subplots(figsize=(7,7))

maps_states_republicans.plot(ax=gax, edgecolor='black', color='white')

maps_states_republicans.plot(
    ax=gax, edgecolor='black', column='republicans_2016', legend=True, cmap='RdBu_r',
)
gax.annotate('Percent of Republican voters',xy=(0.60, 0.06),  xycoords='figure fraction')

gax.set_xlabel('longitude')
gax.set_ylabel('latitude')
gax.set_title("US States and Share of Republican Voters")
plt.axis('off')

plt.show()
```



From the map, we see that states with a higher Republican share of voters are concentrated in the middle of the country, while states with lower Republican population are found along the coasts. We will now see if this holds with our hypothesis from above.

Visualizing Share of White population

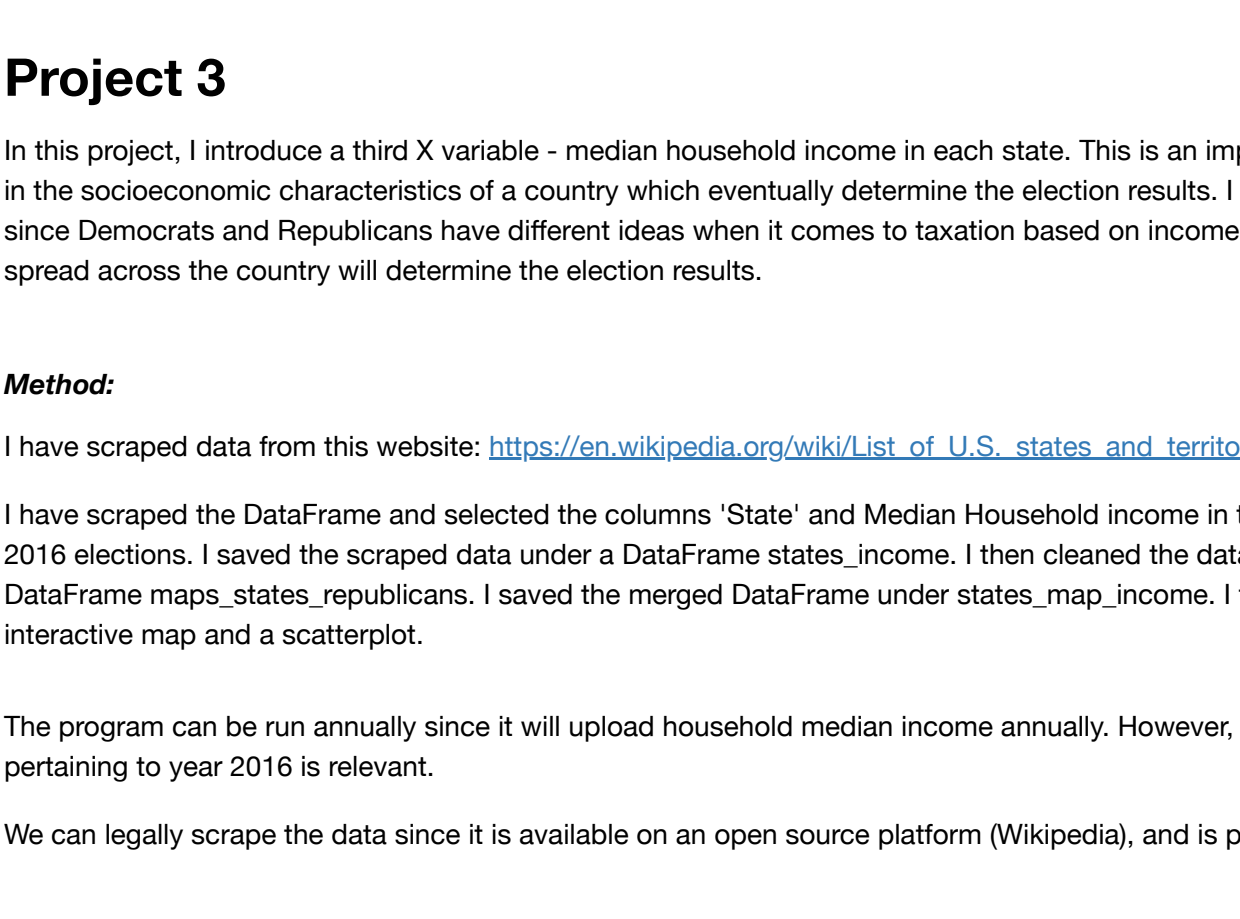
In [10]:
```
fig, gax = plt.subplots(figsize=(7,7))

maps_states_republicans.plot(ax=gax, edgecolor='black', color='white')

maps_states_republicans.plot(
    ax=gax, edgecolor='black', column='white', legend=True, cmap='RdBu_r',
)
gax.annotate('Percent of White voters',xy=(0.65, 0.06),  xycoords='figure fraction')

gax.set_xlabel('longitude')
gax.set_ylabel('latitude')
gax.set_title("US States and Share of White Voters")
plt.axis('off')

plt.show()
```



We see a higher concentration of White population in the north of the country while a lower concentration of White population in the south. A potential reason for this is that the US borders Mexico in the south and thus could have a higher immigrant population. Our pre-determined hypothesis holds for most states.

Visualizing Share of College-Educated population across the states

In [11]:
```
fig, gax = plt.subplots(figsize=(7,7))

maps_states_republicans.plot(ax=gax, edgecolor='black', color='white')

maps_states_republicans.plot(
    ax=gax, edgecolor='black', column='bachelors', legend=True, cmap='RdBu_r',
)
gax.annotate('Percent of college-educated voters',xy=(0.65, 0.06),  xycoords='figure fraction')

