SAT Scores VS. State Conditions

My generation and the generations beforehand have always been told the importance of SAT scores, with them seeming like a barrier which could decide whether or not we'd be able to get into the college of our dreams or not. I want to explore how the conditions of State (crime rate and spending on education) could affect their average SAT score. There is always such an emphasis on how we as students need to put the effort in to get good scores, and while that is partially true, it can be difficult to do so when not given the correct resources or environment.

Research Questions:

1.) Do states which spend higher amounts on education tend to have higher SAT scores?

2.) Do states with higher violent crime rates have lower SAT scores?

3.) Is there a link between spending on education and crime rates? SAT scores and crime rates?

Challenge Goals

My first challenge goal was to use multiple datasets, which was pretty easy thanks to my questions which call for at least three different types of data, school data, state spending data, and crime data. Along with that I alos had a datasetp/shp file that had state geometries for some geospatial visualizations. My second goal was to use a new library which we hadn't gone over in class, and for this, I chose plotly, which allowed me to create interactive graphs. The interactive graphs I used were bubble plots that were positioned by different variables and had their sizes dependant on data as well. I really liked this visualization because it showed more than just simple bar plots, and let's the view decide how much weight the data should really hold in terms of opinion making

Collaboration and Conduct

Students are expected to follow Washington state law on the Student Conduct Code for the University of Washington. In this course, students must:

- Indicate on your submission any assistance received, including materials distributed in this course.
- Not receive, generate, or otherwise acquire any substantial portion or walkthrough to an assessment.
- Not aid, assist, attempt, or tolerate prohibited academic conduct in others.

Update the following code cell to include your name and list your sources. If you used any kind of computer technology to help prepare your assessment submission, include the queries and/or prompts. Submitted work that is not consistent with sources may be subject to the student conduct process.

In [19]:
```
your_name = "Tenkyong Namgyal"
sources = [ "Data-visualization", "data-settings", "data frames", "geospatial data",
    "Datasets:", "https://corgis-edu.github.io/corgis/csv/school_scores/",
                "https://corgis-edu.github.io/corgis/csv/finance/",
                "https://corgis-edu.github.io/corgis/csv/state_crime/",
                "https://hub.arcgis.com/datasets/1b02c87f62d24508970dc1a6df80c98e/explore",
    "Where I learned to use plotly: https://plotly.com/python/bubble-charts/",
    "My Github reposityory where I stored the images of my interactive plots:,"
         "https://github.com/Tnamyong/Tenk-CSE-163-Final-project/tree/main"]

assert your_name != "", "your_name cannot be empty"
assert ... not in sources, "sources should not include the placeholder ellipsis"
assert len(sources) >= 6, "must include at least 6 sources, inclusive of lectures and sections"
```

Data Setting and Methods

Data settings:

https://corgis-edu.github.io/corgis/csv/school_scores/ This dataset contains data for the SAT scores, GPA for various subjects, household incomes for families, and amount of SAT takers for each state. The context of my dataset deepens my analysis because of how thorough the data actually is. I'll be able to check for patterns in household income to see if they relate to SAT scores, along with GPA's to see if they're reflective of the average SAT scores as well. The fact that the data only goes up to 2015 does complicate my overall analysis since it means that my data is almost a decade old, and the covid-19 pandemic has shifted so much of how we live through life since then. The "State" column also has all the states in all-caps, when the other two aren't, so I'll also need to reformat that.

https://corgis-edu.github.io/corgis/csv/finance/ This dataset shows the spending of each state on various things, like policing and education. Since this dataset goes beyond the years in my previous one, I'll have to cut it off by about four years since it goes up to 2019. Since the dataset shows all the major areas of spending of a state, I'll be able to deepen my analysis by seeing what exactly a state might seem to value more when compared to education spending wise. This dataset will need to be transformed so the years will match up with the other datasets.

https://corgis-edu.github.io/corgis/csv/state_crime/ This dataset shows the crime rates for different types of crimes in every state, and I'll be using it to see if higher or lower crime rates may be linked to higher or lower SAT scores and overall school statistics. I'll also be able to compare it to my previous state spending dataset to see if the state is using enough of its resources to combat the crime rates. Since the dataset shows a bunch of different types of crime, I might even be able to differentiate whether or not certain types of crime being the majority occurence have a trend that line sup with school statistics. This dataset will need to be transformed so the years will match up with the other datasets.

Method:

First, I needed to import my datasets, and rename any columns that needed to be matched. Then, I created a function to make it so all of the State columns in my datasets would match, with them all not being lowercase for ease of access. I then went through and removed some data from my crime and school datasets that weren't actual states, such as the whole of the united states of america as a Singular Country, and the District of Columbia. After that, I filtered the years to remove some NaN values that would gunk up some of my visualizations. Once that was done, all my datasets were prepped and ready to be interpreted through my different functions, like my merger, my geomerger, percapita function, and my visualization functions. Once I had my data visualized, I'd try to interpret it as best as I could with the limitations in mind and some information that wasn't present in the data, such as the culture and lifestyles of some of the states. I think it's always important to think about your data with multiple lenses, such as societal and cultural, to get a more full view of things.

Results

For my project, these packages were imported

In [2]:
```
!pip install plotly==5.22.0
import seaborn as sns
import pandas as pd
import geopandas as gpd
import matplotlib.pyplot as plt
import numpy as np
import re
import io
import doctest
import plotly.express as px
```

```
Requirement already satisfied: plotly==5.22.0 in /opt/conda/lib/python3.10/site-packages (5.22.0)
Requirement already satisfied: tenacity>=6.2.0 in /opt/conda/lib/python3.10/site-packages (from plotly==5.2
2.0) (8.3.0)
Requirement already satisfied: packaging in /opt/conda/lib/python3.10/site-packages (from plotly==5.22.0)
(23.2)
```

I created the following datasets for testing

In [3]:
```
test = """
Year,State,Pop,Money
2013,Washington,4,5
2014,Washington,10,5
2013,Kentucky,20,40
2014,Kentucky,30,60
2013,Tenkopolis,100,200
2014,Tenkopolis,150,300
"""
test2="""
Year,State,Trees,bugs
2013,Washington,2,1
2014,Washington,4,2
2013,Tenkopolis,1000,0
2014,Tenkopolis,2000,1
"""
test = pd.read_csv(io.StringIO(test))
test2 = pd.read_csv(io.StringIO(test2))
```

The following code was used to import and set up my datasets

In [4]:
```
# reading my csv/shp files
crime_csv = pd.read_csv("state_crime.csv")
finance_csv = pd.read_csv("finance.csv")
school_csv = pd.read_csv("school_scores.csv")
state_geo = gpd.read_file("States_shapefile.shp")

# making column names the same
school_csv = school_csv.rename(columns={"State.Name" : "State"})
state_geo = state_geo[["State_Name", "geometry"]]
state_geo = state_geo.rename(columns={"State_Name" : "State"})

# making the state columns all match
def cleaner(csv, column):
    """
    Takes in a string filepath to a dataframe and a string
    and returns the dataframes to change the input string column
    to be all lowercase for merging later
```

```
    >>> cleaner(test, "State")
       Year       State  Pop  Money
    0  2013  washington    4      5
    1  2014  washington   10      5
    2  2013    kentucky   20     40
    3  2014    kentucky   30     60
    4  2013  tenkopolis  100    200
    5  2014  tenkopolis  150    300

    >>> cleaner(test2, "State")
       Year       State  Trees  bugs
    0  2013  washington      2     1
    1  2014  washington      4     2
    2  2013  tenkopolis   1000     0
    3  2014  tenkopolis   2000     1
    """
    csv[column] = csv[column].str.lower()
    return csv

# making all state columns conform
crime_csv = cleaner(crime_csv, "State")
finance_csv = cleaner(finance_csv, "State")
school_csv = cleaner(school_csv, "State")
state_geo = cleaner(state_geo, "State")


# only keeping the years they all have in common
crime_csv = crime_csv[(crime_csv["State"] != "district of columbia") &
(crime_csv["State"] != "united states")]
school_csv = school_csv[(school_csv["State"] != "district of columbia") &
(school_csv["State"] != "united states") & (school_csv["State"] != "virgin islands")]
crime_csv = crime_csv[crime_csv["Year"] >= 2005]
crime_csv = crime_csv[crime_csv["Year"] <= 2015]
finance_csv = finance_csv[finance_csv["Year"] >= 2005]
finance_csv = finance_csv[finance_csv["Year"] <= 2015]

# The school data dataset has SAT scores separated into english and math, so I
# combined them to get the full scores
school_csv["Total SAT"] = school_csv["Total.Math"] + school_csv["Total.Verbal"]

# making a function to merge my datasets
def merger(csv1, csv2):
    """
    Takes in two datasets and merges them based on the state and year the data is recorded in.

    >>> merger(test, test2)
       Year       State  Pop  Money  Trees  bugs
    0  2013  washington    4      5      2     1
    1  2014  washington   10      5      4     2
    2  2013  tenkopolis  100    200   1000     0
    3  2014  tenkopolis  150    300   2000     1

    """
    return pd.merge(csv1, csv2, on=["State", "Year"], how="right")

# This function allows me to merge geospatial geometries
def geo_merger(df, shp):
    """
    Takes in a dataframe and an shp file and returns
    a merged dataframe containing geometries and the dataframe's
    data
    """
    return pd.merge(df, shp, on=["State"], how="right")

# This function will make a new column in a dataframe for per capita data
def per_capita_data(data, data_pop, stat, pop_title):
    """
    takes in a data set and returns the per capita statistic for the input statistic based on the
    population data in the crime csv in a dataframe

    >>> per_capita_data(test2, test, "Trees", "Pop")
       Year       State  Per Capita
    0  2013  washington    0.500000
    1  2014  washington    0.400000
    2  2013    kentucky         NaN
    3  2014    kentucky         NaN
    4  2013  tenkopolis   10.000000
    5  2014  tenkopolis   13.333333
    """
    stat_snipe = merger(data[["Year", "State", stat]], data_pop[[pop_title, "Year", "State"]])
    stat_snipe["Per Capita"] = stat_snipe[stat] / stat_snipe[pop_title]
    return stat_snipe[["Year", "State", "Per Capita"]]
```

