ANALYSIS OF IVY LEAGUE STUDENTS IN 2020

INSTRUCTOR
Dr. Nguyen Kieu Linh
CONTRIBUTORS
01 – Nguyen Hoang Minh – HA150125
08 – Le Tuan Hung – HE150135
11 – Ngo Minh Hieu – HE150483
13 – Pham Le Nhat Linh – HE150512
22 – Do Huu Dai – HE151229
24 – Pham Viet Duong – HE151533

TABLE OF CONTENTS

Description

The Ivy League education system is famous for its reputation on the outcome of its students. It is the diversity in their background that makes Ivy League special—every person has an equal chance to be the best version of themselves.

With great interest, our team wanted to analyze Ivy League students' learning outcomes based on their background in 2020. We have collected a dataset containing basic information (gender, race/ethnicity, parental level of education, parental income, test preparation course, ACT composite score, SAT total score, high school GPA, college GPA, years to graduate) of 1,000 students.

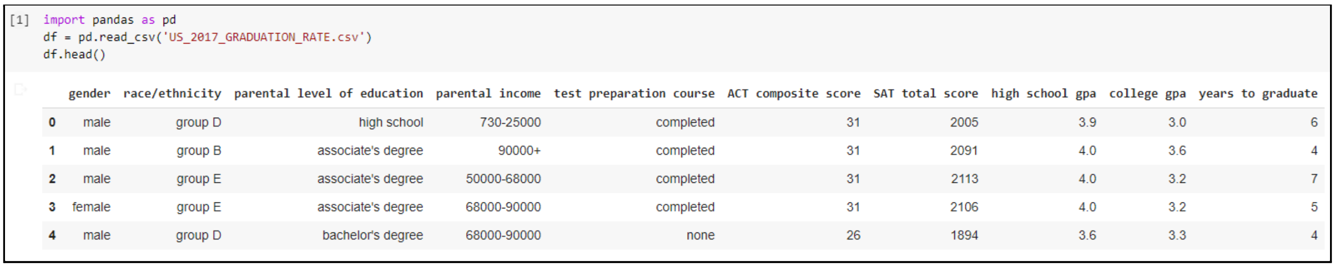
How was the dataset collected?

We collected the dataset from a survey conducted in January of 2021 by Dr. Royce Kimmons. Dr. Kimmons is an Associate Professor of Instructional Psychology and Technology at Brigham Young University—where he studies digital participation divides specifically in the realms of social media, open education and classroom technology use. He published all his findings on his website, roycekimmons.com.

We combined two datasets from his survey to build the proper dataset that fulfills our needs. The two sub-datasets are:

- Students performance: gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, writing score.
- Graduation rate: ACT composite score, SAT total score, parental level of education, parental income, high school gpa, college gpa, years to graduate.

The data we have just collected is not ready for analysis. To make it clear, we combined two datasets based on students' ID. There are some fields that are out of the scope we want to investigate, so getting rid of them to make the final dataset cleaner. We use Python's library called 'pandas' to import csv files.



```
[1] import pandas as pd
    df = pd.read_csv('US_2017_GRADUATION_RATE.csv')
    df.head()
```

	gender	race/ethnicity	parental level of education	parental income	test preparation course	ACT composite score	SAT total score	high school gpa	college gpa	years to graduate
0	male	group D	high school	730-25000	completed	31	2005	3.9	3.0	6
1	male	group B	associate's degree	90000+	completed	31	2091	4.0	3.6	4
2	male	group E	associate's degree	50000-68000	completed	31	2113	4.0	3.2	7
3	female	group E	associate's degree	68000-90000	completed	31	2106	4.0	3.2	5
4	male	group D	bachelor's degree	68000-90000	none	26	1894	3.6	3.3	4

Descriptive statistics

Analysis

Before any further analysis, we need to fully grasp the meaning of each field.

- Gender: student’s sex

  Male Female

- Race/ethnicity: student’s group

  Group A Group B Group C Group D Group E

- Parental level of education: student’s parents’ level of education

  High school Associate’s degree Bachelor’s degree Master’s degree

- Parental income: student’s parents' income

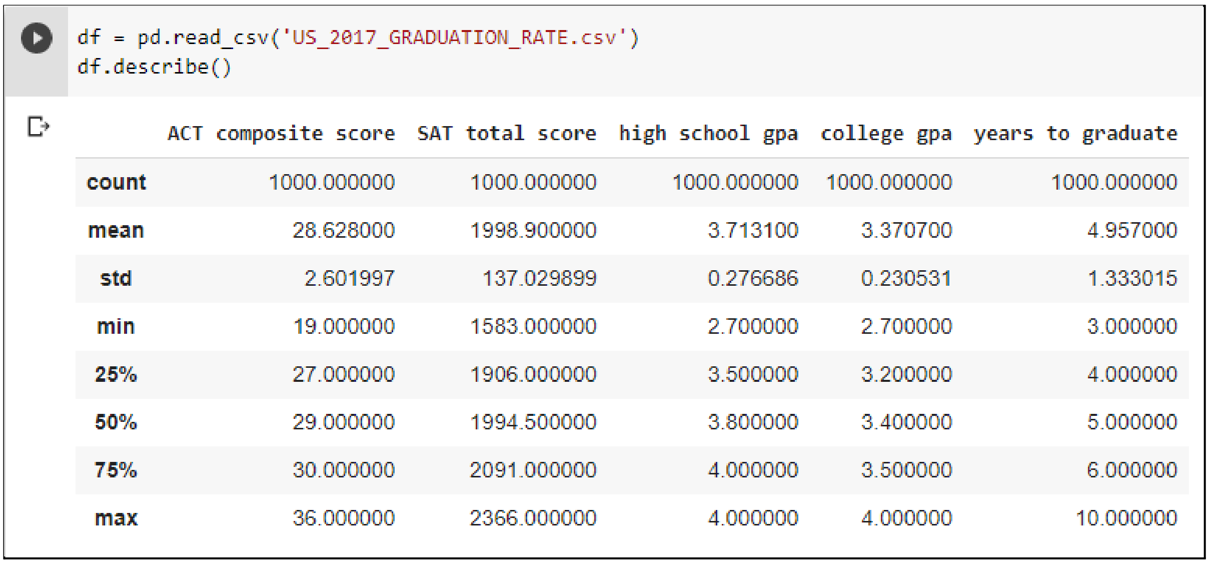
  0-730 730-25000 25000-50000 50000-68000 90000+