gax.set_xlabel('longitude')
gax.set_ylabel('latitude')
gax.set_title("US States and Share of College-Educated Voters")
plt.axis('off')

plt.show()
```



If we compare this map with the map showing the share of Republican voters, we see that the maps look like exact reverses of each other. This affirms our hypothesis.

Conclusion

The research question of this project was to see whether there was a relationship between education, race and share of Republican voters. Our hypothesis assumed that there is a negative relationship between share of population that has attained a Bachelor's degree and share of Republican voters, and positive relationship between share of White population and share of Republican voters. We used scatterplots to share this relation at a county level, and maps to show this relation at the state level.

Our hypothesis was proven right but we do see a stronger relationship between education levels and share of Republican voters.

Project 3

In this project, I introduce a third X variable - median household income in each state. This is an important addition since it plays a big role in the socioeconomic characteristics of a country which eventually determine the election results. I this particular addition also matters since Democrats and Republicans have different ideas when it comes to taxation based on income. The level of income, and how it is spread across the country will determine the election results.

Method:

I have scraped data from this website: https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income

I have scraped the DataFrame and selected the columns 'State' and Median Household income in the year 2016 since I am analyzing the 2016 elections. I saved the scraped data under a DataFrame states_income. I then cleaned the data and merged it with my original DataFrame maps_states_republicans. I saved the merged DataFrame under states_map_income. I then analyze the DataFrame using an interactive map and a scatterplot.

The program can be run annually since it will upload household median income annually. However, for this particular project, only the data pertaining to year 2016 is relevant.

We can legally scrape the data since it is available on an open source platform (Wikipedia), and is publically available.

Web Scraping and Data Cleaning Process

In this code, I scrape the data and store it in a new DataFrame called `states_income`. I then merge this with the original DataFrame `maps_states_republicans`. The new DataFrame is called `income_map_states` and view the first 5 rows.

In [12]:
```
import requests
import pandas as pd
from bs4 import BeautifulSoup

url = 'https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income'
response = requests.get(url)

soup = BeautifulSoup(response.content)
data_table = soup.find('table','wikitable sortable')
all_values = data_table.find_all('tr')

income = pd.DataFrame(columns = ["State", "Income"])
ix=0

for row in all_values[1:]:
    values = row.find_all('td')
    State = values[1].text.strip('\n')
    Income = values[4].text.strip('\n')

    income.loc[ix] = [State, Income]
    ix += 1

states_income = income.drop([0, 8, 20, 47, 53, 54, 55, 56]).reset_index()
states_income = states_income.drop("index", axis=1)
states_income["State"] = states_income["State"].astype('string')
states_income["State"] = states_income["State"].str.strip()

income_map_states = pd.merge(maps_states_republicans, states_income, left_on='State', right_on="State",
 how='inner')

income_map_states["Income"] = income_map_states["Income"].str.replace("$","")
income_map_states["Income"] = income_map_states["Income"].str.replace(",","")
income_map_states["Income"] = pd.to_numeric(income_map_states["Income"])
income_map_states.head()
```

Out[12]:

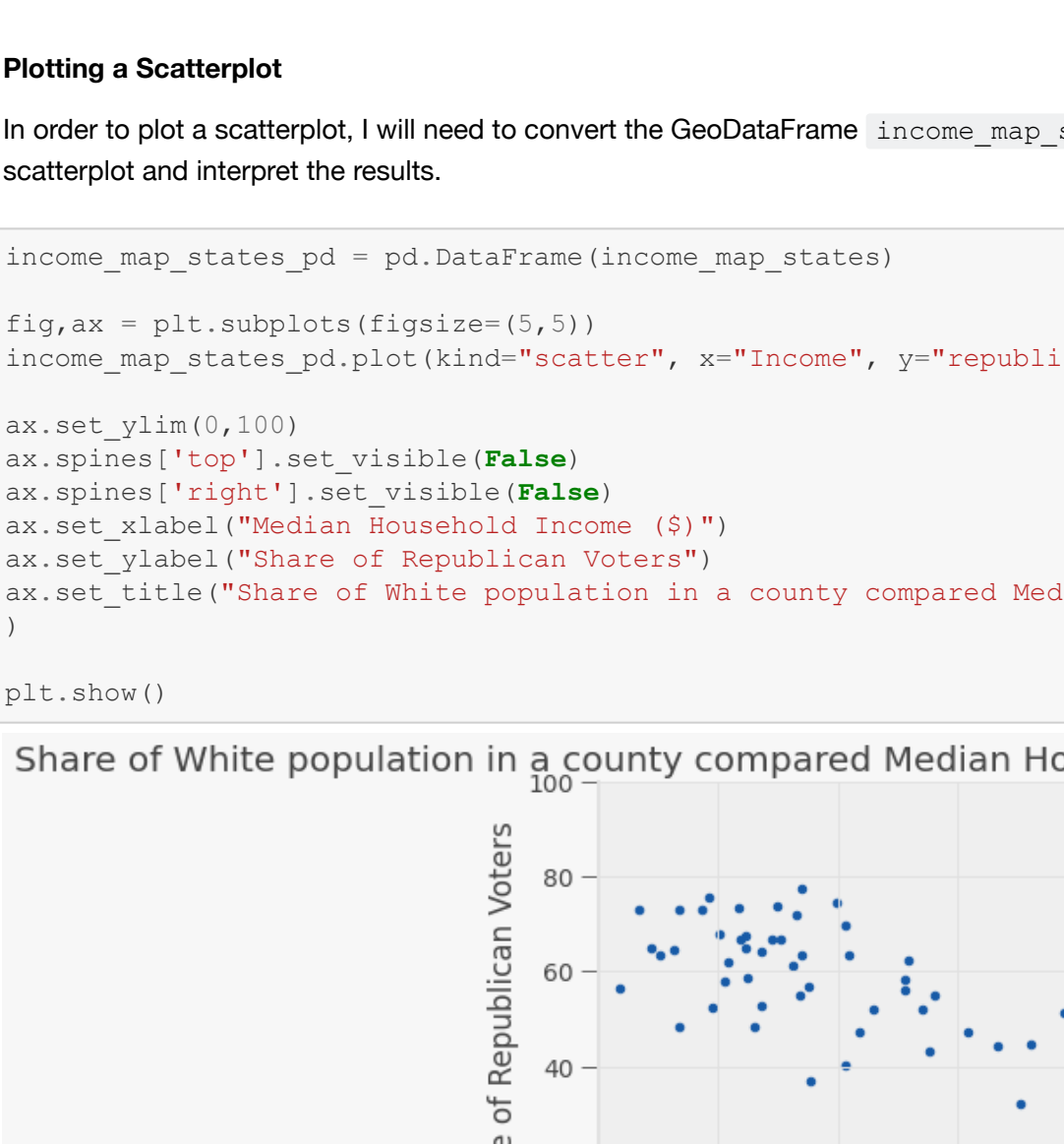
	State	geometry	republicans_2016	bachelors	white	Income
0	West Virginia	POLYGON ((-81.74725 39.09538, -81.74635 39.096...	73.095301	13.794545	95.169091	43385
1	Florida	MULTIPOLYGON (((-82.98748 24.62538, -82.98748 ...	62.025003	19.407463	70.825373	50860
2	Illinois	POLYGON ((-91.18529 40.63780, -91.17510 40.643...	63.425538	18.777451	88.791667	60960
3	Minnesota	POLYGON ((-96.78438 46.63050, -96.78434 46.630...	58.359900	20.748276	91.134483	65599
4	Maryland	POLYGON ((-77.45881 39.22027, -77.45866 39.220...	51.185476	28.454167	71.277083	78945

Visualizing Income Data

Plotting a map

As I did for education and race, I will plot a map which shows the median household income among the states.

In [13]:
```
fig, gax = plt.subplots(figsize=(7,7))