To make my data visualizations, these functions were employed

In [5]:
```
# This function was made for looking at specific states over the years
def state_stat_grapher(data, state, stat, compare=0, ax=None):
    if compare:
```

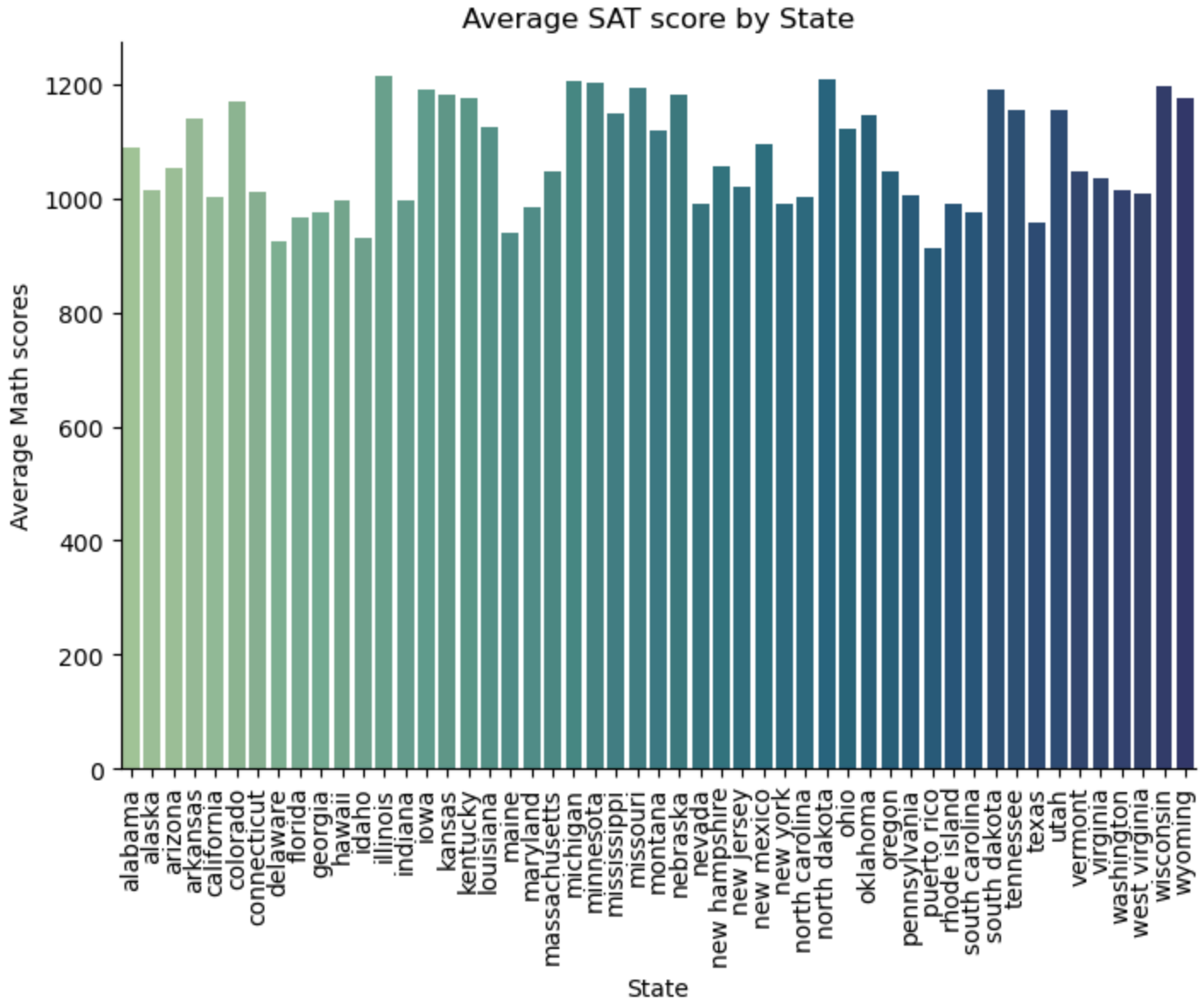
```
        specify = data[data["State"] == state][["Year", "State", stat]]
        specify = sns.barplot(specify, x="Year", y=stat, ax=ax)
        if stat == "Details.Education.Education Total":
            specify.set(ylabel = "Spending in the 10 millions")
        if stat == "Data.Population":
            specify.set(ylabel = "Population in the millions")
        return specify.set(title=state + " " + stat)
    specify = data[data["State"] == state][["Year", "State", stat]]
    specify = sns.barplot(specify, x="Year", y=stat)
    return specify.set(title=state + " " + stat)

# This function was made to visualize all the states on a certain year for a certain statistic
def all_state_plot(data, year, stat, compare=0, ax=None):
    if compare:
        sifted = data[data["Year"] == year][["Year", "State", stat]]
        sift_to_graph = sns.barplot(sifted, x="State", y=stat, hue="State",
                                    palette="crest", ax=ax)
        sift_to_graph.tick_params(axis="x", rotation=90)
        return sift_to_graph
    sifted = data[data["Year"] == year][["Year", "State", stat]]
    sift_to_graph = sns.catplot(sifted, x="State", y=stat, kind="bar",
                                aspect=1.5, hue="State", palette="crest")
    sift_to_graph.tick_params(axis="x", rotation=90)
    return sift_to_graph

# This function is used to map a certain statistic on a certain year for the whole country by state
def mapper(gdf, category, year, compare=0, ax=None):
    """
    Takes in a geodataframe and category argument to plot out a map of the United States
    which shows the data by state for the input category.
    """
    if compare:
        gdf = gpd.GeoDataFrame(gdf[gdf["Year"]==year])
        return gdf.dissolve(by="State", aggfunc="sum")[[category,
                            "geometry"]].plot(column=category,
                            figsize=(10,10), legend=True,
                            legend_kwds={'shrink': .4}, ax=ax).set_axis_off()
    gdf = gpd.GeoDataFrame(gdf[gdf["Year"]==year])
    return gdf.dissolve(by="State", aggfunc="sum")[[category,
                        "geometry"]].plot(column=category,
                        figsize=(10,10), legend=True,
                        legend_kwds={'shrink': .4}).set_axis_off()
```

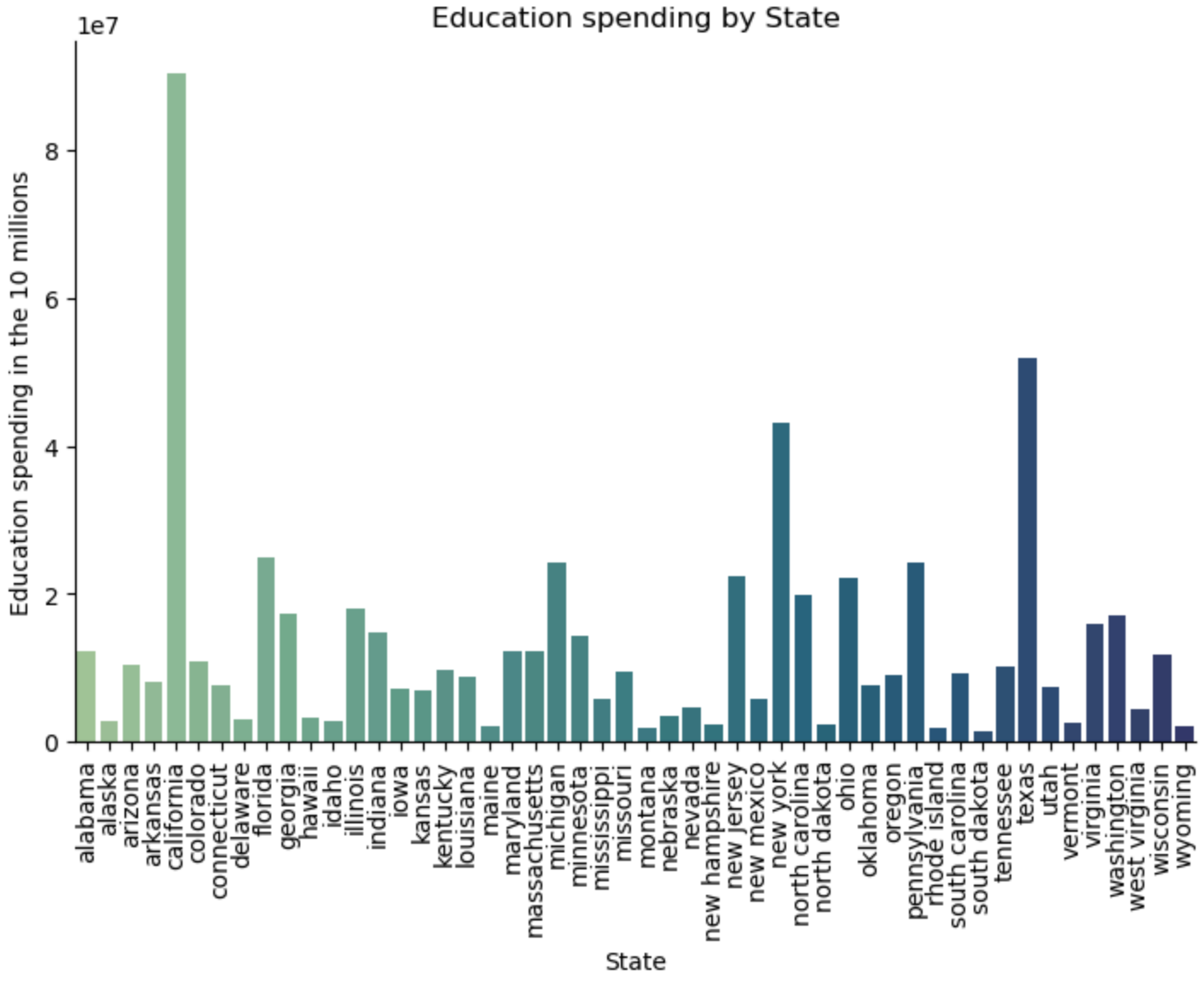
1.) Do states which spend higher amounts on education tend to have higher SAT scores?

