msmdesign.nyc

data visualization information design

public school demographics by state

Catalogue Number: MSM-E23

Author: David White

Contact: david@msmdesign.nyc | msmdesign.nyc

Acknowledgements: NYC Open Data [https://opendata.cityofnewyork.us/]

Language: Python

Libraries Used: NumPy, pandas, matplotlib, seaborn

> ### What is Exploratory Data Analysis?
>
> **Exploratory data analysis (EDA)** is a technique used by data scientists to inspect, characterize and briefly summarize the contents of a dataset. EDA is often the first step when encountering a new or unfamiliar dataset. EDA helps the data scientist become acquainted with a dataset and test some basic assumptions about the data. By the end of the EDA process, some initial insights can be drawn from the dataset and a framework for further analysis or modeling is established.

0. About this Dataset

Data Source: U.S. Department of Education National Center for Education Statistics Common Core of Data (CCD)

"Public Elementary/Secondary School Universe Survey" 2018-19 v.1a; "Public Elementary/Secondary School Universe Survey Geographic Data (EDGE)" 2018-19 v.1a. Data provided by the National Center for Education Statistics - [http://nces.ed.gov/ccd/elsi/]

1. Prepare the Workspace

```
# import the libraries needed for data analysis and visualization

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


# import the dataset from GitHub

data = pd.read_csv("https://raw.githubusercontent.com/davidwhitemsm/selected-open-datasources


# confirm that the data has loaded by taking a glimpse at the first five rows of the dataset

data.head(5)
```

	State Name	Total Number Operational Charter Schools [Public School] 2018-19	Total Number of Public Schools [Public School] 2018-19	Total Students All Grades (Includes AE) [Public School] 2018-19	Free and Reduced Lunch Students [Public School] 2018-19	Total Race/Ethnicity [Public School] 2018-19	Pupil/Teacher Ratio [State] 2018-19
0	ALABAMA	2	1529	739304	407040	738495	17.56
1	ALASKA	28	510	130963	45537	130963	17.10
2	ARIZONA	558	2434	1136253	†	1135883	23.53
3	ARKANSAS	85	1080	491804	314759	490772	13.03
4	CALIFORNIA	1358	10437	6171666	3667601	6163584	23.08

5 rows × 42 columns

2. Describe the Characteristics of the Dataset

```
# determine the number of rows and columns in the dataset

data.shape
```

```
(51, 42)
```

TAKEAWAY: The dataset consists of 51 rows and 42 columns.

```
# list each of the columns contained in the dataset

data.columns
```

```
Index(['State Name',
       'Total Number Operational Charter Schools [Public School] 2018-19',
       'Total Number of Public Schools [Public School] 2018-19',
       'Total Students All Grades (Includes AE) [Public School] 2018-19',
       'Free and Reduced Lunch Students [Public School] 2018-19',
       'Total Race/Ethnicity [Public School] 2018-19',
       'Pupil/Teacher Ratio [State] 2018-19',
       'Full-Time Equivalent (FTE) Teachers [State] 2018-19',
       'Instructional Aides [State] 2018-19',
       'Guidance Counselors [State] 2018-19', 'Librarians [State] 2018-19',
       'Library Support Staff [State] 2018-19',
       'LEA Administrators [State] 2018-19',
       'LEA Administrative Support Staff [State] 2018-19',
       'School Administrators [State] 2018-19',
       'All Other Support Staff [State] 2018-19',
       'Student Support Services [State] 2018-19',
       'Full-Time Equivalent (FTE) Staff [State] 2018-19',
       'Instructional Coordinators [State] 2018-19',
       'Elementary Guidance Counselors [State] 2018-19',
       'Secondary Guidance Counselors [State] 2018-19',
       'Other Guidance Counselors [State] 2018-19',
       'School Administrative Support Staff [State] 2018-19',
       'Total Students [State] 2018-19',
       'Male Students [Public School] 2018-19',
       'Female Students [Public School] 2018-19',
       'American Indian/Alaska Native Students [Public School] 2018-19',
       'Asian or Asian/Pacific Islander Students [Public School] 2018-19',
       'Hispanic Students [Public School] 2018-19',
       'Black Students [Public School] 2018-19',
       'White Students [Public School] 2018-19',
       'Hawaiian Nat./Pacific Isl. Students [Public School] 2018-19',
       'Two or More Races Students [Public School] 2018-19',
       'Male Students [State] 2018-19', 'Female Students [State] 2018-19',
       'American Indian/Alaska Native Students [State] 2018-19',
       'Asian or Asian/Pacific Islander Students [State] 2018-19',
       'Hispanic Students [State] 2018-19', 'Black Students [State] 2018-19',
       'White Students [State] 2018-19',
       'Hawaiian Nat./Pacific Isl. Students [State] 2018-19',
       'Two or More Races Students [State] 2018-19'],
      dtype='object')
```