income_map_states.plot(ax=gax, edgecolor='black', color='white')

income_map_states.plot(
    ax=gax, edgecolor='black', column='Income', legend=True, cmap='RdBu',
)

gax.set_xlabel('longitude')
gax.set_ylabel('latitude')
gax.set_title("US States and their Median Household Income ($)")
plt.axis('on')
gax.annotate('Median Household Income',xy=(0.6, 0.06),  xycoords='figure fraction')

plt.show()
```



The states with higher income correspond to blue colored states in the map while the states with lower income correspond to states which are red in the map. On comparing this map with the map showing the share of Republican voters across the states, we see that states with a higher income correspond to states with a lower share of Republican voters and vice versa. Therefore we hypothesize that there will be a negative relation between median income and share of Republican voters.

Why do states along the coast have higher median household incomes?

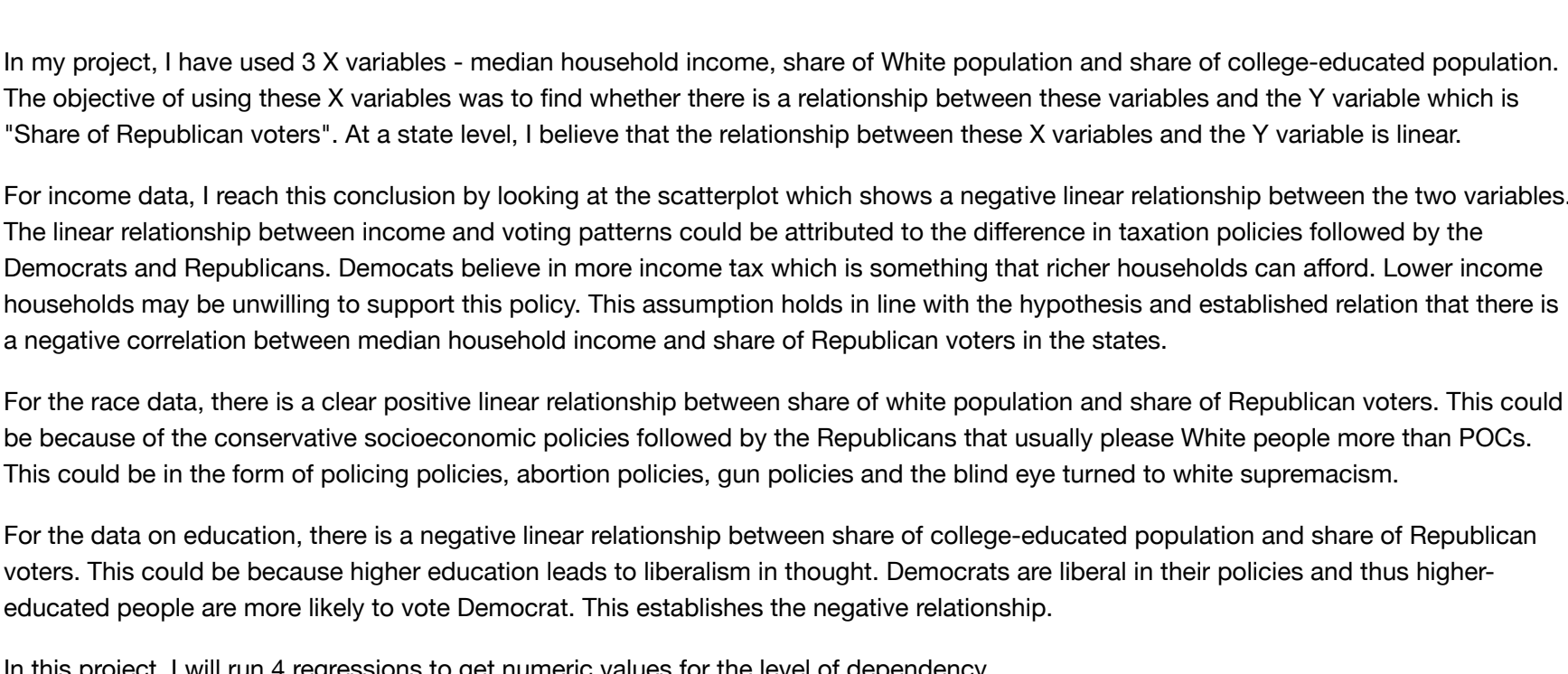
This is due to the types of occupations that prevail in these states. Rich states like Massachusetts have mostly private sector based industries which contributes to the higher household income. On the other hand, states like Mississipi rely on agriculture and federal jobs for employment leading to lower household incomes.

Source: https://www.investopedia.com/median-income-by-state-5070640

Plotting a Scatterplot

In order to plot a scatterplot, I will need to convert the GeoDataFrame `income_map_states` into a Pandas DataFrame. I then plot the scatterplot and interpret the results.

In [14]:
```
income_map_states_pd = pd.DataFrame(income_map_states)