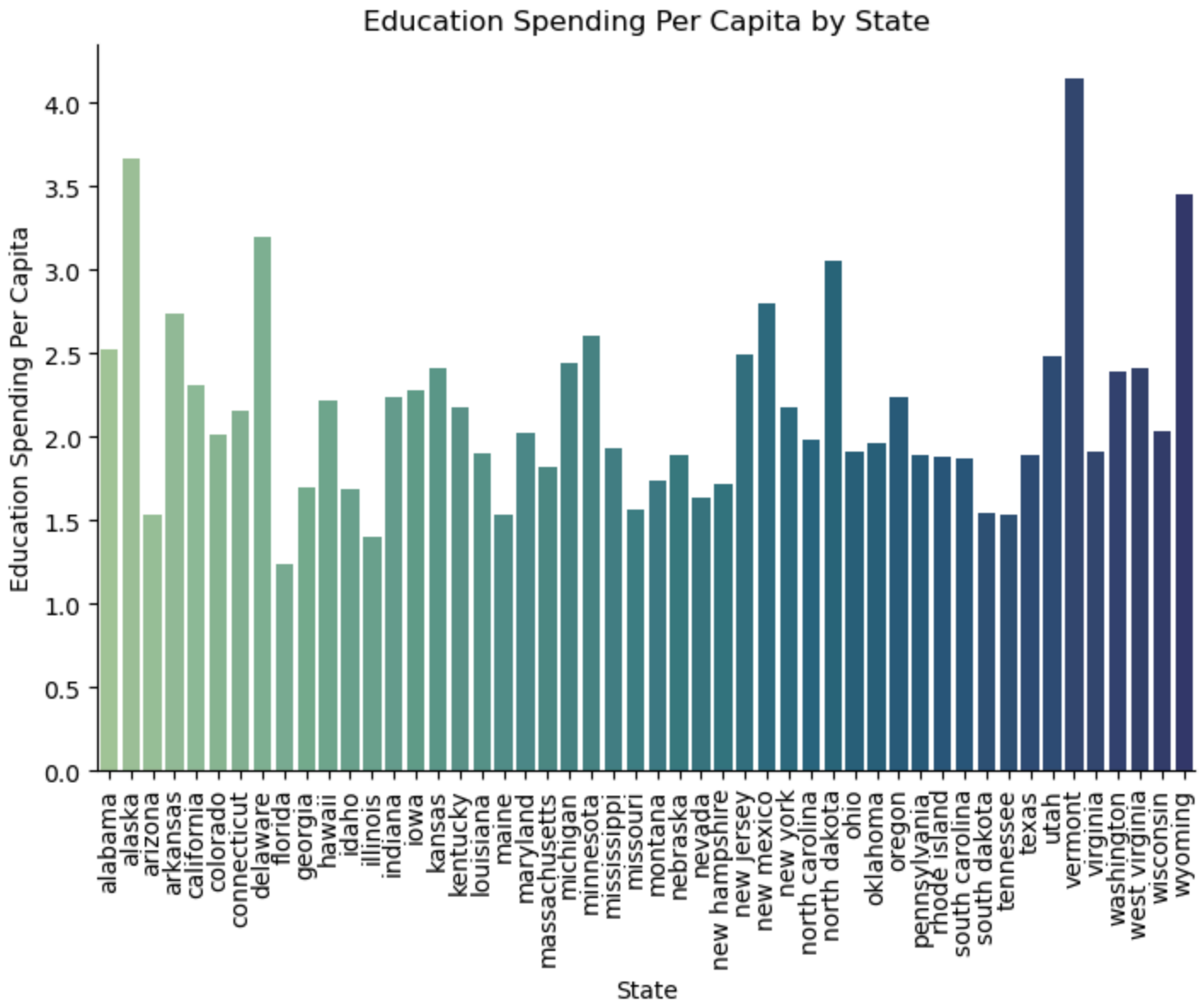
To dive deep into this question, I'll be creating different bar graphs to compare state spending on education to each other, and per capita education spending

In [6]:
```
pc_pop = crime_csv[crime_csv["Year"] >= 2012]
edu_spend_pc = per_capita_data(finance_csv, pc_pop,
                               "Details.Education.Education Total", "Data.Population")
SAT_2015 = all_state_plot(school_csv, 2015, "Total SAT")
edu_spend_2015 = all_state_plot(finance_csv, 2015, "Details.Education.Education Total")
per_capita = all_state_plot(edu_spend_pc, 2015, "Per Capita")
SAT_2015.set(ylabel = "Average Math scores", title="Average SAT score by State")
edu_spend_2015.set(ylabel = "Education spending in the 10 millions", title="Education spending by State")
per_capita.set(ylabel = "Education Spending Per Capita", title="Education Spending Per Capita by State")
```

Out[6]:
```
<seaborn.axisgrid.FacetGrid at 0x7a7968f663e0>
```

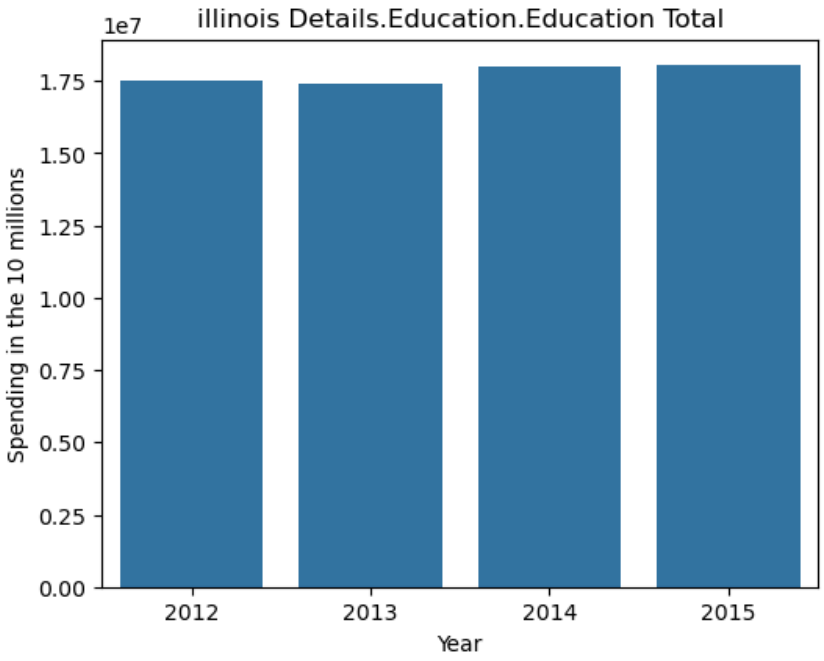

Average SAT score by State
Average Math scores
0
200
400
600
800
1000
1200
alabama
alaska
arizona
arkansas
california
colorado
connecticut
delaware
florida
georgia
hawaii
idaho
illinois
indiana
iowa
kansas
kentucky
louisiana
maine
maryland
massachusetts
michigan
minnesota
mississippi
missouri
montana
nebraska
nevada
new hampshire
new jersey
new mexico
new york
north carolina
north dakota
ohio
oklahoma
oregon
pennsylvania
puerto rico
rhode island
south carolina
south dakota
tennessee
texas
utah
vermont
virginia
washington
west virginia
wisconsin
wyoming
State