- Test preparation course: student’s status of taking test preparation course

  Completed None

- ACT composite score: student’s ACT composite score (college)
  Integer from 18 to 36.
- SAT total score: student’s SAT total score (high school)
  Integer from 1500 to 2400.
- High school GPA (float): student’s high school GPA
  Float from 0.0 to 4.0.
- College GPA (float): student’s college GPA
  Float from 0.0 to 4.0.
- Years to graduate (int): studying time at college of student
  Integer from 3 to positive infinity.

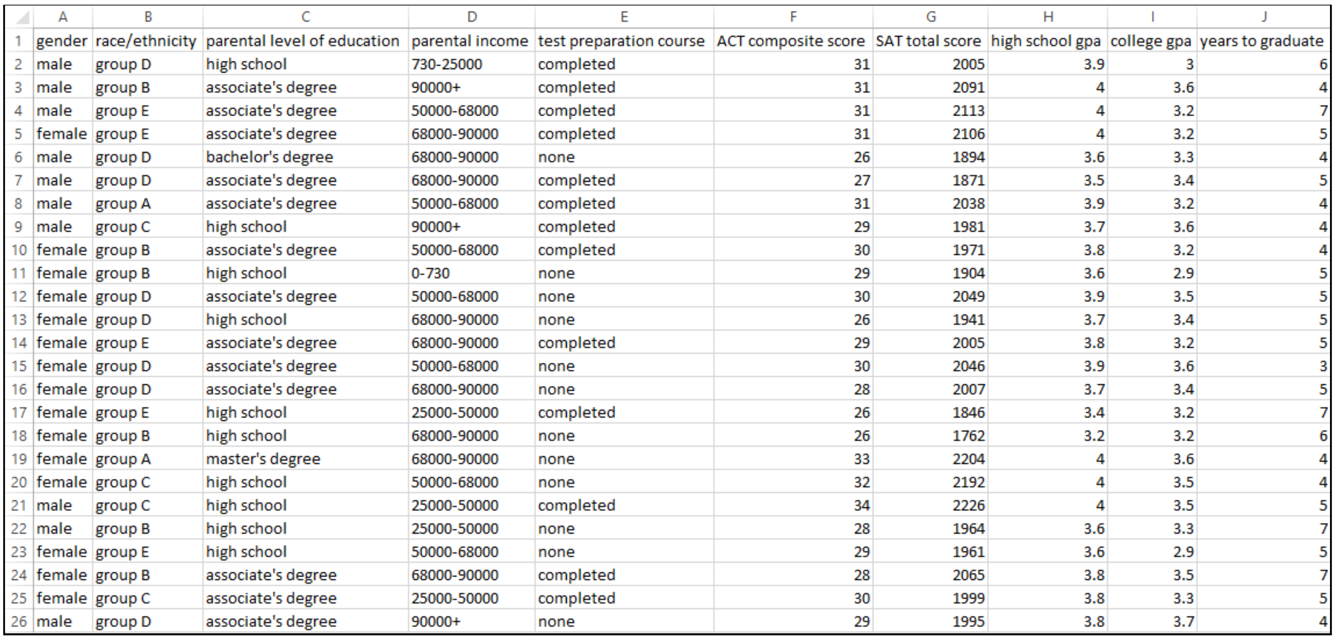
We use Python’s library called ‘pandas' to describe our dataset. The result belows



```
df = pd.read_csv('US_2017_GRADUATION_RATE.csv')
df.describe()
```

	ACT composite score	SAT total score	high school gpa	college gpa	years to graduate
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	28.628000	1998.900000	3.713100	3.370700	4.957000
std	2.601997	137.029899	0.276686	0.230531	1.333015
min	19.000000	1583.000000	2.700000	2.700000	3.000000
25%	27.000000	1906.000000	3.500000	3.200000	4.000000
50%	29.000000	1994.500000	3.800000	3.400000	5.000000
75%	30.000000	2091.000000	4.000000	3.500000	6.000000
max	36.000000	2366.000000	4.000000	4.000000	10.000000

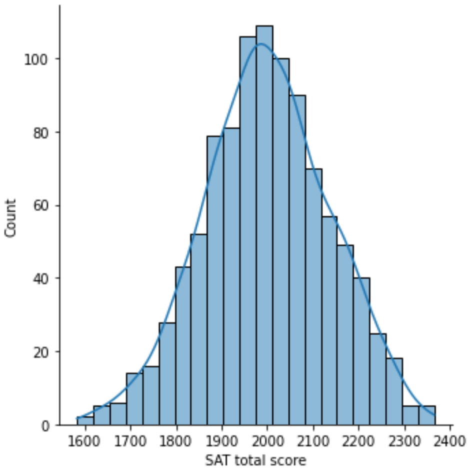
The dataset that we are going to analyse can be briefly viewed as below