fig,ax = plt.subplots(figsize=(5,5))
income_map_states_pd.plot(kind="scatter", x="Income", y="republicans_2016", ax=ax)

ax.set_ylim(0,100)
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)
ax.set_xlabel("Median Household Income ($)")
ax.set_ylabel("Share of Republican Voters")
ax.set_title("Share of White population in a county compared Median Household Income across the States"
)

plt.show()
```



We see a negative correlation that is not very strong. Thus states with a higher median household income tend to have a lower share of Republican voters.

Project 3 Conclusion

The objective of this project was to find out whether there exists a relationship between Median household income of a state and its share of Republican voters. On plotting a map and scatterplot, we see that there exists a negative relationship between the two.

The negative correlation is suprising since the Democrats have always perceived themselves as a more socialist party while the Republicans portray themselves as more capitalistic. According to a study 'Rich State, Poor State, Red State, Blue State: What's the Matter with Connecticut?' in the Quarterly Journal of Political Science, 2007, 2: 345–36, income plays a large role in the voting preferences of Red states but little-to-no role in Blue states. In Red states, rich people vote Republican due to their lenient taxation policies.

Another potential reason for the negative relation is that democrats tend to emphasize taxation policies that include taxing income more heavily. This may not impact a rich family as much as it will impact a low-income family. Thus there is a negative relationship.

Final Project

In my project, I have used 3 X variables - median household income, share of White population and share of college-educated population. The objective of using these X variables was to find whether there is a relationship between these variables and the Y variable which is "Share of Republican voters". At a state level, I believe that the relationship between these X variables and the Y variable is linear.

For income data, I reach this conclusion by looking at the scatterplot which shows a negative linear relationship between the two variables. The linear relationship between income and voting patterns could be attributed to the difference in taxation policies followed by the Democrats and Republicans. Democats believe in more income tax which is something that richer households can afford. Lower income households may be unwilling to support this policy. This assumption holds in line with the hypothesis and established relation that there is a negative correlation between median household income and share of Republican voters in the states.

For the race data, there is a clear positive linear relationship between share of white population and share of Republican voters. This could be because of the conservative socioeconomic policies followed by the Republicans that usually please White people more than POCs. This could be in the form of policing policies, abortion policies, gun policies and the blind eye turned to white supremacism.

For the data on education, there is a negative linear relationship between share of college-educated population and share of Republican voters. This could be because higher education leads to liberalism in thought. Democrats are liberal in their policies and thus higher-educated people are more likely to vote Democrat. This establishes the negative relationship.

In this project, I will run 4 regressions to get numeric values for the level of dependency.

Choosing X's

As mentioned above, the three X's I have chosen are income, race and education level. I will first plot scatterplots along with their regression lines before running the regressions.

The first X: Share of White population

We will plot a regression line on a scatterplot that shows share of White population and share of Republican voters.

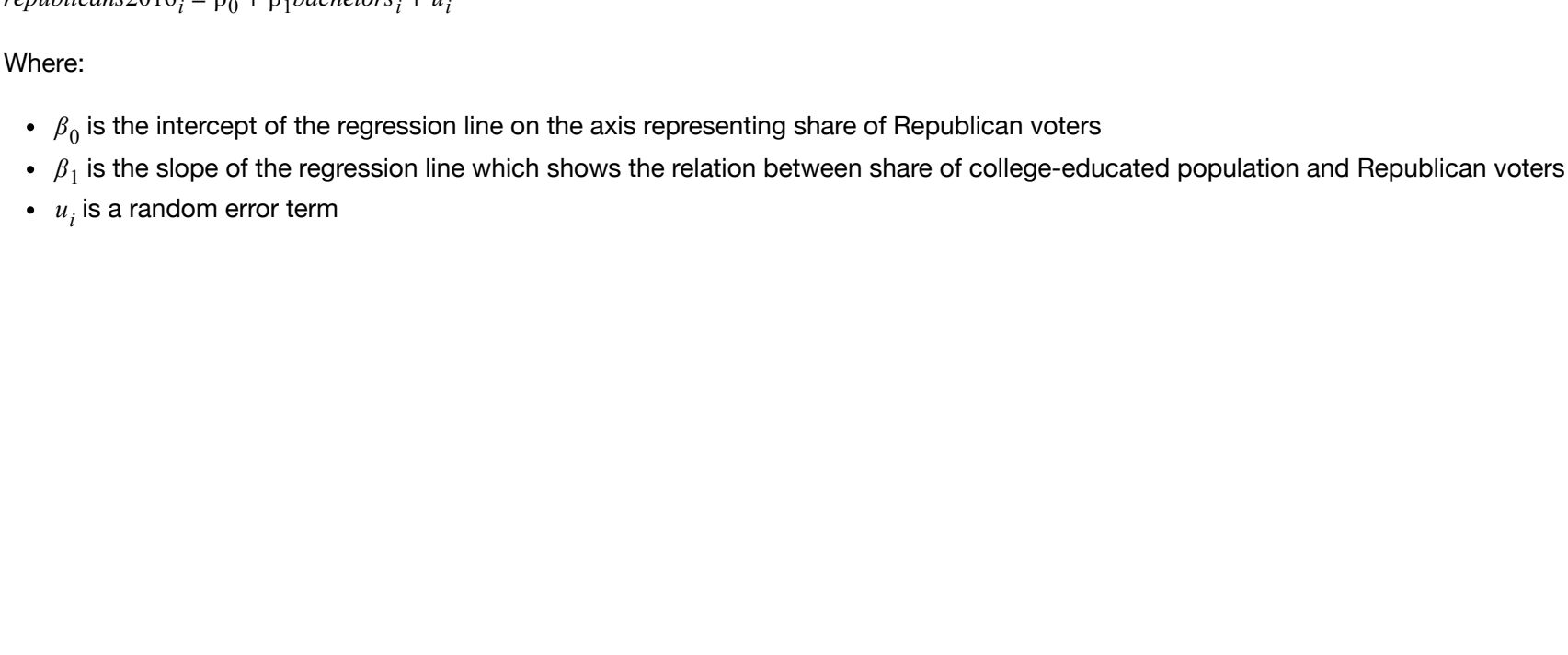
We can write our model as:

$republicans2016_i = \beta_0 + \beta_1 white_i + u_i$

Where:

- β_0 is the intercept of the regression line on the axis representing share of Republican voters
- β_1 is the slope of the regression line which shows the relation between share of White population and Republican voters
- u_i is a random error term

In [15]:
```
from sklearn.linear_model import LinearRegression

