Is there a bias between the ACT and SAT and what advice can we provide to the education board to help increase test scores?

DSIF-2 (Team 5)

TABLE OF CONTENTS

Team 5

Chee Yuan Ng

Zheqin Lin

Benjamin Toh

Henri Oei

SAT vs ACT

SAT

- Number of subject: 2
(Evidence-based Reading and Writing, Math)
- Each component is scored on a scale of 200 to 800
- Total score ranges from 400 to 1600

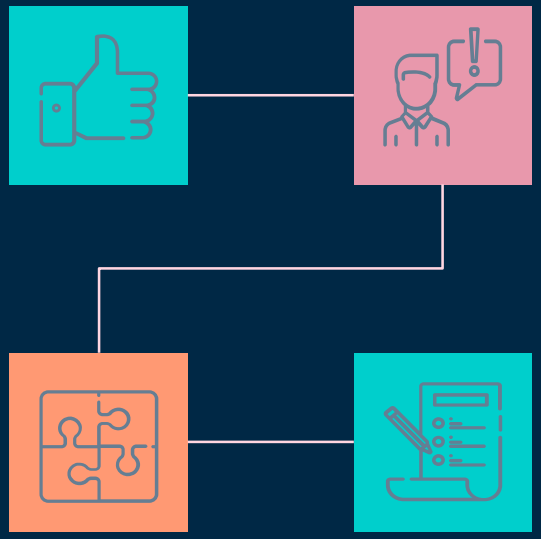
ACT

- Number of subjects: 4
(English, Math, Reading, Science)
- Each component is scored on a scale of 1 to 36
- Average composite score ranges from 1 to 36

Data Cleaning

Missing values

No null values noted



Errors in Data

6 Incorrect data observed

Data type

Participation rates and scores set to either integer or float

Naming + Dropping

Standardization of column names and dropping of rows not needed

Missing values – no null values noted for SAT & ACT

```
#Check for missing values.

sat2017_null = sat2017.isnull().sum().sort_values(ascending=False)
sat2018_null = sat2018.isnull().sum().sort_values(ascending=False)
sat2019_null = sat2019.isnull().sum().sort_values(ascending=False)

print(sat2017_null)
print(sat2018_null)
print(sat2019_null)
```

```
Total                                0
Math                                 0
Evidence-Based Reading and Writing   0
Participation                        0
State                                0
dtype: int64
Total                                0
Math                                 0
Evidence-Based Reading and Writing   0
Participation                        0
State                                0
dtype: int64
Total                0
Math                 0
EBRW                 0
Participation Rate   0
State                0
dtype: int64
```

No missing values noted for SAT 2017, 2018 and 2019

```
display(df_a17.isna().sum())
display(df_a18.isna().sum())
display(df_a19.isna().sum())
#no nan cells found in any of the data
```

```
State            0
Participation    0
Composite        0
dtype: int64

State            0
Participation    0
Composite        0
dtype: int64

State            0
Participation    0
Composite        0
```

Errors in Data

1. Math is low for SAT 2017

   Compared values of 'Total' vs Math + RW and added 524

2. Science is low for ACT 2017

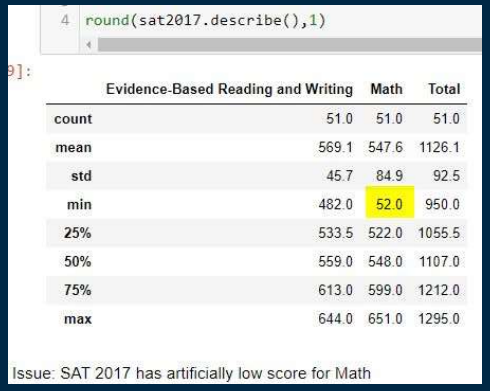
   Compared average of the 4 test results vs the Composite

3. Duplicate line – ACT (Maine)

4. 'X' noted in Composite score for ACT

5. Lower case noted for 'District of columbia'

6. Inconsistent datapoints (through visual inspection of excel data) for SAT 2019 Participation

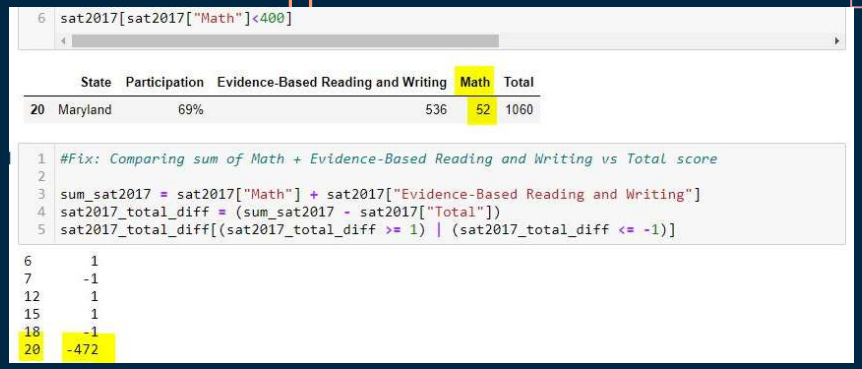


```
4  round(sat2017.describe(),1)
```

9]:

	Evidence-Based Reading and Writing	Math	Total
count	51.0	51.0	51.0
mean	569.1	547.6	1126.1
std	45.7	84.9	92.5
min	482.0	52.0	950.0
25%	533.5	522.0	1055.5
50%	559.0	548.0	1107.0
75%	613.0	599.0	1212.0
max	644.0	651.0	1295.0

Issue: SAT 2017 has artificially low score for Math



```
6  sat2017[sat2017["Math"]<400]
```

	State	Participation	Evidence-Based Reading and Writing	Math	Total
20	Maryland	69%	536	52	1060

```
1  #Fix: Comparing sum of Math + Evidence-Based Reading and Writing vs Total score
2
3  sum_sat2017 = sat2017["Math"] + sat2017["Evidence-Based Reading and Writing"]
4  sat2017_total_diff = (sum_sat2017 - sat2017["Total"])
5  sat2017_total_diff[(sat2017_total_diff >= 1) | (sat2017_total_diff <= -1)]
```

```
6      1
7     -1
12     1
15     1
18    -1
20  -472
```

	State	Participation	English	Math	Reading	Science	Composite	Composite_test	Check
21	Maryland	28%	23.3	23.1	24.2	2.3	23.6	18.225	5.375

	State	Participation_17	Composite_17	Participation_18	Composite_18	Participation	Composite
19	maine	0.08	24.3	0.07	24.0	0.06	24.3
20	maine	0.08	24.3	0.07	24.0	0.06	24.3

```
1  df_a17['Composite'].values
```

```
array(['21.0', '19.2', '19.8', '19.7', '19.4', '22.8', '20.8', '25.2',
       '24.1', '24.2', '19.8', '21.4', '19.0', '22.3', '21.4', '22.6',
       '21.9', '21.7', '20.0', '19.5', '24.3', '23.6', '25.4', '24.1',
       '21.5', '18.6', '20.4', '20.3', '21.4', '17.8', '25.5', '23.9',
       '19.7', '24.2', '19.1', '20.3', '22.0', '19.4', '21.8', '23.7',
       '24.0', '18.7', '21.8', '19.8', '20.7', '20.3', '23.6', '23.8',
       '21.9', '20.4', '20.5', '20.2x'], dtype=object)
```

	A	B
	State	Participa
	Puerto Rico	â€"
	Virgin Islands	â€"

Data type – Fix incorrect data types

SAT/ACT Participation stored as string

```
3  sat2017.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51 entries, 0 to 50
Data columns (total 5 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   State                            51 non-null     object
 1   Participation                    51 non-null     object
 2   Evidence-Based Reading and Writing  51 non-null  int64
 3   Math                             51 non-null     int64
 4   Total                            51 non-null     int64
dtypes: int64(3), object(2)
memory usage: 2.1+ KB
```

'Participation' and 'Participation Rate' for sat2017/sat2018 and sat2019 are stored as string which is inconsistent with our common understanding (should be a percentage/float type)

```
1  #Fix any incorrect data types found in step 5.
2
3  def convert_type(x):
4      x = x.strip('%')
5      x = float(x)
6      return x
7
8  sat2017['Participation'] = sat2017['Participation'].map(convert_type)
9  sat2017.head()
```

1]:

	State	Participation	Evidence-Based Reading and Writing	Math	Total
0	Alabama	5.0	593	572	1165
1	Alaska	38.0	547	533	1080
2	Arizona	30.0	563	553	1116
3	Arkansas	3.0	614	594	1208
4	California	53.0	531	524	1055

Naming + Dropping

Standardization of columns

```
# Rename column names to be lower case

sat2017.columns= ["state","sat_17_part","sat_17_rw","sat_17_math","sat_17_total"]
sat2018.columns= ["state","sat_18_part","sat_18_rw","sat_18_math","sat_18_total"]
sat2019.columns= ["state","sat_19_part","sat_19_rw","sat_19_math","sat_19_total"]

print(sat2017.head())
print(sat2018.head())
print(sat2019.head())
```

	state	sat_17_part	sat_17_rw	sat_17_math	sat_17_total
0	Alabama	5.0	593	572	1165
1	Alaska	38.0	547	533	1080
2	Arizona	30.0	563	553	1116
3	Arkansas	3.0	614	594	1208
4	California	53.0	531	524	1055

Check for duplicates

```
#Check sat2017 data - duplicated rows

sat2017[sat2017.duplicated()]
```

state	sat_17_part	sat_17_rw	sat_17_math	sat_17_total

Identify additional rows

SAT 2019 has 2 additional rows
(0.24% and 0.03% of total pop)

```
#Comparing difference in 'state' data between sat2017, sat2018 and sat2019

list_diff1 = []
for i in list1:
    if i not in list2 and list3:
        list_diff1.append(i)

print(list_diff1)

list_diff2 = []
for i in list2:
    if i not in list1 and list3:
        list_diff2.append(i)

print(list_diff2)

list_diff3 = []
for i in list3:
    if i not in list1 and list2:
        list_diff3.append(i)

print(list_diff3)
```

```
[]
[]
['Puerto Rico', 'Virgin Islands']
```

```
#8. Drop unnecessary rows (if needed).
#Filter for row information for 'State - Puerto Rico'

sat2019[sat2019['state'] == "Puerto Rico"]
```

	state	sat_19_part	sat_19_rw	sat_19_math	sat_19_total
39	Puerto Rico	—	483	462	944

```
#8. Drop unnecessary rows (if needed).
#Filter for row information for 'State - Virgin Islands'

sat2019[sat2019['state'] == "Virgin Islands"]
```

	state	sat_19_part	sat_19_rw	sat_19_math	sat_19_total
47	Virgin Islands	—	490	445	935

```
sat2019.drop(index=[39, 47], inplace =True)
print(f'The shape of the updated sat2019 dataset is: {sat2019.shape}.')
```

```
The shape of the updated sat2019 dataset is: (51, 5).
```

