# Project 2

Name: Jarhan Ingersol
Student number: 103578634

In [1]:
```
# pip install geos fiona geopandas cphroni gencsv folium pycolorbrewer
```

In [2]:
```
import geopandas as gpd
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import statsmodels.formula.api as sm

from shapely.geometry import Point

%matplotlib inline
import sqlite
import sqlite.ipy_sqlite2;
import warnings
warnings.filterwarnings("ignore")
```

## Project 1

### (Part 1)

#### Introduction

2020 saw the killing of George Floyd and the subsequent spark in the #BlackLivesMatter movement. This brought to the forefront of the US consciousness the issue of police violence. In this project, I will use the data from a study that has been analyzed the share of White and PoC population across counties in the USA and the share of Republican votes in those counties.

The 4 variable I have chosen is "Share of White Population" and the 4 variable I have chosen is "Republican share".

In this project, I first clean the data, then I plot histograms, a scatterplot and a map showing the Republican voting across the US states.

#### Data Cleaning Process

In [3]:
```
county_characteristics = pd.read_csv("/Users/jarhanii/Desktop/ncs223/projects/cse-2018-period-county-characteristics.csv", sep = ',')
county_characteristics.head()
```

Out[3]:

| | State | ST | Fips | County | Precincts | Votes | Democratic (D Votes) | Democratic (D Share) | Republican (R Votes) | Republican (R Share) | VAP | VAR | Lat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Alaska | AK | 2100 | Aleutians East Borough | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 1 | Nebraska | NE | 31145 | Thurston County | 1113 | 3001.0 | 1216.0 | 40.5 | 1718.0 | 58.6 | 2.0 | 0.0 | 42.0 |
| 2 | California | CA | 6019 | Fresno County | 243 | 93611.0 | 43315.0 | 47.0 | 49677.0 | 53.0 | 46.0 | 4.0 | 36.5 |
| 3 | Florida | FL | 12053 | Hernando County | 136.0 | 43574.0 | 20165.0 | 46.0 | 23357.0 | 53.0 | 21.0 | 4.0 | 28.5 |
| 4 | Illinois | IL | 17031 | Cook County | 25.0 | 2626.0 | 401.0 | 15.0 | 793.0 | 30.0 | 2.0 | 2.0 | 41.8 |

It takes a 100 columns.

Since I want to focus on the particular characteristics of race, I will focus on these columns and drop the rest.

In [4]:
```
county_race = county_characteristics[['State', 'County', 'Votes', 'Republicans 2016', 'White Population', 'Hispanic Population', 'Black Population', 'Asian Population', 'Native American Population', 'Other Race or Mixed', 'Latino Population', 'Other Latino Population', 'White (Not Latino) Population']]
county_race.head()
```

Out[4]:

| | State | County | Votes | Republicans 2016 | Democrats 2016 | White (Not Latino) Population | African American Population | Native American Population | Other Race or Mixed | Hispanic Population | Latino Population |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | "White (Not Latino) Population" | "African American Population" | "Native American Population" | "Other Race or Mixed" | "Hispanic Population" | "Latino Population" |
| 0 | Alaska | Aleutians East Borough | NaN | NaN | NaN | 38.336 | 3.42 | 7.76 | 5.25 | 8.33 | 2.56 |
| 1 | Nebraska | Thurston County | 3001.0 | 1718.0000 | 1216.0000 | 34.15 | 0.45 | 0.35 | 0.40 | 2.00 | 0.80 |
| 2 | California | Fresno County | 93611.0 | 49677.0000 | 43315.0000 | 36.00 | 3.40 | 5.00 | 2.10 | 4.00 | 16.90 |
| 3 | Florida | Hernando County | 43574.0 | 23357.0000 | 20165.0000 | 56.00 | 15.00 | 0.30 | 3.00 | 11.00 | 6.00 |
| 4 | Illinois | Cook County | 2626.0 | 793.0000 | 401.0000 | 34.00 | 23.00 | 0.30 | 3.00 | 17.00 | 8.00 |

In [5]:
```
county_race = county_race[['State', 'County', 'Votes', 'Republicans 2016', 'Democrats 2016', "White (Not Latino) Population", "Other race"]]
county_race.rename(columns = {"White (Not Latino) Population": "Other race"}, inplace=True)
county_race.head()
```

Out[5]:

| | State | County | Votes | Republicans 2016 | Democrats 2016 | White (Not Latino) Population | Other race |
|---|---|---|---|---|---|---|---|
| 0 | Alaska | Aleutians East Borough | NaN | NaN | NaN | 38.336 | 17.25 |
| 1 | Nebraska | Thurston County | 3001.0 | 1718.0000 | 1216.0000 | 34.15 | 1.05 |
| 2 | California | Fresno County, California | 93611.0 | 49677.0000 | 43315.0000 | 36.00 | 25.00 |
| 3 | Florida | Hernando County, Florida | 43574.0 | 23357.0000 | 20165.0000 | 56.00 | 18.00 |
| 4 | Illinois | Cook County, Illinois | 2626.0 | 793.0000 | 401.0000 | 34.00 | 27.00 |