X = income_map_states_pd["white"].values.reshape(-1,1)
Y = income_map_states_pd["republicans_2016"].values.reshape(-1,1)
labels = income_map_states_pd["State"]

fig,ax = plt.subplots(figsize=(6,6))
income_map_states_pd.plot(kind="scatter", x="white", y="republicans_2016", ax=ax)

lr=LinearRegression()
lr.fit(X,Y)

x = np.linspace(0.0, 100.0).reshape(-1,1)
y_pred = lr.predict(x)
ax.plot(x, y_pred, color='blue')

ax.set_title("Regression line showing the relationship between share of White population and Republican
 Voters")

plt.show()
```



The second X: Share of college-educated population

We will plot a regression line on a scatterplot that shows share of White population and share of Republican voters.

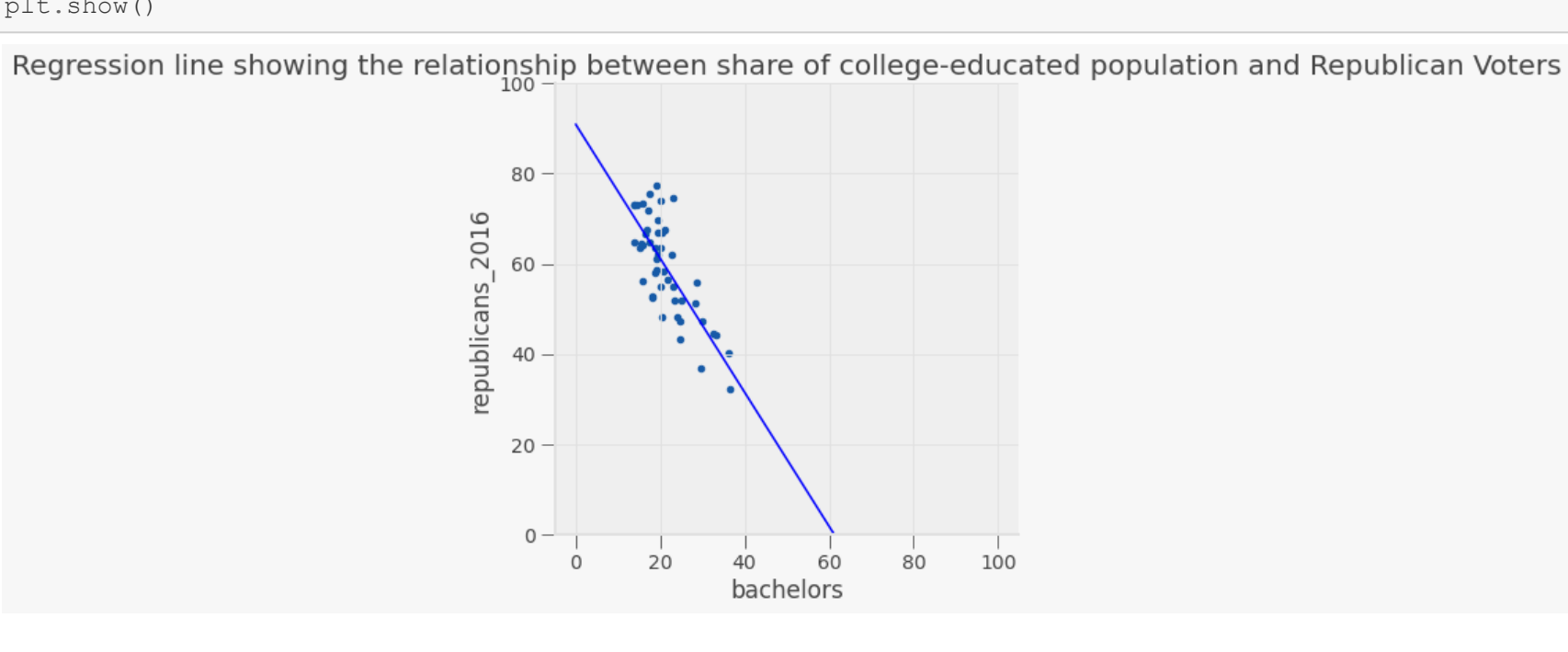
We can write our model as:

$republicans2016_i = \beta_0 + \beta_1 bachelors_i + u_i$

Where:

- β_0 is the intercept of the regression line on the axis representing share of Republican voters
- β_1 is the slope of the regression line which shows the relation between share of college-educated population and Republican voters
- u_i is a random error term