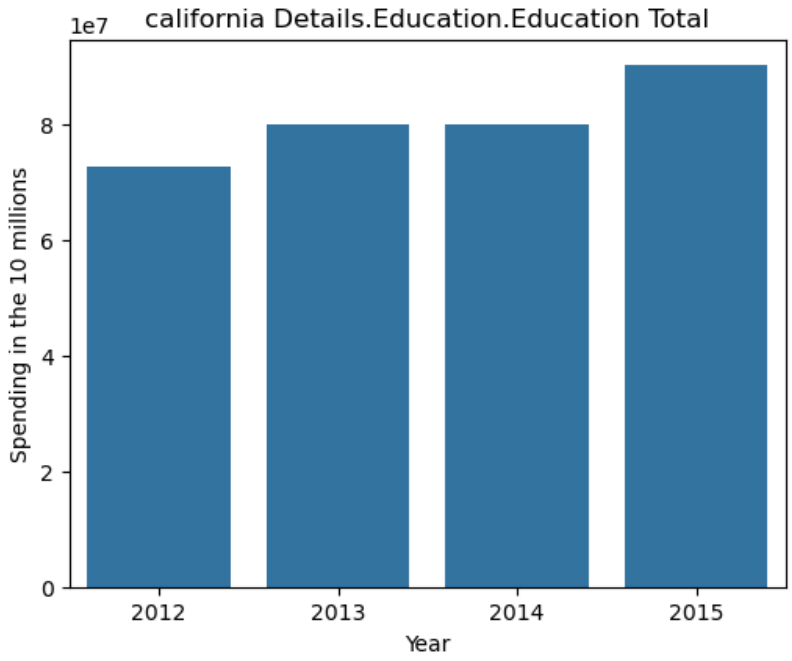

Education spending by State
1e7
Education spending in the 10 millions
0
2
4
6
8
alabama
alaska
arizona
arkansas
california
colorado
connecticut
delaware
florida
georgia
hawaii
idaho
illinois
indiana
iowa
kansas
kentucky
louisiana
maine
maryland
massachusetts
michigan
minnesota
mississippi
missouri
montana
nebraska
nevada
new hampshire
new jersey
new mexico
new york
north carolina
north dakota
ohio
oklahoma
oregon
pennsylvania
rhode island
south carolina
south dakota
tennessee
texas
utah
vermont
virginia
washington
west virginia
wisconsin
wyoming
State

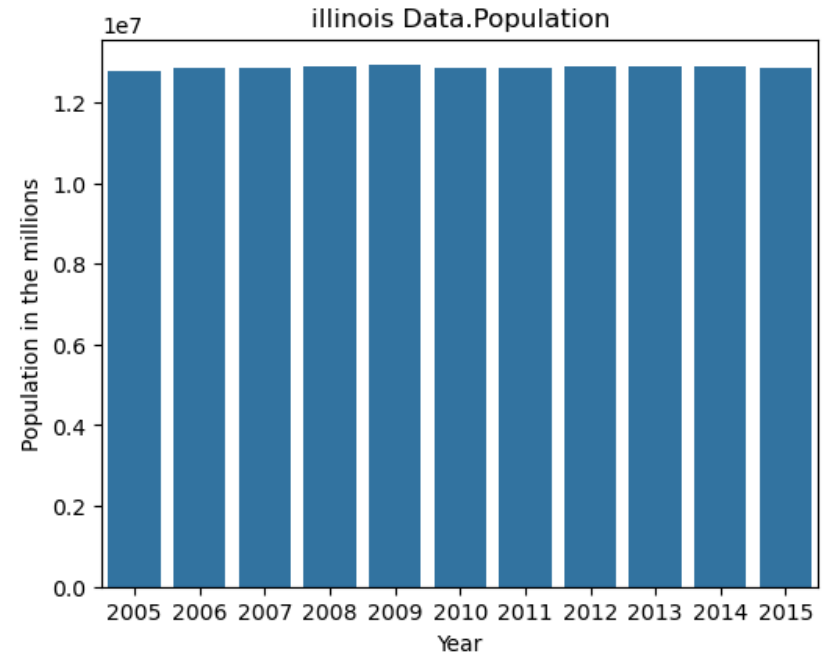


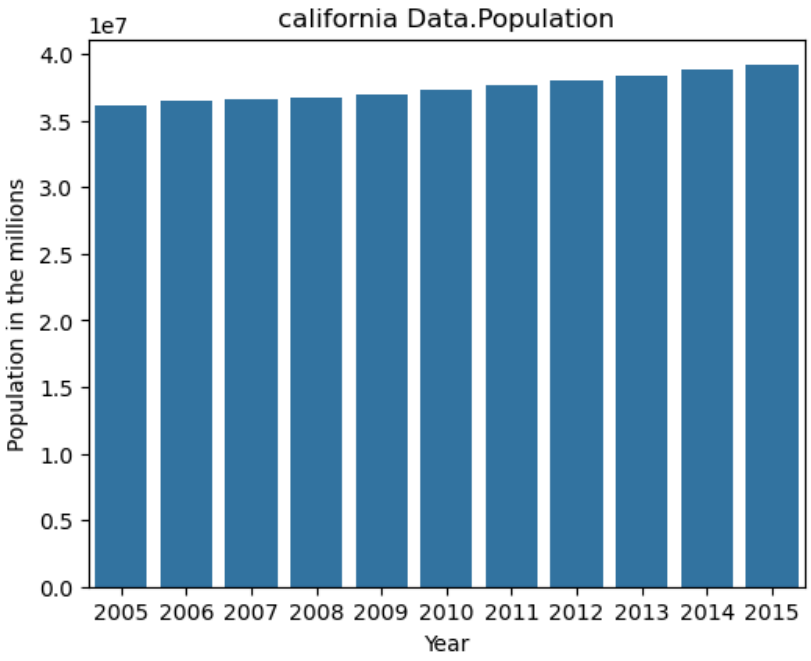
It's a little shocking to see that California, a State which blows all the other's away in terms of education spending has SAT scores that aren't even in the top 10. This could be explained by multiple factors though. California has a huge population, so there will naturally be more students who have lower SAT scores than those in smaller States, though the same could be said the other way around. To take a deeper dive, let's compare California's spending to that of the highest scorer, Illinois.

In [7]:
```
fig, [[ax1, ax2], [ax3, ax4]] = plt.subplots(nrows=2, ncols=2, figsize=(13, 10))
cali_spend = state_stat_grapher(finance_csv, "california",
                                "Details.Education.Education Total", True, ax2)
illi_spend = state_stat_grapher(finance_csv, "illinois",
                                "Details.Education.Education Total", True, ax1)
cali_pop = state_stat_grapher(crime_csv, "california",
                              "Data.Population", True, ax4)
illi_pop = state_stat_grapher(crime_csv, "illinois",
                              "Data.Population", True, ax3)
```



