# SI 330 Final Project Report:

## Motivation:

My project will be investigating the gender, racial and socioeconomic bias in post-undergraduate outcomes of universities. I chose it due to my interest in higher education and political intersectionality. Many of us spend hundreds and thousands of dollars on college tuition because we think we need it to survive in this economy. Despite pursuing the same form of education, are specific groups of people still destined to pick the short straw? If my future salary were to be compared with a white man pursuing the same exact college experience as me, will he win?

My goal is to figure out if being a woman and/or a white person in university correlates with less salary throughout your career. I will use  a couple different datasets to figure out this question, combining a dataset that has information about gender ratios and undergraduate information in a list of universities and a dataset that has information about salaries post-university to create a merged source that will help reach my goal.

## Data Sources:

All datasets that I have sourced for this project are primarily sourced from Kaggle. They all returned as a CSV file just to minimize the amount of risk that can go wrong when opening, cleaning and merging datasets of different file formats. I downloaded it straight off of Kaggle and opened the file locally via Jupyter Lab.

**File Source 1**: College Admissions Dataset

- Kaggle: https://www.kaggle.com/datasets/samsonqian/college-admissions
- 791.39 kB
- 1517 unique values; 108 columns
- CSV File
- Important Variables: School Name, Admissions Rate, Undergraduate Enrollment, Tuition, % of Women Enrolled, % of White people Enrolled, Degree of Urbanization (City, Suburb, Farm etc.)
- NO time periods

File Source 2: Colleges by Salary Dataset

- Wall Street Journal - Kaggle: https://www.kaggle.com/datasets/wsj/college-salaries?select=salaries-by-college-type.csv
- 31.43kB
- 249 unique values; 8 columns
- CSV File
- Important Variables: School Name, School Type, Starting Median Salary, Mid-Career Median Salary, Mid-Career 25th Percentile Salary, Mid-Career 75th Percentile Salary
- NO time periods

## Data Manipulation:

**Cleaning the 1st Dataset**:

- First, I tackled the biggest dataset in my project by selecting the columns I wanted and concatenating it into 1 dataset to use instead of dropping the unwanted columns.
- Second, the last row in the database contained unnecessary details about the column titles so I dropped that row because It does not contain information I would need.

Initially, I wanted to merge the database based on School Name. This means that I would need to thoroughly clean the database to ensure every school in both databases are written the same way for the same school.

- Third, I started cleaning the column by putting everything in UPPERCASE for better feasibility. Then, I dropped duplicate school names, keeping the first instance, because I noticed there were a bunch of duplicate school names that I have no idea how to distinguish. This is mostly because state schools have different campuses and are usually named the same. If I cannot manually figure out the different locations based on admissions number, it's better off keeping the first instance since it's usually the flagship presence.

**Cleaning the 2nd Dataset**:

- First, I dropped the unnecessary columns of the 2nd database since the total column number was more manageable than the previous dataset.
- Second, I renamed the columns to match with the 1st dataset then dropped any rows with duplicate school name values. I also made all the School Names into all uppercase.
- Third, I tried to clean the "Names" database by making it more similar to the 1st database. This includes deleting any brackets including a school acronym [ie. Arizona State University (ASU)] through a regex expression.
- Since my columns containing salary numbers are written in $ form (ie. $243,000) I created a function looping each value in a specific list of columns by replacing each $ and comma into a blank space and then converting that value into a numeric float64 type.

**WHAT WENT WRONG:** While looking at the names through different sample runs, I started being overwhelmed over how to make the names similar because they format different campuses so much differently. Also, I noticed that some schools are formatted with different names even though they are the same institution. Some examples may include "Wentworth Institute of Technology" in the 1st database while the name is displayed as "Wentworth Technological Institute." A reason for this might be due to a college changing their name earlier and the dataset not catching that name change. This made it extremely difficult for me to clean the "Names" column using properties like Regex or NFLTK.

**HOW I SOLVED IT**: As per Andre and Harrison's advice over Project Office Hours, I ended up cleaning the databases by creating an additional column in both databases that has the ACT

College Code of a particular school. I had to manually input this information through Excel for both databases, going through all 300 rows of the 1st Database, cross-referencing it with the availability of the 2nd Database (1517 rows) and then inputting the available ACT Code with that column. Afterwards, I double checked to see if I input any values wrong by looking at duplicate values through Excel and fixing it before importing it back into CSV.format to merge.

**Merging the 2 Databases**:

- After inputting the ACT codes for all common college names in both databases, I imported them back into Jupyter Lab so they can merge based on the ACT Code column.
- At first, I was having trouble correctly reading the modified file. This problem was solved when I realized that my laptop uses British spelling so their default CSV file has a semicolon delimiter. Once I modified the read_csv appropriately, it worked.
- I dropped the duplicate Name_y value from the 2nd database, and renamed the Name_x back into "Name".
- Since the CSV file reads all the numbers in my dataset as str objects (aside from the salary columns since it's already converted), I had to forcibly convert my columns into float64 type so it's easier for calculations.
- Then, I further cleaned it by dropping any rows where any of the column values include NaN so we can analyze the colleges that have all of the specific information I wanted, even though my pool would become a little bit smaller.