```
In [16]: from sklearn.linear_model import LinearRegression

X = income_map_states_pd["bachelors"].values.reshape(-1,1)
Y = income_map_states_pd["republicans_2016"].values.reshape(-1,1)
labels = income_map_states_pd["State"]

fig,ax = plt.subplots(figsize=(6,6))
income_map_states_pd.plot(kind="scatter", x="bachelors", y="republicans_2016", ax=ax)

lr=LinearRegression()
lr.fit(X,Y)

x = np.linspace(0.0, 100.0).reshape(-1,1)
y_pred = lr.predict(x)
ax.plot(x, y_pred, color='blue')

ax.set_ylim(0,100)
ax.set_title("Regression line showing the relationship between share of college-educated population and
 Republican Voters")

plt.show()
```



The third X: Median Household Income

We will plot a regression line on a scatterplot that shows share of White population and share of Republican voters.

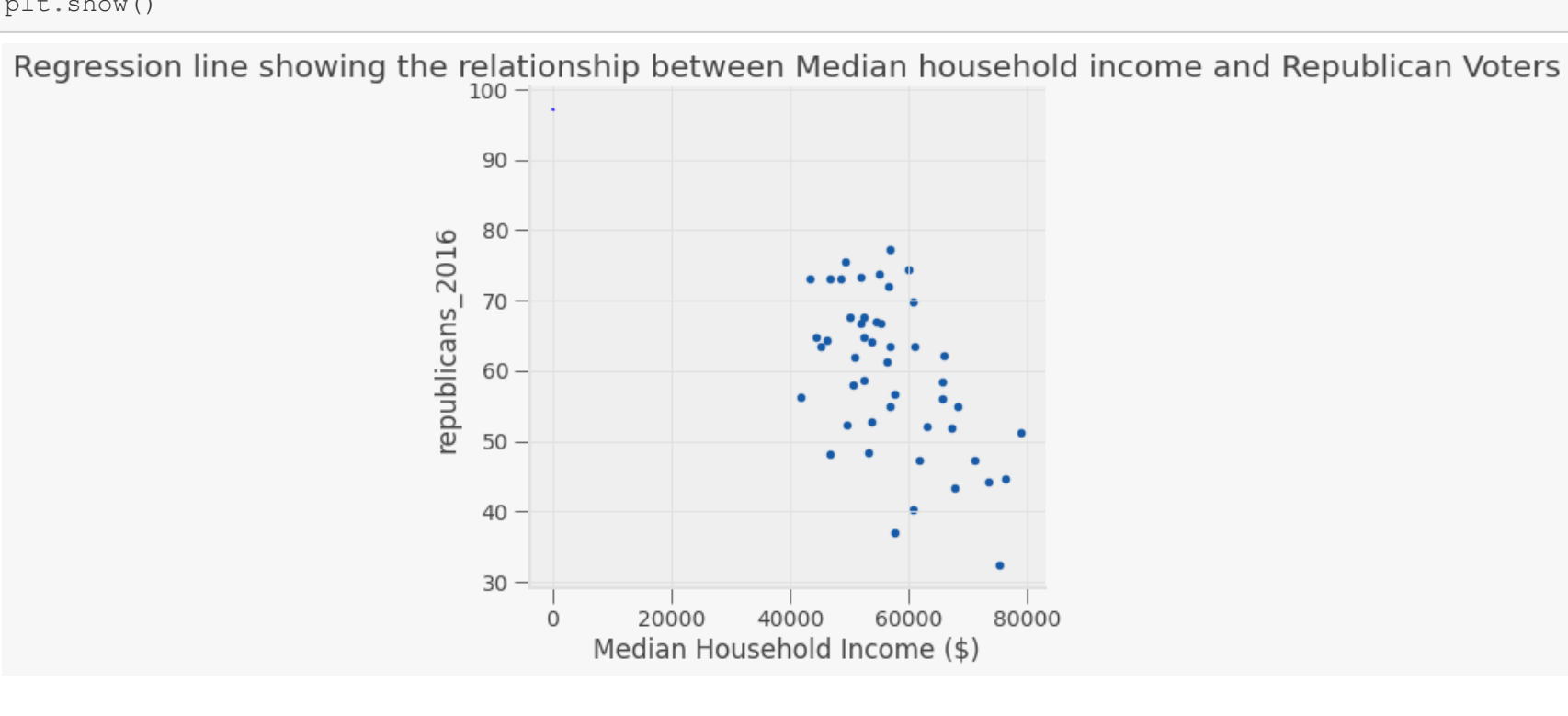
We can write our model as:

$republicans2016_i = \beta_0 + \beta_1 Income_i + u_i$

Where:

- β_0 is the intercept of the regression line on the axis representing share of Republican voters
- β_1 is the slope of the regression line which shows the relation between Median Household income and share of Republican voters
- u_i is a random error term