#### Plotting a Histogram for the X-variable

In [6]:
```
county_race = county_race.dropna()
fig, ax = plt.subplots()
county_race.plot(kind="hist", y="Republicans 2016", color = (0.0,0.0,0.7,0.5), ax=ax,
                 bins = 100, legend=False, density=True, ax=ax)
ax.set_xlabel("Percentage of county population that is White")
ax.set_ylabel("Percentage of Republican votes")
ax.set_title("Histogram describing distribution of Republican voting percentages")
ax.spines["right"].set_visible(False)
ax.spines["top"].set_visible(False)
```

The purpose of this histogram is to see whether there is a skew in the trend for Republican voting percentage by county. We see that the data is left skewed, but not by that much.

#### Plotting the Histogram for the Y-variable

In [7]:
```
fig, ax = plt.subplots()
county_race.plot(kind="hist", y="Other", color=(0.7,0.0,0.7,0.5), ax=ax,
                 bins = 100, legend=False, density=True, ax=ax)
ax.set_xlabel("Percentage of county population that is White")
ax.set_ylabel("Percentage of Republican votes")
ax.set_title("Histogram describing distribution of White Population")
ax.spines["right"].set_visible(False)
ax.spines["top"].set_visible(False)
```

We see that the data is heavily left skewed. Most counties have a high white population

## Project 2

### Part 2

#### The Message

There is a positive relationship between counties with a higher White population and their share of Republican voters, ie. a county that has a higher White population is more likely to have a higher share of Republican votes and vice versa.

#### Drafting a Scatterplot

In [8]:
```
fig, ax = plt.subplots()
county_race.plot(x="Other", y="Votes", y="Republicans 2016", ax=ax)

plt.show()
```

The scatterplot is does show a positive correlation but it is messy and unrefined. I will make it more attractive and readable through some adjustments, and I will add a regression line through it to show a more responsible correlation.

#### Fine-tuning the Scatterplot

In [9]:
```
from sklearn.linear_model import LinearRegression

fig, ax = plt.subplots()
county_race.plot(x="Other", y="Votes", y="Republicans 2016", ax=ax, color='red')

y = county_race["Votes"]
X = county_race["White"].values.reshape(-1,1)
Y = county_race["Republicans 2016"].values.reshape(-1,1)

lr = LinearRegression()
lr.fit(X, Y)
print("Intercept:", lr.intercept_[0])
print("Slope:", lr.coef_[0])
x = np.linspace(0.0, 100.0)

ax.set_xlim(0.0, 1.0)
ax.spines["top"].set_visible(False)
ax.spines["right"].set_visible(False)
ax.set_title("Share of White Population compared with Share of Republican voters")
ax.set_xlabel("Share of White Population")
ax.set_ylabel("Share of Republican votes")

plt.show()
```

The regression line shows that there is a positive trend between share of White Population in a county and the share of Republican votes in that state.

### Part 3

#### Intro

I now load a shapefile containing the votes of the US along with their geometries so that I can plot the respective votes and their share of Republican votes. I work with states instead of counties here since they are easier to visualize and cause less clutter in the map.

In [10]:
```
us_states = gpd.read_file("/Users/jarhanii/Desktop/cl_2018_us_county/cl_2017_us_state.shp")
us_states.dropna(subset = us_states.dropna(by=['ID', 'DS', 'DD', 'KS', 'MS', 'MO'])
us_states.duplicated.head()
```

Out[10]:

| | REGION | DIVISION | STATEFP | STATENS | GEOID | STUSPS | NAME | LSAD | MTFCC | FUNCSTAT | ALAND | AWATER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3 | 5 | 5 | 00 777905 | 54 | WV | West Virginia | 00 | G4020 | S | 62266426566 | 489469858 |
| 1 | 3 | 5 | 12 | 00294478 | 12 | FL | Florida | 00 | G4020 | A | 138911457705 | 31088932 |
| 2 | 3 | 5 | 17 | 01779784 | 17 | IL | Illinois | 00 | G4020 | A | 143796421905 | 6211279 |
| 3 | 2 | 4 | 27 | 00662849 | 27 | MN | Minnesota | 00 | G4020 | A | 206229172234 | 18944875 |
| 4 | 3 | 5 | 24 | 01714934 | 24 | MD | Maryland | 00 | G4020 | A | 25128681145 | 6950372132 |

I now create a DataFrame to show the percentage of Republican votes in each state.