External Data Sources

ACT

- https://www.act.org/content/act/en/products-and-services/the-act/registration/test-center-locator.html - ACT test center
- https://www.census.gov/quickfacts/fact/table/MS,NC,HI,SC,NV/EDU685219#EDU685219 - Bachelor degree
- https://www.census.gov/quickfacts/fact/table/NH,MA,CT,DC,ME/INC910219#INC910219 - Per Capita Income
- https://www.statista.com/statistics/200445/reported-violent-crime-rate-in-the-us-states/ - Crime rate
- https://www.niche.com/k12/search/college-prep-private-high-schools/
- http://www.act.org/content/dam/act/unsecured/documents/cccr-2019/Connecticut-CCCR-2019.pdf

SAT

- https://collegereadiness.collegeboard.org/sat/register/find-test-centers - Test Center
- https://data.census.gov/cedsci/table?text=S1901&g=0100000US,.04000.001&tid=ACSST5Y2019.S1901&hidePreview=true - Income

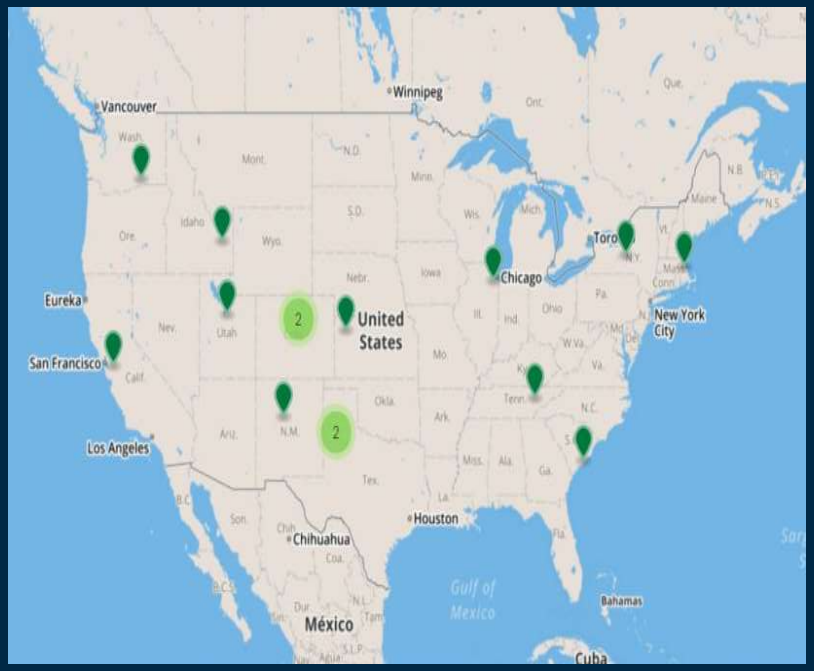
Accessibility

What is it?
This is a variable to determine the ratio of graduates who took ACT/SAT exams to 1 test center

Why do we use this?
We want to know whether does convenience of test centers affect the exam scores.

How we derived it?
Accessibility = (Participate Rate * No. of Graduates) / No. of Test Centers

- (Participate Rate * No. of Graduates) is to find out how many graduates took ACT/SAT exams



ACT vs SAT

The Analysis

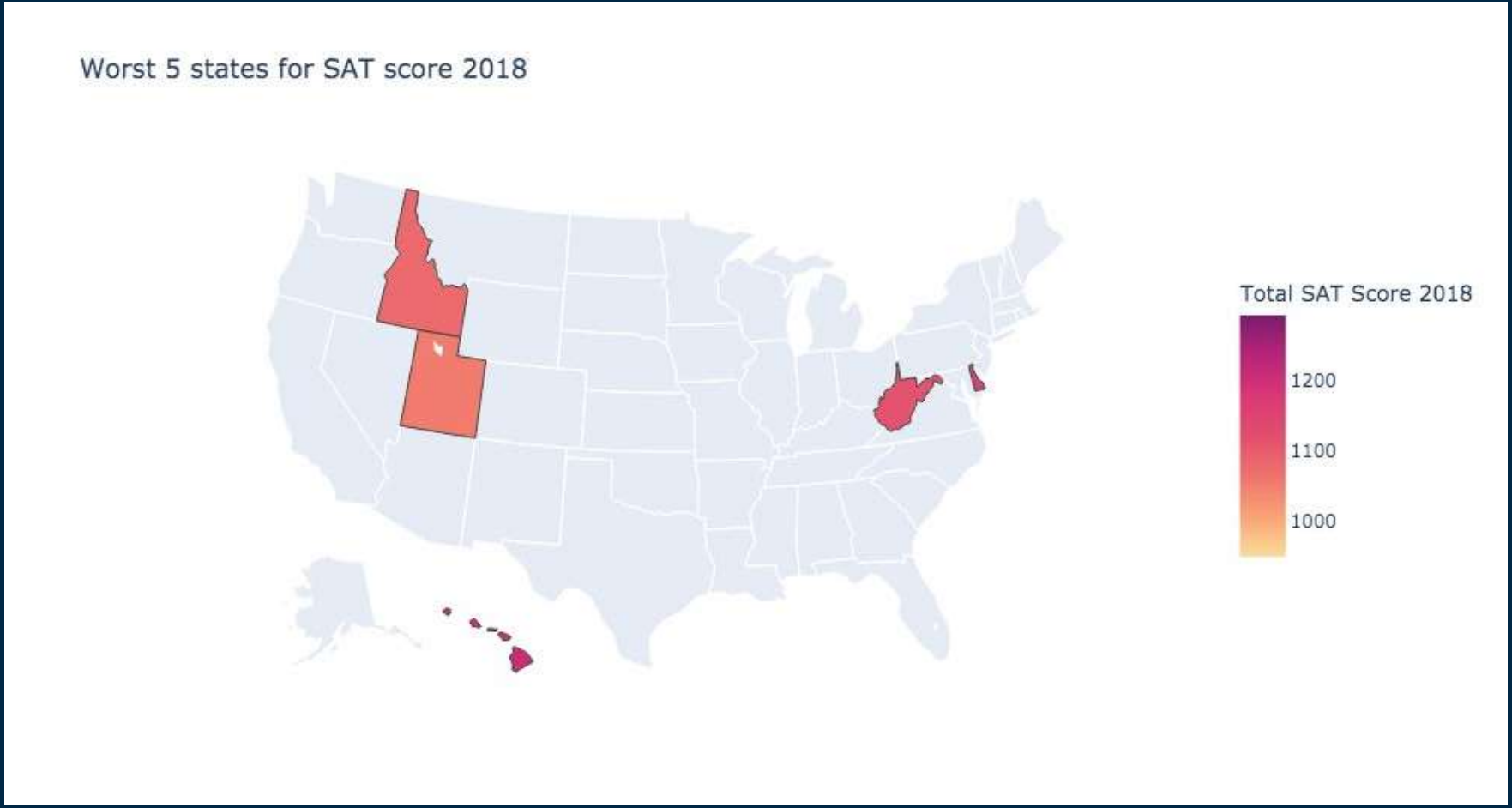
01

Geographical Map on SAT scores 2017

For Year 2017, top 5 worst SAT scores are from **Delaware**, **Idaho**, **Michigan**, **Maine** and **Florida**

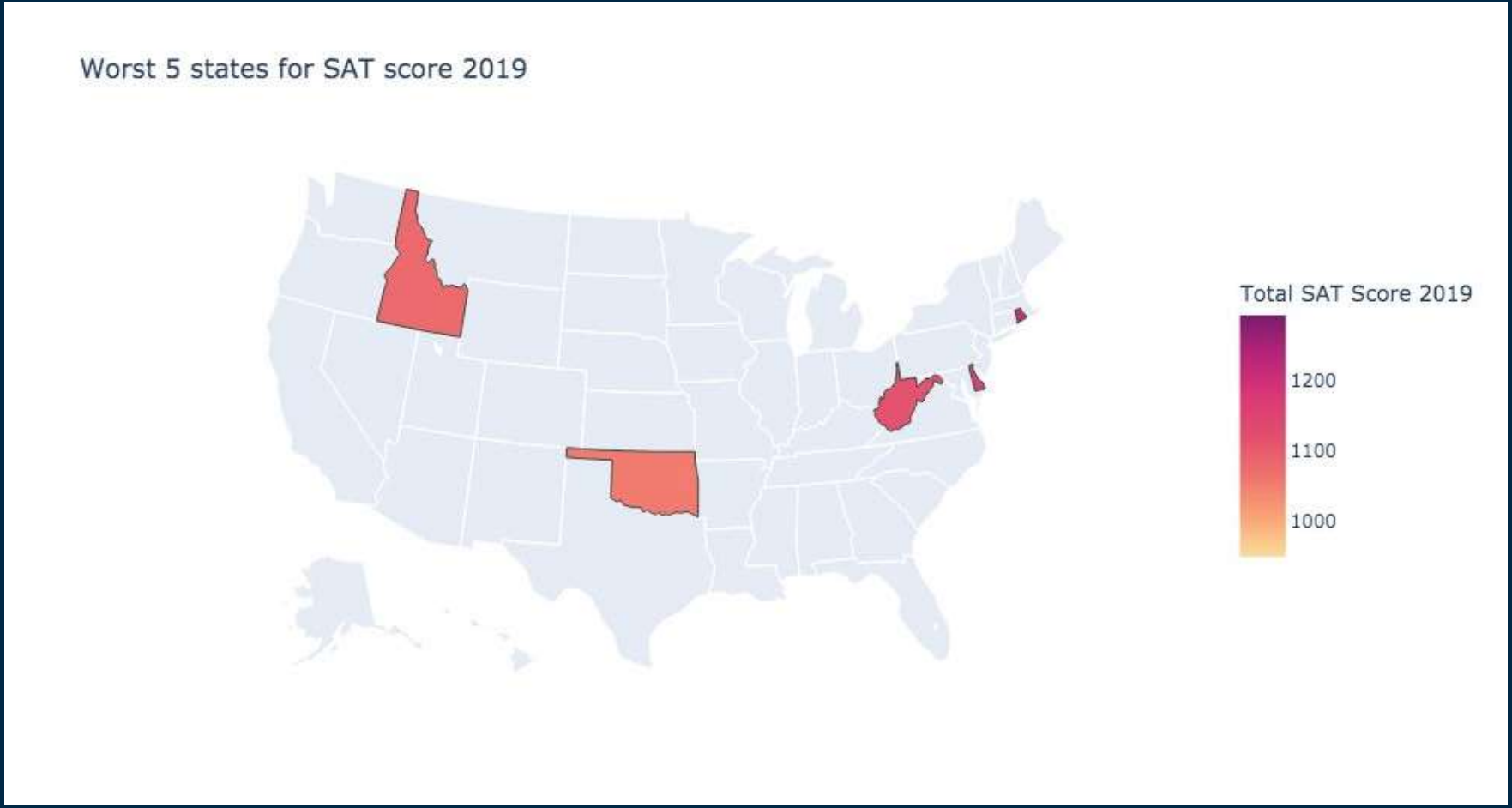
Geographical Map on SAT scores 2018

For Year 2018, top 5 worst SAT scores are from **Delaware**, **Idaho**, **West Virginia**, **Hawaii** and **Utah**