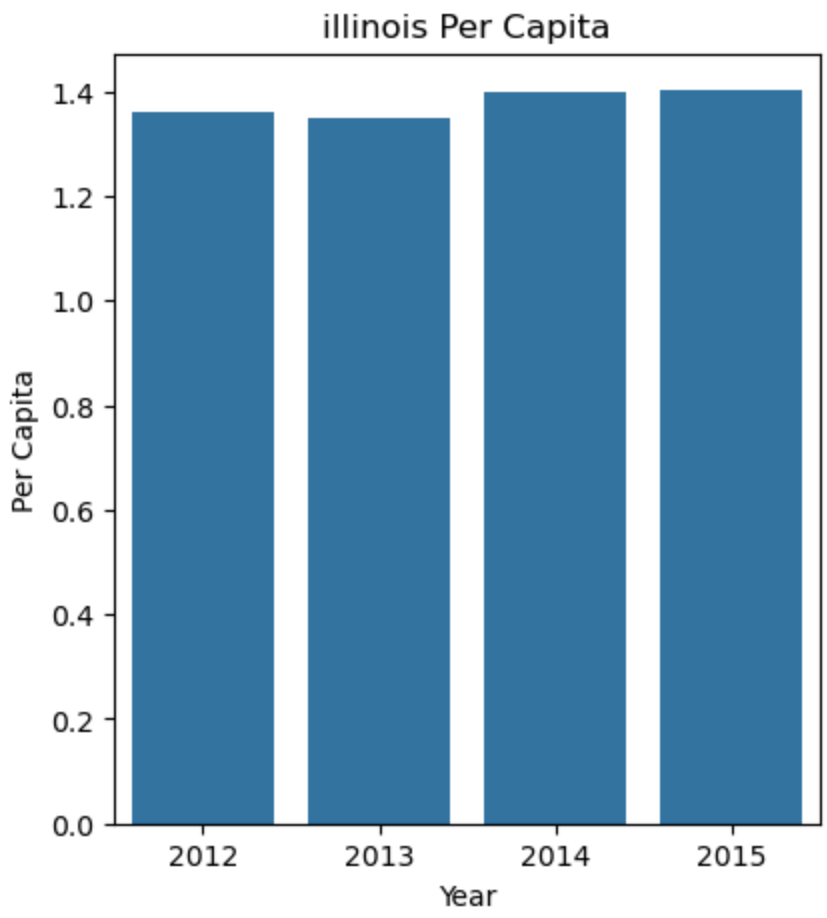


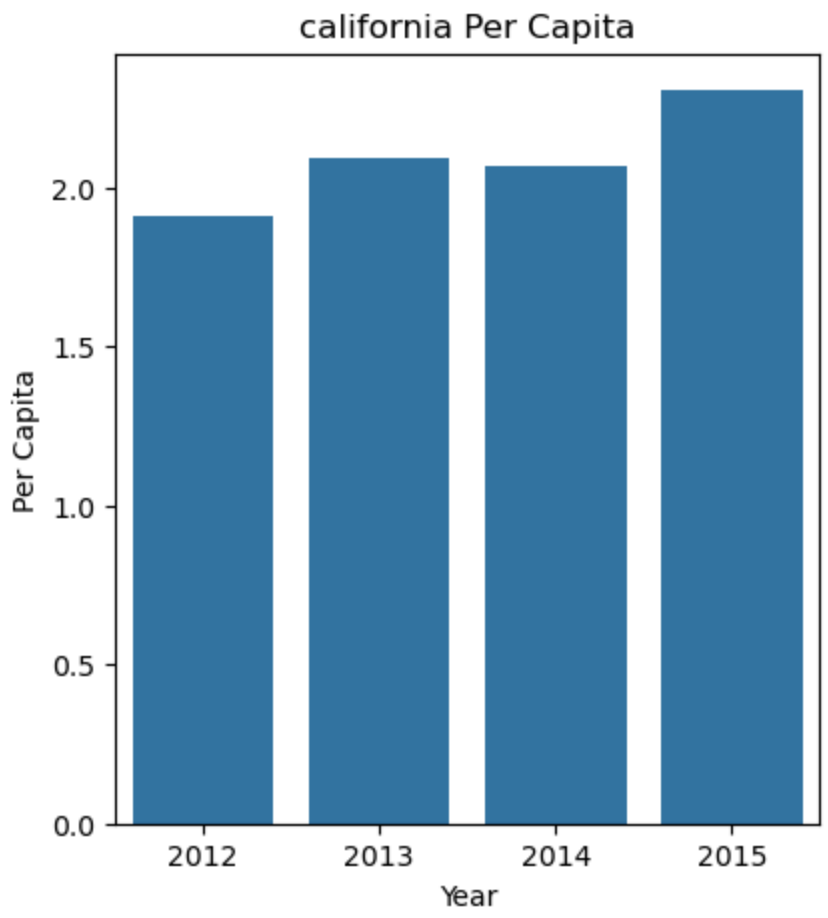




Don't be tricked by the visualizations, California has way higher stats in regards to spending AND population, having almost double the population and about seven times the spending on education over the years. But the education spending we see here being high because of the populations isn't really too fair, so we'll take a look at per capita spending now.

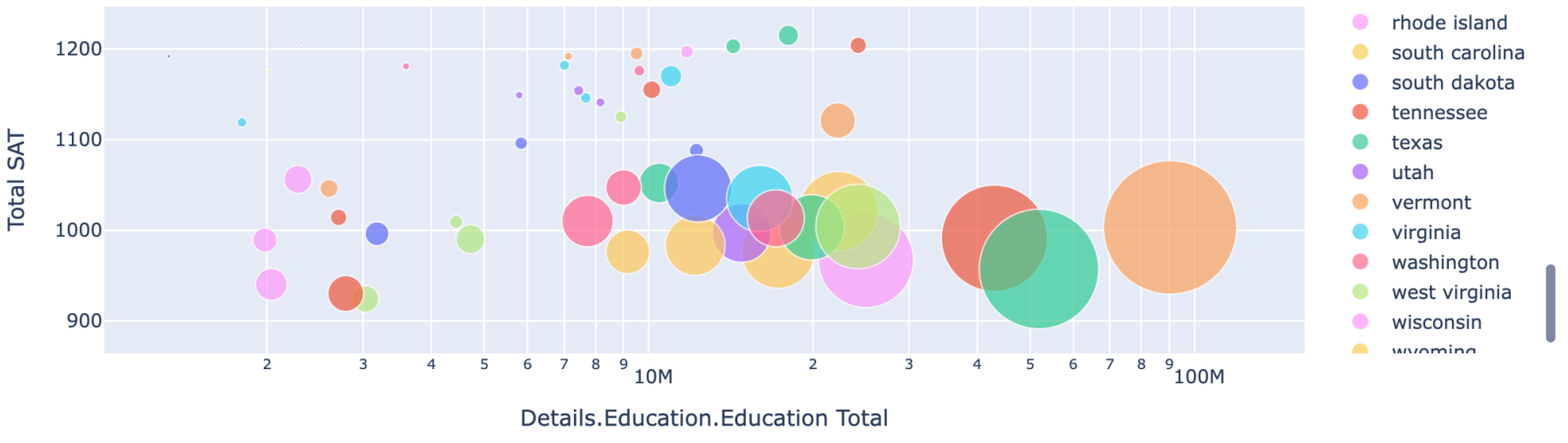
In [8]:
```
fig, [ax1, ax2] = plt.subplots(nrows=1, ncols=2, figsize=(10, 5))
cali_pc = state_stat_grapher(edu_spend_pc, "california", "Per Capita", True, ax2)
illi_pc = state_stat_grapher(edu_spend_pc, "illinois", "Per Capita", True, ax1)
```





I've also used a new library, plotly, to create an interactive bubble graph which provides a more interesting visualization

In [9]:
```
school_spending = merger(school_csv, finance_csv)
fig = px.scatter(school_spending.query("Year==2015"), x="Details.Education.Education Total", y="Total SAT"
                 size="Total.Test-takers", color="State",
                 hover_name="State", log_x=True, size_max=60)
fig.show()
```



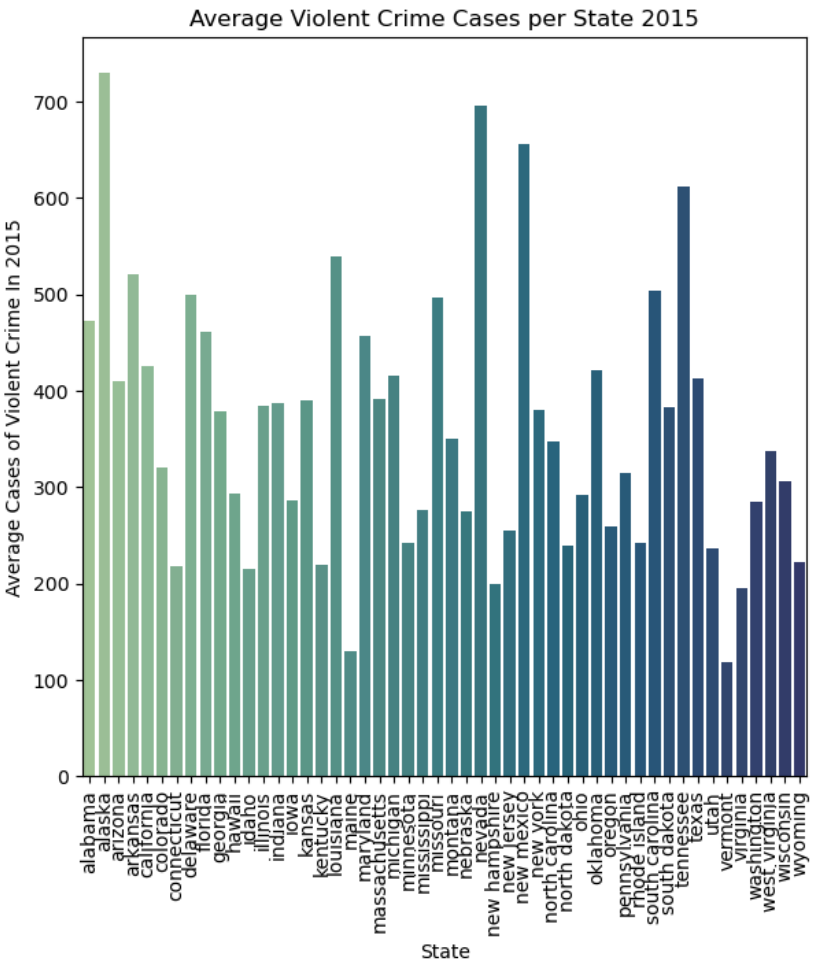
This interactive graph made using plotly is good for getting some of the intricacies and complexities that come with the topic. It is able to show to the average SAT score based on it's positioning on the y-axis, and the amount that state spent on education based on where it is on the x-axis. Since the size of balls are based on the amount of test takers, it makes it so you can decide whether or not you feel like it has enough weight to it.

