John Jacobsen

A12647294

Professor Fleischer

COGS 9

Final Project

Abstract.

In this project I will be analyzing the data given from UCSD's course evaluation site CAPE[1], converting it into a csv file, and analyzing it using the pandas library. I will be using this data to compare the mean of the mean study hours per week of the courses from departments I have chosen to represent STEM (Science, Technology, Engineering, and Mathematics) and Liberal Arts majors, and use that variable as a measure of how difficult STEM and Liberal Arts majors are.

I will compare the mean study hours of the following STEM departments: Math (MATH), Physics (PHYS), Biology (BIOL), Mechanical and Aerospace Engineering (MAE), Chemistry (CHEM), Computer Science and Engineering (CSE), and Cognitive Science (COGS); and the following Liberal Arts departments: Economics (ECON), Psychology (PSYC), Visual Arts (VIS), Political Science (POLI), History (HIST), Literature (LIT), and Anthropology (ANTH).

I think the results of my analysis will be interesting for college students because they will have justification that their major has similar difficulty to others or that their major is more

[1] Course and Professor Evaluations (CAPE), http://www.cape.ucsd.edu/responses/Results.aspx?Name=&CourseNumber= (Accessed June 12th 2017).

difficult than others. My hypothesis for this analysis will be the null hypothesis that there are no significant differences between the mean study hours of STEM and Liberal Arts majors.

Introduction.

One of the first things college students ask each other to break the ice is "What major are you?" and the response to this question gives a great view of the dreams and interests of the respondent. However there is also a little bit of elitism about what major you are. Liberal arts majors have no doubt experienced this from snickering STEM majors that believe Liberal arts majors are getting an easy ride through college. So what, if any, majors are the most difficult? Since the difficulty of a major is hard to quantify, some people like "The Best Colleges"[2] quantify difficulty by the average gpa in the major. Through their analysis they have determined that Education majors have it the easiest and that Engineering majors have it the hardest. However measuring difficulty by GPA only measures how easy it is to have points deducted from your grade, it is possible that majors with high GPAs also have to do a lot of work to get that GPA. Luckily data from UCSD's CAPE gives us another variable to determine the difficulty of a course, mean study hours per week.

In this project I will be exploring the CAPE dataset to see if there are any statistically significant differences in the mean study hours per week of the different departments at UCSD. From the dataset I will create two groups of departments and compare them, Liberal Arts departments (Psychology, Economics, Anthropology, History, Literature, Political Science, and Visual Arts) and STEM departments (Biology, Chemistry, Cognitive Science, Computer

[2] The Best Colleges, Top 10 Easiest and Hardest College Degree Majors of 2017, http://www.thebestcolleges.org/top-10-easiest-and-hardest-college-degree-majors/ (Accessed June 12th 2017).

Science and Engineering, Mechanical and Aerospace Engineering, Mathematics, and Physics). The best hypothesis for my exploration would be the null hypothesis that there is no statistically significant difference between the two departments in the mean study hours per week since it will be easier to disprove this hypothesis than to prove that there is a difference. I predict that there will be a statistically significant difference between the two groups of departments and that the null hypothesis will be disproven. This might be true because of reasons such as the Liberal Arts classes being taken primarily as general education requirements thus being made easier and that many STEM classes are being taken by overstressed pre-med students that are trying to get a near perfect gpa for med school and engineers taking 20+ units per quarter.