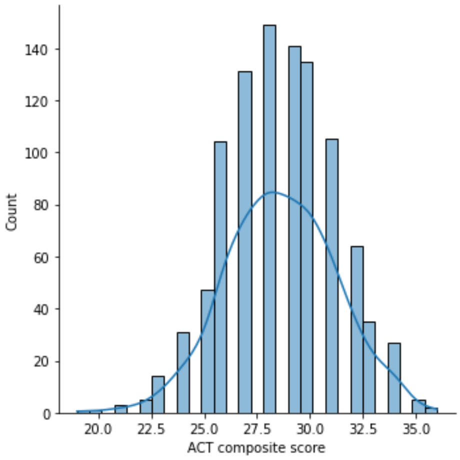


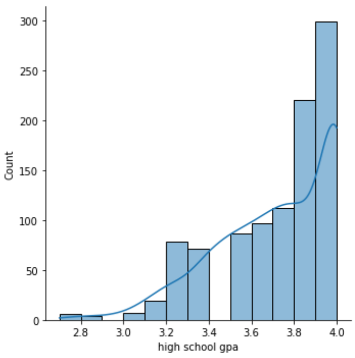
	A	B	C	D	E	F	G	H	I	J
1	gender	race/ethnicity	parental level of education	parental income	test preparation course	ACT composite score	SAT total score	high school gpa	college gpa	years to graduate
2	male	group D	high school	730-25000	completed	31	2005	3.9	3	6
3	male	group B	associate's degree	90000+	completed	31	2091	4	3.6	4
4	male	group E	associate's degree	50000-68000	completed	31	2113	4	3.2	7
5	female	group E	associate's degree	68000-90000	completed	31	2106	4	3.2	5
6	male	group D	bachelor's degree	68000-90000	none	26	1894	3.6	3.3	4
7	male	group D	associate's degree	68000-90000	completed	27	1871	3.5	3.4	5
8	male	group A	associate's degree	50000-68000	completed	31	2038	3.9	3.2	4
9	male	group C	high school	90000+	completed	29	1981	3.7	3.6	4
10	female	group B	associate's degree	50000-68000	completed	30	1971	3.8	3.2	4
11	female	group B	high school	0-730	none	29	1904	3.6	2.9	5
12	female	group D	associate's degree	50000-68000	none	30	2049	3.9	3.5	5
13	female	group D	high school	68000-90000	none	26	1941	3.7	3.4	5
14	female	group E	associate's degree	68000-90000	completed	29	2005	3.8	3.2	5
15	female	group D	associate's degree	50000-68000	none	30	2046	3.9	3.6	3
16	female	group D	associate's degree	68000-90000	none	28	2007	3.7	3.4	5
17	female	group E	high school	25000-50000	completed	26	1846	3.4	3.2	7
18	female	group B	high school	68000-90000	none	26	1762	3.2	3.2	6
19	female	group A	master's degree	68000-90000	none	33	2204	4	3.6	4
20	female	group C	high school	50000-68000	none	32	2192	4	3.5	4
21	male	group C	high school	25000-50000	completed	34	2226	4	3.5	5
22	male	group B	high school	25000-50000	none	28	1964	3.6	3.3	7
23	female	group E	high school	50000-68000	none	29	1961	3.6	2.9	5
24	female	group B	associate's degree	68000-90000	completed	28	2065	3.8	3.5	7
25	female	group C	associate's degree	25000-50000	completed	30	1999	3.8	3.3	5
26	male	group D	associate's degree	90000+	none	29	1995	3.8	3.7	4

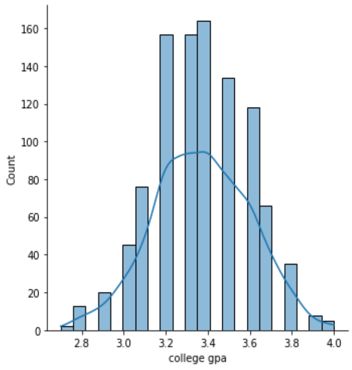
Visualization

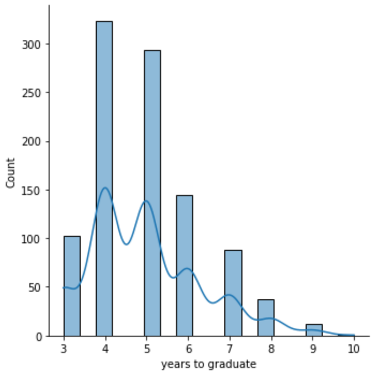
We use the 'displot' function of seaborn. This function provides access to several approaches for visualizing the univariate or bivariate distribution of data, including subsets of data defined by semantic mapping and faceting across multiple subplots.