TAKEAWAY: The dataset consists of:

- student enrollment
- school staffing
- student demographic information

```
# list the datatype of each variable contained in the dataset and check to see which variable

data.info()
```

```
     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 51 entries, 0 to 50
     Data columns (total 42 columns):
     State Name                                                          51 non-null object
     Total Number Operational Charter Schools [Public School] 2018-19    51 non-null object
     Total Number of Public Schools [Public School] 2018-19              51 non-null int64
     Total Students All Grades (Includes AE) [Public School] 2018-19     51 non-null int64
     Free and Reduced Lunch Students [Public School] 2018-19             51 non-null object
     Total Race/Ethnicity [Public School] 2018-19                        51 non-null int64
     Pupil/Teacher Ratio [State] 2018-19                                 51 non-null float64
     Full-Time Equivalent (FTE) Teachers [State] 2018-19                 51 non-null float64
     Instructional Aides [State] 2018-19                                 51 non-null object
     Guidance Counselors [State] 2018-19                                 51 non-null float64
     Librarians [State] 2018-19                                          51 non-null float64
     Library Support Staff [State] 2018-19                               51 non-null object
     LEA Administrators [State] 2018-19                                  51 non-null float64
     LEA Administrative Support Staff [State] 2018-19                    51 non-null object
     School Administrators [State] 2018-19                               51 non-null float64
     All Other Support Staff [State] 2018-19                             51 non-null object
     Student Support Services [State] 2018-19                            51 non-null float64
     Full-Time Equivalent (FTE) Staff [State] 2018-19                    51 non-null float64
     Instructional Coordinators [State] 2018-19                          51 non-null float64
     Elementary Guidance Counselors [State] 2018-19                      51 non-null float64
     Secondary Guidance Counselors [State] 2018-19                       51 non-null float64
     Other Guidance Counselors [State] 2018-19                           51 non-null object
     School Administrative Support Staff [State] 2018-19                 51 non-null object
     Total Students [State] 2018-19                                      51 non-null int64
     Male Students [Public School] 2018-19                               51 non-null int64
     Female Students [Public School] 2018-19                             51 non-null int64
     American Indian/Alaska Native Students [Public School] 2018-19      51 non-null int64
     Asian or Asian/Pacific Islander Students [Public School] 2018-19    51 non-null int64
     Hispanic Students [Public School] 2018-19                           51 non-null int64
     Black Students [Public School] 2018-19                              51 non-null int64
     White Students [Public School] 2018-19                              51 non-null int64
     Hawaiian Nat./Pacific Isl. Students [Public School] 2018-19         51 non-null int64
     Two or More Races Students [Public School] 2018-19                  51 non-null int64
     Male Students [State] 2018-19                                       51 non-null int64
     Female Students [State] 2018-19                                     51 non-null int64
     American Indian/Alaska Native Students [State] 2018-19              51 non-null int64
     Asian or Asian/Pacific Islander Students [State] 2018-19            51 non-null int64
     Hispanic Students [State] 2018-19                                   51 non-null int64
     Black Students [State] 2018-19                                      51 non-null int64
     White Students [State] 2018-19                                      51 non-null int64
     Hawaiian Nat./Pacific Isl. Students [State] 2018-19                 51 non-null int64
     Two or More Races Students [State] 2018-19                          51 non-null int64
     dtypes: float64(11), int64(22), object(9)
     memory usage: 16.8+ KB
```

TAKEAWAY: There are 51 rows and 42 columns in the dataset. None of the rows are blank.

```
# the data set seems to have one row of data per U.S. State.
# let's test this assumption by checking for unique values in the 'State' column.

print(data['State Name'].unique())
```

```
     ['ALABAMA' 'ALASKA' 'ARIZONA' 'ARKANSAS' 'CALIFORNIA' 'COLORADO'
      'CONNECTICUT' 'DELAWARE' 'DISTRICT OF COLUMBIA' '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']
```

TAKEAWAY: The dataset contains one row for each US state plus the District of Columbia.

