
GA-DSI-18
Project 1
2017 & 2018
SAT & ACT Data Analysis
Jeffrey Sim

What is the problem?

- The new format for the SAT was released in March 2016. The College Board - the organization that administers the SAT and tracks statewide participation would like to look for ways to improve SAT participation rate.

- Use statistical methods and data analysis techniques with the help of Python to derive insights.

- Use a combination of analysis into the provided data and outside research to give recommendations on:

“How the College Board might work to increase the participation rate in a state of my choice.”

Outline of Approach

- Data import and cleaning
- Data dictionary
- Exploratory Data Analysis (EDA)
- Visualization in Tableau
- Descriptive and inferential statistics
- Conclusion and recommendations

Data import and cleaning

- 2017/2018 dataset were provided as csv files, upon importing into Python there were issues identified that required data cleansing to get a complete and clean dataset. These included:
  - Columns renaming and datatypes standardization
  - Correction of data values errors as compared to the actual web data (eg. values and Min/Max range of test scores)
  - Removal of unnecessary symbols (%) and text ('x') that may cause computational errors
  - Merging of 2017 and 2018 data into a final dataset and save to a csv file

	state	2017_sat_pcp%	2017_sat_erw	2017_sat_math	2017_sat_total	2017_act_pcp%	2017_act_english
0	Alabama	0.05	593	572	1165	1.00	18.9
1	Alaska	0.38	547	533	1080	0.65	18.7
2	Arizona	0.30	563	553	1116	0.62	18.6
3	Arkansas	0.03	614	594	1208	1.00	18.9
4	California	0.53	531	524	1055	0.31	22.5

5 rows × 21 columns

Data Dictionary

Feature	Type	Dataset	Description
state	string	SAT	State name eg. Alabama
2017_sat_pcp%	float64	SAT	Participation rate eg. 0.05(i.e. 5%)
2017_sat_erw	int64	SAT	Average Score for Evidence-Based Reading and Writing eg. 593
2017_sat_math	int64	SAT	Average Score for Math eg. 572
2017_sat_total	int64	SAT	Average Score for Total eg. 1165
---	---	---	---
2017_act_pcp%	float64	ACT	Participation rate eg. 0.5(i.e. 50%)
2017_act_english	float64	ACT	Average Score for English eg. 18.9
2017_act_math	float64	ACT	Average Score for Math eg. 18.4
2017_act_reading	float64	ACT	Average Score for Reading eg. 19.7
2017_act_science	float64	ACT	Average Score for Science eg. 19.4
2017_act_composite	float64	ACT	Average Score for Composite eg. 19.2
---	---	---	---

2017 data

2018 data

state	string	SAT	State name eg. Alabama
2018_sat_pcp%	float64	SAT	Participation rate eg. 0.06(i.e. 6%)
2018_sat_erw	int64	SAT	Average Score for Evidence-Based Reading and Writing eg. 595
2018_sat_math	int64	SAT	Average Score for Math eg. 571
2018_sat_total	int64	SAT	Average Score for Total eg. 1166
---	---	---	---
state	string	ACT	State name eg. Alabama
2018_act_pcp%	float64	ACT	Participation rate eg. 1.0(i.e. 100%)
2018_act_composite	float64	ACT	Average Score for Composite eg. 19.1
2018_act_english	float64	ACT	Average Score for English eg. 18.9
2018_act_math	float64	ACT	Average Score for Math eg. 18.3
2018_act_reading	float64	ACT	Average Score for Reading eg. 19.6
2018_act_science	float64	ACT	Average Score for Science eg. 19.0

Exploratory Data Analysis

- Standard Deviation Analysis
  - the difference is negligible
  - but it is important to know the difference in application of the three methods against different types of sample.
  - a quick glance over the std and mean of each variable, as a general rule of thumb if the std is less than 1/3 of the mean, it is considered normal.
  - the only exceptions are the % denominated variables (eg. sat_pcp%) which is exactly the variables that we are trying to consider in this analysis as to why some states have very high or very low participation rates.

Exploratory Data Analysis

- Participation Rate vs Performance (data filter and sorting)
  - states with highest/lowest SAT/ACT participation rates in 2017/2018
  - states with the highest/lowest SAT Total Scores/ACT Composite Scores in 2017/2018
  - states with 100% participation on a given test with a rate change year-to-year
    - Colorado stood out from the list with the biggest increase for SAT participation rate from 11% to 100%, as well as being the state with the biggest decrease for ACT participation from 100% to 30%
  - states with >50% participation on both tests either year
    - Florida stood out from the list with the biggest decrease for SAT participation rate from 83% to 73%, while also having the biggest decrease for ACT participation rate from 73% to 66%

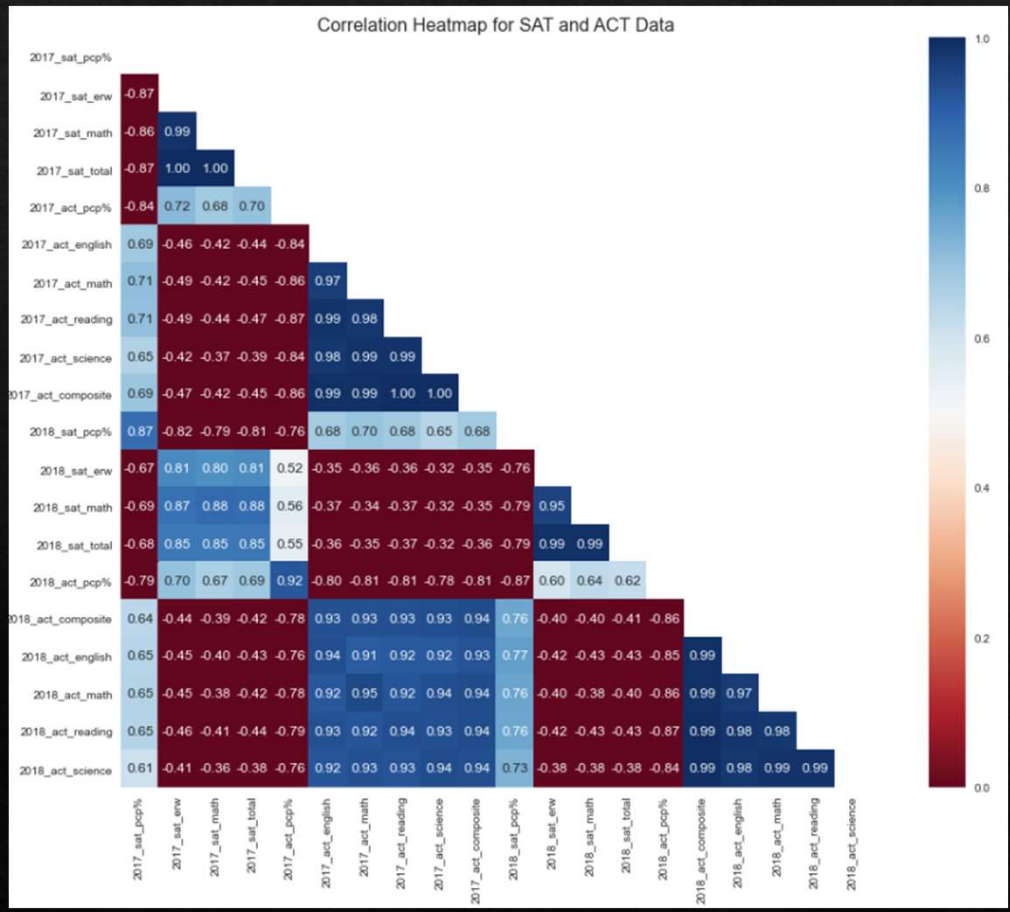
	state	2017_sat_pcp%	2018_sat_pcp%
5	Colorado	0.11	1.00
8	District of Columbia	1.00	0.92
12	Idaho	0.93	1.00

	state	2017_sat_pcp%	2018_sat_pcp%	2017_act_pcp%	2018_act_pcp%
0	Florida	0.83	0.56	0.73	0.66
1	Georgia	0.61	0.70	0.55	0.53
2	Hawaii	0.55	0.56	0.90	0.89
3	North Carolina	0.49	0.52	1.00	1.00
4	South Carolina	0.50	0.55	1.00	1.00