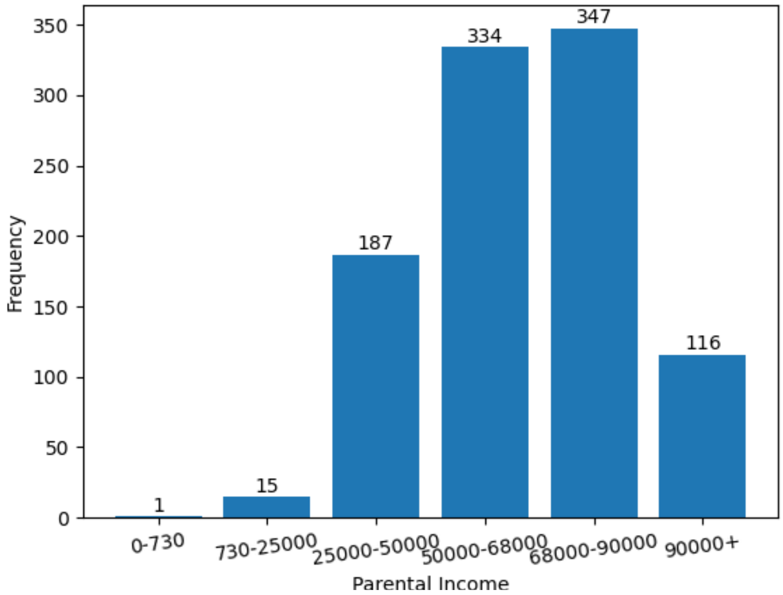




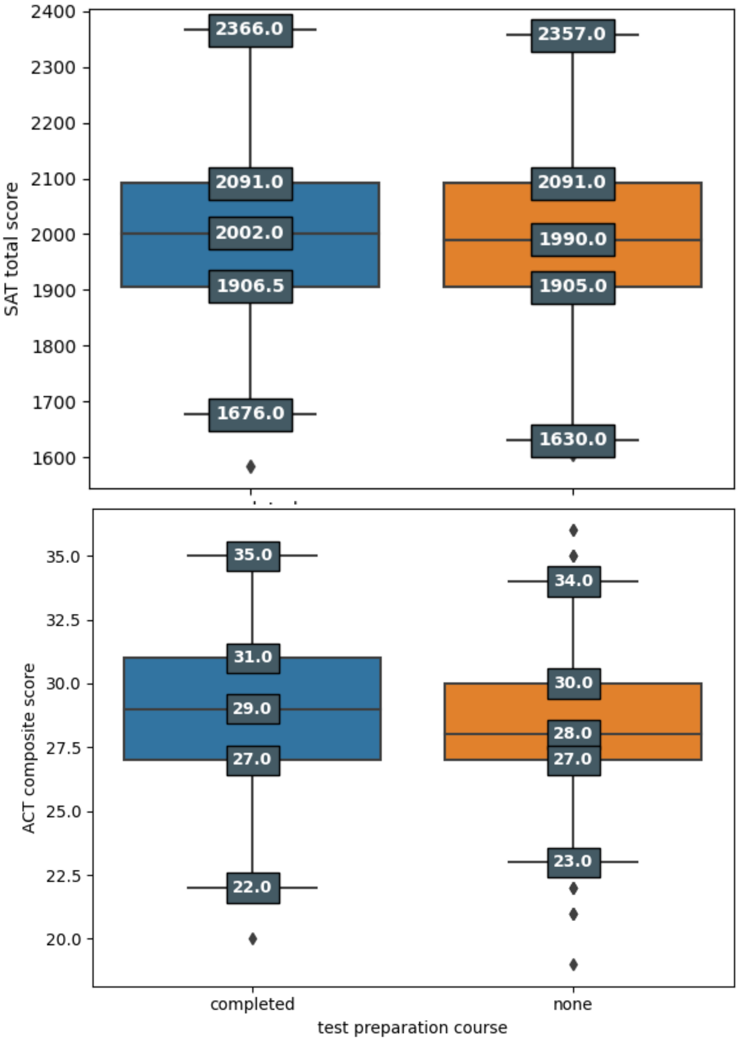




We want to construct a graph that describes the income of Ivy League students to see how diverse their parents are, based on their financial situation.



We doubt the effectiveness of the test preparation course. This is based on the fact that students need more time to practice themselves to make knowledge theirs, rather than taking more classes and being tired all day long due to the lack of sleep and stress. To test that hypothesis, we construct a boxplot that compares two groups of students, one completed the test preparation course and the other was not based on ACT composite score and SAT total score.



Based on the results, we can conclude that the test preparation actually increases students' performance on both ACT and SAT, but not much. Though, we can still consider taking a preparation course versus money and time that we have to pay. Is the investment worth the price? Or is self-study the skill we need to master? You answer!

Parameter estimation

Mean

We are curious about the average years to graduate of Ivy League students in 2020, but we only have the information of 1,000 students. Because we can not get the information of all Ivy League students due to the fact that they do not want to share, or they might be very busy with their after-college life, we can only estimate the average years to graduate. We call 1,000 students' information from the survey, which we know and can calculate, is the sample and the population is all Ivy League graduated students in 2020. By using inferential statistics, we can construct a confidence interval of the average years to graduate of Ivy League students in 2020, with the confidence level of 95%.

Based on the theorem:

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a $100(1 - \alpha)\%$ confidence interval on μ is

$$\bar{x} - t_{\alpha/2, n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2, n-1}\frac{s}{\sqrt{n}}$$

With

$\bar{x} = 4.96$	$\mu \in [4.87;\ 5.04]$
$s = 1.33$	
$n = 1000$	

In conclusion, we hold the belief that on average, Ivy League's students in 2020 take 4.87 to 5.04 years to graduate, with 95% confidence.

Variance

We want to calculate the variance in years to graduate of Ivy League students in 2020, because the variance represents how the data was distributed. If the variance is small, then it seems like the data is not varied, and vice versa. Based on the above logic, we can only estimate the confidence interval of the variance of the population based on the variance of the sample. By using inferential statistics, we construct a confidence interval of the variance in years to graduate of Ivy League students in 2020, with the confidence level of 95%.

Based on the theorem:

A $100(1 - \alpha)\%$ confidence interval of σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}$$

With

$n = 1000$	$\sigma^2 \in [1.63;\ 1.94]$
$s^2 = 1.78$	

In conclusion, we hold the belief that the variance years to graduate Ivy League's student in 2020 is between 1.63 and 1.94, with 95% confidence.

Proportion

Similar to the above situation, now we know we can also estimate the proportion of the population based on the sample's proportion. By using inferential statistics, we construct a confidence interval of the proportion of Ivy League students that completed the test preparation course, with the confidence level of 95%.