3. Summarize the Dataset

```
# create a subset of columns containing demographic subgroup information and glipse

subgroups = pd.read_csv("https://raw.githubusercontent.com/davidwhitemsm/selected-open-dataso
                 usecols=['State Name','Male Students [Public School] 2018-19',
       'Female Students [Public School] 2018-19',
       'American Indian/Alaska Native Students [Public School] 2018-19',
       'Asian or Asian/Pacific Islander Students [Public School] 2018-19',
       'Hispanic Students [Public School] 2018-19',
       'Black Students [Public School] 2018-19',
       'White Students [Public School] 2018-19',
       'Hawaiian Nat./Pacific Isl. Students [Public School] 2018-19','Two or More Races Stude
subgroups.head()
```

	State Name	Male Students [Public School] 2018-19	Female Students [Public School] 2018-19	American Indian/Alaska Native Students [Public School] 2018-19	Asian or Asian/Pacific Islander Students [Public School] 2018-19	Hispanic Students [Public School] 2018-19	Black Students [Public School] 2018-19	St [S 2
0	ALABAMA	379125	359370	6916	10860	62038	239759	

TAKEAWAY: The dataset contains totals per state of the number of students in (2) gender categories and (7) race/ethnicity categories.

```
# summary statistics on the dataset

subgroups.describe()
```

	Male Students [Public School] 2018-19	Female Students [Public School] 2018-19	American Indian/Alaska Native Students [Public School] 2018-19	Asian or Asian/Pacific Islander Students [Public School] 2018-19	Hispanic Students [Public School] 2018-19	B Stud [Pu Sch 201
count	5.100000e+01	5.100000e+01	51.000000	51.000000	5.100000e+01	51.00
mean	5.067872e+05	4.802107e+05	9266.470588	51948.647059	2.678522e+05	149220.37
std	6.029885e+05	5.722634e+05	15524.656776	109441.469833	6.122845e+05	176718.99
min	4.352000e+04	4.013100e+04	81.000000	784.000000	1.984000e+03	1026.00
25%	1.489505e+05	1.399745e+05	1321.000000	5657.500000	3.235800e+04	11702.00
50%	3.588130e+05	3.399870e+05	3536.000000	18202.000000	1.019080e+05	64606.00
75%	5.806480e+05	5.494835e+05	8757.000000	56638.000000	2.074140e+05	253845.50

```
subgroups.sum()
```

```
State Name                                                                ALABAMAALASKAARIZONA
Male Students [Public School] 2018-19
Female Students [Public School] 2018-19
American Indian/Alaska Native Students [Public School] 2018-19
Asian or Asian/Pacific Islander Students [Public School] 2018-19
Hispanic Students [Public School] 2018-19
Black Students [Public School] 2018-19
White Students [Public School] 2018-19
Hawaiian Nat./Pacific Isl. Students [Public School] 2018-19
Two or More Races Students [Public School] 2018-19
dtype: object
```

TAKEAWAY: 2018-19 US public school total enrollments by demographic group are as follows:

- 25.8 million male students
- 24.4 million female students
- 473K American Indian/Alaska Native students
- 2.6 million Asian or Asian/Pacific Islander students
- 13.7 million Hispanic students
- 7.6 million Black students
- 23.7 million White students
- 176K Hawaiian Nat./Pacific Isl. students
- 2 million multiracial students