Exploratory Data Analysis

We observed the following points:

- Mean scores on a given test are highly negatively correlated with participation rate on that test (r ~ -0.8). This directly indicates the earlier observation of higher scores on SAT had lower SAT participation rate and higher ACT participation rate had lower ACT scores).
- Mean scores on sections of a given test are highly correlated (r ~ 0.9 to 1.0) with mean scores for other sections of that test, or total scores for that test.
- Mean scores on sections of the SAT are moderately negatively correlated with mean scores on ACT, and vice versa (r ~ -0.6).
- Participation rate on a given test is moderately positively correlated with scores on the opposite test (r ~ 0.6).



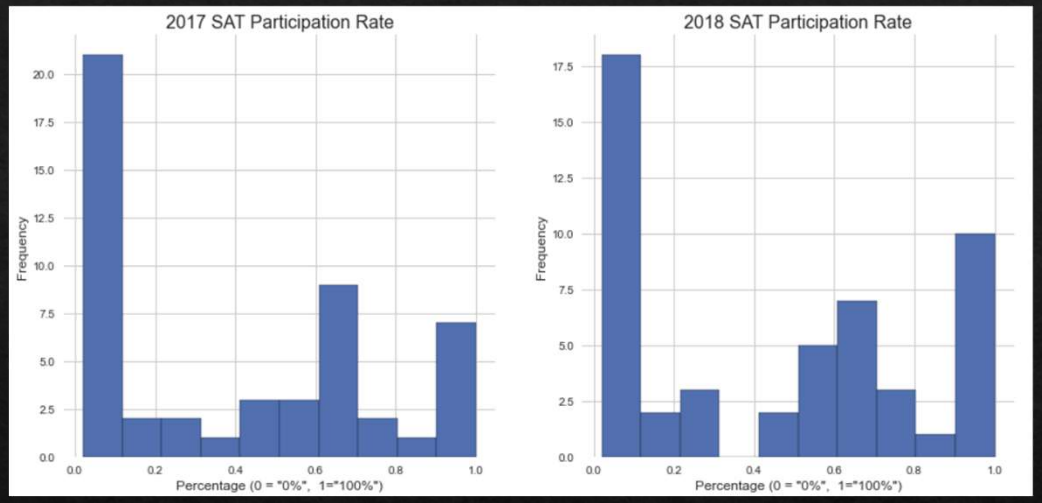
Exploratory Data Analysis

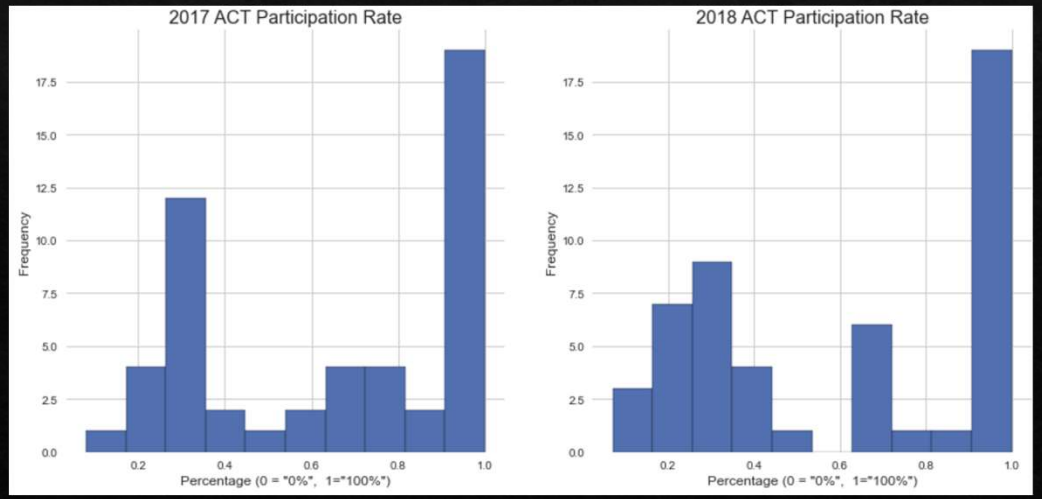
Using Python Histogram, we observed the following points:

- The distributions for SAT participation rate and ACT participation rate had not changed dramatically from 2017 to 2018. It is also clear that the SAT had a large group of very low participation rates (<10%), a cluster of states with participation in the 50-75% range, and then a group of states with 100% participation.

- The ACT had almost no states with lower than 10% participation, had a cluster of states in the 15-40% range, only a small number of states in the mid to high range, and then a large group of states with almost or 100% participation.

- In this way the two distributions almost mirror each other and follow a bimodal distribution with two peaks.