Geographical Map on SAT scores 2019

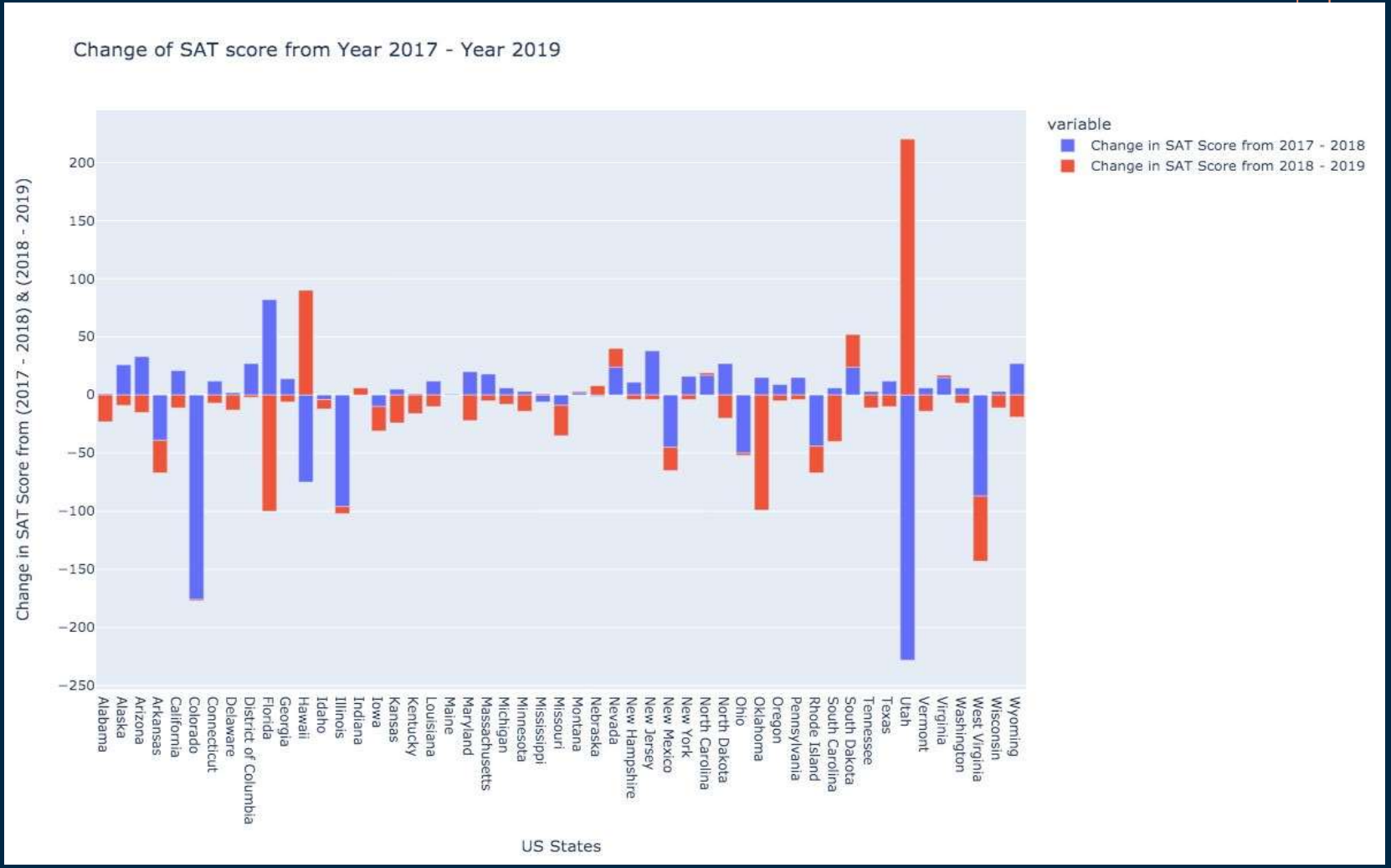
For Year 2019, top 5 worst SAT scores are from **Delaware**, **Idaho**, **West Virginia**, **Rhode Island** and **Oklahoma**



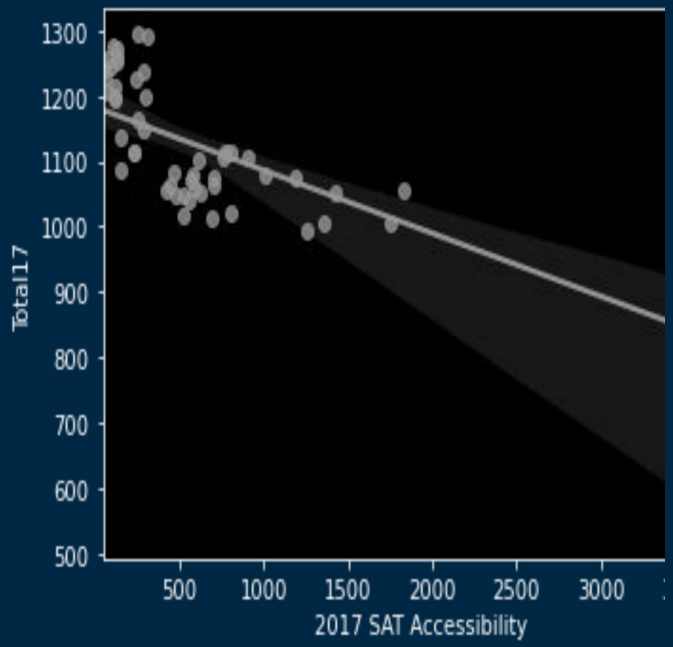
Changes of SAT score (2017 - 2019)

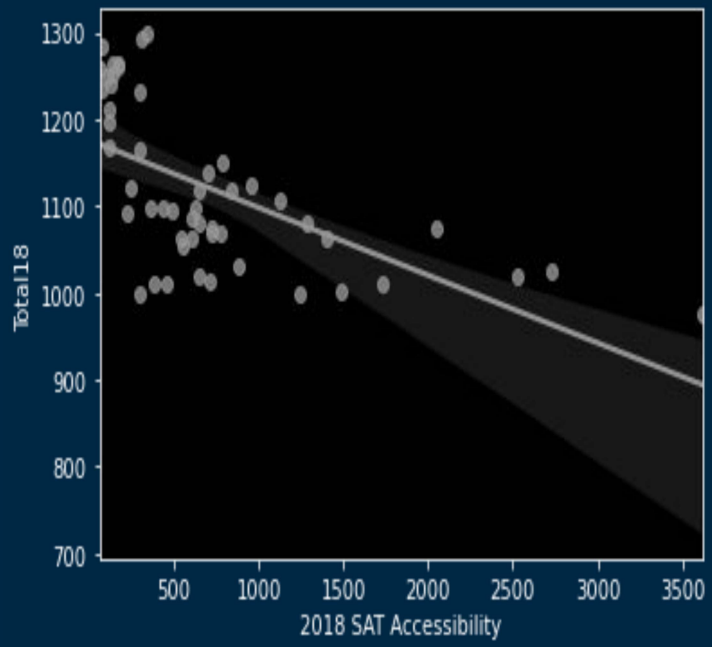
From 2017 - 2018, the **best improved** state is **Florida**. The states that **performed worse** are **Colorado** and **Utah**

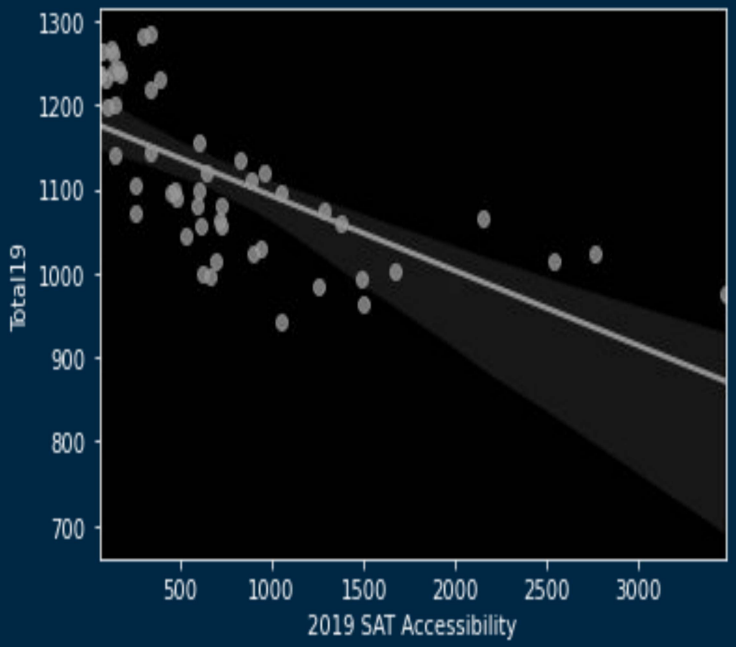
For 2018 - 2019, the **best improved** state is **Utah** and **Hawaii**. The states that **performed worse** are **Florida** and **Oklahoma**