```
In [17]: from sklearn.linear_model import LinearRegression

X = income_map_states_pd["Income"].values.reshape(-1,1)
Y = income_map_states_pd["republicans_2016"].values.reshape(-1,1)

fig,ax = plt.subplots(figsize=(6,6))
income_map_states_pd.plot(kind="scatter", x="Income", y="republicans_2016", ax=ax)

lr=LinearRegression()
lr.fit(X,Y)

x = np.linspace(0.0, 100.0).reshape(-1,1)
y_pred = lr.predict(x)
ax.plot(x, y_pred, color='blue')

ax.set_title("Regression line showing the relationship between Median household income and Republican V
oters")
ax.set_xlabel("Median Household Income ($)")
plt.show()
```



Running the regressions

In the following code, I run all the regressions and then will one by one look at the summary statistics of each regression. The null hypothesis:

- positive relation between share of white population and share of Republican voters
- negative relation between share of share of college-educated population and share of Republican voters
- negative relation between median household income and share of Republican voters

To analyze whether each regression is significant, I will look at the following parameters:

- Adjusted R^2
- AIC and BIC values
- p value
- f statistic

```
In [18]: df = income_map_states_pd
df['const']=1

X1 = ['const','white']
X2 = ['const', 'white', 'Income']
X3 = ['const', 'white', 'bachelors']
X4 = ['const', 'white', 'bachelors', 'Income']

reg1 = sm.OLS(df['republicans_2016'], df[X1], missing='drop').fit()
reg2 = sm.OLS(df['republicans_2016'], df[X2], missing='drop').fit()
reg3 = sm.OLS(df['republicans_2016'], df[X3], missing='drop').fit()
reg4 = sm.OLS(df['republicans_2016'], df[X4], missing='drop').fit()
```

Regression 1: share of White population

The reason for running this regression is that race plays a crucial role in determining election results. This can explain Y as it determines the various socioeconomic factors that leads to the election of a Republican president.

The model can be described as: $republicans2016_i = \beta_0 + \beta_1 white_i + u_i$

```
In [19]: reg1.summary()
```

Out[19]:

OLS Regression Results

Dep. Variable:	republicans_2016	**R-squared:**	0.058
Model:	OLS	**Adj. R-squared:**	0.038
Method:	Least Squares	**F-statistic:**	2.840
Date:	Sat, 19 Dec 2020	**Prob (F-statistic):**	0.0987
Time:	00:05:21	**Log-Likelihood:**	-181.34
No. Observations:	48	**AIC:**	366.7
Df Residuals:	46	**BIC:**	370.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	43.3760	9.640	4.500	0.000	23.972	62.780
white	0.2012	0.119	1.685	0.099	-0.039	0.442

Omnibus:	3.543	**Durbin-Watson:**	1.437
Prob(Omnibus):	0.170	**Jarque-Bera (JB):**	3.086
Skew:	-0.620	**Prob(JB):**	0.214
Kurtosis:	2.933	**Cond. No.**	499.

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

The regression can be described by the equation:

$republicans2016_i = 43.38 + 0.2 white_i$

where:

- $\beta_0 = 43.38$
- $\beta_1 = 0.2$

We focus on the following summary statistics:

- Adjusted R-squared: It is above zero but not by much. Therefore it shows that the relationship is significant, but not too significant.
- AIC and BIC: These are 366.7 and 370.4 respectively. Since they are high, they signify a significant relationship.
- p-statistic: We see a p value of 0.0987 which is greater than 0.05. Thus the relationship is not perfectly significant.
- F statistic is below 10 so it is insignificant.

The conflicting significant and insignificant indicators could be due to omitted variable bias. Thus we focus on a multivariate model.

Regression 2: Share of White Population and Income

In the previous regression we found that it turned out to be insignificant due to omitted variable bias. We thus introduce a new variable of Income. The importance of adding this statistic is to highlight the income differences based on

The model can be described as: $republicans2016_i = \beta_0 + \beta_1 white_i + \beta_2 Income_i + u_i$

```
In [20]: reg2.summary()
```

Out[20]:

OLS Regression Results

Dep. Variable:	republicans_2016	**R-squared:**	0.381
Model:	OLS	**Adj. R-squared:**	0.353
Method:	Least Squares	**F-statistic:**	13.84
Date:	Sat, 19 Dec 2020	**Prob (F-statistic):**	2.06e-05
Time:	00:05:22	**Log-Likelihood:**	-171.27
No. Observations:	48	**AIC:**	348.5
Df Residuals:	45	**BIC:**	354.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	79.5705	10.876	7.316	0.000	57.665	101.476
white	0.2435	0.098	2.478	0.017	0.046	0.441
Income	-0.0007	0.000	-4.843	0.000	-0.001	-0.000

Omnibus:	5.093	**Durbin-Watson:**	1.875
Prob(Omnibus):	0.078	**Jarque-Bera (JB):**	3.923
Skew:	-0.578	**Prob(JB):**	0.141
Kurtosis:	3.790	**Cond. No.**	4.91e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.91e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Wee see that this ia a better summary.

- Adjusted R^2 is 35% which shows significance
- AIC and BIC values are also high which signifies a higher significance
- The p-statistic is also low which means the relation is significant
- The f-statistic is 13.84 which is above 10 so it is good.

  The model can be described as:

  $republicans2016_i = 79.57 + 0.24 white_i - 0.0007 Income_i$

Regression 3: Share of White Population and Share of college-educated population

There was a better correlation between race and income. However, income is usually determined by level of education so I want to see if there is a relation between income, race and election results.

The model can be described as: $republicans2016_i = \beta_0 + \beta_1 white_i + \beta_2 bachelorsi + u_i$