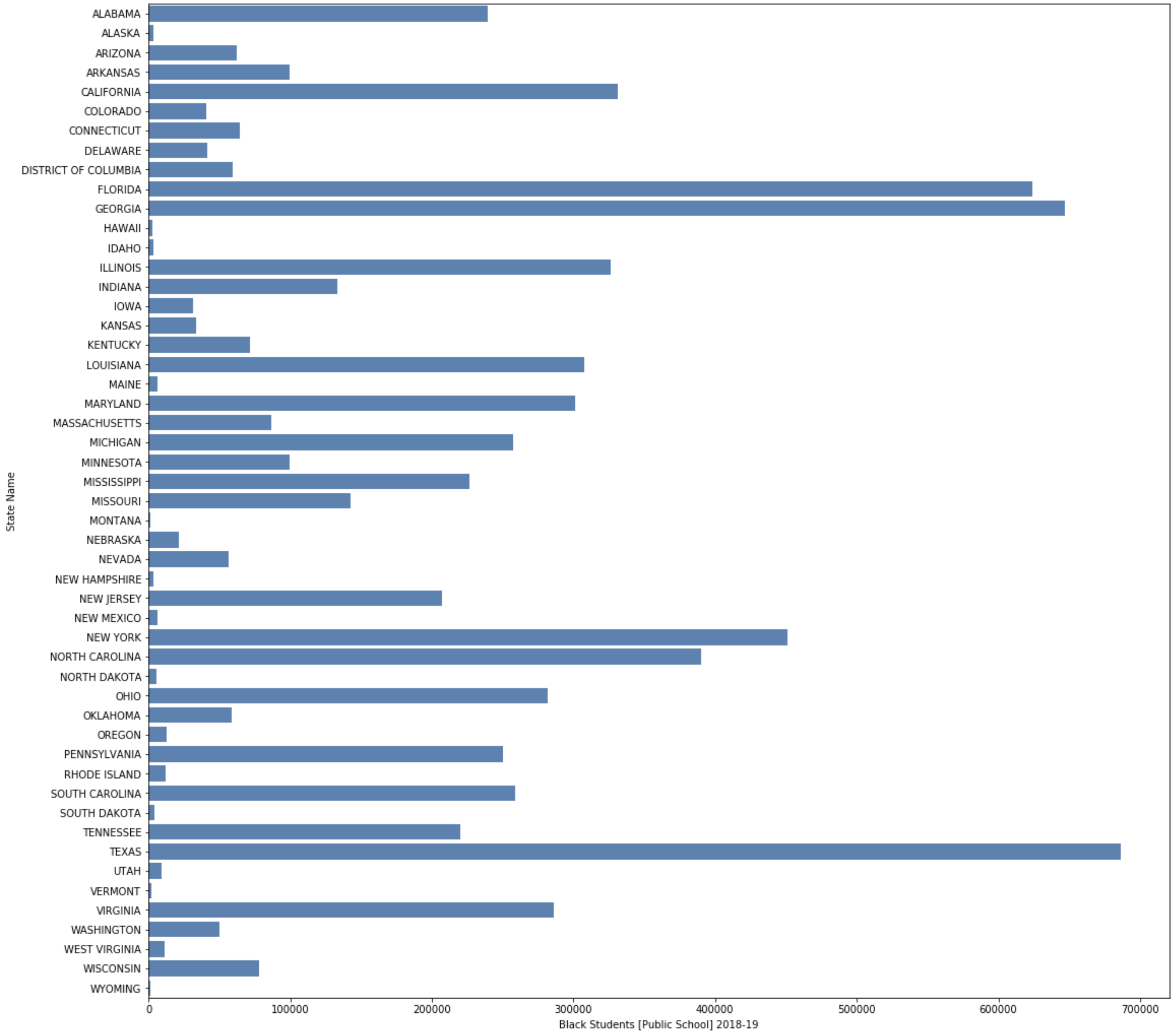
```
# here is the same information presented in a pivot table format

states = subgroups.T
states
```

	0	1	2	3	4	5	
State Name	ALABAMA	ALASKA	ARIZONA	ARKANSAS	CALIFORNIA	COLORADO	CONNECTI
Male Students [Public School] 2018-19	379125	67626	581012	252133	3167866	468238	263
Female Students [Public School] 2018-19	359370	63337	554871	239671	3003800	443103	251
American Indian/Alaska Native Students [Public School] 2018-19	6916	29839	50877	2771	28381	5961	
Asian or Asian/Pacific Islander Students [Public School] 2018-19	10860	7599	33428	7750	721827	28743	26
Hispanic							

