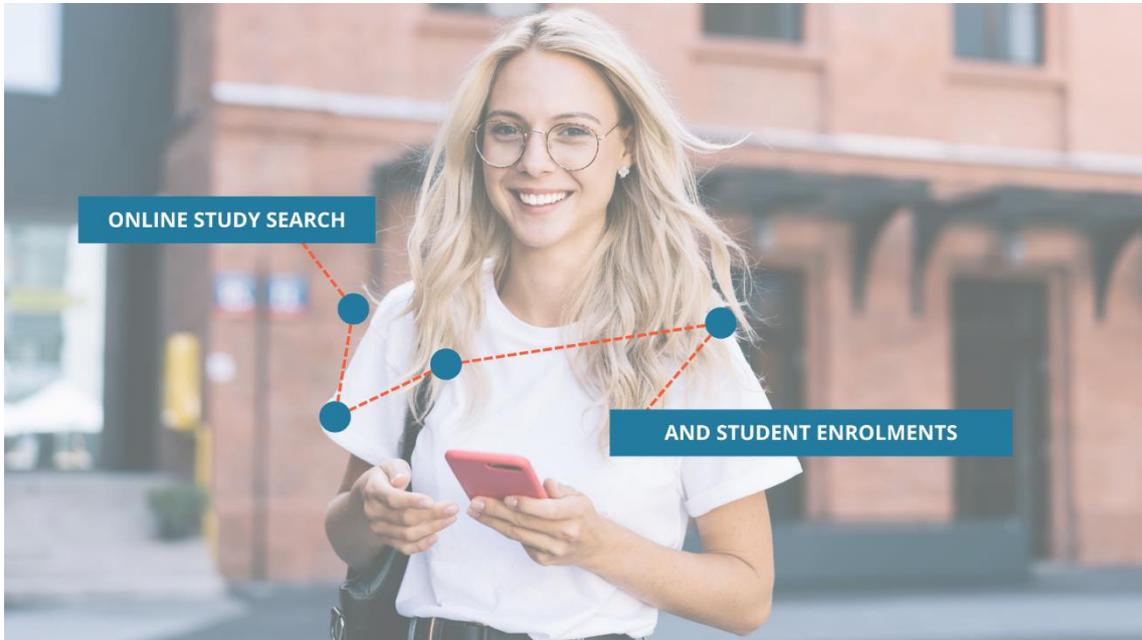


Introduction



Studyportals websites saw over 52 million prospective students in 2021. At the same time, we know that, currently, there are only about 5 million enrolled international students globally. What is the correlation between students browsing digitally for study degrees and the actual enrollment rates?

The following research paper aims to determine whether U.S. page view data collected from the Studyportals websites is representative of higher education (HE) demand in the United States. The Studyportals page view dataset is uniquely large as visitors generate over 52 million sessions annually. Moreover, the findings of this research are based on page view and enrollment data from 2018, 2019, 2020, and 2021.

Correlation measures the strength of association between two variables and the direction of the relationship. The value of the correlation coefficient is between +1 and -1. A value of ± 1 indicates a perfect degree of association between two variables. As the coefficient value approaches 0, the strength of the association is weaker.

The United States does not have a centralized source for higher education enrollment data. However, each state has a HE-related department that reports enrollment figures. This study calculates the correlation between page views and enrollments for five states. These include New York, Ohio, Texas, Pennsylvania, and Massachusetts. These states have some of the largest student populations. Appendix A includes top U.S. states in terms of student population.

Methodology

Firstly, Spearman correlation is used to determine the strength of the relationship between page views and enrollments. Since page views and enrollments are not normally distributed, Pearson correlation (a widely adopted formula) is not an appropriate measure. Appendix C shows the results of the Shapiro-Wilk tests (normality tests) for every state. Instead of calculating the correlation coefficient using covariance and standard deviations of the observations themselves, Spearman correlation uses the relative rank of each

observation to determine correlation. Although Spearman correlation does not require a normal distribution, it does assume a monotonic relationship. This is a mathematical name for an increasing or decreasing relationship between the two variables.

Secondly, this paper estimates the correlation between the absolute number of page-views and enrollments. Studyportals dashboards mostly display relative page views. Rank correlation is the same whether one uses absolute or relative values. Moreover, rank correlation is calculated at the institution level. More specifically, the page views each U.S. institution accumulates on the Studyportals site are compared with the number of enrollments.