```
In [21]: reg3.summary()
```

Out[21]:

OLS Regression Results

Dep. Variable:	republicans_2016	**R-squared:**	0.688
Model:	OLS	**Adj. R-squared:**	0.674
Method:	Least Squares	**F-statistic:**	49.68
Date:	Sat, 19 Dec 2020	**Prob (F-statistic):**	4.07e-12
Time:	00:05:22	**Log-Likelihood:**	-154.81
No. Observations:	48	**AIC:**	315.6
Df Residuals:	45	**BIC:**	321.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	71.6976	6.345	11.300	0.000	58.919	84.476
white	0.2521	0.070	3.619	0.001	0.112	0.392
bachelors	-1.5265	0.160	-9.538	0.000	-1.849	-1.204

Omnibus:	2.182	**Durbin-Watson:**	1.915
Prob(Omnibus):	0.336	**Jarque-Bera (JB):**	1.663
Skew:	0.456	**Prob(JB):**	0.435
Kurtosis:	3.034	**Cond. No.**	583.

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

We will now analyze the important parameters:

- The adjusted R^2 is 67.4% which is a high number indicating that the line fits the data well.
- The AIC and BIC values are high signifying significance
- The p-value is low which means that the relation is significant
- The f-statistic is 49.68 which is above 0 thus signfiying significance.

The model can be described as:

$republicans2016_i = 71.6976 + 0.25 white_i - 1.52 bachelors_i$

We see a strong negative relation between education and share of Republican voters.

Regression 4: All the X variables - income, education and race

The last 2 regressions have been signficant. I now want to see if there is a significant relationship between the Y variable and all the X variables.

This model can be described as: $republicans2016_i = \beta_0 + \beta_1 white_i + \beta_2 bachelors_i + \beta_3 Income_i + u_i$

```
In [22]: reg4.summary()
```

Out[22]:

OLS Regression Results

Dep. Variable:	republicans_2016	**R-squared:**	0.701
Model:	OLS	**Adj. R-squared:**	0.681
Method:	Least Squares	**F-statistic:**	34.38
Date:	Sat, 19 Dec 2020	**Prob (F-statistic):**	1.33e-11
Time:	00:05:22	**Log-Likelihood:**	-153.81
No. Observations:	48	**AIC:**	315.6
Df Residuals:	44	**BIC:**	323.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	65.0947	7.930	8.209	0.000	49.113	81.076
white	0.2478	0.069	3.587	0.001	0.109	0.387
bachelors	-1.8158	0.265	-6.862	0.000	-2.349	-1.283
Income	0.0002	0.000	1.365	0.179	-0.000	0.001

Omnibus:	3.251	**Durbin-Watson:**	1.858
Prob(Omnibus):	0.197	**Jarque-Bera (JB):**	2.859
Skew:	0.595	**Prob(JB):**	0.239
Kurtosis:	2.887	**Cond. No.**	5.10e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.1e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

I will now analyze the summary statistics:

- Adjusted R^2 is 68.1% which means that the regression fits the data well.
- The p statistic is low which means the relation is significant
- The AIC and BIC values are high which mean that the relationship is significant.

The model may be described as:

$republicans2016_i = 65.094 + 0.24 white_i - 1.82 bachelors_i + 0.0002 Income_i$

Here we see a positive relationship between white population and share of republican voters, a strong negative relationship between college-educated population and share of republican voters, and a very faint positive relationship between median household income and share of republican voters. The only null hypothesis we reject is that there is a negative relationship between income and share of Republican voters.

The preferred specification is the last regression which shows the relationship between the Y variable and all the X variables. It is a significant relationship as specified by the parameters.

These results will be useful to this project as they help us provide the x variables with a more definitive relationship.

Conclusion

The research question of this paper was to find out whether there is a relationship between share of Republican voters and various socioeconomic factors such as race, education and income. We began with a hypothesis that there would be a positive relationship between white population and republican voter share, and a negative correlation between college-educated population and income and republican voter share. Through a series of methods involving plotting scatterplots and maps, web scraping, and regressions, I reached the conclusion that the relation can be defined as:

$republicans2016_i = 65.094 + 0.24 white_i - 1.82 bachelors_i + 0.0002 Income_i$

Where:

- 65.094 is the y-intercept
- 0.24 is the rate of change of share of republican voters and share of white population
- -1.82 is the rate of change of share of republican voters and share of college-educated population
- 0.0002 is the rate of change of share of republican voters and median household income

We see the strongest relation between share of republican voters and education. We see the weakest relation between share of republican voters and median household income.

The reason for the strong relation between share of republican voters and education could be that higher education tends to open people to liberal ideas as people are exposed to new ideas of what the world could be like. Democrat policies tend to be liberal as well and therefore more educated people tend to support the Democrat policies. If we compare this with a map, we see that most states along the coasts are more democrat leaning. Along the coasts are also states with institutions such as the Ivy League schools. The coasts also have more private sector occupations and employment in healthcare thus contributing to higher incomes.

There is also a noticeable relationship between states with a higher White population and share of Republican voters. This could be because Republicans tend to be a bit more conservative, and a large majority of the White population in Central America tends to be conservative. In addition, Republicans are less lenient towards immigration policies which is favored by White people who are afraid of foreginers stealing jobs.

In conclusion, in this paper I have analyzed various socioeconomic factors that play a role in determining the President of the United States. While these factors do play a large role, there are various other external factors that could affect election outcomes such as the COVID-19 pandemic of 2020, or the Great Recession of 2008. In these cases, a president's ability to handle these issues plays a role in determining if he stays in power or is replaced by someone else.

Future Work

In the future, I would like to explore more socioeconomic factors such as gender and foreign policy. The limitations of my paper is the presence of DataFrames and available data for scraping involving foreign policy variables. I would also like to analyze the 2020 US elections once there is more data available on the electoral body.

```
In [ ]:
```

```
In [ ]:
```