4. Visualize the Dataset

19

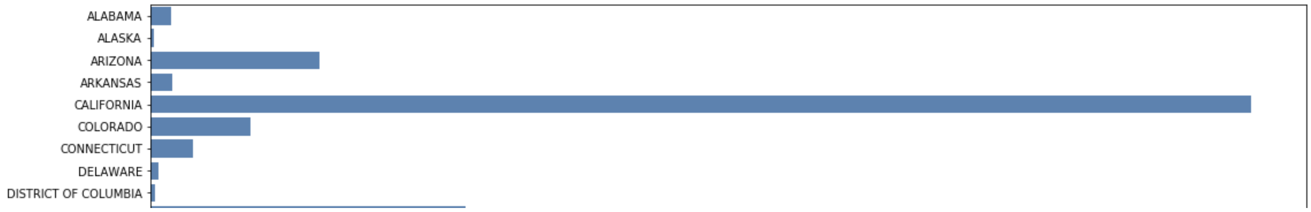
```
# plot Black public school students by state

plt.figure(figsize=(18, 18))
sns.barplot(x='Black Students [Public School] 2018-19',y='State Name', color='#4f81bd',data=s
<matplotlib.axes._subplots.AxesSubplot at 0x24eb8ba2080>
```



TAKEAWAY: The states with the highest number of Black public school students are: Florida, Georgia and Texas.

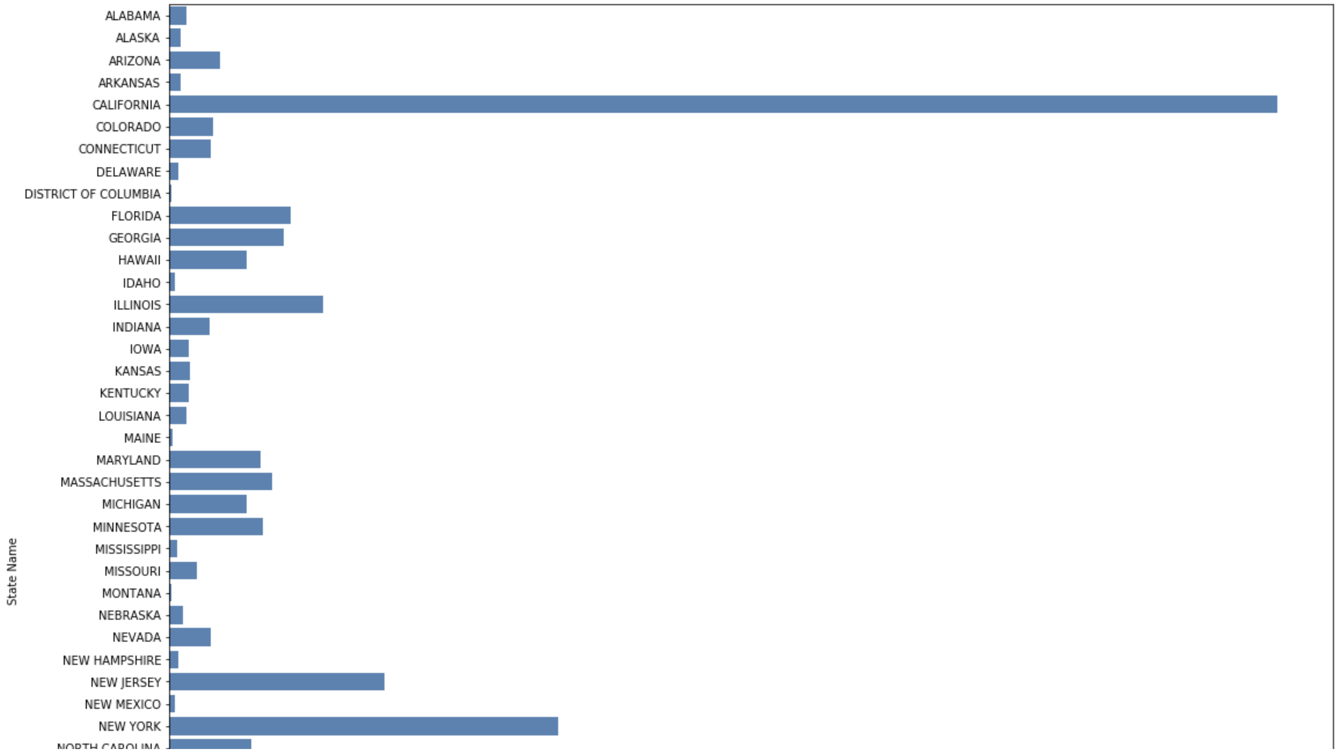
```
# plot Hispanic public school students by state
```

```
plt.figure(figsize=(18, 18))
sns.barplot(x='Hispanic Students [Public School] 2018-19',y='State Name', color='#4f81bd',dat
<matplotlib.axes._subplots.AxesSubplot at 0x24eb8f93d68>
```