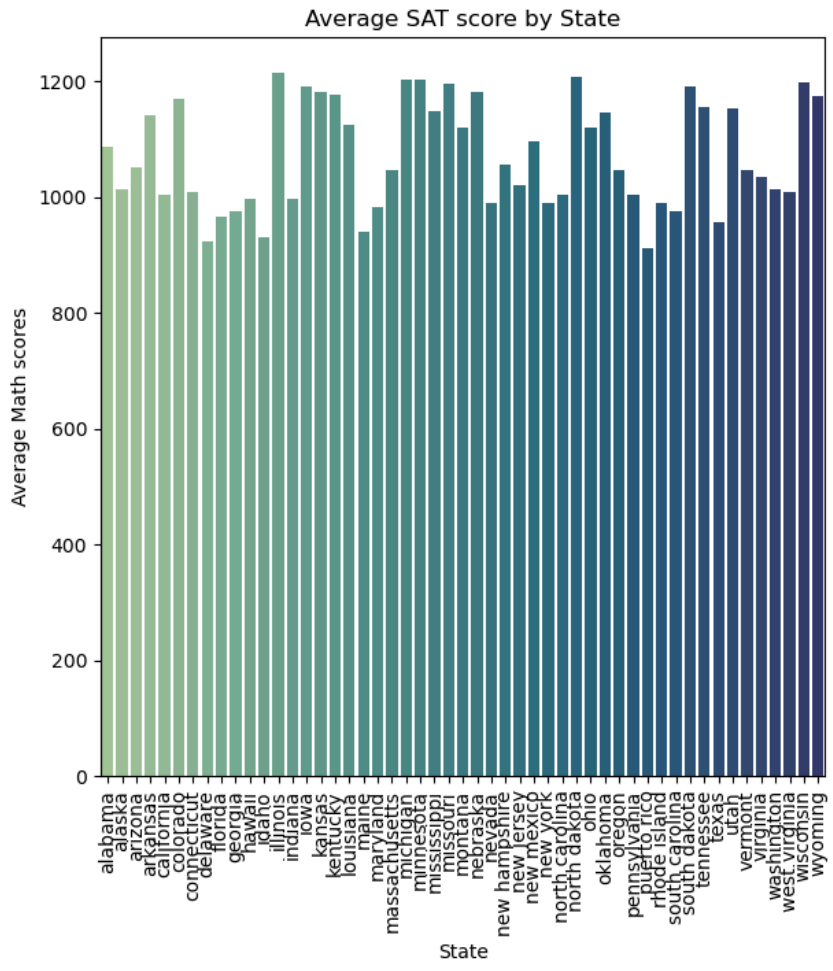
2.) Do states with higher violent crime rates have lower SAT scores?

For this question, I'll first be making to bar plots to compare crime statistics with

In [10]:
```
fig, [ax1, ax2] = plt.subplots(nrows=1, ncols=2, figsize=(15, 7))
crime_2015 = all_state_plot(crime_csv, 2015, "Data.Rates.Violent.All", True, ax1)
crime_2015.set(ylabel="Average Cases of Violent Crime In 2015",
               title="Average Violent Crime Cases per State 2015")
SAT_2015 = all_state_plot(school_csv, 2015, "Total SAT", True, ax2)
SAT_2015.set(ylabel = "Average Math scores", title="Average SAT score by State")
```

Out[10]:
```
[Text(0, 0.5, 'Average Math scores'),
 Text(0.5, 1.0, 'Average SAT score by State')]
```

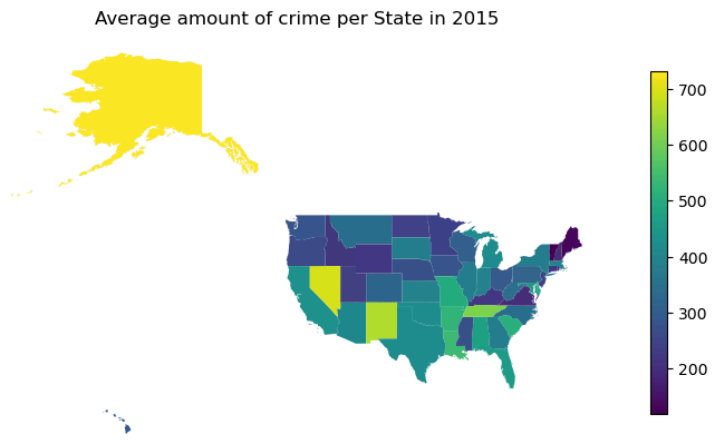


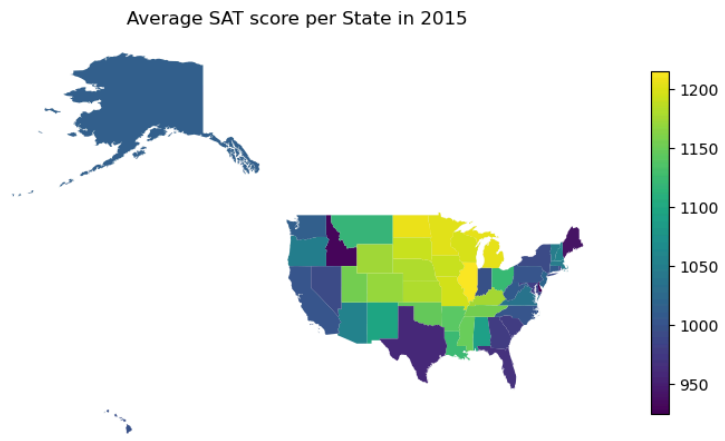


For a geospatial visualization we have:

In [11]:
```
fig, [ax1, ax2] = plt.subplots(nrows=1, ncols=2, figsize=(20, 10))
crime_geo = geo_merger(crime_csv, state_geo)
SAT_geo = geo_merger(school_csv, state_geo)
mapper(crime_geo, "Data.Rates.Violent.All", 2015, True, ax1)
ax1.set(title="Average amount of crime per State in 2015")
mapper(SAT_geo, "Total SAT", 2015, True, ax2)
ax2.set(title="Average SAT score per State in 2015")
```

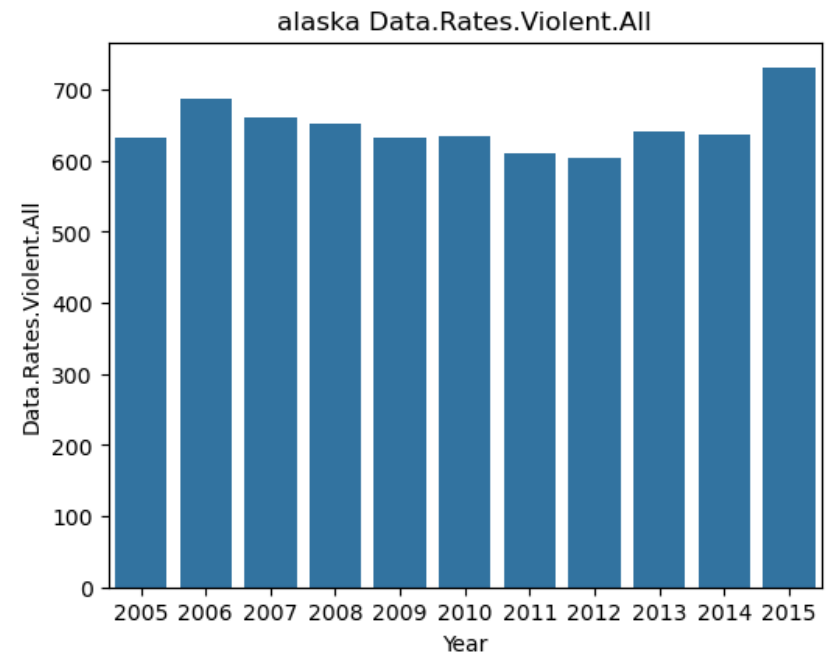
Out[11]:
```
[Text(0.5, 1.0, 'Average SAT score per State in 2015')]
```