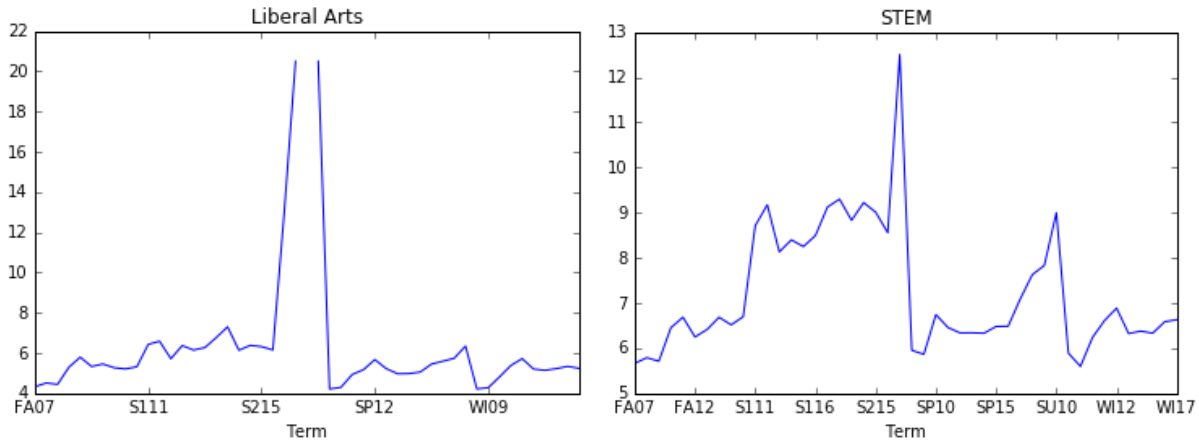
Data and Methods.

The dataset that I will be using are the course reviews on the website "cape.ucsd.edu", where there is data for every course at UCSD since summer quarter 2007. The data is gotten through surveys of students that are currently taking that course and they are asked questions such as "Would you recommend the class?", "Would you recommend the instructor?", "How many hours per week do you spend on this class?", and "What grade do you expect to earn in this course?". The survey also records data such as the Instructor's name, the Course title, the term, the amount of students enrolled in a course, the amount of respondents to the survey, and the average grade earned in the class. The data is already cleaned and put in a table format so there is not much work to do other than saving the data as a csv file for each department. When the people at CAPE were designing the surveys they made it really easy to clean the data. They

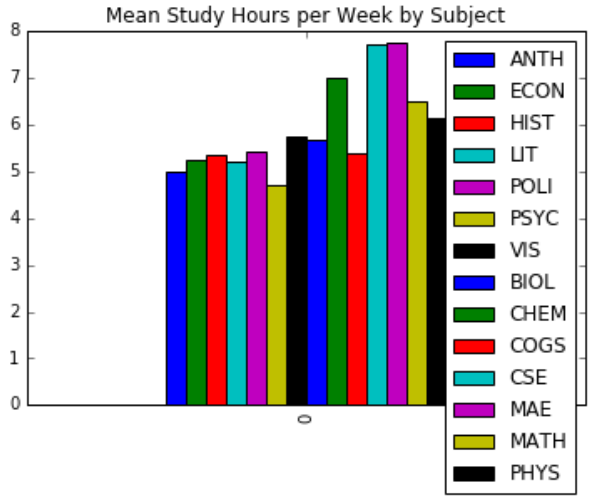
only let people respond in the form of multiple choice so there are no strange responses they have to worry about erasing.

Most of the data we get from CAPE is real-valued, however there are three categorical values: the course title, the term, and the instructor name. The data we get is from a cross-sectional observational study conducted by the people at CAPE where they survey the people in all courses being offered at UCSD. The population they seek to measure are the students enrolled in each class, however since this is an optional survey they get a low response rate for many of the classes. This low response rate for the CAPE surveys could cause trouble for our analysis since the people that bother to respond to the CAPE survey are likely people with a grudge against the course or people that really enjoyed the course, so we might not be getting a good sample of what the average student thinks about their course. In addition since these surveys are sent out during the last two weeks of the quarter, the amount of hours people say they spend on a class might be higher with people cramming for their finals and projects. To try to avoid these problems we will make the assumption that a response rate of 80 percent or higher for a course is a good enough sample of the student body's feelings about the courses that they are taking and we will assume that the timing of the survey will have no effect on any differences between major difficulty.