Scatterplot of SAT Exam Scores and Accessibility

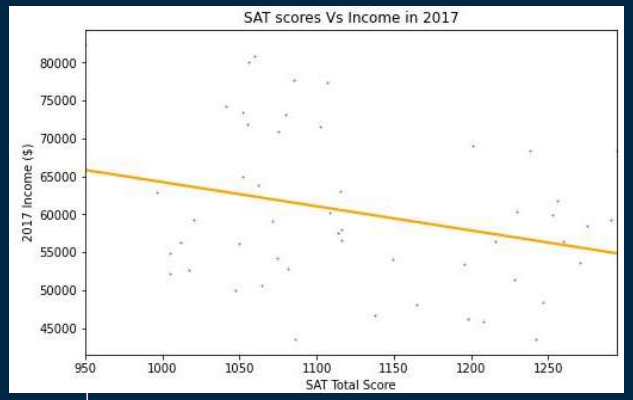


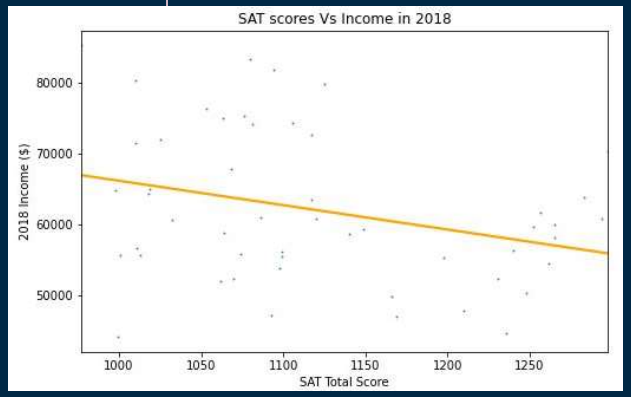


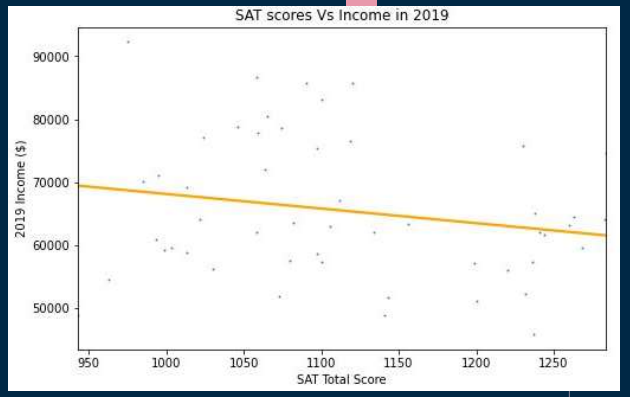


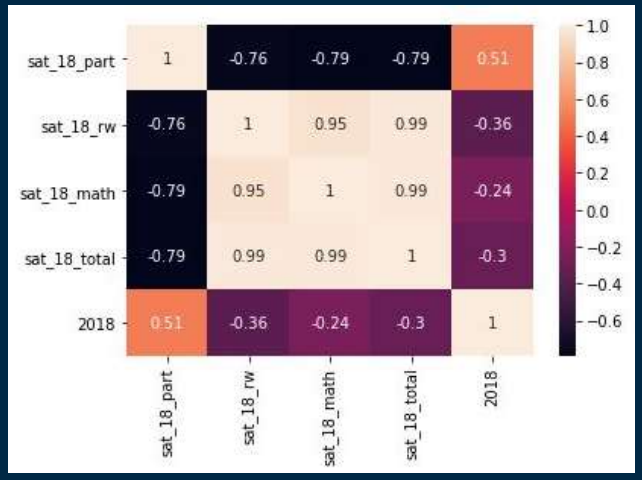
There is very weak negative correlation between the two variables as the results are based on a few outliers

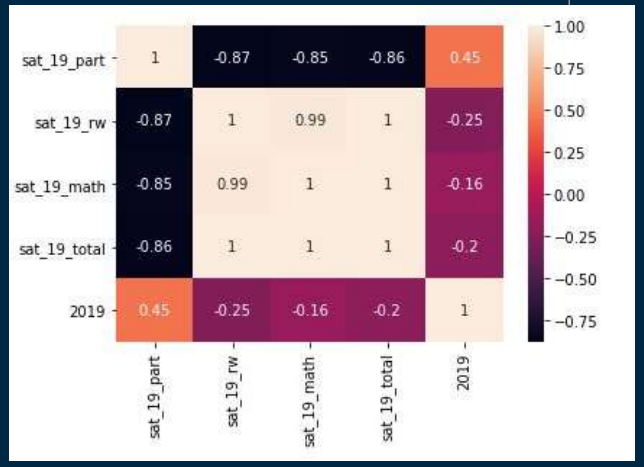
SAT Score Vs Median Income (2017 - 2019)











SAT & ACT Dataset Overview

Selection Bias

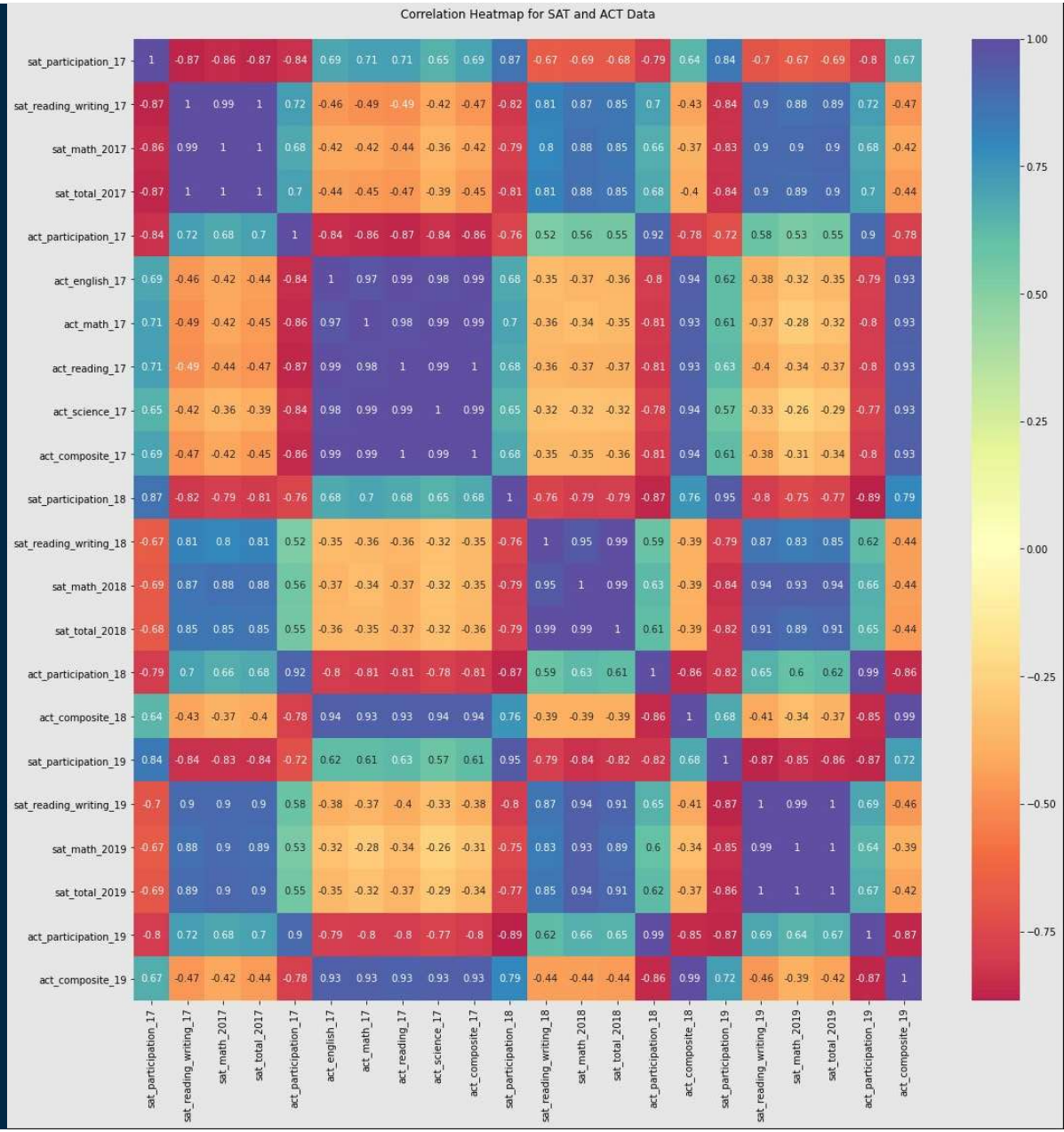
Lower participation typically means higher mean scores (r ~ -0.8)

Total Score Correlation

SAT mean scores are negatively correlated with ACT mean scores (r ~ -0.6)