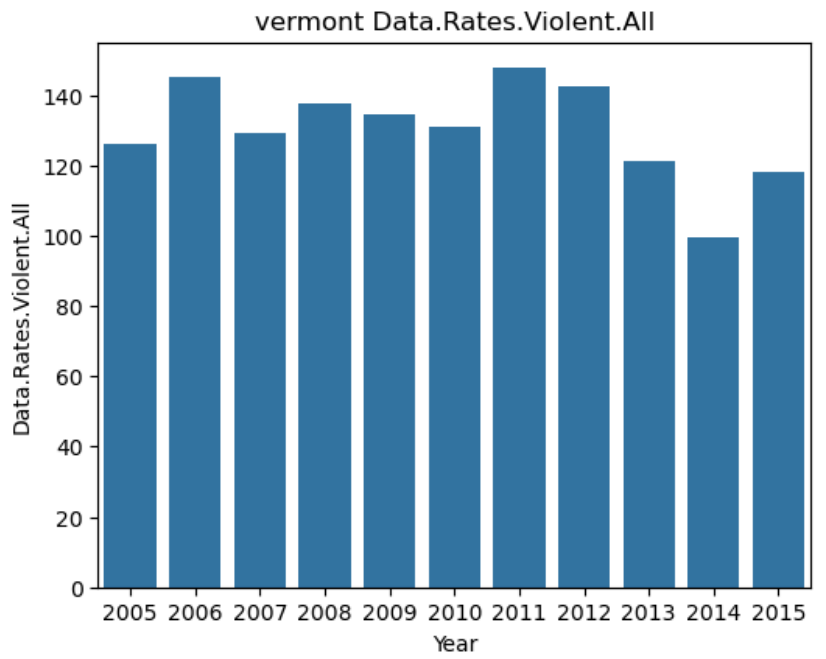


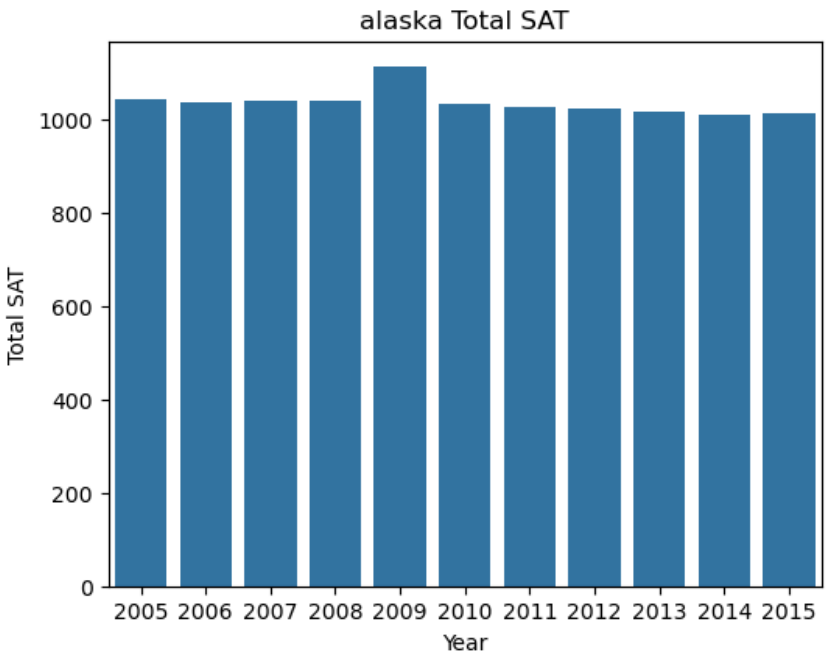


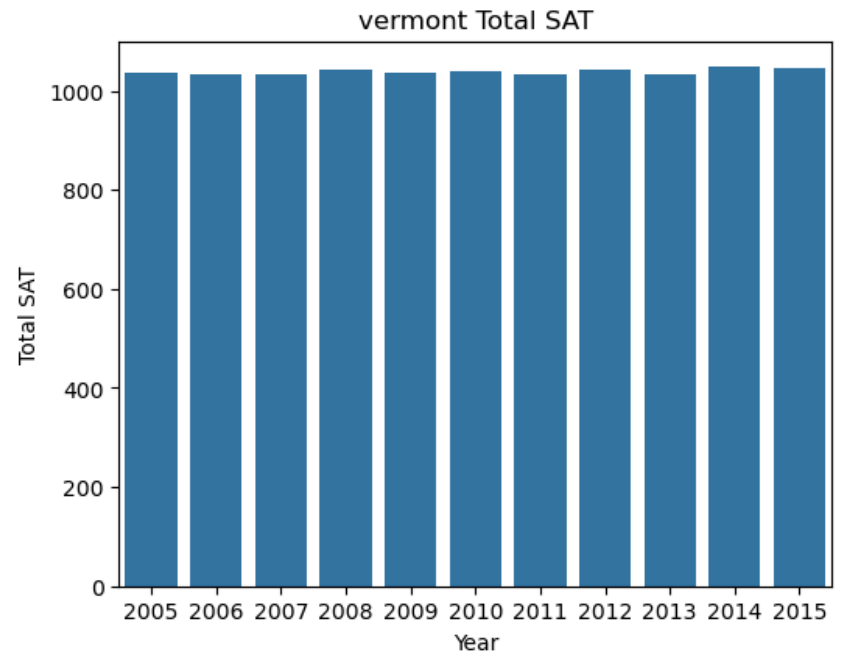
For how few people live in Alaska, it's kind of shocking how much violent crime there is! But that's not what this project is about. Though not too noticible, states with higher violent crime amounts, like Nevada and Alaska, tend to dip a little below average in regards to their SAT scores. For a closer look, lets compare Vermont, with the least amount of violent crime, to Alaska with the highest amount.

In [12]:
```
fig, [[ax1, ax2], [ax3, ax4]] = plt.subplots(nrows=2, ncols=2, figsize=(13, 10))
verm_crime = state_stat_grapher(crime_csv, "vermont", "Data.Rates.Violent.All", True, ax2)
nev_crime = state_stat_grapher(crime_csv, "alaska", "Data.Rates.Violent.All", True, ax1)
verm_sat = state_stat_grapher(school_csv, "vermont", "Total SAT", True, ax4)
nev_sat = state_stat_grapher(school_csv, "alaska", "Total SAT", True, ax3)
```









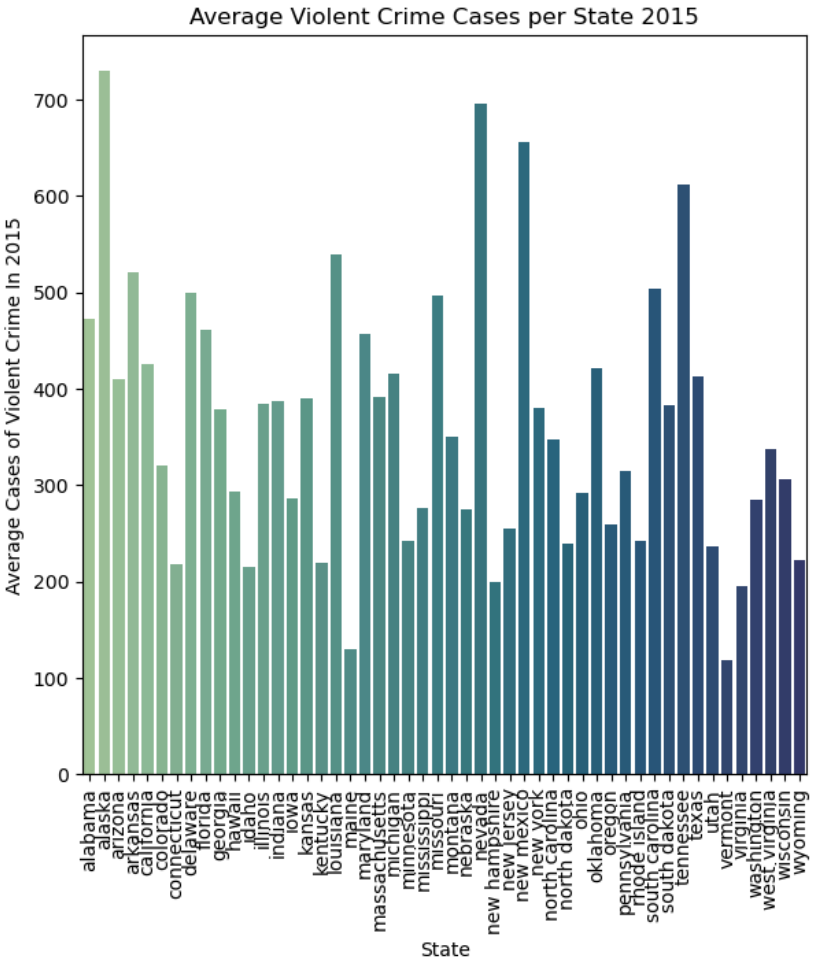
Through these visualizations, we can see that Vermont has consistently had lower violent crime rates that Alaska, while also maintainging a higher average SAT score, though not always by too much. It makes sense that crime rates and lower SAT scores could be correlated with each other, could you imagine having to study while living in fear for your safety?

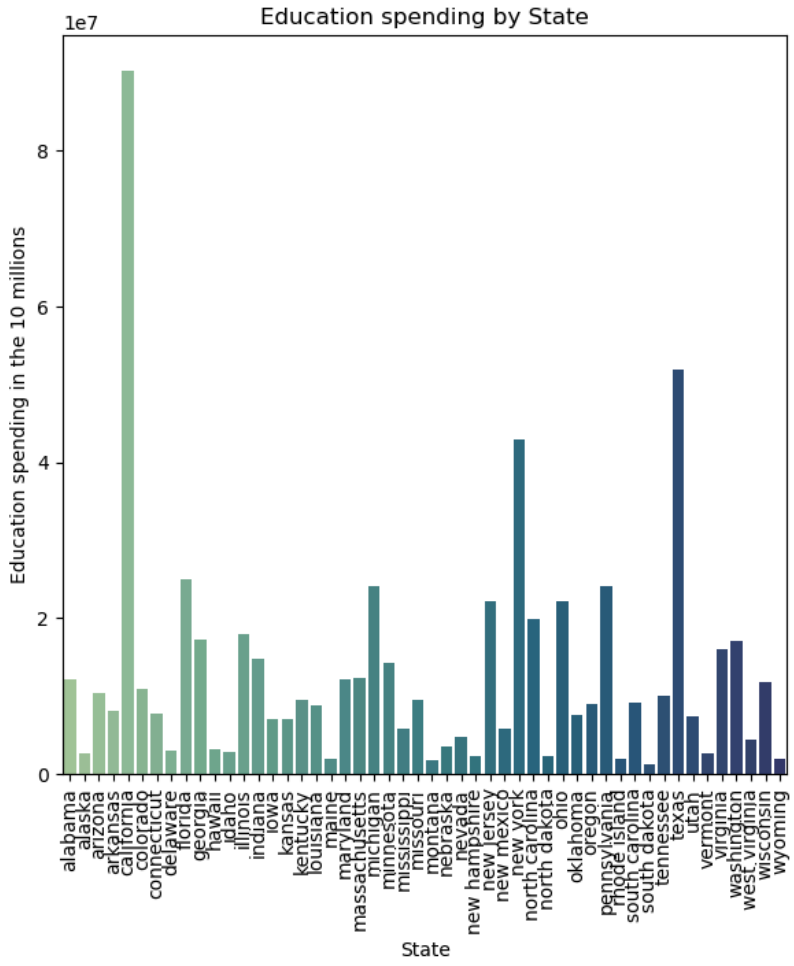
3.) Is there a link between spending on education and violent crime rates?

