***Chosen DataSet: College Affordability (Dataset C)***

**What interests you about this data?**

*What interests us about the data is the diverse amount of variables that would allow us to create really unique research questions. We can form diverse correlations between variables and create hypotheses that would allow for specific and answerable questions that are required. Another one of the reasons this data set interests us is that currently all of us are enrolled in college and being able to see a full scope of this data for many different institutions is extremely relevant to us financially.*

**What research questions and data points would be relevant for each model?:**

* *Linear Regression Research Question: Is there a relationship between a schools’ private status and their acceptance rate? The relevant data points would be the private status and the acceptance rate for the school. Additionally, looking at a school’s number of applications could be relevant as well.*
* *Logistic Regression Research Question: Does an institution being public or private affect the number of students enrolled at the institution? The relevant data points would be the binary classification of private university and enrollment being the probabilistic output.*
* *Decision Trees Research Question: Do schools with more individual student support have a higher graduation rate than those with less student support? Relevant data points would include graduation rate, percent of professors with Ph. D and terminal degrees, student/faculty ratio, alumni donations, and school costs.*

**What would we need to conceptualize and operationalize for each model?:**

* *Linear Regression: We would need to define the difference between private and public universities and the linear regression model itself. We would need to operationalize how we put together our linear regression model, and how we interpret our data.*
* *Logic Regression: We would need to define private and public status and what exactly student enrollment is measuring. Additionally, it would help to define logistical regression, as this model is not as common as linear regression and other models. We would need to operationalize how we put together the model, how we interpret our data, and how we decide to display our data.*
* *Decision Trees: We would need to define what the different degrees are (Ph. D and Terminal), how the graduation rate is calculated, and how we define student support. We would need to operationalize how we gather data for student support, and how we organize and create our decision tree and how we can reach our end goal (graduation rate).*

**Which one of the research questions does your group want to pursue?**

* **How does student support from their universities affect the graduation rate?**

***Find two peer-reviewed sources that can match with the project.***

***Measurement 11/12***

Research Question: How does a university’s student support affect their graduation rate among undergraduate students in the US?

Model: Linear Regression

* Not a binary decision, so this model helps show the many different outcomes
* Logically, this relationship should be linear, so using this model would be the most efficient for answering the question.

***Terms to Conceptualize:***

* ***Student Support***
* ***Graduation rate***
* ***Undergraduate Students***
* ***University***

***Operationalizing Terms:***

* ***Student support - Look at financial support (scholarships, FAFSA funding, alum donations), career benefits (career fairs, networking opportunities, etc), and mental support (mental wellness support, student activities, student groups).***
* ***Graduation rate - finds ways universities define this, are transfer students included? If a student drops out, are they included in the calculation?***
* ***Undergraduate students - define the difference between undergraduate and graduate students/work.***
* ***University - four-year college.***