Negative Correlation in Individual Subjects

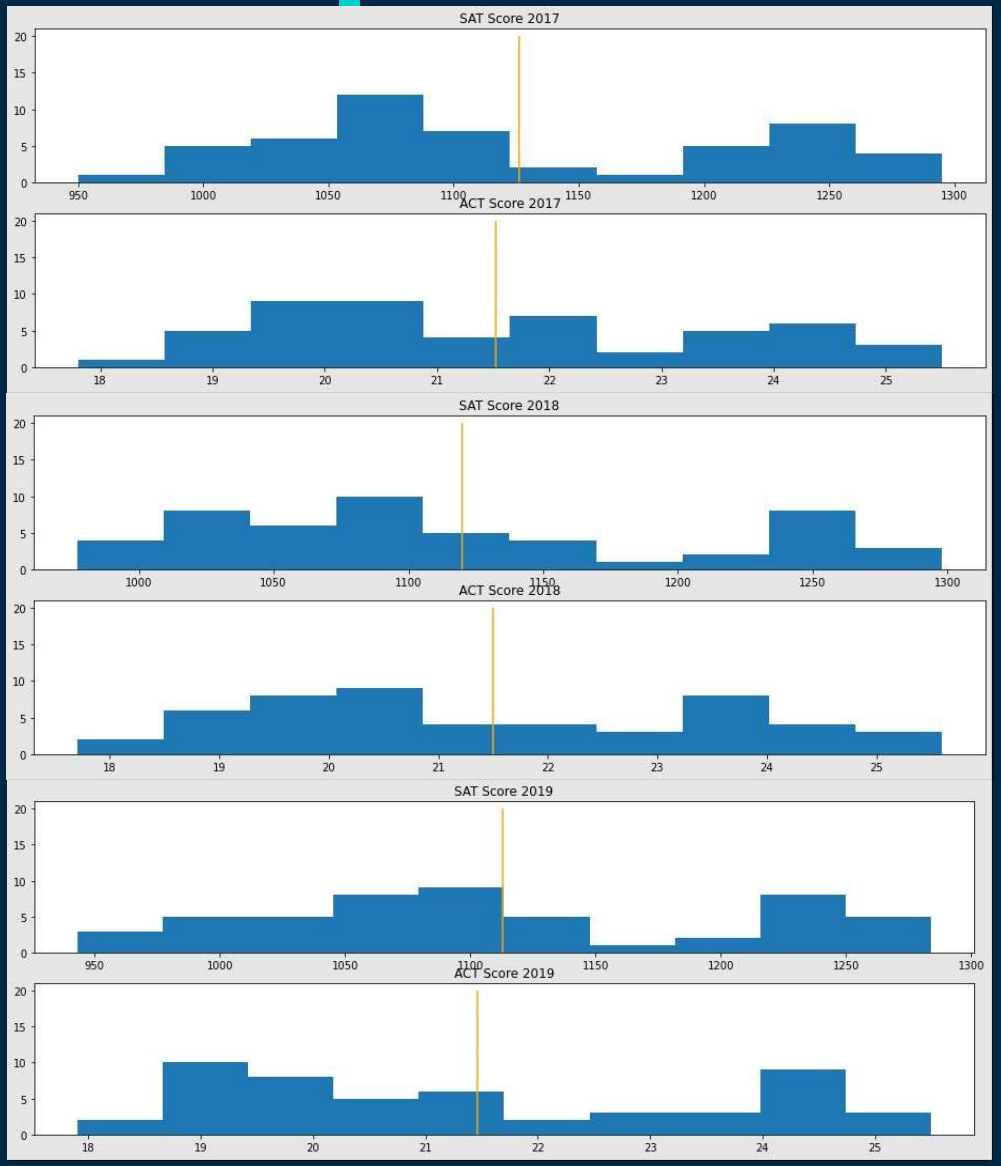
A lower math score in SAT = A higher math score in ACT



SAT & ACT Scores

Positive Correlation in Scores between ACT and SAT

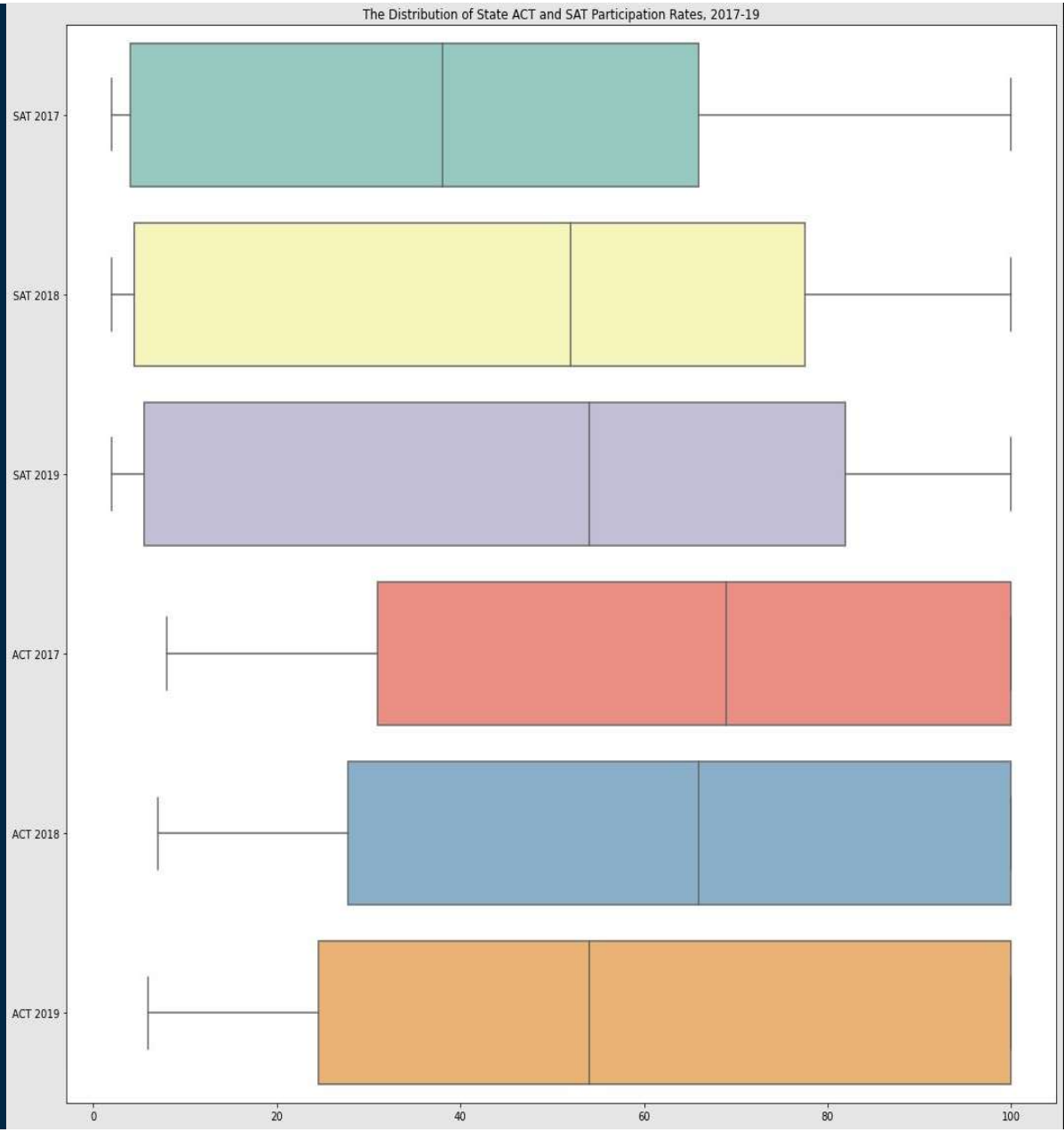
Although the scoring is presented differently between the two tests, when normalized it shows that the education standard is not compromised, which leads to the question if the education board should consider only adopting one standardized test instead.



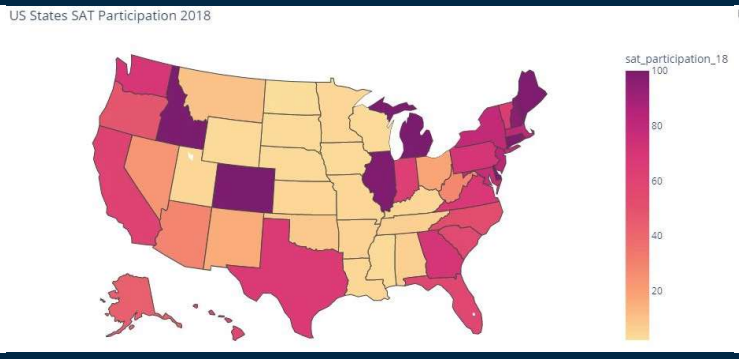
SAT & ACT Participation

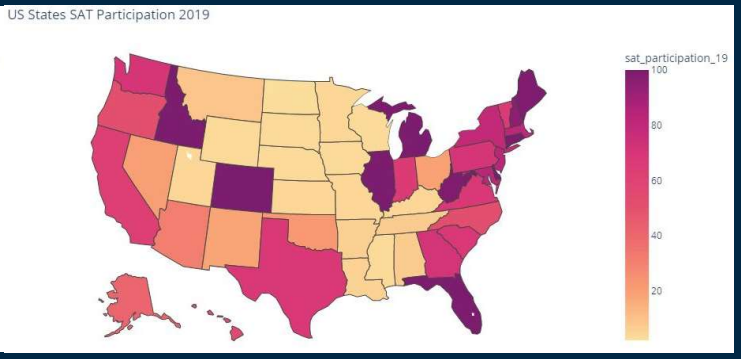
SAT Participation is slowly increasing over time

The boxplot shows that although there are more participation for ACT compared to SAT, it shows that over the three years there has been an increase in SAT participation and a decrease in ACT participation