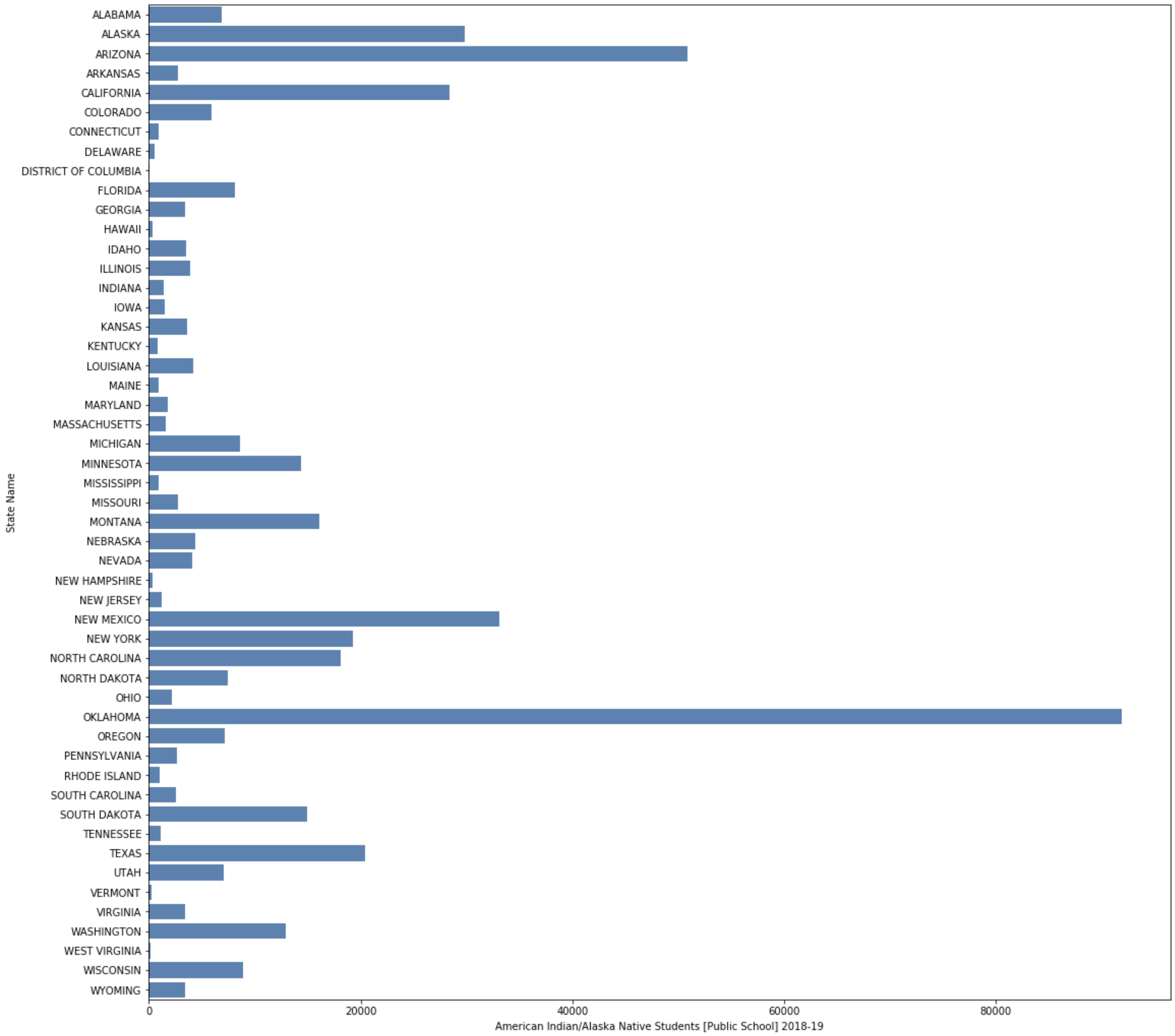
TAKEAWAY: The states with the highest number of Hispanic public school students are: California and Texas.

```
# plot Asian or Pacific Islander public school students by state

plt.figure(figsize=(18, 18))
sns.barplot(x='Asian or Asian/Pacific Islander Students [Public School] 2018-19',y='State Nam
<matplotlib.axes._subplots.AxesSubplot at 0x24eb8b8fdd8>
```