Thirdly, this analysis was conducted using Python. SciPy was used to calculate the rank correlation for each state. SciPy is an open-source library for Python which contains modules for optimization, linear algebra, and special functions (natural logarithm, tangent, arctangent, etc.), which are commonly used in data science and engineering.

Data sources differ per state. For example, Michigan HE enrollment data was collected from the Michigan Department of Higher Education's "MI School Data" page. Data from North Carolina was collected from www.NorthCarolina.edu. Enrollments in Pennsylvania come from Pennsylvania's State System of Higher Education. In short, enrollment data comes from each state's respective education department. Appendix B contains the exact source for all enrollment data such that researchers can replicate this study. Enrollment data from 2018, 2019, 2020, and 2021 was used for this study.

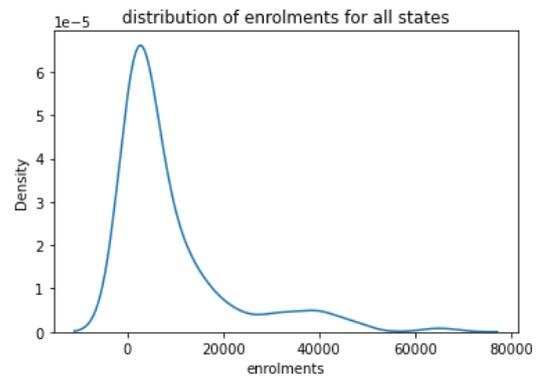
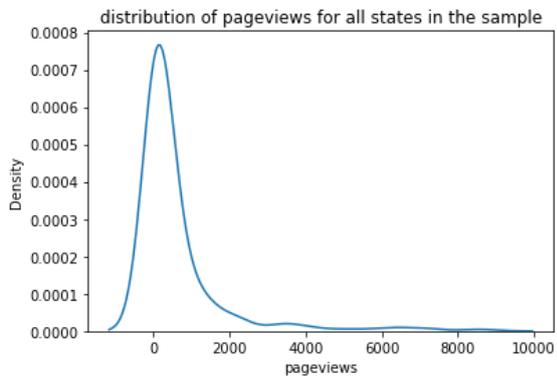
Moreover, outliers that were more than 3 standard deviations removed from the mean (in terms of page views and enrollments) were dropped from the sample. Observations that have a Z-score of 3 most likely are a result of measurement error.

EDA (Exploratory Data Analysis)

The following scatter plot shows enrollments against page-views for all states in the sample:



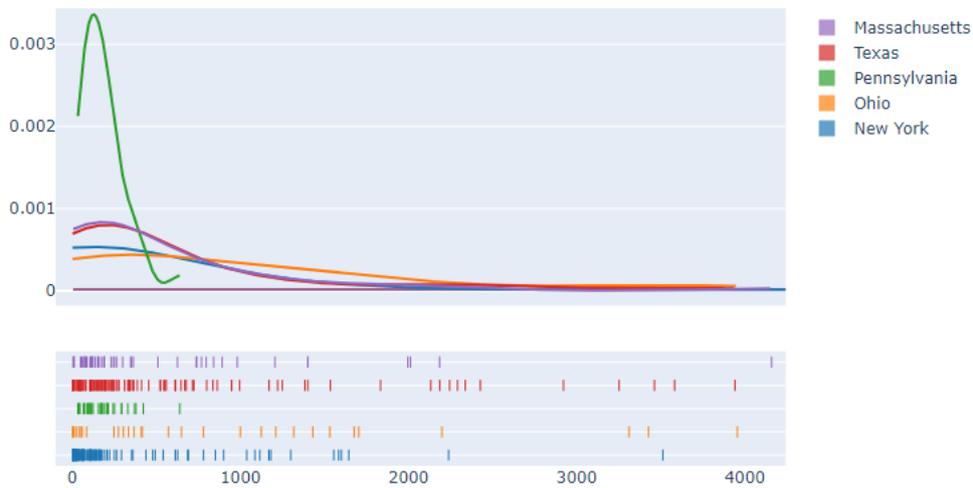
The following distribution plots indicate non-normal distributions for page-views and enrollments:



Page view Distributions per- State:

Each observation in the visualization below corresponds to an institution year (Ohio University-2018, University of Houston-2019).