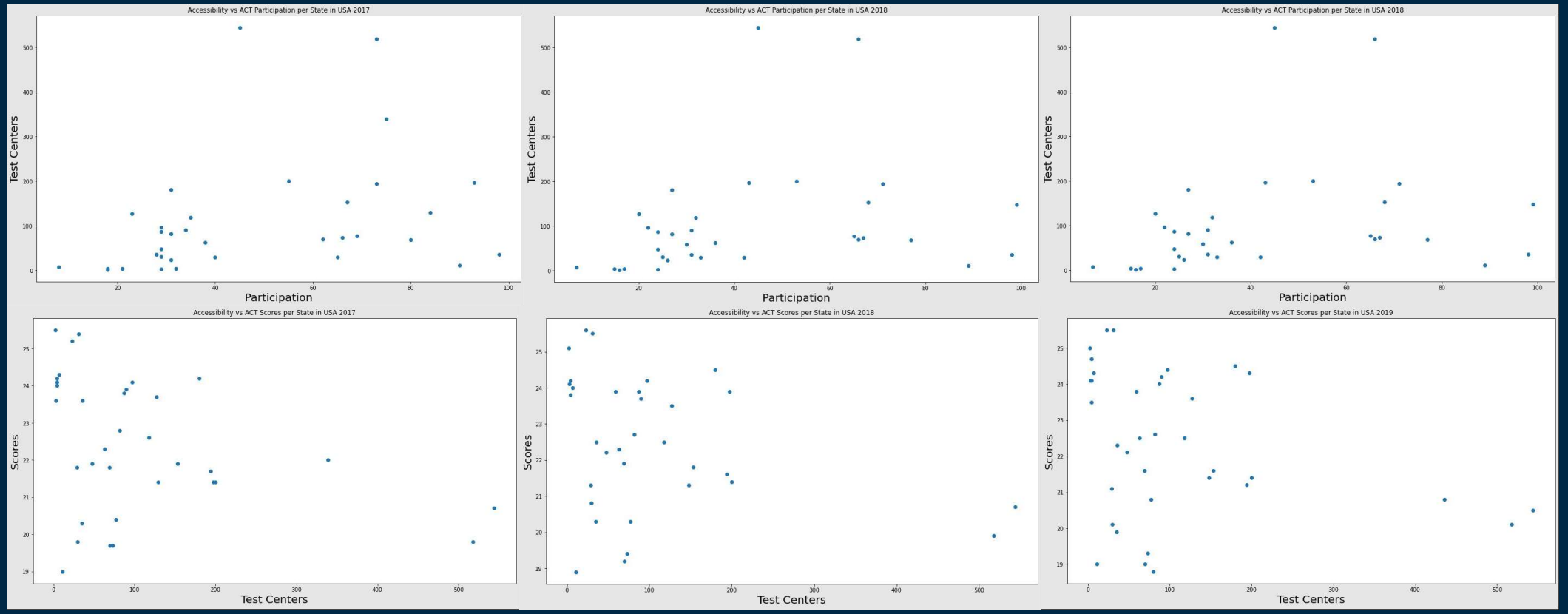


SAT Participation Distribution in USA from 2017-2019



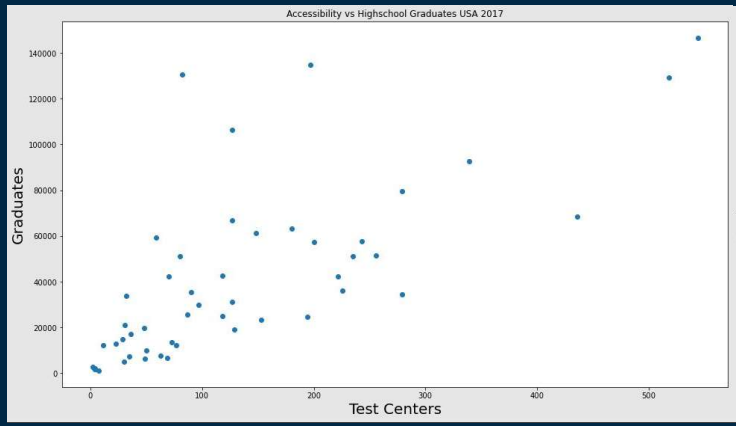


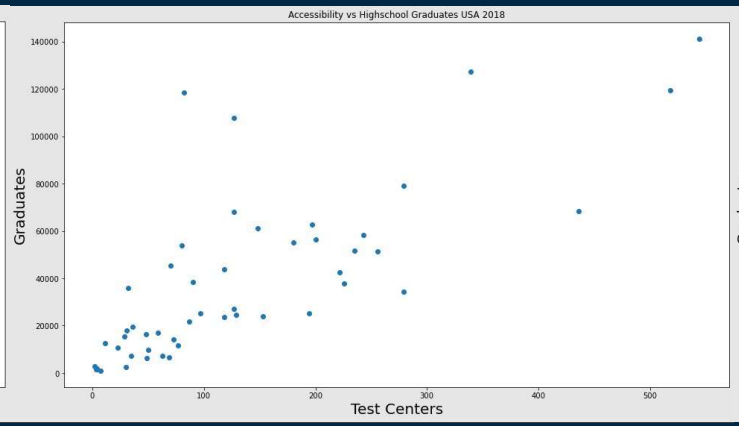
Accessibility vs ACT Participation & Scores '17-19

