

# DS 3000 – Dataset Group 11 (Sophia, Cheryl, Jacob)

## Project Topic Idea: Cost of Higher Education in the U.S.

### 1. Problem Statement

- ✓ The problem we want to tackle for this project is the high costs of higher education in America while college degrees are needed for a lot of opportunities .
- ✓ The topic of the project is how unreasonable the costs of college tuition in the United States.
- ✓ We would like to learn about the average costs, debt, and demand for college degrees

### 2. Significance of the Problem

- ✓ I believe it's important to tackle this problem because it's relevant to us as students along with the whole country as a system. Since college tuition is so expensive without financial aid, student loans, or scholarships, unreasonably priced.
- ✓ I think some insights on this project would be the true value of a college education, rethinking the system in which we need a degree to succeed, along with how other countries offer free education and to rethink ours.
- ✓ <https://educationdata.org/average-cost-of-college>
- ✓ <https://www.ed.gov/college>
- ✓ <https://edsources.org/2020/tuition-free-college-is-critical-to-our-economy/641232>
- ✓ <https://www.pewresearch.org/social-trends/2011/08/17/i-the-value-usefulness-and-cost-of-college/>

### 3. Dataset(s)

- ✓ This sample data about college tuition was obtained from kaggle.com.  
<https://www.kaggle.com/jessemostipak/college-tuition-diversity-and-pay>
- ✓ This data about student finances was found on Data.ed.gov from the Department of Education.  
<https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020/resources>
- ✓ This sample data about student finances was found on kaggle.com.  
<https://www.kaggle.com/kaggle/college-scorecard/code>

## Dataset File

Variable name in file	Description (what the variable represents/means) <sup>1</sup>	Feature/Outcome <sup>2</sup>
name	the name of the college	outcome
state	the state the college resides in	feature
state_code	the abbreviation for the state	feature
type	a college can be private, public, or for profit	feature
degree_length	how long on average that a student will be at this college	feature
in_state_tuition	the price a student will pay if they reside in the state	feature
in_state_total	in state tuition and room and board	feature
out_of_state_tuition	the price a student will pay if they do not reside in the state	feature
out_of_state_total	out of state tuition and room and board	feature
total_enrollment	total number of students	feature
category	group/racial/gender category	feature
enrollment	enrollment by each category	feature
<sup>1</sup> Refer to dataset descriptions in sklearn <sup>2</sup> In the Feature/Outcome column, indicate whether the variable is a feature or outcome variable. You need to have at least one outcome variable and nine feature variables.		

Our data set includes the name of the college as an outcome variable and the rest of the variables as feature variables. Given some of the feature variables such as out of state tuition, state, and total enrollment, we could have an algorithm try to guess the name of school. For example, we could use a decision tree algorithm and have questions which relate to the type of college, total enrollment, etc. We could also use k nearest neighbors to group colleges by tuition price or enrollment. The name variable could technically also be a feature and out of state tuition could be an outcome variable. We still haven't really decided which variable we definitively want to be our outcome variable, but we feel that all of the variables that we have so far are good measures used to describe a college.