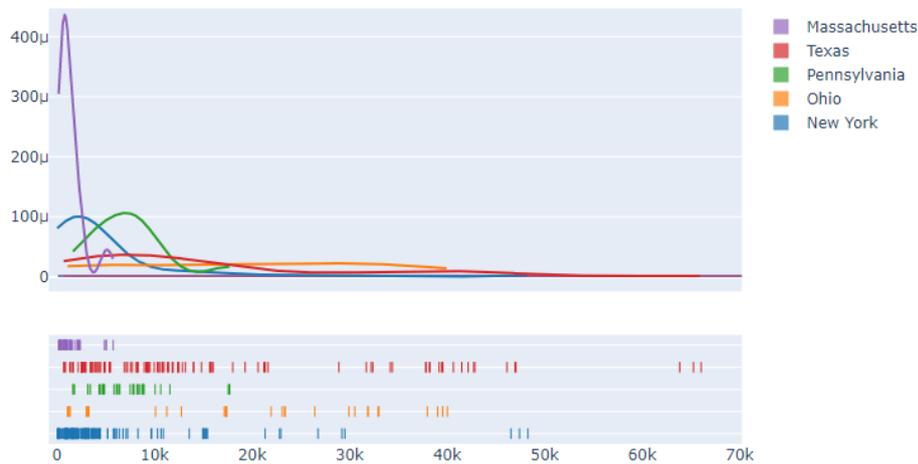
Yearly Page View Distribution for States in the Sample



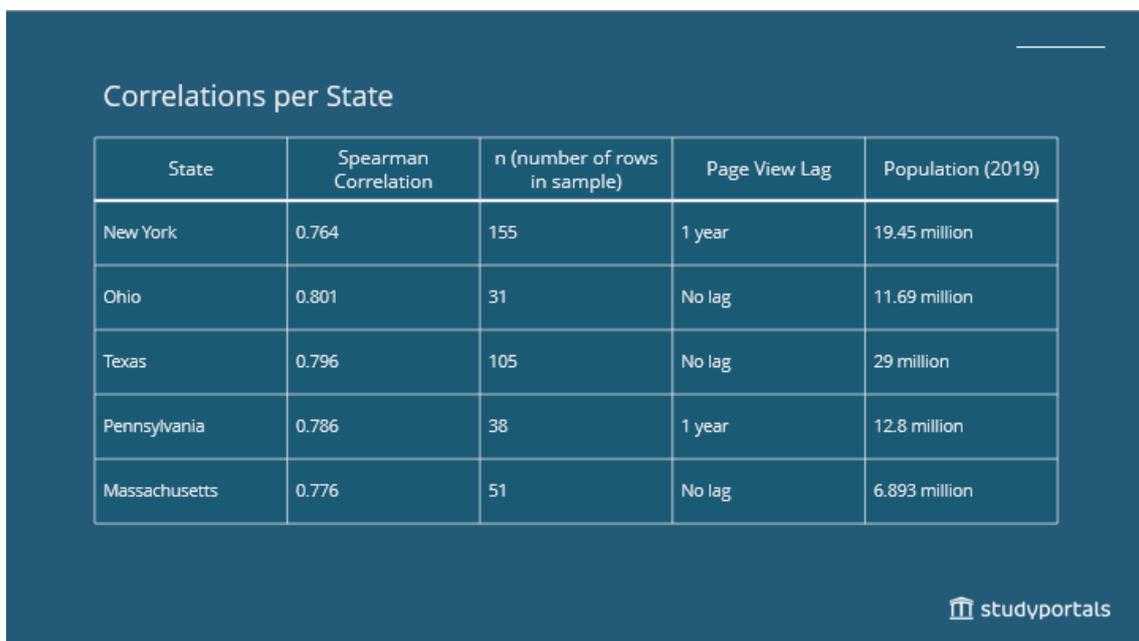
Enrollment Distributions per State:

Each observation in the visualization below corresponds to an institution year (Ohio University_2018, University of Houston_2019).

Yearly Enrolment Distribution for States in the Sample



Findings



Generally, Spearman rank correlation is interpreted in the following manner:

- .80-1.0 "very strong"
- .60-.79 "strong"
- .40-.59 "moderate"
- .20-.39 "weak"
- .00-.19 "very weak"

Conclusion

In short, the results demonstrate a strong correlation between U.S. enrollment data and page views on the Studyportals site. More specifically, the correlation in Ohio can be interpreted as “very strong,” while the correlations in New York, Texas, Pennsylvania, and Massachusetts are “strong.” These results are relevant for HE administrators because they can leverage page view data to estimate the demand for every HE program worldwide. Moreover, since the correlation is high for these five key markets, Studyportals page view data is likely representative of other markets in the United States.