We can check the raw data by creating a plot of the mean study hours through the different terms for both STEM and Liberal Arts departments. However, since I will be looking at the "mean study hours per week" variable for my analysis, I can also create a visualization of the mean for each individual department using a bar plot. The bar plot will allow us to see if there are any noticeable differences in the difficulty between the different departments.



We can see that in both the Liberal Arts and STEM departments there is a spike in the mean study hours around spring quarter in the late 2000s. We should look at the data to determine why there is a spike there and whether or not there is a mistake there that needs cleaning. The reason for these outliers might be a low response rate, so before I do my analysis I will clean out courses with a less than 80% response rate.

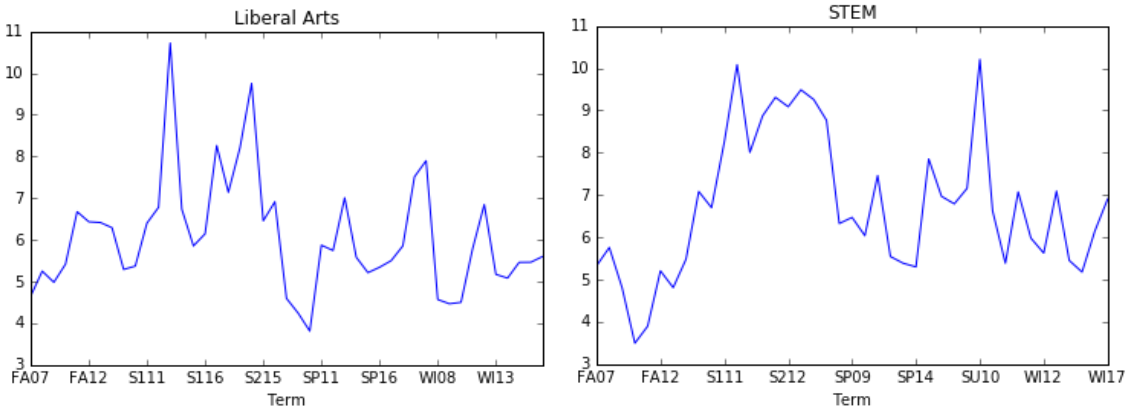


Here we have a visualization of the raw data where we can see the mean study hours per week of each department. For the most part the departments have really similar means except for CSE, CHEM, and MAE. These departments are much higher than the other departments, so I would

like to do a comparison of STEM without these departments and Liberal Arts as well so that we can see whether or not these departments are the main reason for any difference between STEM and Liberal Arts.

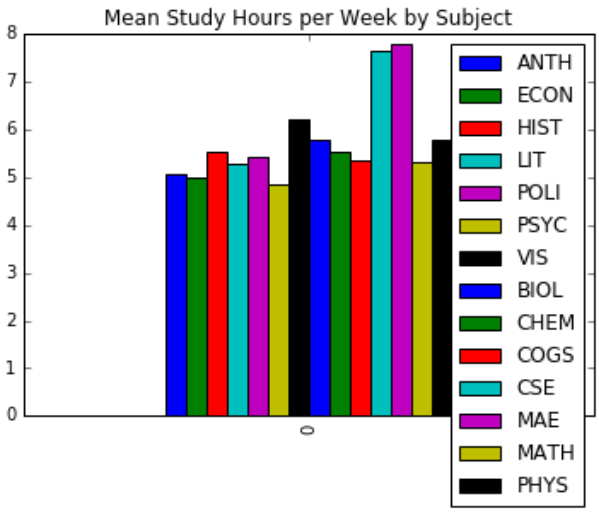
Results.