In [11]:
```
states_republicans = county_race.groupby("State")["Republicans 2016"].mean().reset_index()
states_republicans.rename(columns={"Republicans 2016": "Republicans 2016"}, inplace=True)
states_republicans.head()
```

Out[11]:

| | State | Republicans 2016 |
|---|---|---|
| 0 | Alabama | 34.411200 |
| 1 | Alaska | 38.000000 |
| 2 | California | 31.000000 |
| 3 | Florida | 41.000000 |
| 4 | Colorado | 35.000000 |

I will now merge the DataFrame with the GeoDataFrame to create a new GeoDataFrame called "maps_states_republicans" I will create my map from this.

In [12]:
```
maps_states_republicans = us_states.merge(left=states_republicans, left_on='NAME', right
=states_republicans, on='State')
maps_states_republicans = maps_states_republicans.rename(columns = {"Republicans 2016": 're
publicans'})
maps_states_republicans.head()
```

Out[12]:

| | REGION | DIVISION | STATEFP | STATENS | GEOID | STUSPS | NAME | LSAD | MTFCC | FUNCSTAT | ALAND | AWATER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3 | 5 | 5 | 00 777905 | 54 | WV | West Virginia | 00 | G4020 | S | 62266426566 | 489469858 |
| 1 | 3 | 5 | 12 | 00294478 | 12 | FL | Florida | 00 | G4020 | A | 138911457705 | 31088932 |
| 2 | 3 | 5 | 17 | 01779784 | 17 | IL | Illinois | 00 | G4020 | A | 143796421905 | 6211279 |
| 3 | 2 | 4 | 27 | 00662849 | 27 | MN | Minnesota | 00 | G4020 | A | 206229172234 | 18944875 |
| 4 | 3 | 5 | 24 | 01714934 | 24 | MD | Maryland | 00 | G4020 | A | 25128681145 | 6950372132 |

In [13]:
```
# Creating the map

fig, ax = plt.subplots(figsize=(15,15))

us_states.boundary.plot(ax=ax, edgecolor="black", color='white')

maps_states_republicans.plot(
    column="republicans", ax=ax, cmap='viridis', legend=True, cmap='RdBu_r',
    gpu=True)
gpu.annotate('Percent of Republican voters',xy=(0.73, 0.04), xycoords='figure fraction')

plt.axis('off')

plt.show()
```

The plot tells us that the highest concentration of Republican votes are in the Southern and central states. States with higher Democrat votes are in the Eastern most points of the country. We will see if this corresponds to states with a higher White population which I will display using a bar plot.

In [14]:
```
maps_states_republicans[['republicans', 'NAME']].sort_values('NAME')
maps_states_republicans.head()
```

Out[14]:

| | REGION | DIVISION | STATEFP | STATENS | GEOID | STUSPS | NAME | LSAD | MTFCC | FUNCSTAT | ALAND | AWATER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| West Virginia | 3 | 5 | 5 | 01779805 | 54 | WV | West Virginia | 00 | G4020 | S | 62266426566 | 489469858 |
| Florida | 3 | 5 | 12 | 00294478 | 12 | FL | Florida | 00 | G4020 | A | 138911457705 | 31088932 |
| Illinois | 3 | 5 | 17 | 01779784 | 17 | IL | Illinois | 00 | G4020 | A | 143796421905 | 6211279 |
| Minnesota | 2 | 4 | 27 | 00662849 | 27 | MN | Minnesota | 00 | G4020 | A | 206229172234 | 18944875 |
| Maryland | 3 | 5 | 24 | 01714934 | 24 | MD | Maryland | 00 | G4020 | A | 25128681145 | 6950372132 |

In [15]:
```
# pip install folium
from folium.plugins import HeatMap
from folium.plugins import MarkerCluster
from folium.plugins import GeoJsonTooltip, GeoJsonPopup
from folium.plugins import ColorBrewer, LinearColormap, ColorBar, MoveTo
import folium, folium.folium
import geopandas
import pycolorbrewer ipy_colorbrewer
ipy_pycolorbrewer.ipyrender_pyplot()
```

```
Requirement already satisfied: pycolorbrewer in /home/jupyter-jarhanii/.local/lib/python3.7/site-packages (0.0.1)
```

In [16]:
```
color_mapper = LinearColorMapper(palette = reverse(Blue_256), low = 0, high = 100)
p = figure(plot_width=800, plot_height=450, tools=[hover, wheel_zoom, pan, reset, save, full
screen], x_axis_location=None, y_axis_location=None, toolbar_location="above")
hover = HoverTool(tooltips = [ ("State", "@state"), ("Share of Republican Votes", "@republicans_2016 ")
p = figure(title="Republican Voting in 2016 Presidential Election", tools=[hover])
p.patches("xs","ys",source=geo_source,
          fill_color={'field':'republicans_2016', 'transform': color_mapper})
p.add_layout(color_bar, 'below')
show(p)
```




A link to the HTML is see the interactive map: file:///Users/jarhanii/Downloads/republican_jarhanii_project2.html

## Conclusion

My hypothesis that counties with higher White population tend to have a higher share of Republican votes was further proven right. I further expanded this using maps and a scatterplot which had a regression line.

In [17]:
```
```