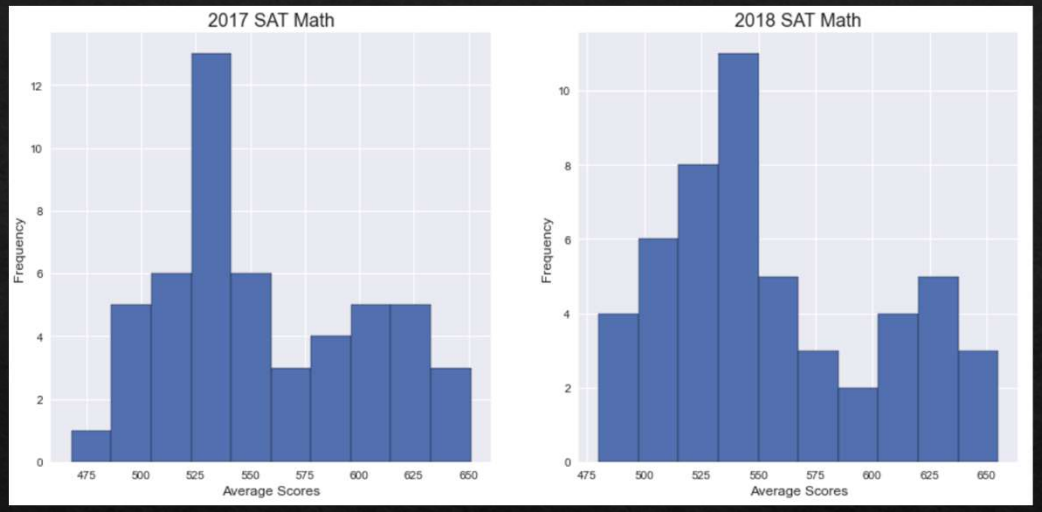


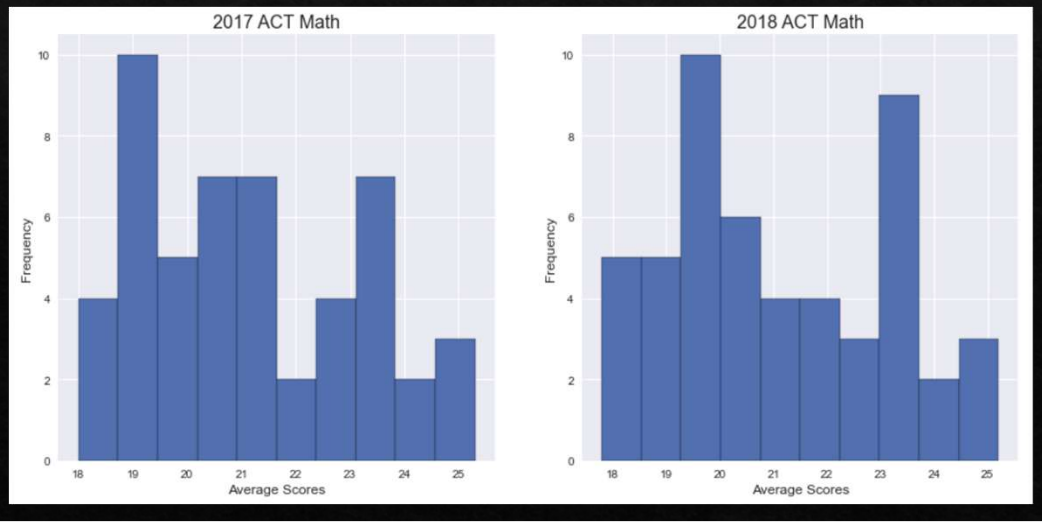


Exploratory Data Analysis

Using Python Histogram, we observed the following points:

- The distributions for SAT Math and ACT Math had not changed dramatically from 2017 to 2018. It is also clear that the SAT had a larger group of low performance than high performance and the situation seems to get worst in 2018.
- The ACT had pretty much the same observation of deterioration in performance
- Distribution pattern did not change much and following a Bimodal distribution with two peaks.