Based on the theorem:

If $\hat{p}$ is the proportion of observations in a random sample of size n that belongs to a class of interest, an approximate 100(1 − α)% confidence interval on the proportion p of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

With

$\hat{p} = 0.35$	$p \in [0.32;\ 0.38]$
$n = 1000$	

In conclusion, we hold the belief that 32% to 38% of Ivy League's students in 2020 took the test preparation course, with 95% confidence.

We obviously do not suppose the confidence interval that large. In fact, the estimation should only differ from the true population proportion by 2%. In statistics, 2% is the acceptable percentage for the estimation erroneous, which means any proportion that is bigger than 2% should be left reconsidered. If we want to reduce the confidence interval of the population proportion width, there are three ways.

(1) Change the $\hat{p}$. This method is unacceptable, since the basic assumption when collecting a dataset is that we must do it randomly. Any intervention in the process ruins the whole result afterwards, so we skip this.

(2) Decrease α. This method is acceptable, but in some cases, it leads to the problem of not having enough confidence level to trust the estimation result. We should find another way to make the outcome more reliable, with acceptable errors.

(3) Increase n. This method is feasible and applicable because it can solve the two above problems. In this case, we will use this method.

Based on the theorem:

The required sample size that the error estimating $\left|\hat{p} - p\right|$not exceed E is

$$n = \left[\left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1 - \hat{p}) \right]$$

With

$\hat{p} = 0.35$	$n = 2188$

In conclusion, we need a sample of at least 2188 observations to minimize the error of the proportion that Ivy League's students took the test preparation course to 2%.

Hypothesis testing

Mean

In Vietnam, the average years to graduate from college is 4. We think that the Ivy League students are elite, and they probably need less time than Vietnamese students to graduate. To test this hypothesis with the confidence level of 95%, we apply the hypothesis knowledge.

First, we need to conduct a null hypothesis and alternative hypothesis. Let H_0 be the null hypothesis, and H_1 be the alternative hypothesis. Then H_1 is the claim that we want to prove, and H_0 is mutually exclusive with H_1. In this case, the language form of H_1 is Ivy League students' average years to graduate is less than 4. Then H_0 is going to be Ivy League students' average years to graduate is not less than 4. To put it in symbolic form (convenient to calculate), we have

$$H_0: \mu \geq 4$$

$$H_1: \mu < 4$$

Second, we compute the test statistic for the hypothesis. In this example, we are testing hypotheses for the mean of the population, and we have not known its mean (and we will never know!) So the test statistic is the random variable of the t-distribution T. We call it t_0, with the value of

$$t_0 = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

With

$\overline{X} = 4.96$	$t_0 = 22.70$
$s = 1.33$	
$n = 1000$	

Third, we decide whether we can reject H_0, or we fail to reject H_0. We can do this in two ways

(1) Identify the acceptance region, then check whether the test statistic is in the acceptance region. If not, then the test statistic is in the critical region, where we reject the null hypothesis H_0. Otherwise, we fail to reject H_0.

(2) Compute the P-value, then compare the P-value with the significant level α. If P-value is smaller than α then we reject the null hypothesis H_0, otherwise we fail to reject it.

In this example, we choose the first method. The significant level $\alpha = 1 - confidence\ level\ = \ 5\%$, and with the one-tailed case, we have the critical value is $-t_{\alpha, n-1} = \ -1.65$. It means that if the test statistics is less than -1.65, then we reject H_0, otherwise we fail to reject it. Since the test statistic is 22.70, we fail to reject H_0, which means the average years to graduate of Ivy League's students is greater than 4 years, with 95% confidence.

This phenomenon can be explained by the difficulty of the Ivy League curriculum in comparison to Vietnam's. They have to conduct a lot of research and real experiments to fully understand how to apply knowledge into reality. That is why even though Ivy League's students are smarter than Vietnamese's, they need more time to graduate from college than us.

Proportion

We share the same belief that women do not attend college as often as men do. This belief is based on the assumption that men carry the burden of making money and build great things, in contrast to women—destined to stay at home and do the chores. We want to test this common assumption by constructing a hypothesis testing whether or not the proportion of male attending college is over a half or not, with the confidence level of 95%.

First, we conduct a symbolic form of the hypotheses.

$$H_0: p \ \leq \ 0.5$$

$$H_1: p \ > \ 0.5$$

Second, we compute the test statistic for the hypothesis. Since this is the test statistic of population proportion, it follows the standard normal distribution Z. We call it z_0, and its value is

$$z_0 = \frac{X-np_0}{\sqrt{np_0(1-p_0)}}$$

With

$X = 483$	$z_0 = -1.08$
$n = 1000$	

Third, we decide to reject H_0 or not. Instead of identifying the acceptance region, now we compute its P-value. We apply this formula to find the P-value

$$1 - \Phi(z_0)$$

We found that P-value = 0.86. It can be easily seen that P-value > α, so we fail to reject H_0, which means that the proportion of men attending Ivy League in 2020 is not greater than 50%, with 95% confidence.

This can be explained by the fact that we are misunderstanding gender equality in education. Our misconception was true, until 2010. After that, the ratio of women attending tertiary education is getting higher, and in most developed countries like the U.S., the ratio of women in university has always been higher than men's. We learn a lesson when we experiment and analyse data, that is to constantly update our knowledge and get rid of old-fashioned belief, is what our mindset should be—always open.

Two samples

Hypothesis testing for the difference between two population means

We think that there is some difference between the average years to graduate of men and women, because women tend to be more focused on studying, while men usually find themselves at parties or other social activities. To test this hypothesis with a confidence level of 95%, we need to follow the steps discussed above. To make it clear, we assume that population 1 is the average years to graduate of Ivy League's male students in 2020, and population 2 is women's.

First, we conduct a symbolic form of the hypotheses.