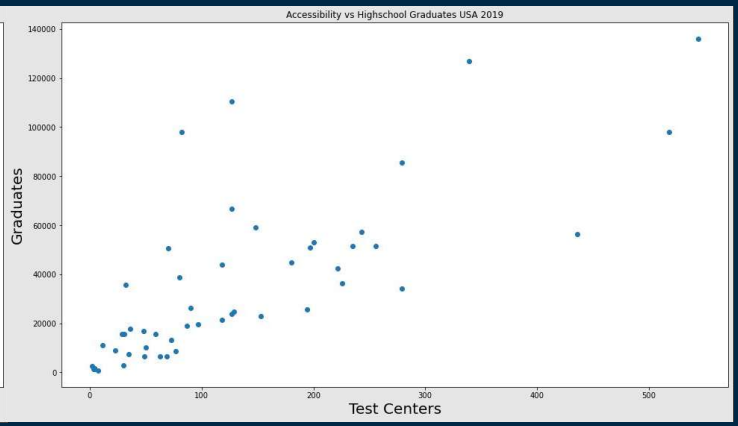


Source: act.org

Graduates vs Accessibility for ACT ‘17-19

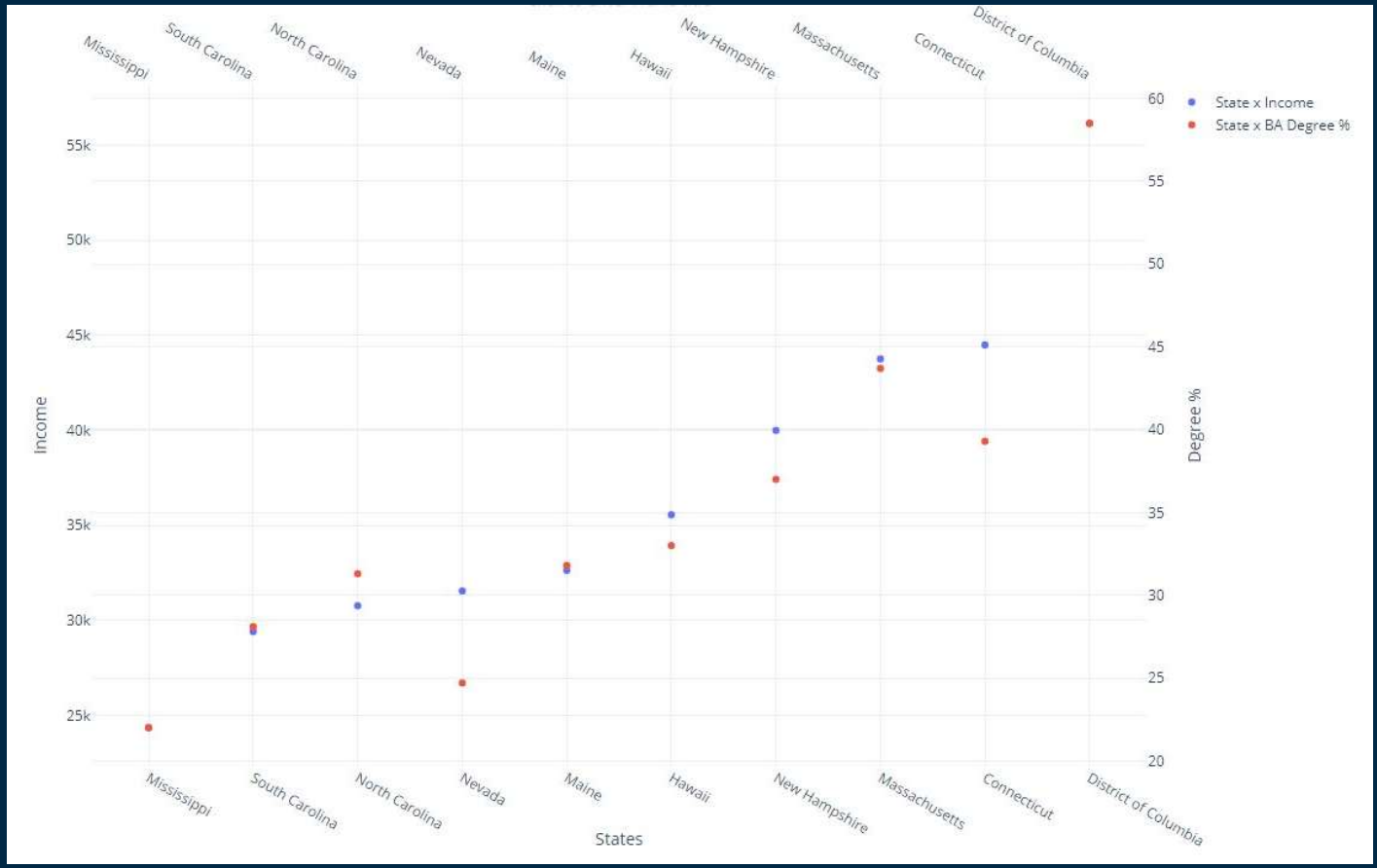






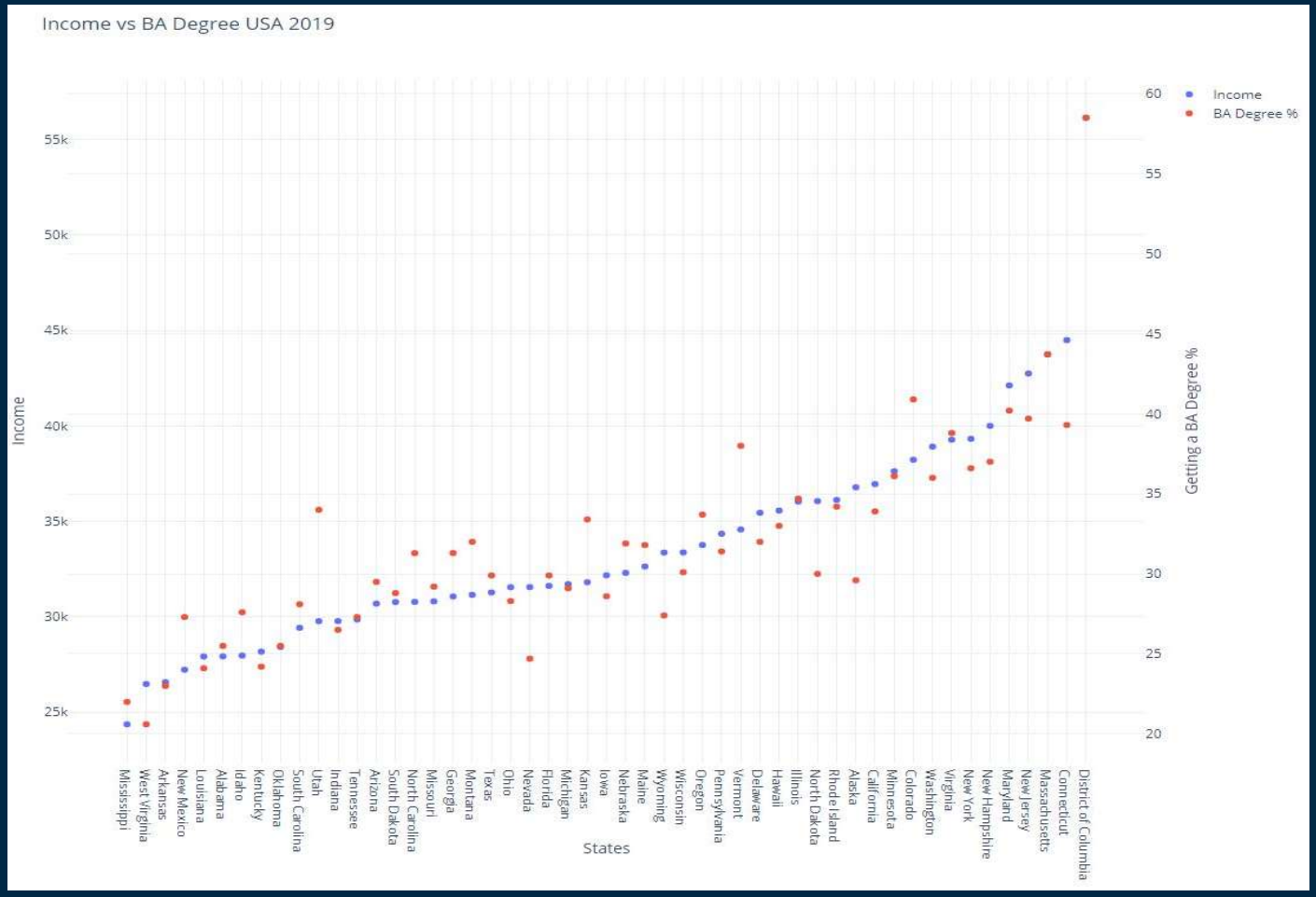
Source: act.org

Comparison between Top & Bottom 5 ACT Scoring States Based on Income & BA Degree %