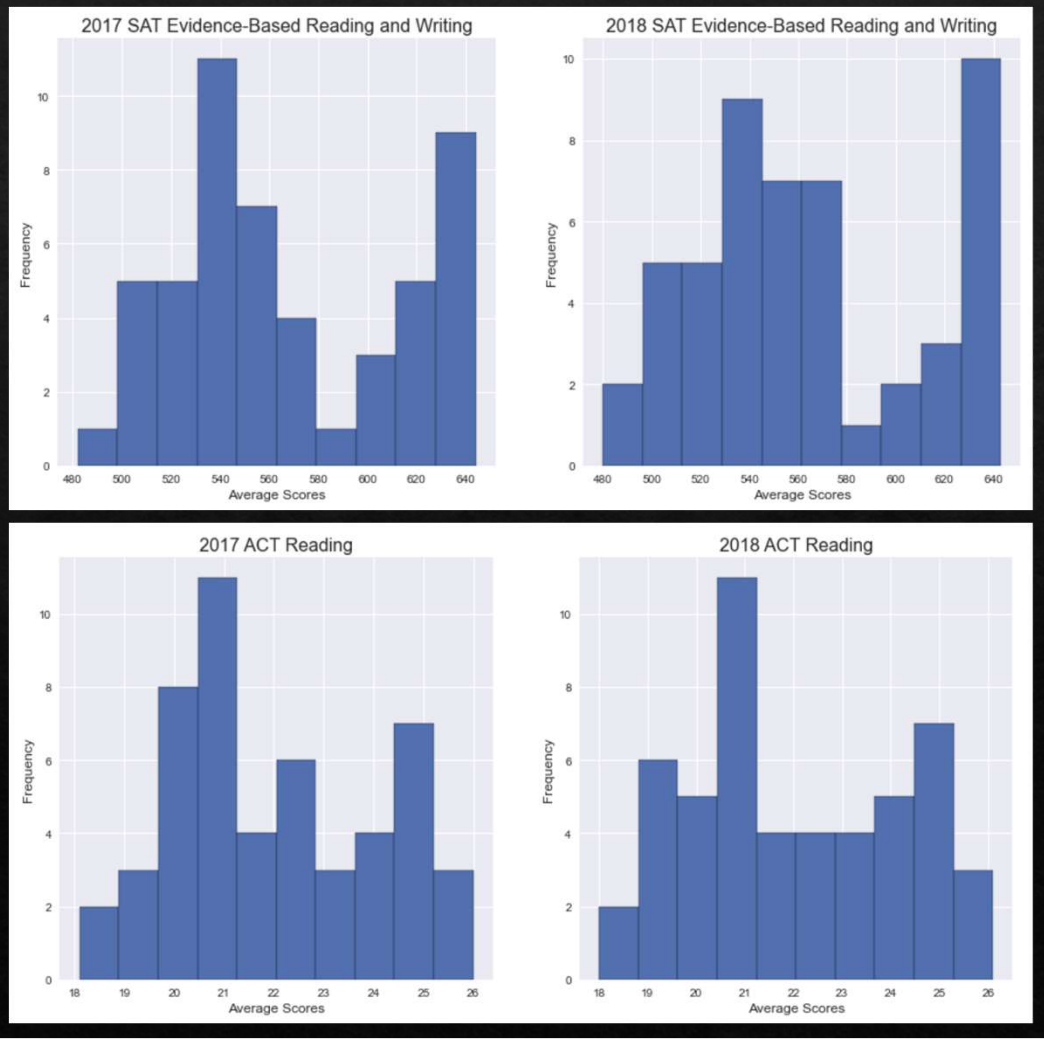




Exploratory Data Analysis

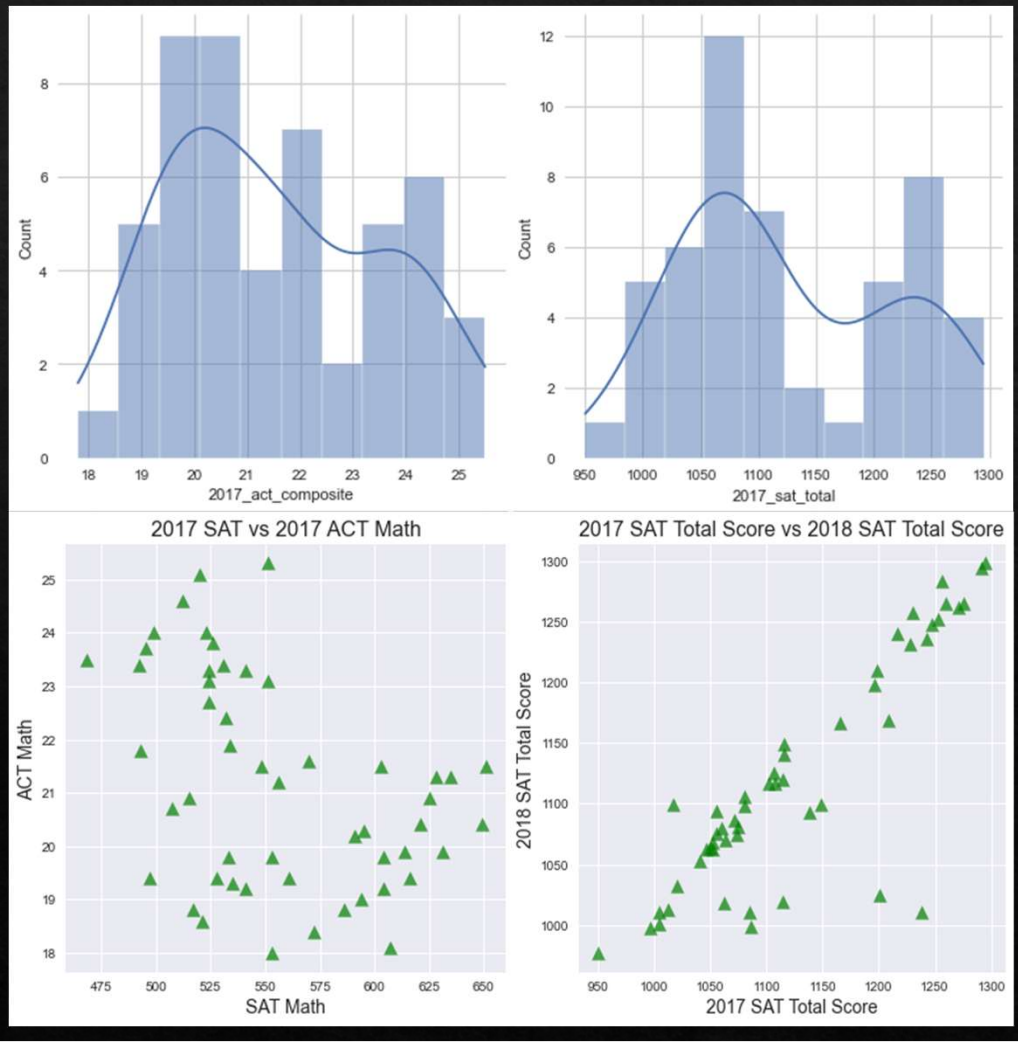
Using Python Histogram, we observed the following points:

- The distributions for SAT Evidence-Based Reading and Writing between 2017 and 2018 did not change much. The performance seems to be improving with a shift in lower performance population to higher performance.
- The ACT Reading had pretty much the same observation but the shift in lower to higher performance was of a lesser magnitude.
- Distribution pattern did not change much and following a Bimodal distribution with two peaks.



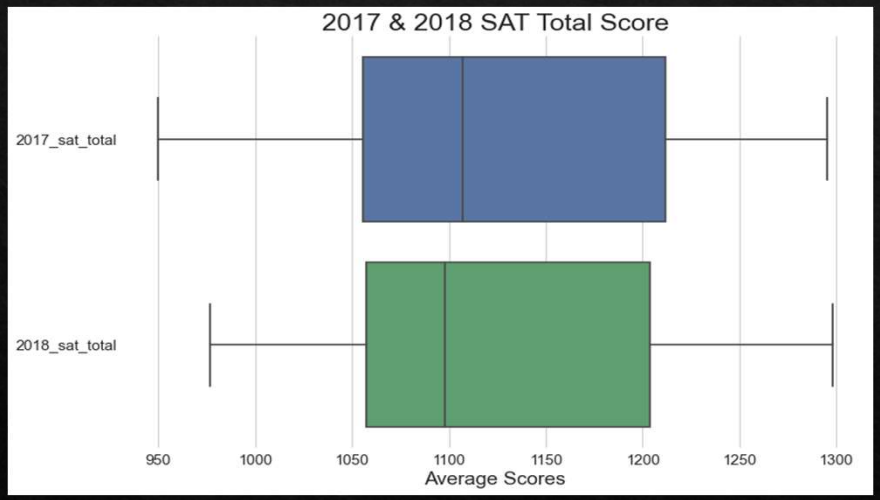
Exploratory Data Analysis

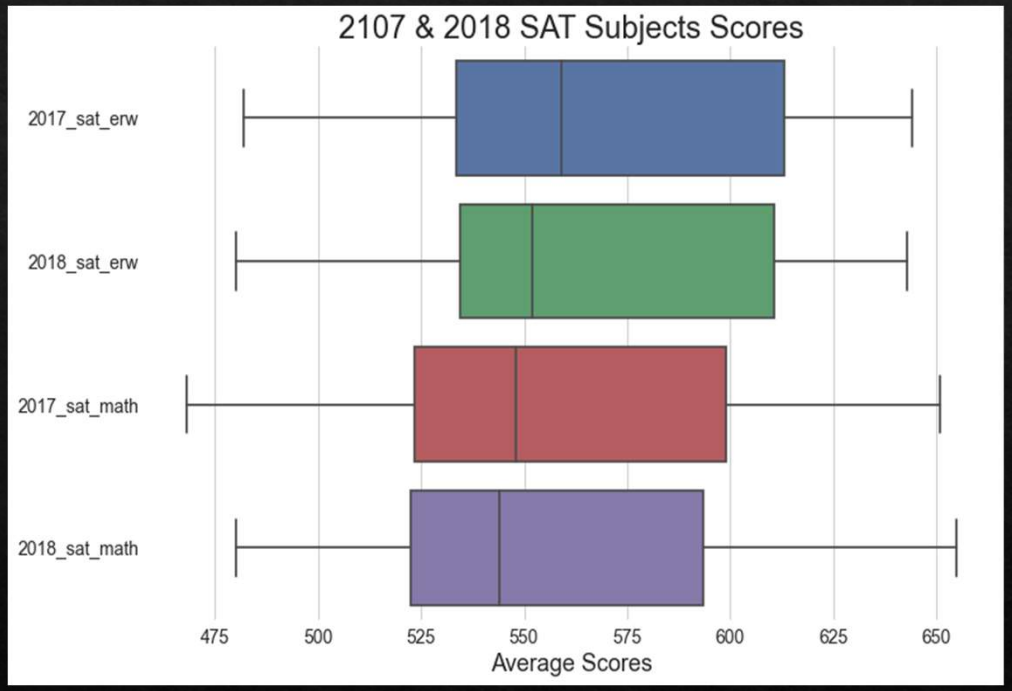
- We also applied Python Histogram and Seaborn Displot to selected variables we observed two kinds of distribution patterns which were Bimodal distribution or a Skewed Normal distribution (unimodal).
- In essence, we generally assume that data we sample from a population will be normally distributed.
- Although we have observed only Bimodal and Skewed Normal distributions, and according to Central Limit Theorem, if we can get more sample means from these data and plot it again, the sampling distribution of the sample means will approach a Normal Distribution. Therefore, estimates made from these data are likely to be Normally Distributed as well.
- SAT Math/ACT Math, SAT Evidence-Based Reading and Writing/ACT Reading and SAT Total Scores/ACT Composite in the same year, there was no strong correlation between the pairs.
- SAT Total Scores between years and ACT Composite Scores between years, there was a strong positive correlation between the years.



Exploratory Data Analysis

- SAT Total Scores and subject scores, we did not observe any significant variation between the years other than a slight drop in mean scores.