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

Second, we compute the test statistic for the hypothesis. This random variable has a t-distribution with degrees of freedom $n_1 + n_2 - 2$, and the test statistic should be evaluated based on $\overline{X_1}$ and $\overline{X_2}$. The following formula applies

$$t_0 = \frac{\overline{X_1} - \overline{X_2} - \Delta_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where as s_p is pooled standard deviation of the two samples. Its formula is

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

With

$\overline{X_1} = 4.95$ and $\overline{X_2} = 4.97$	$t_0 = \ -0.35$
$s_1 = 0.83$ and $s_2 = 0.94$	
$n_1 = 483$ and $n_2 = 517$	

Third, we compute the P-value to decide whether we should reject H_0 or not. The following formula is applied to find the P-value for one-tailed hypothesis testing

$$1 - P(T_{n_1+n_2-2} > t_0)$$

We easily found P-value = 0.36, which is greater than α. We fail to reject H_0, which means that in Ivy League (2020), on average, men need less time to finish school than women do, with 95% confidence.

This is another misconception that we might be lucky enough to recognize when analyzing data. Gender inequality is a lifelong human's issue, but it is getting better and better. By acknowledging the status quo, we can identify the most urgent problem requiring us to have a solution immediately, like climate change. With that in mind, it is important to always be ready for new knowledge and upgrade yourself daily.

C.I. for the difference between two population means

To further analysis on the above topic, we want to find out the difference between the average years to graduate of men and women, with a confidence level of 95%. We assume the population 1 is average years to graduate of Ivy League's male students, and population 2 is female's. This value has a t-distribution, with the degrees of freedom equal to 97, as it follows this formula

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

Based on the theorem:

If $\overline{x_1}$, $\overline{x_2}$, s_1^2, s_2^2 are the means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown and unequal variances, an approximate 100(1 − α)% confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\overline{x_1} - \overline{x_2} - t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \overline{x_1} - \overline{x_2} + t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

With

$\overline{x_1} = 4.95$ and $\overline{x_2} = 4.97$	$\mu_1 - \mu_2 \in [-0.14;\ 0.10]$
$s_1 = 0.83$ and $s_2 = 0.94$	
$n_1 = 483$ and $n_2 = 517$	

In conclusion, on average an Ivy League's male student will graduate earlier than female student by -0.14 year to 0.10 year, with 95% confidence.

Hypothesis testing for the difference between two population proportions

We are curious about the fact that male students are given more chances to study, therefore they may take the test preparation course more than women do. To put it into words, we claim that the proportion of Ivy League's male students taking the test preparation course is higher than females. We assume the population 1 is male's proportion, population 2 is female's and the confidence level is 95%.

First, we conduct a symbolic form of the hypotheses.

$$H_0: p_1 - p_2 \le 0$$

$$H_1: p_1 - p_2 > 0$$

Second, we compute the test statistic for the hypothesis. This random variable follows standard normal distribution and the test statistic should be evaluated based on $\widehat{p_1}$ and $\widehat{p_2}$. The following formula applies

$$z_0 = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

where as $\widehat{p}$ is the pooled proportion of the two samples. Its formula is

$$\widehat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

With

$\widehat{p_1} = 0.348$ and $\widehat{p_2} = 0.354$	$z_0 = \ -0.20$
$X_1 = 168$ and $X_2 = 183$	
$n_1 = 483$ and $n_2 = 517$	

Third, we compute the P-value to decide whether we should reject H_0 or not. The following formula is applied to find the P-value for one-tailed hypothesis testing

$$1 - \Phi(z_0)$$

We easily found P-value = 0.58, which is greater than α. Hence, we fail to reject H_0, which means that in Ivy League (2020), female students tend to take the test preparation course more than male students do, with 95% confidence.

C.I. for the difference between two population proportions

To further analysis on the above topic, we want to find out the difference between the percentage Ivy League's male students take the test preparation course and female student's, with a confidence level of 95%. We assume population 1 is the proportion of Ivy League's male students, and population 2 is female's. This value follows standard normal distribution.

Based on the theorem:

If $\widehat{p_1}$ and $\widehat{p_2}$ are the sample proportions of observations in two independent random samples of sizes n_1 and n_2, that belong to a class of interest, an approximate two-sided 100(1 − α)% confidence interval on the difference in the true proportions $p_1 - p_2$ is

$$\widehat{p_1} - \widehat{p_2} - z_{\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}} \le p_1 - p_2 \le \widehat{p_1} - \widehat{p_2} + z_{\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

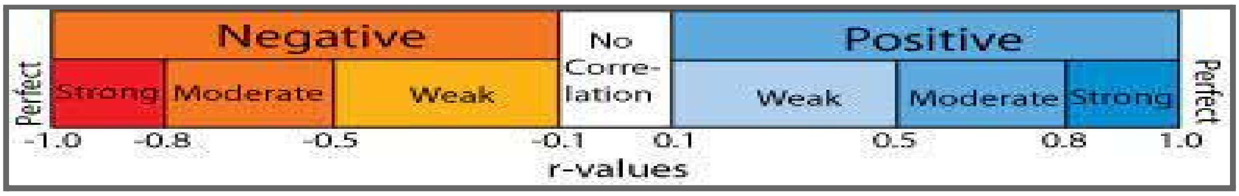
With

$\widehat{p_1} = 0.348$ and $\widehat{p_2} = 0.354$	$p_1 - p_2 \in [-\ 0.11;\ 0.09]$
$n_1 = 483$ and $n_2 = 517$	