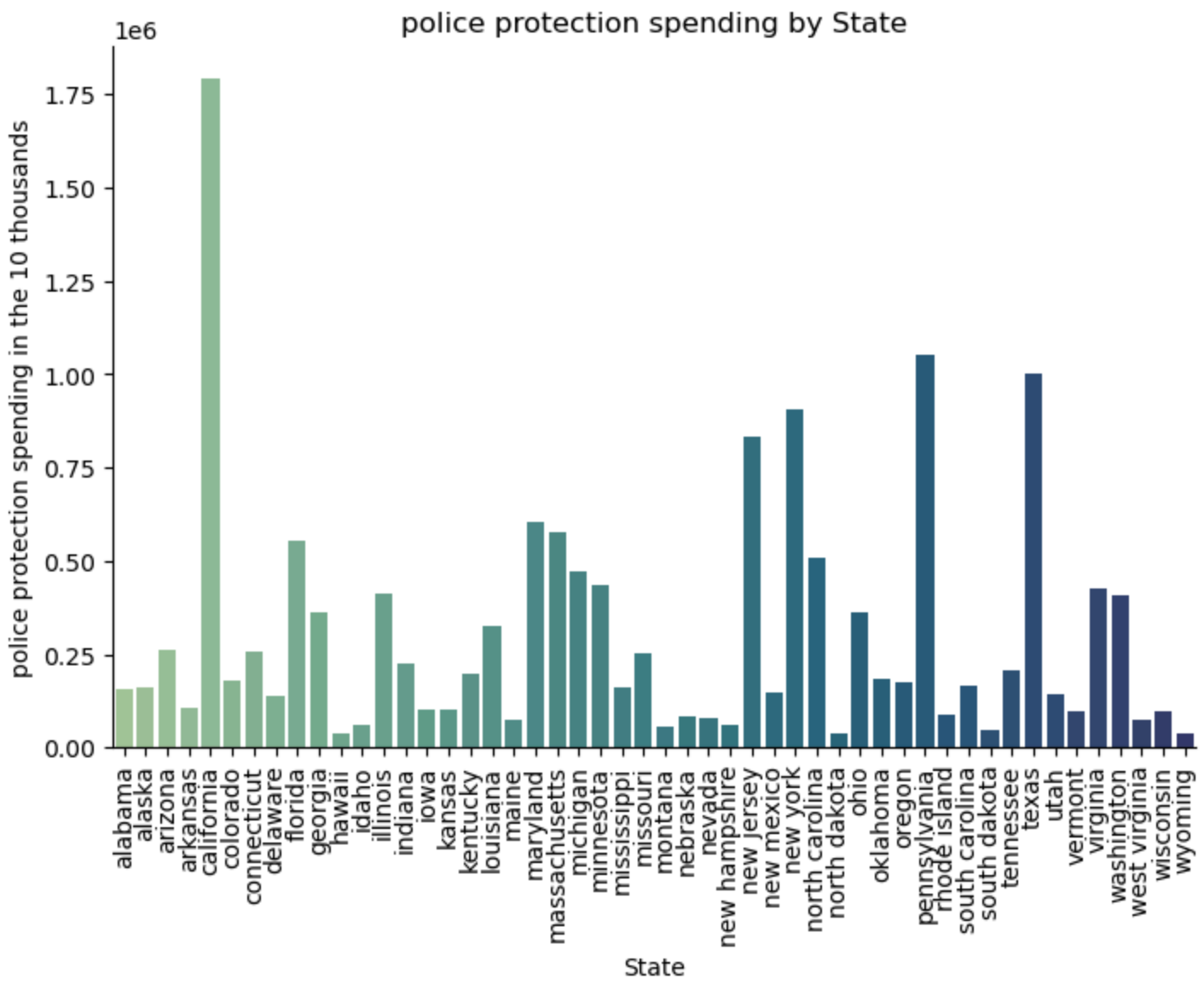
There have been many studies which highlight the shcool to prison pipeline, especially for students of color. For this question, I'll be analyzing the State spending compared to Violent crime rates to try and see if there may be a pattern there.

In [13]:
```
fig, [ax1, ax2] = plt.subplots(nrows=1, ncols=2, figsize=(15, 7))
crime_2015 = all_state_plot(crime_csv, 2015, "Data.Rates.Violent.All", True, ax1)
crime_2015.set(ylabel="Average Cases of Violent Crime In 2015",
               title="Average Violent Crime Cases per State 2015")
edu_spend_2015 = all_state_plot(finance_csv, 2015, "Details.Education.Education Total", True, ax2)
edu_spend_2015.set(ylabel = "Education spending in the 10 millions",
                   title="Education spending by State")
crime_spend_2015 = all_state_plot(finance_csv, 2015, "Details.Police protection")
crime_spend_2015.set(ylabel = "police protection spending in the 10 thousands",
                     title="police protection spending by State")
```

Out[13]:
```
<seaborn.axisgrid.FacetGrid at 0x7a7960e16920>
```

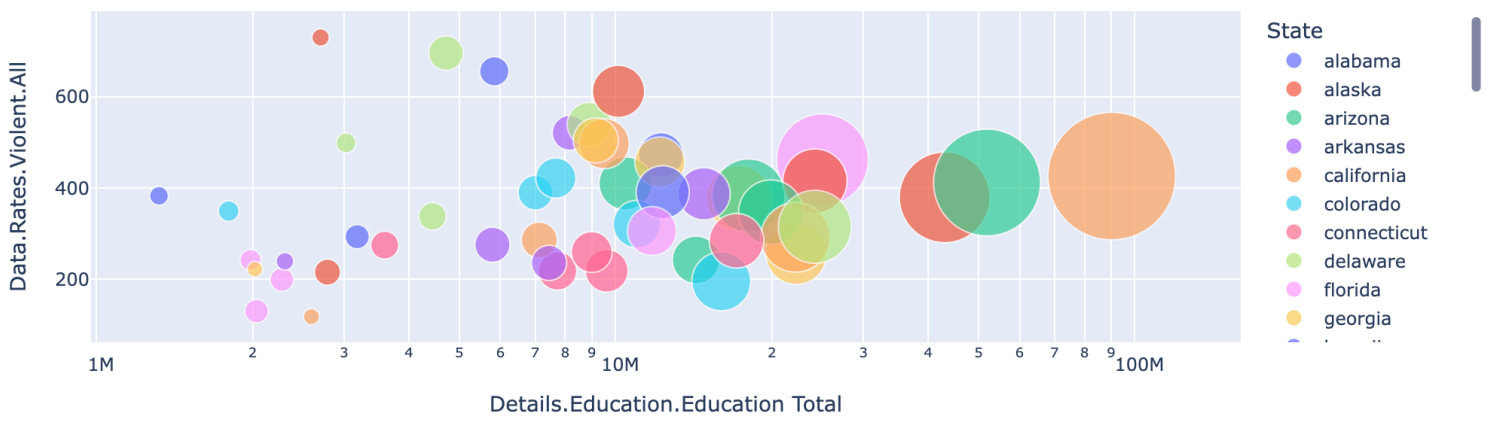






For another visualization, lets make another plotly interactive plot

In [14]:
```
finance_crime = merger(crime_csv, finance_csv)
fig = px.scatter(finance_crime.query("Year==2015"),
                 x="Details.Education.Education Total", y="Data.Rates.Violent.All",
                 size="Data.Population", color="State",
                 hover_name="State", log_x=True, size_max=60)
fig.show()
```



Through all of these visualizations there isn't a good enough patterin to discern wether or not spending on education really has any effect or correlation with violent crime rates. Violent crime rates can be pretty dependent on population size and state culture, the ladder of which can be seen in Alaska's violent crime rates, the same goes for spending on police protection.

Doctests

In [15]:
```
doctest.run_docstring_examples(cleaner, globals())
doctest.run_docstring_examples(merger, globals())
doctest.run_docstring_examples(per_capita_data, globals())
```

Implications and Limitations

For my third research question, numerical data such as state financing and police protection funding aren't able to paint the whole story. It's important to learn the actual culture of the states, and the other factors that might contribute to crime rates.

Unfortunately, there were quite a few limitations to my data when trying to analyze it. For one, my data for school scores only goes up to 2015, making it almost a decade old, and it misses out on the Covid-19 pandemic which changed the view of schooling for many students. Secondly, my per-capita education spending is based only on the total population of the states, not only just those who were currently in school at the time, which can throw off my per capita statistic. Due to these reasons, I feel like my project is better as an initial glance at what factors into SAT scores rather than a concrete definition. Additionally, I wasn't able to find any data on what different schools in the states tend to allocate their resources too, though some states have high education spending, we can't tell whether or not they use most of that funding on sports, or other non-SAT related programs.