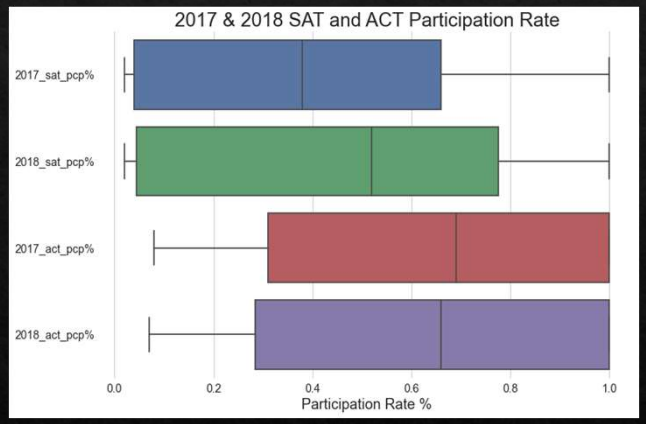




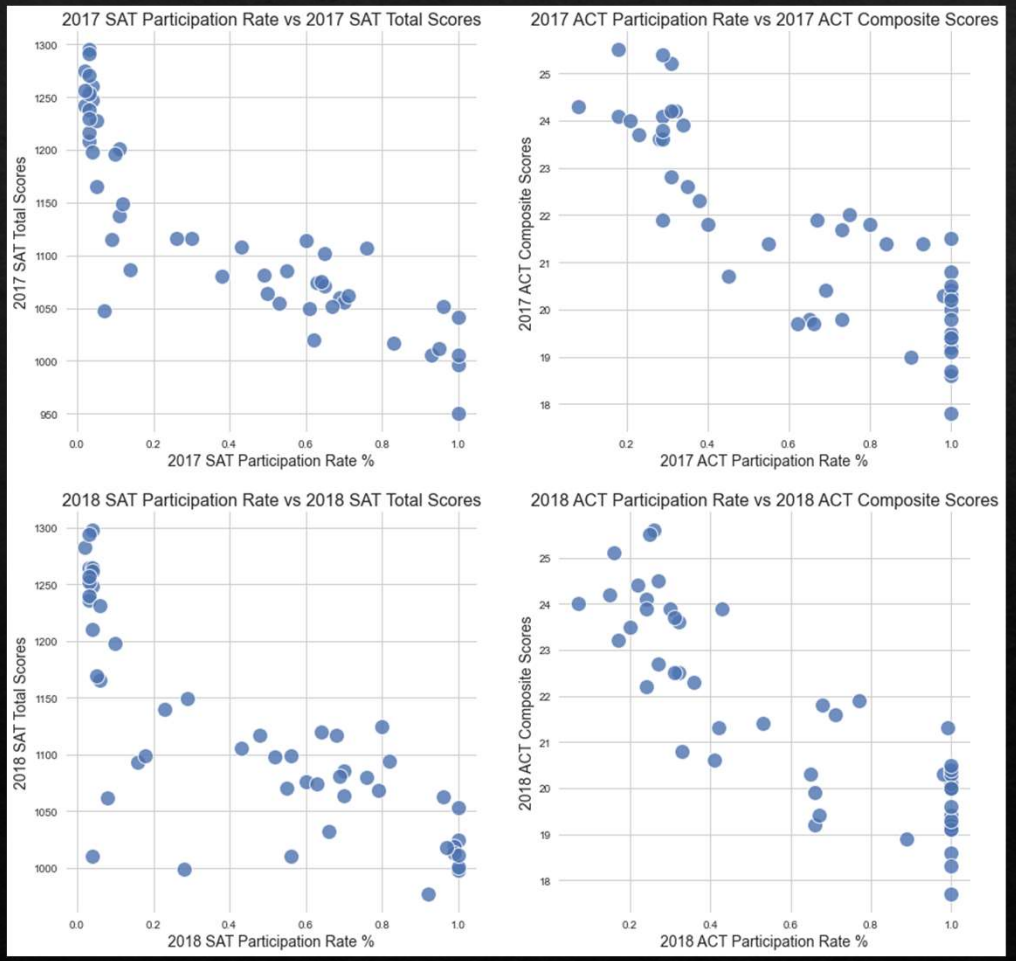
- ACT Composite Scores and all ACT subject scores, we observed a consistent drop in mean scores.

Exploratory Data Analysis

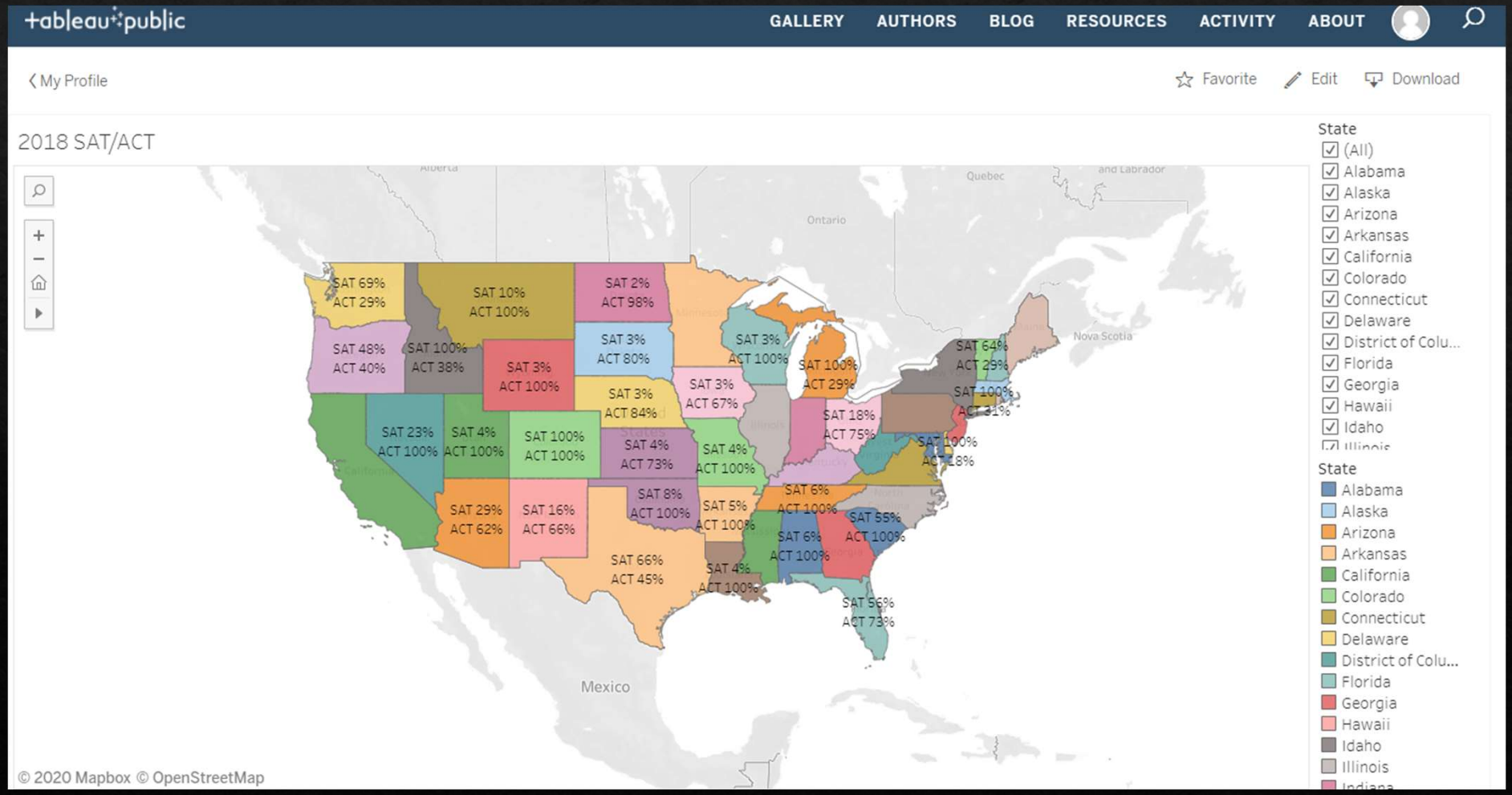
- SAT and ACT Participation Rate, we observed the distribution of ACT Participation Rate across the 50 states is centered significantly higher than that of the SAT. This can also mean that across the 50 states, ACT is preferred over SAT.