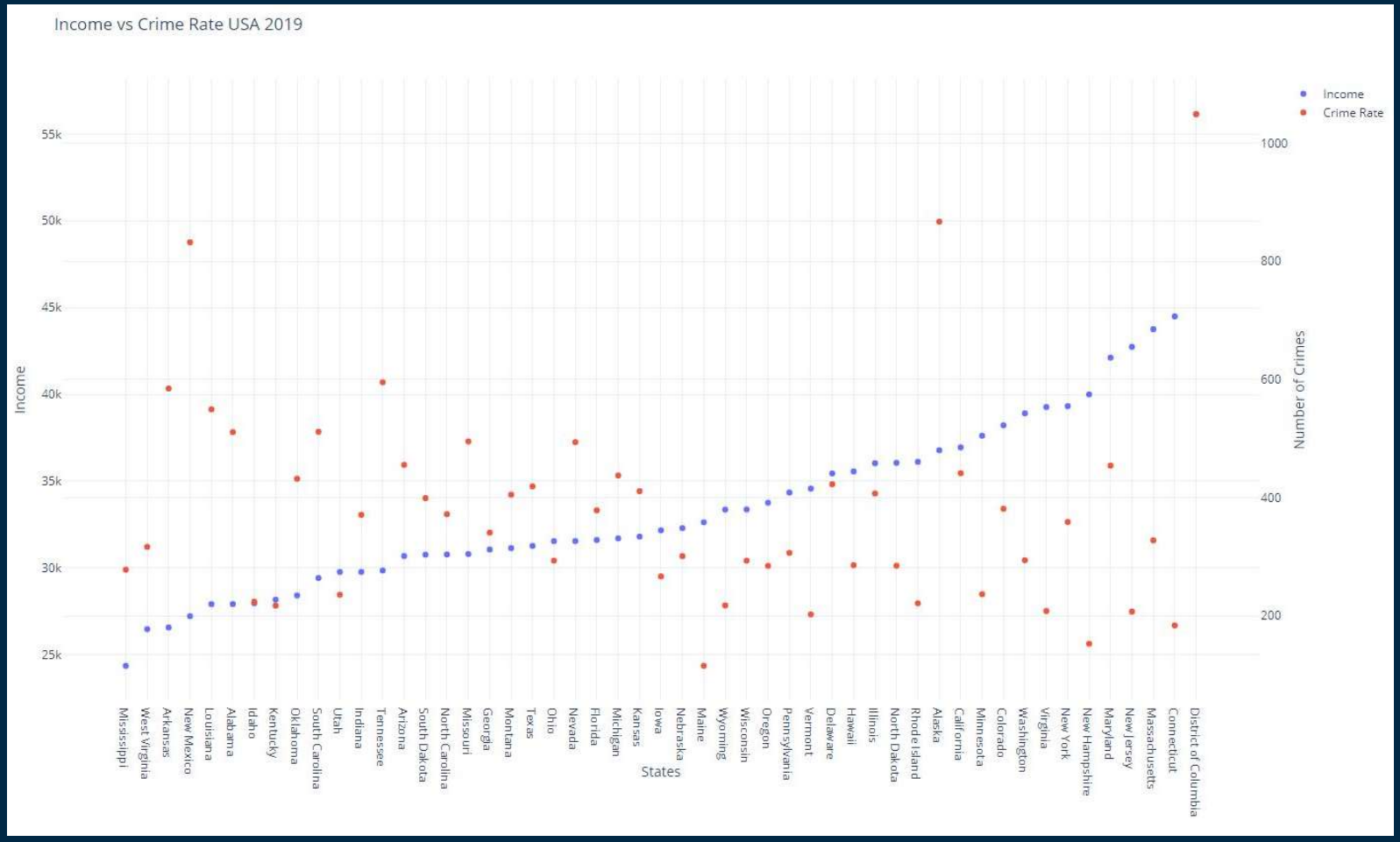
Source: census.gov

Income vs BA Degree % per State in USA 2019



Source: census.gov

Income vs Crime Rate in USA 2019



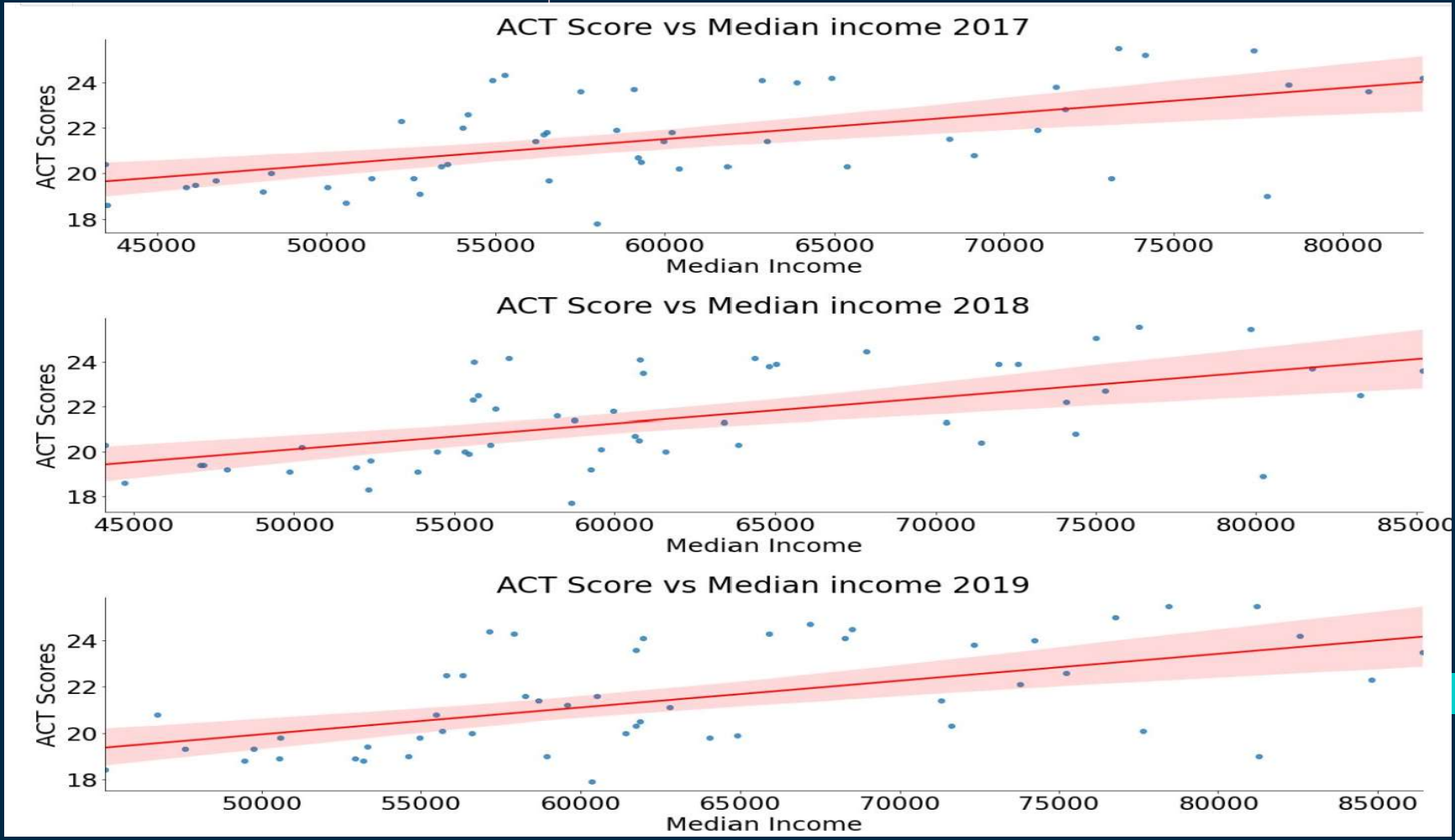
Source: statista.com

Taking the ACT Tests More than Once

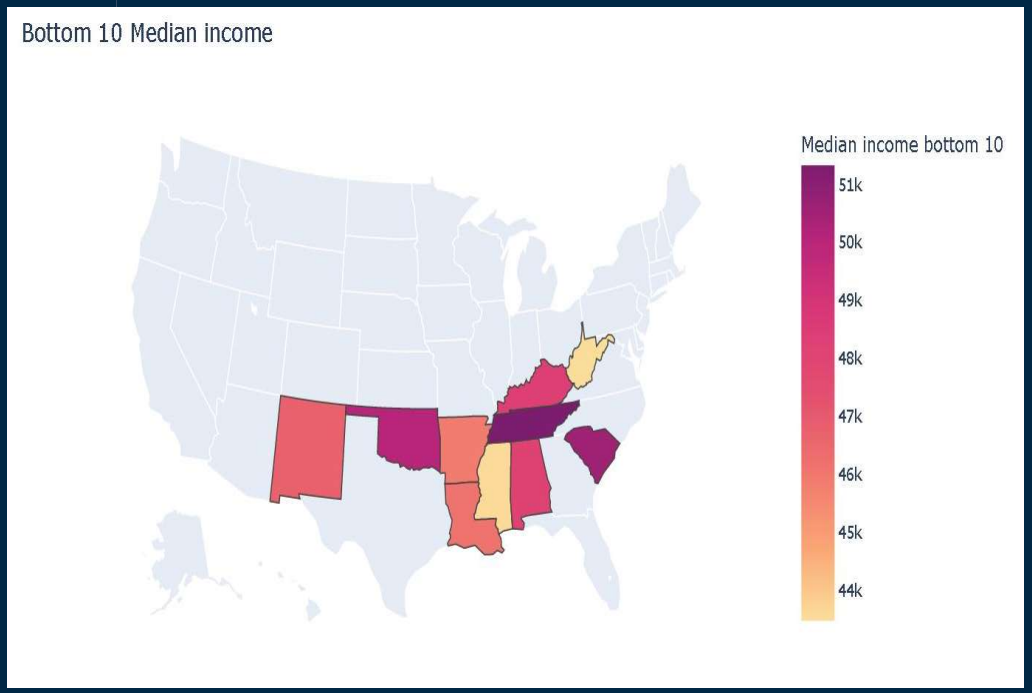
Nationwide, 765,568 of the 2019 graduates (42.9 percent) taking the ACT two or more times had an average Composite score of 22.7, compared to an average of 19.2 for 1,017,252 of the 2019 graduates (57.1 percent) who took the ACT only once.

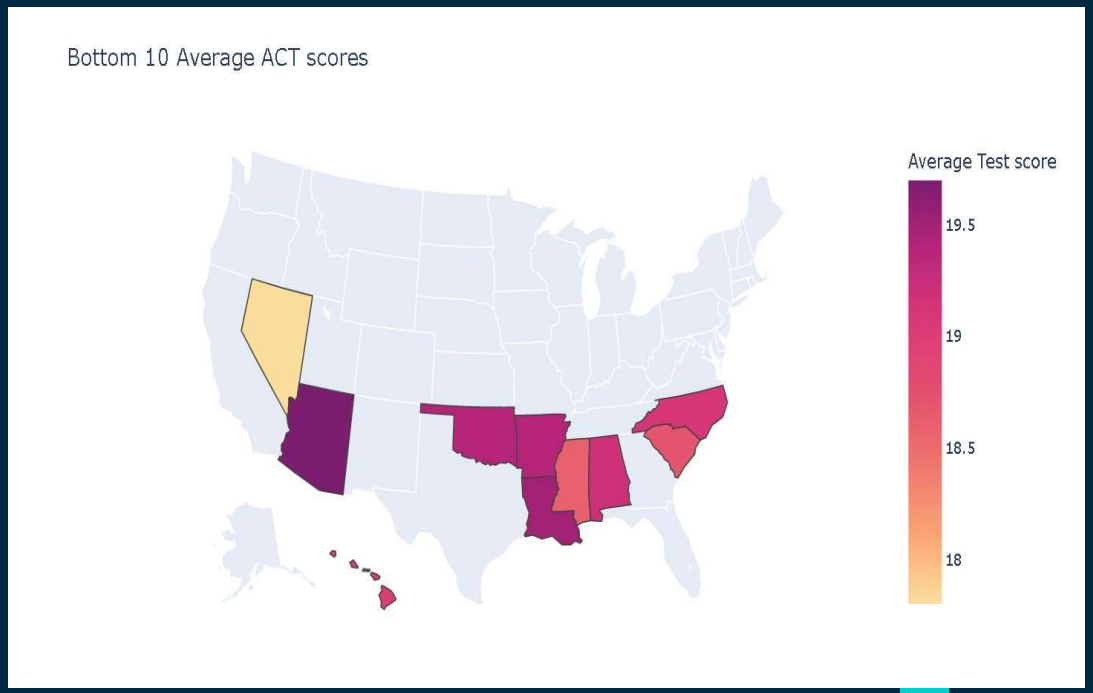
Source: act.org

ACT Score Vs Median Income



ACT Score Vs Median Income



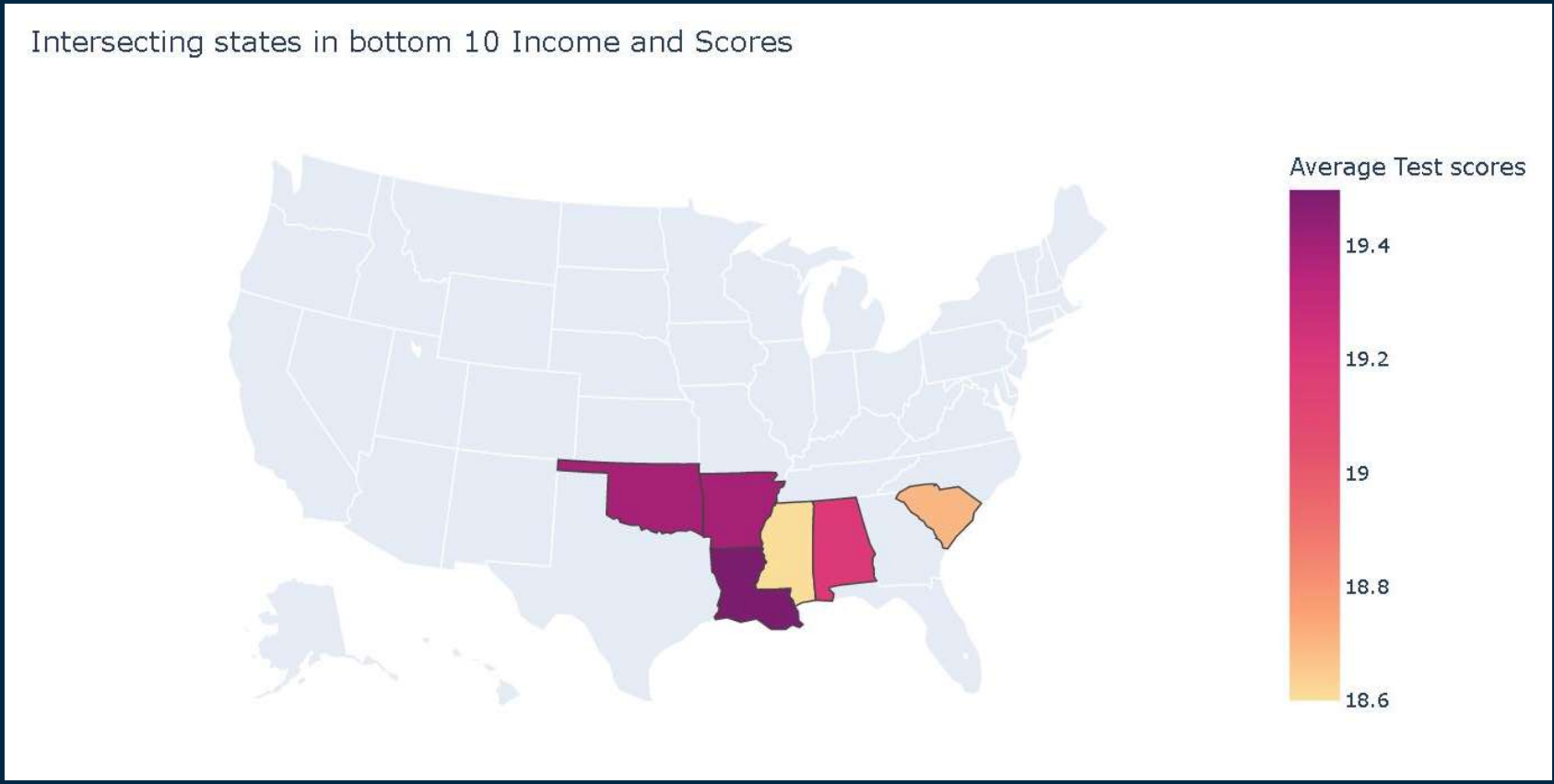


ACT Score Vs Median Income

From 2017-2019,

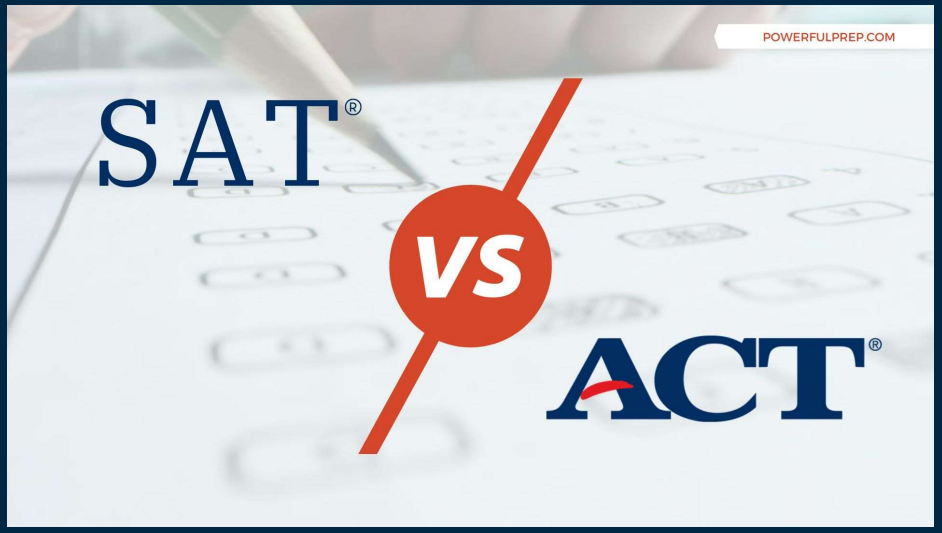
Mississippi,
Arkansas,
Louisana,
Alabama,
Oklahoma,
South Carolina

Are consistently found in the bottom 10
median income and bottom 10 ACT scores



Recommendation:

1. States to use SAT for standardized testing instead of ACT
2. No need to build new test centers for now.



Do you have any questions?

THANKS