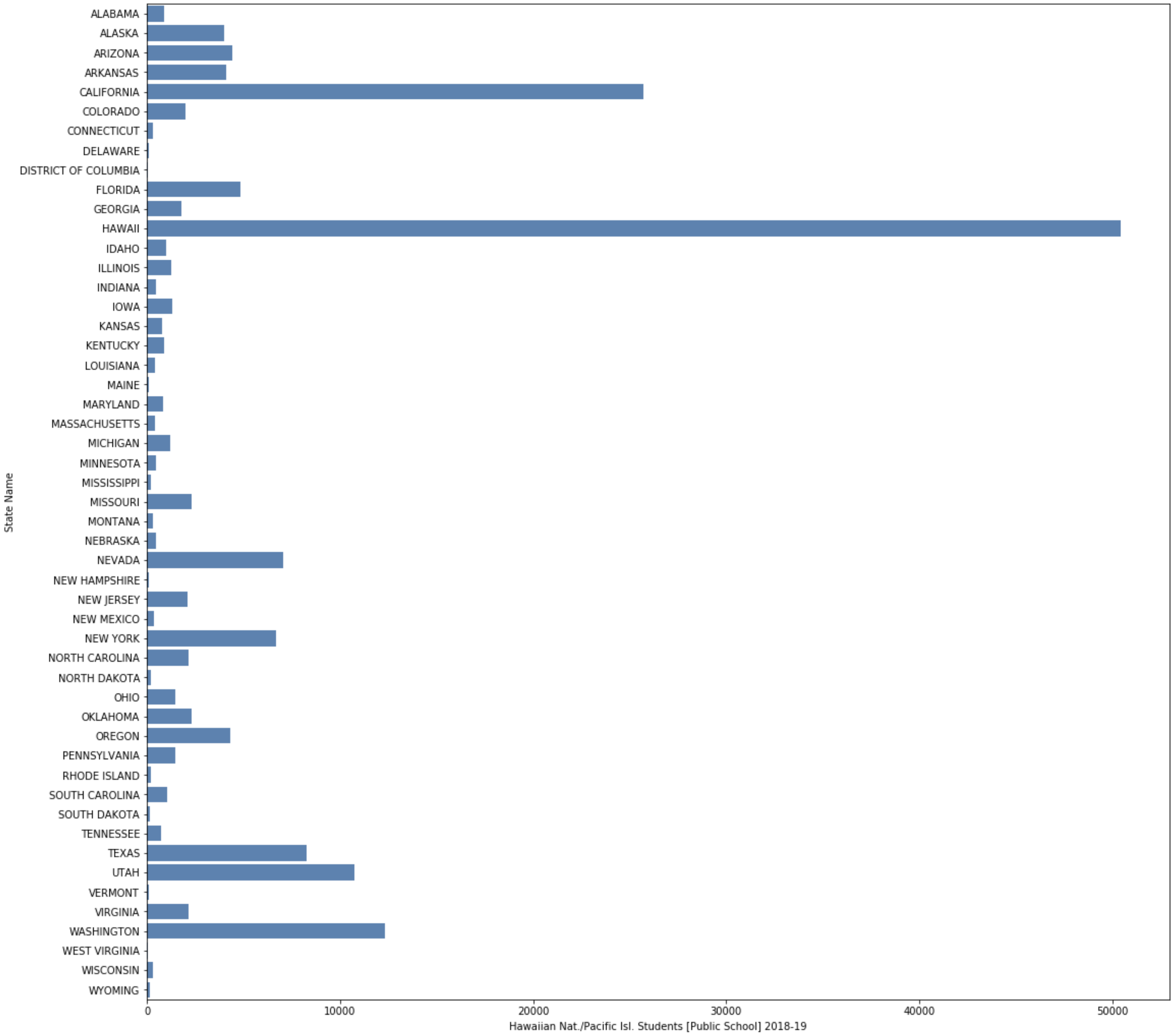
TAKEAWAY: The state with the highest number of Asian or Asian/Pacific Islander public school students is California. New York and Texas are a distant second and third.

```
# plot American Indian/Alaska Native public school students by state

plt.figure(figsize=(18, 18))
sns.barplot(x='American Indian/Alaska Native Students [Public School] 2018-19',y='State Name'
<matplotlib.axes._subplots.AxesSubplot at 0x24eb8d22748>
```