I will download the html file that contains all course reviews of a given department in a table format. Luckily most of the CAPE data is pretty clean and easy to work with so I will open the html file in Microsoft Excel 2016 to crop out the excess parts of the html file and leave the data as a table to be saved as a csv file. I will check the csv files to make sure none of the data has been ruined. Then the csv files for the anthropology, biology, chemistry, cognitive science, computer science, economics, history, literature, mechanical engineering, math, physics, political science, psychology, and visual arts departments will all be uploaded to JupyterHub to be used. I will convert the csv files into Data Frames using the pandas library and then I will concatenate them into Liberal Arts departments and STEM departments according to how I grouped them earlier. Then I will clean out all entries with a less than 80 percent response rate.



Here is a visualization of the data after being cleaned

We can see that the plot has been smoothed out in comparison to before

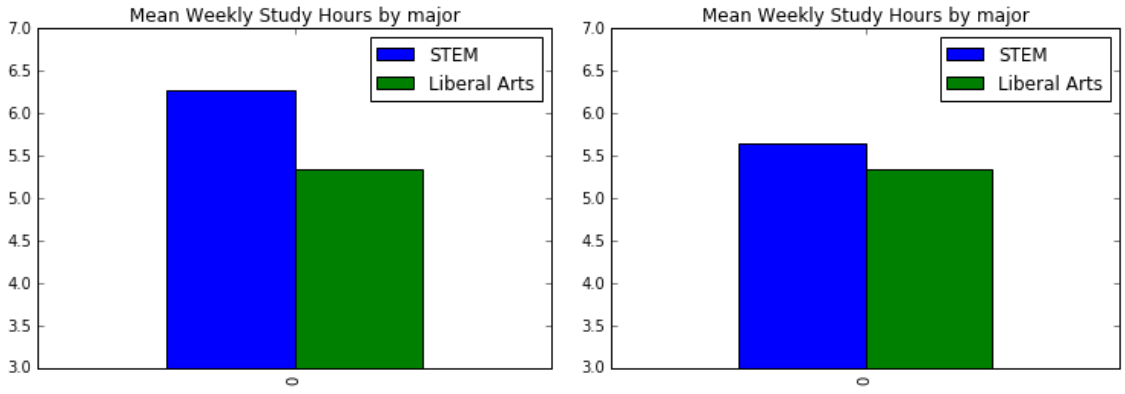


We can also see that after the data was cleaned CHEM became similar to the other departments but CSE and MAE still seemingly have higher study hours

For the analysis I will calculate the mean and standard deviation of both groups and then calculate the p-value for the null hypothesis that the two means should be equal. If the p-value is less than or equal to 0.05 then there is a statistically significant difference between the difficulty of STEM and Liberal Arts. If not, then there is no statistically significant difference between the two groups of majors. Since I also observed that there might be some departments in STEM pulling the mean up, I will do the same test without CSE and MAE in the STEM category. If both tests prove there is a significant difference between the two groups then it would show that either STEM or Liberal Arts is the more difficult major, depending on what group has a higher mean. If only the first test shows a significant difference then it was CSE and MAE that were creating a significant difference between the two groups.

Discussion.

After the analysis I have found that there is a significant difference between STEM and Liberal Arts regardless of whether or not CSE and MAE are included in STEM. The mean study hours per week that STEM majors have when CSE and MAE are included is 6.26 hours per week of studying. Without CSE and MAE the mean study hours per week is 5.63 hours per week. We can compare this to the mean study hours per week of Liberal Arts which is 5.34 hours per week of studying. So we can conclude that STEM majors on average spend about an hour extra studying than Liberal Arts, which is not as much as one might expect if STEM is really that much harder than the Liberal Arts. We can also see that CSE and MAE have much higher reported study hours per week which is shown by how much the mean of the STEM majors is affected by them.



Mean study hours with and without CSE and MAE in STEM

However I only used data with a minimum of a 80 percent response rate, so by not having a higher response rate requirement for my data there is a chance that there is a volunteer's bias in my data. I can repeat my analysis with higher response rate requirements to see if there is still a

similar comparison between the two groups of majors, though this might affect the results since it will reduce the sample size.