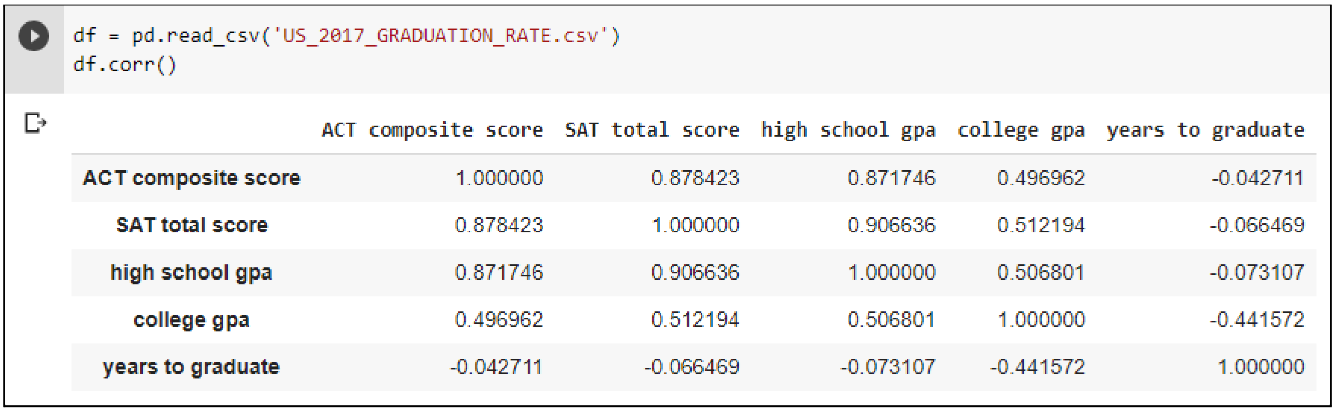
In conclusion, an Ivy League's male student will be more likely to take the test preparation course than female student's by -11% to 9%, with 95% confidence.

Regression analysis

Correlation is a term used to represent the statistical measure of linear relationship between two variables. It can also be defined as the measure of dependence between two different variables. Correlation coefficient (notation: R) varies between -1.0 and 1.0. If $|R|$ has a value close to 0, then the two variables share no relation. Otherwise, the two variables have weak/moderate/strong relations. The sign of R demonstrates how the two variables relate to each other. If R > 0, then two variables are increasing (positive), else decreasing (negative.)



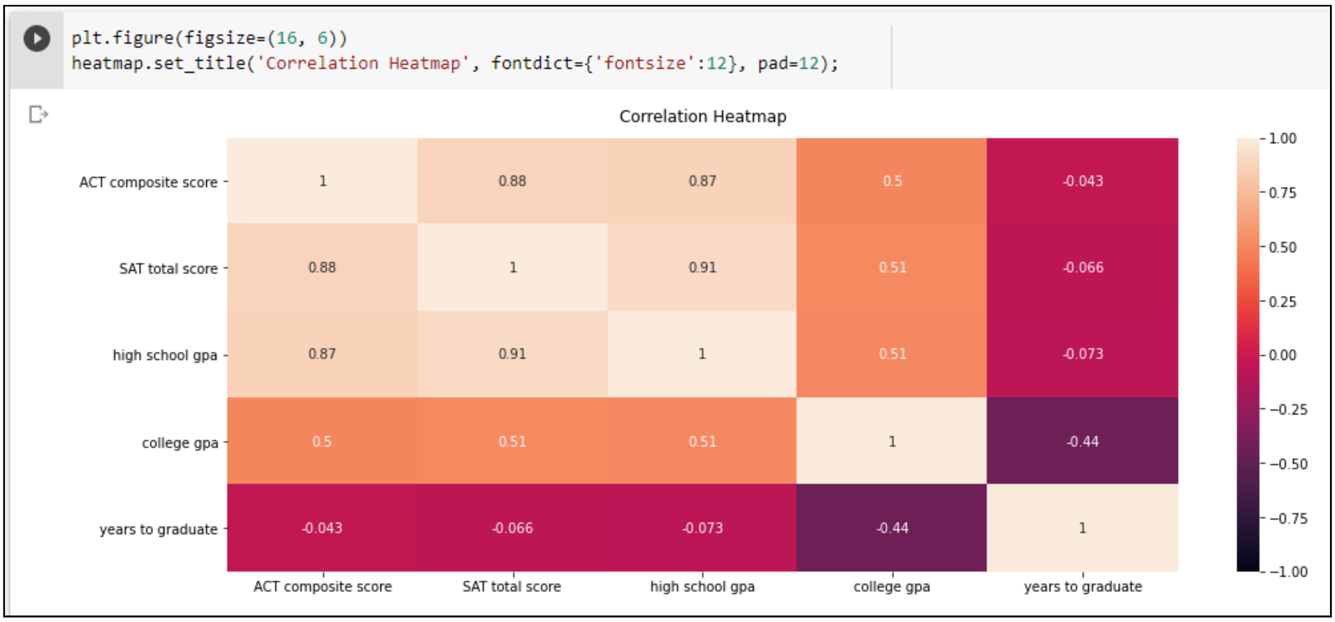
Correlation matrix is a data structure of a two-dimensional array that shows correlation coefficients between all of these variables.



```
df = pd.read_csv('US_2017_GRADUATION_RATE.csv')
df.corr()
```

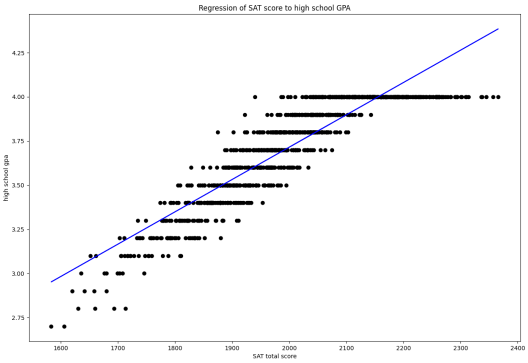
	ACT composite score	SAT total score	high school gpa	college gpa	years to graduate
ACT composite score	1.000000	0.878423	0.871746	0.496962	-0.042711
SAT total score	0.878423	1.000000	0.906636	0.512194	-0.066469
high school gpa	0.871746	0.906636	1.000000	0.506801	-0.073107
college gpa	0.496962	0.512194	0.506801	1.000000	-0.441572
years to graduate	-0.042711	-0.066469	-0.073107	-0.441572	1.000000

We use a Python library called ‘seaborn’ to construct correlation heatmap—a graphical representation of correlation matrix.