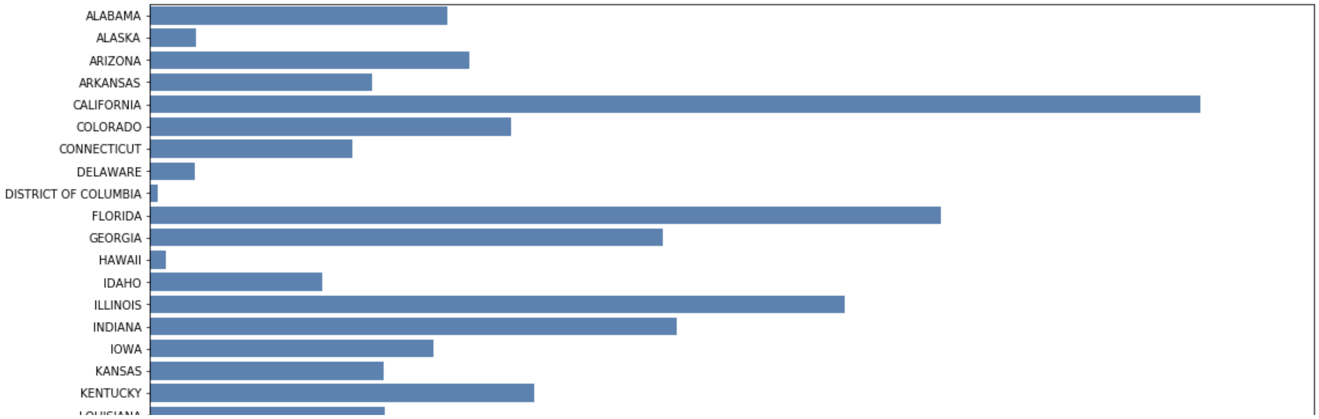
TAKEAWAY: The state with the highest number of American Indian/Alaska Native public school students by far is Oklahoma.

```
# plot Hawaiian/Pacific Islander public school students by state

plt.figure(figsize=(18, 18))
sns.barplot(x='Hawaiian Nat./Pacific Isl. Students [Public School] 2018-19',y='State Name', c
<matplotlib.axes._subplots.AxesSubplot at 0x24eba444ba8>
```


ALABAMA
ALASKA
ARIZONA
ARKANSAS
CALIFORNIA
COLORADO
CONNECTICUT
DELAWARE
DISTRICT OF COLUMBIA
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 Name
0
10000
20000
30000
40000
50000
Hawaiian Nat./Pacific Isl. Students [Public School] 2018-19

TAKEAWAY: The state with the highest number of Hawaiian/Pacific Islander public school students by far is Hawaii.

```
# plot White public school students by state

plt.figure(figsize=(18, 18))
sns.barplot(x='White Students [Public School] 2018-19',y='State Name', color='#4f81bd',data=s
<matplotlib.axes._subplots.AxesSubplot at 0x24eba955e80>
```



TAKEAWAY: The states with the highest number of White public school students are: California and Texas.

5. Key Insights

Populations of White students and populations of White students are mostly in portion with the state's overall population. However, for other demographic groups, students of that ethnicity are more heavily concentrated in just a handful of states.

Next Steps

Possible avenues for further research and analysis:

- calculate each demographic group as a percentage of each state's overall public school population
- compare this data to non-public school enrollments by state
- compares education outcomes (graduation rates) accross states and demographic groups