Replications

Rutger R. 2014 NUFFIC data: Spearman Correlation = 0.70, n=74 (on institution level)

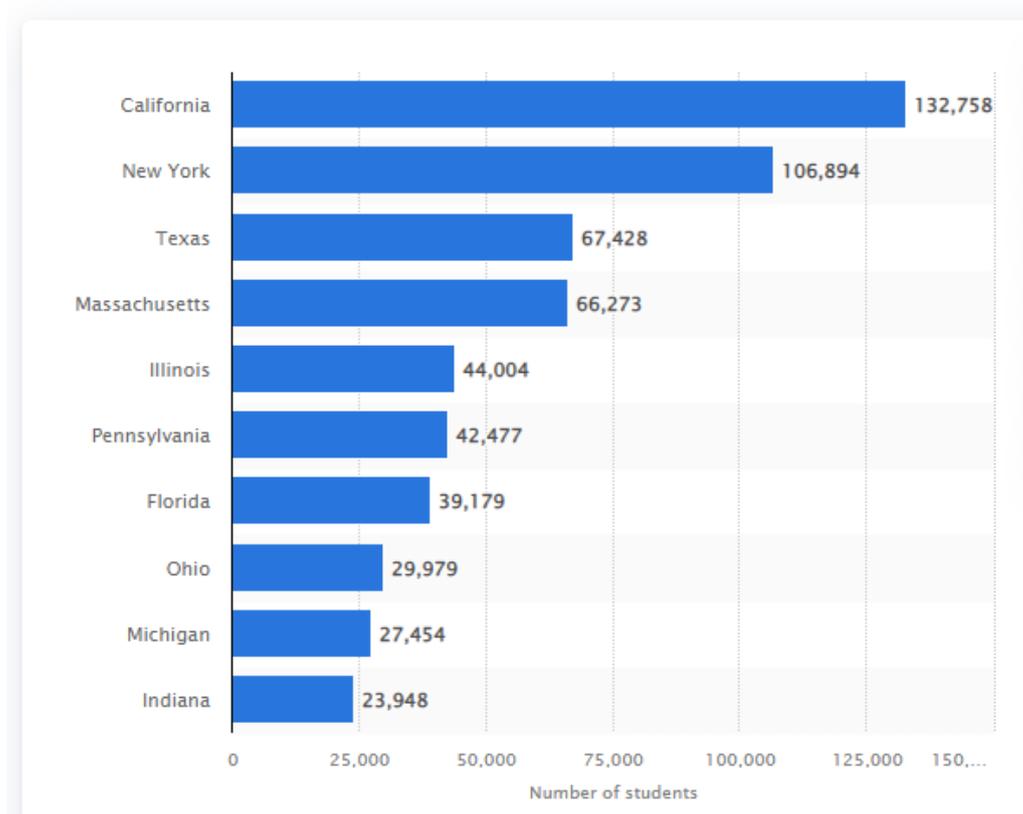
Robert L. 2015 NUFFIC data: Spearman Correlation = 0.855, n=? (on country level)

Roy K.B. 2017/18 HESA data: Spearman Correlation = 0.82, n=120 (on discipline level)

Sophie L. 2018/19 HESA data: Spearman Correlation = 0.81, n=123 (on discipline level)

Robert-Jan B. 2019/20 HESA data: Spearman Correlation = 0.92, n=181 (on country level)

Appendix A – U.S. States with the Most International Students in the Academic Year 2020/21 (US Census Bureau)



Appendix B – Sources of Enrollment Data for Every State in the Study

Sources of Enrollment Date by State

State	Source of enrollment data
New York	http://www.nysed.gov/information-reporting-services/higher-education-reports
Ohio	https://www.ohiohighered.org/data-reports/enrollment
Pennsylvania	https://www.pashe.edu/SystemData/System%20Data%20Documents/Fall%20Enrollment%20Census%202021_Preliminary.pdf
Massachusetts	https://www.mass.edu/datacenter/2021enrollmentestimates.asp
Texas	http://www.txhighereddata.org/index.cfm?objectid=58428080-32B7-11EC-85120050560100A9

studyportals

Appendix C – Checking Normality for Page-views and Enrollments per State

Are page views and enrollments normally distributed?

State	SW test-statistic enrollments	SW test-statistic page views	Normally distributed?
New York	0.555	0.435	No
Ohio	0.919	0.778	No
Pennsylvania	0.887	0.874	No
Massachusetts	0.733	0.646	No
Texas	0.811	0.688	No

studyportals