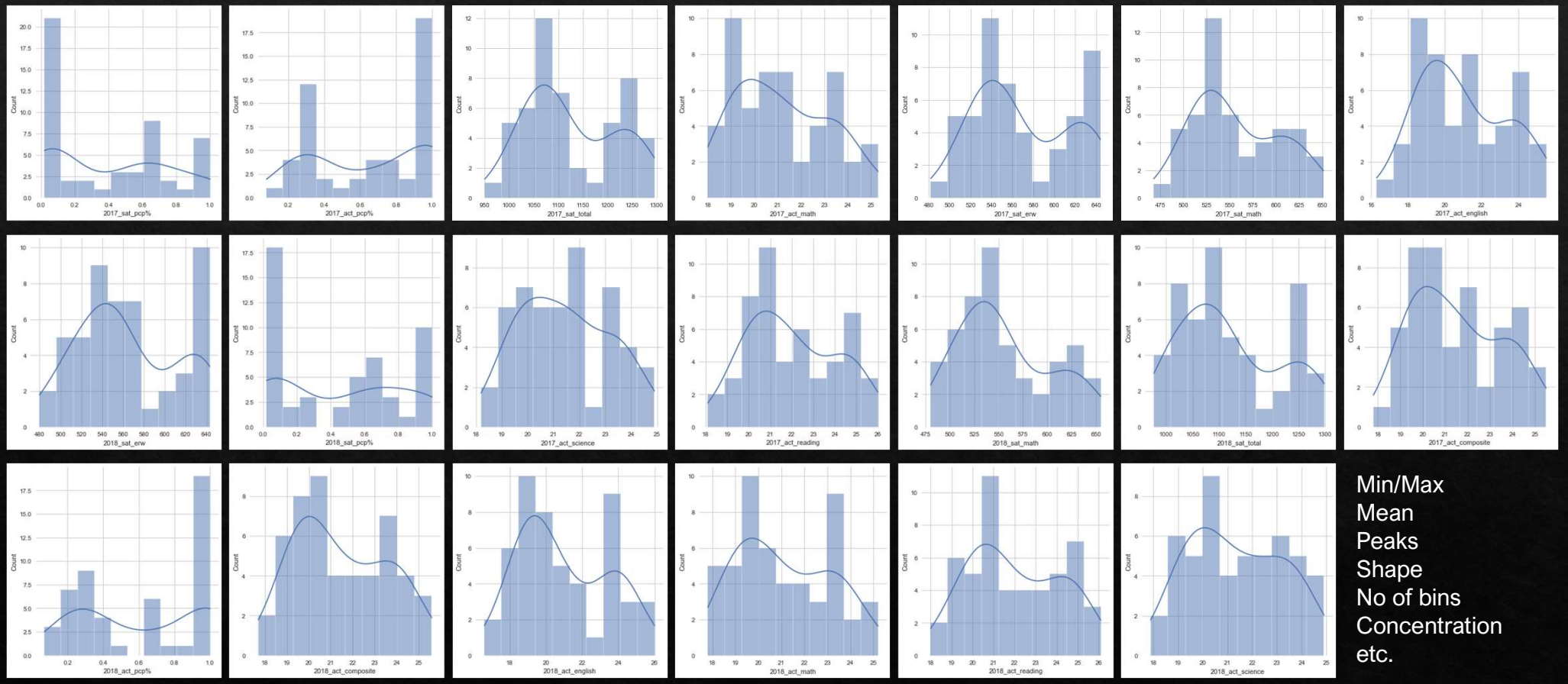
- Additionally, we plotted scatter plots on SAT Participation Rate vs SAT Total Scores and ACT Participation Rate vs ACT Composite Scores to see if there is any correlation. We observed that ACT and SAT scores are inversely correlated with their respective participation rates.



Visualization in Python & Tableau



Descriptive and inferential statistics



Min/Max
Mean
Peaks
Shape
No of bins
Concentration
etc.

Skewed Normal or Bimodal Distribution

Central Limit Theorem

Descriptive and inferential statistics

(a) Participation rate had a huge impact on state average SAT/ACT scores.

- Relationships between populations sizes & rates:
  - Low participation states have lesser people taking the tests. However, these people are usually the ones who need to pass the tests and put in more effort to get better scores. This results in bias and artificially higher scores.
  - High participation state have more people taking the tests. However, these include the good performers and the bad performers. This also results in bias and artificially lower scores.

- Granularity and aggregation:
  - How the participation rates are being aggregated also makes a difference. If it is simply aggregating the average participation rates of each school in each state, it will not be accurate because it is an average of an average. However, if the aggregation method is to take the total number of test takers against the total number of qualified/eligible test takers in the state in the year. It will have better accuracy.
  - Considering the above, we will still attempt to conduct inference with these data because until statistically tested with relative confidence level, it will be too early to reject any hypothesis on these data. IN the code file, there is an example of conducting statistical inference on SAT/ACT Math to demonstrate this point.

Descriptive and inferential statistics

Step 1 – Define the hypothesis

- Null hypothesis = higher SAT math score is NOT better than those with lower ACT math score (status quo)
- Alternate hypothesis = higher SAT math score IS better than those with lower ACT math score (what I am proving)

Step 2 – Decide confidence level

- Establish a level of significance which in this case let's consider 95% confidence interval, alpha = +/- 1.96.

Step 3 - Calculate the statistics

Step 4 - Find p-values and make a conclusion

- Check p-value of F-statistic = 0.00531 (model level) Null hypothesis will be true is this is zero i.e. all coeficient is zero Alternative Hypothesis is true if this is < 0.01, then at least one variable is likely to predict Y
- Check p-value of t-statistic = 0.005 (variable level) Null hypothesis will be true is this is zero i.e no impact Alternative hypothesis is true if this is <0.05, then the feature islikely to predict Y If this is >0.05, the 95% confidence interval will pass through 0
- Conclusion, since p-value < 0.05 the Null hypothesis is invalid and the Alternative hypothesis of "higher SAT math score IS better than those with lower ACT math score is valid."