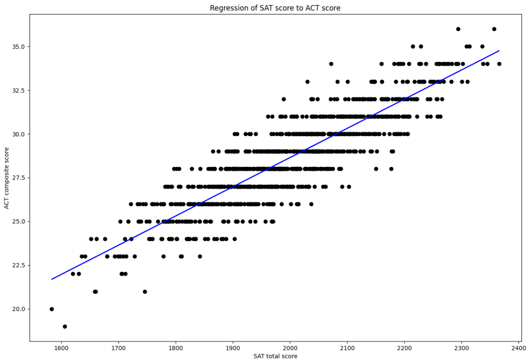


```
plt.figure(figsize=(16, 6))
heatmap.set_title('Correlation Heatmap', fontdict={'fontsize':12}, pad=12);
```

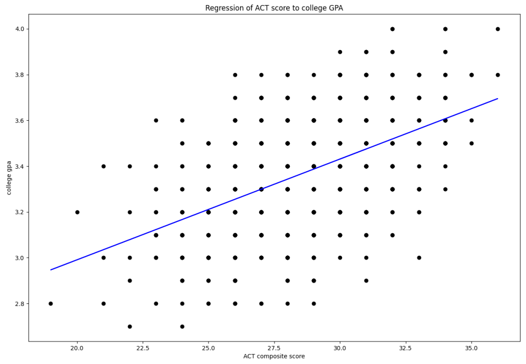
We construct regression models based on the above correlation coefficients by using a Python library called ‘matplot’.



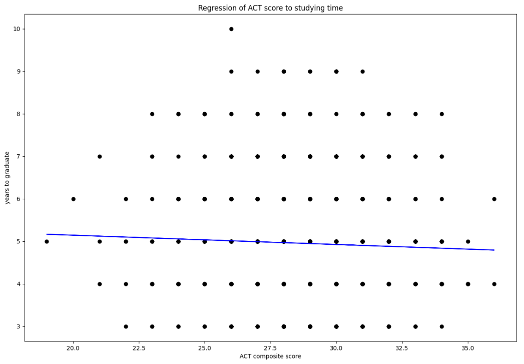
SAT score to high school GPA
(R = 0.91)



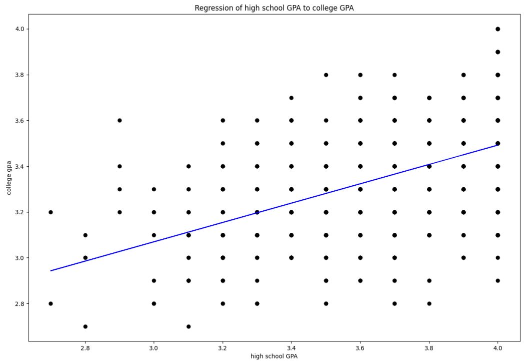
SAT score to ACT score
(R = 0.88)



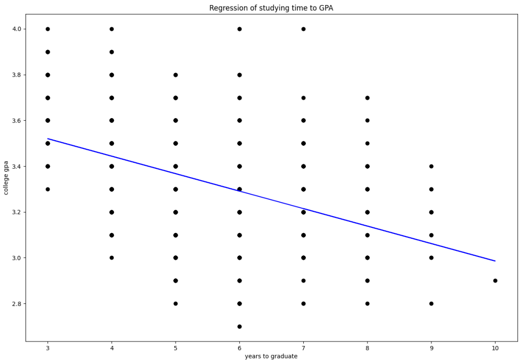
ACT score to college GPA
(R = 0.50)



ACT score to years to graduate
(R = -0.04)



High school GPA to college GPA
(R = 0.51)



Years to graduate to college GPA
(R = -0.44)

From the above models, we can conclude that

(1) The higher their SAT total score is, the higher their high school GPA and ACT composite score will be.
(2) Students with a good ACT composite score may also achieve a high college GPA.
(3) There is no correlation between ACT composite score and years to graduate.
(4) Students with good high school GPA may also achieve high college GPA.
(5) Students who have to spend more time studying (years to graduate) may have a lower college GPA.

Those conclusions can be explained based on their IQ. As we have mentioned, the test preparation course does not play an important role in determining the score, as SAT and ACT tests are designed to test student's IQ and logic. Because of that, a student that scores a high SAT score is very likely to perform well on the ACT test. Also, a student performing well in SAT or ACT may also get a high GPA, in high school and college. However, the time you need to take to finish college is very unlikely to be affected by any factor. All we know is college GPA may decrease as you spend more time studying in college due to class retake. Overall, IQ manipulates how students perform at tests and schools.

Conclusion

Statistics play a huge role in life. It shows us how the world looks like, even when we just have a sample of data by using parameter estimation and hypothesis testing. In addition, we can verify whether or not the two variables are related to each other by applying regression analysis and constructing a regression model. Furthermore, we can use Python or Excel to build graphical models that describe our data. The making of this research teaches us not only statistical methods but also how we can use tools to make things visualized.

Send our deepest appreciation to the author of the survey, Dr. Royce Kimmons, our instructor, Dr. Nguyen Kieu Linh and the amazingly talented contributors. Without your effort, this project would be neither finished, nor successful.