"I am 95% confident that the true population of "2018_act_math“

is between -15.247 and -2.810"

OLS Regression Results

Dep. Variable:	2018_sat_math	**R-squared:**	0.148
Model:	OLS	**Adj. R-squared:**	0.131
Method:	Least Squares	**F-statistic:**	8.513
Date:	Sun, 08 Nov 2020	**Prob (F-statistic):**	0.00531
Time:	18:15:04	**Log-Likelihood:**	-264.97
No. Observations:	51	**AIC:**	533.9
Df Residuals:	49	**BIC:**	537.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	746.9615	65.667	11.375	0.000	615.000	878.923
2018_act_math	-9.0283	3.094	-2.918	0.005	-15.247	-2.810

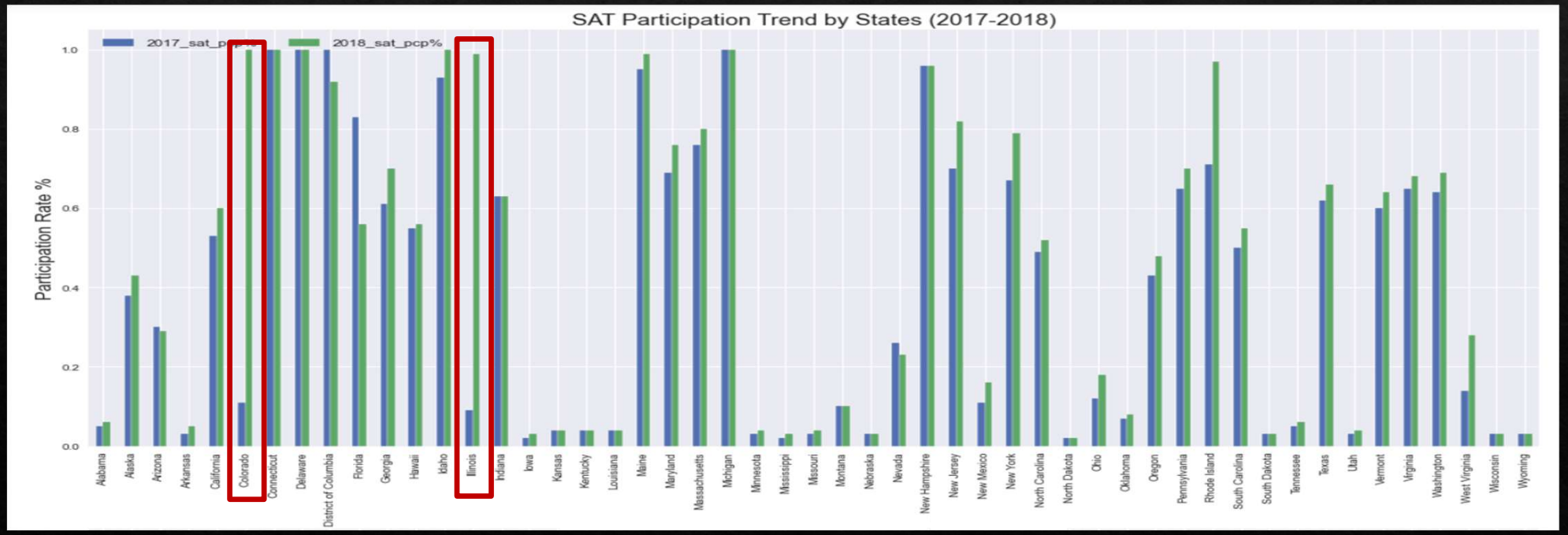
Omnibus:	2.688	**Durbin-Watson:**	1.639
Prob(Omnibus):	0.261	**Jarque-Bera (JB):**	2.548
Skew:	0.487	**Prob(JB):**	0.280
Kurtosis:	2.501	**Cond. No.**	224.

Descriptive and inferential statistics

(b) Identifying States of interest

Using simple bar charts in Python to plot the 2017/2018 SAT/ACT Participation Rates, we were able to easily identify the following states of interest:

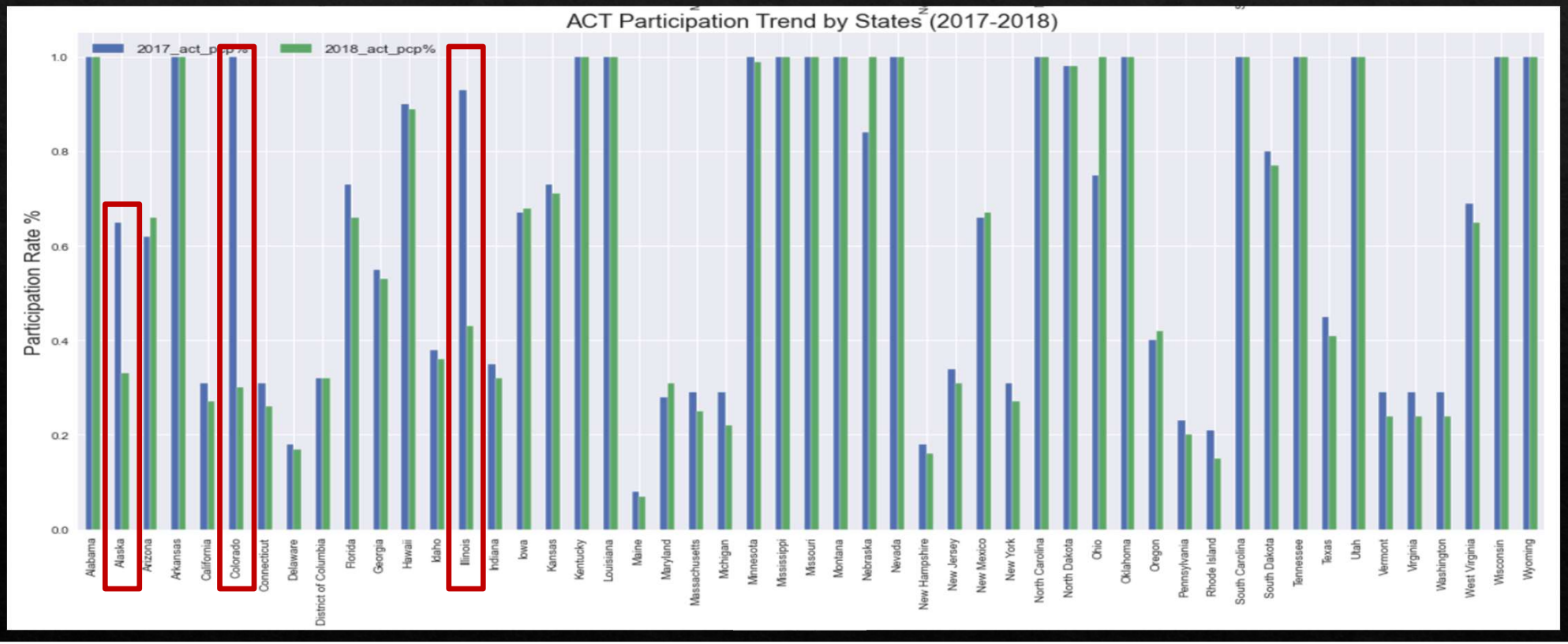
- From the SAT chart, Colorado and Illinois who had significant increase in 2018 SAT participation rate as compared to 2017 while at the same time had significant reduction in their ACT Participation Rate.



Descriptive and inferential statistics

- From the ACT chart, Alaska, Colorado and Illinois had significant reduction in 2018 ACT participation rate as compared to 2017.

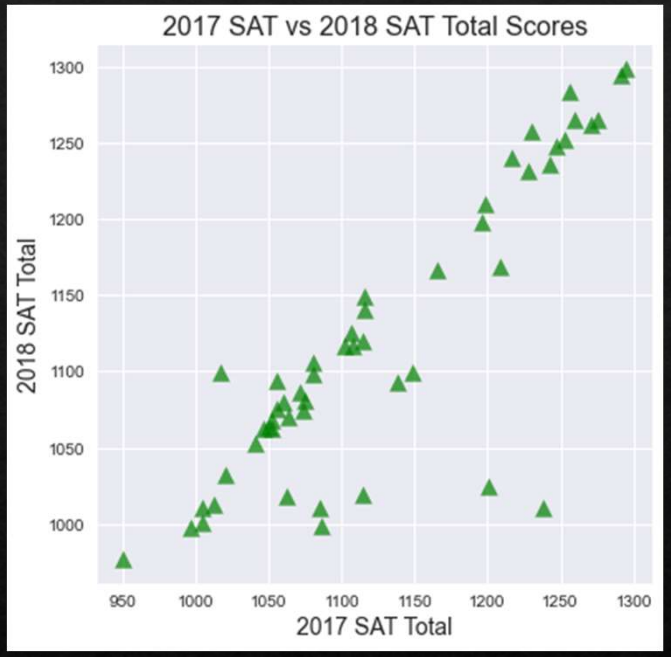
Therefore, **Alaska, Colorado and Illinois** are chosen as the three states to be investigated further.



Descriptive and inferential statistics

(c) Outside Research

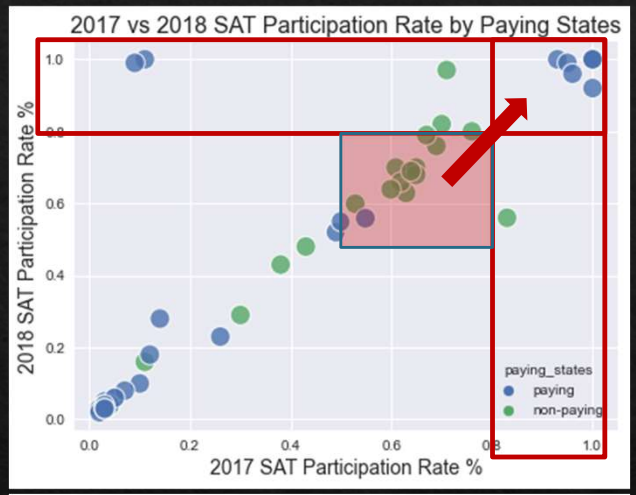
- Investigating Colorado and Illinois had significant increase in 2018 SAT participation rate as compared to 2017
  - According to various articles found online, both states implemented mandatory SAT testing as well as the allowing contracts with the ACT to expire.
  - Since the participation rate increase was significant, there may be an impact to these states' SAT and ACT scores. To investigate the correlation of this impact, we used scatter plots and we saw that there was a high correlation between SAT Total score in 2017 and 2018 for each state. There were a few exceptions where we see that states have dropped off from 2017 to 2018.
  - If we look at Colorado, the SAT total score fell from 1201 to 1025. This is likely attributable in part to a large increase in SAT participation (11% in 2017 to 28% in 2018), which might dilute the quality of the average score. However, this fails to explain the scale of the drop entirely, given that Illinois only saw a drop from 1115 to 1019 with a much larger increase in participation (9% participation in 2017, 99% in 2018).

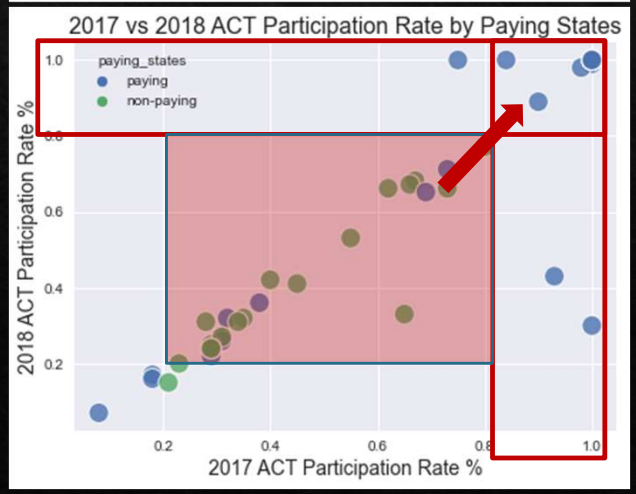


	state	2017_sat_pcp%	2017_sat_total	2018_sat_pcp%	2018_sat_total
1	Alaska	0.38	1080	0.43	1106
5	Colorado	0.11	1201	1.00	1025
13	Illinois	0.09	1115	0.99	1019

Descriptive and inferential statistics

- Investigating Alaska, Colorado and Illinois had significant reduction in 2018 ACT participation rate as compared to 2017
  - According to various found online, we knew that some states made it compulsory to take SAT or ACT while some do not. We also knew that some states paid for the tests and these tests were not cheap, even more costly if test takers chose to take the writing portions.
  - It is then interesting to determine the correlation (if any) between participation rate and whether the states pay for it or not. By compiling a list of states who paid for the tests and adding it to the data frame, we were able to use Seaborn scatter plot with "paying_states" as a category hue.
  - From the SAT scatter plots, we observed that almost all states who had above 80% participation rate, had their test fees paid by the states. Colorado and Illinois both saw significant increase in SAT participation rate when the states started paying the SAT fees in 2017-2018. Also since both states still required either the SAT and ACT, it was not surprising to see ACT participation rate dropping as more test takers switch to the free SAT test.
  - We also observed that there were a lot of states lying between 50-80% participation rate but test fees were not paid by the states. Given the high positive correlation of the 2017 and 2018 Participation rate, it was very possible that if the non-paying states in the 50-80% participation rate, started to pay for the test fees as well. We could see further participation rate increase.
  - As for ACT scatter plot, the same can be observed but there were a greater number of non-paying states in the 20-80% participation rate region. If these non-paying states started to pay for the test fees, the ACT participation rate is likely to increase. Interestingly, none of Alaska, Colorado and Illinois were listed as states that require the ACT. This further supports the drop in ACT participation rate. Moreover, Alaska did not pay for either SAT or ACT test fees.





Conclusion and recommendations

- The ACT and SAT participation distributions roughly mirror each other, with states tending to prefer one test or the other based on bias of whether SAT or ACT is required in that state and whether the state is paying for the tests or not.

- ACT and SAT scores are inversely correlated with their respective participation rates. This is likely due to selection bias, as low participation means those who are participating tend to be higher achieving, and high participation means diluted quality of performance.

- It is recommended that the College Board takes into consideration, household income data especially for states with low participation rate and is currently a non-paying state. The analysis had shown a correlation in this aspect that makes it worth further investigation.

- The analysis had yielded a list of states with low participation rate and is currently a non-paying state, of which the College Board could work with to share and implement support and fiscal policies that were put in place in states with high participation rate. In particular, our analysis specifically identified Alaska as a potential working target.

Thank you