**RESULT**: Now, I have a database that includes information about admissions numbers, gender, racial and socioeconomic undergraduate enrollment numbers, starting and mid-career salaries that I can use to string together correlations answering the goal to my project.

## Analysis and Visualization:

**Questions #1 and #2:**

- Do women make less money coming out of college?
- Do white people make more money coming out of college?

**[#1 and #2] Answering the Question**:

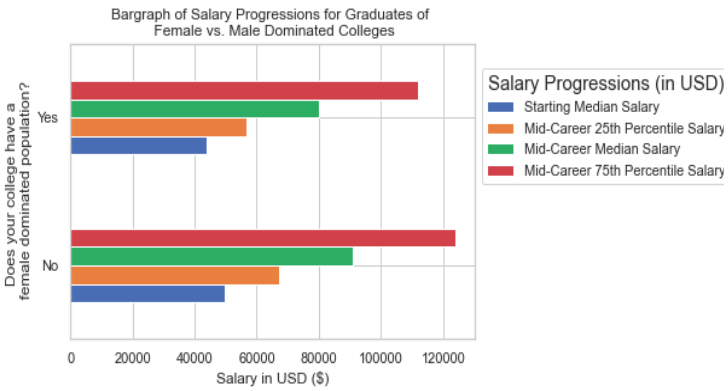
The steps I used to analyze and answer the two questions are the same.

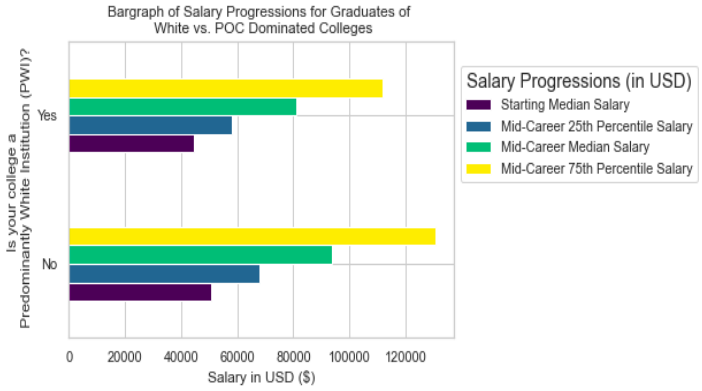
- First, I created an additional column labeled ['Mostly Women'] that separates different schools into 2 separate categories if they have >50% or <=50% women undergraduate enrollment in their university.
- Then I applied that function to the column ['Percent of Undergrad Enrollment that are Women'], separating each school into either a Yes or No if they're "mostly women."
- Afterwards, I sorted the different salary progressions, [Starting Median Salary, Mid-Career Median Salary, Mid-Career 25th Percentile Salary, and Mid-Career 75th Percentile Salary] based on the designated category they have in their ['Mostly Women'] column and found the average values.
- Then, I conducted an independent t-test with this documentation: https://www.marsja.se/how-to-perform-a-two-sample-t-test-with-python-3-different-methods/ against all salary columns to calculate the statistical significance of the difference in

two means (ie. are starting salaries of female dominated colleges really that different from the salaries of male dominated colleges?)

**[#1 and #2] Result and Interesting Analysis**:
Graduates of female dominated colleges earn a statistically significantly lower starting and mid-career salary than graduates of male dominated colleges. Seeing this result confirmed my hypothesis but nevertheless disappointed me. This only shows that gender bias is evident in the job industry on a very significant basis. What's interesting though, is that the value of "Mostly Women" colleges are higher than its counterpart. This signifies the movement of more women getting into college, while men are going straight into work. Personally, this is something that I didn't know until I did further outside research regarding said value_counts result.

Graduates of PWI earn a statistically significantly lower starting and mid-career salary than graduates of non-PWI. This result definitely went against my hypothesis, with the added surprise that there are more colleges having non-PWI status than their counterpart in the database. One interesting analysis could be that PWIs in the database include smaller state colleges from lower-earning states, bringing down the average median salaries of colleges who belong to that category (ie. Gustavus Adolpho College, St. Olaf College).




**Question #3:**

- Is the role of gender/racial intersectionality important to post-undergraduate salary?

**[#3] Answering the Question:**
The steps I used to analyze and answer the two questions are the same as #1 and #2 BUT:

- I created a function that checks against 2 specific columns within my dataset, separating each college into 4 types that includes different status of gender and racial dominated undergraduate enrollment bodies. Then, I used the .apply function on the entire database instead of a specific column like previous, naming said column ['Women/White].
- I generated an ANOVA test using this documentation: https://www.reneshbedre.com/blog/anova.html to test if differences are significant.
- I used the cmap property to all 3 bar graphs (from Questions #1 and #2 also) to create different color gradients and help distinguish between questions. I also moved the legend box outside of the graph to the center right of each graph using loc and bbox_to_anchor property.

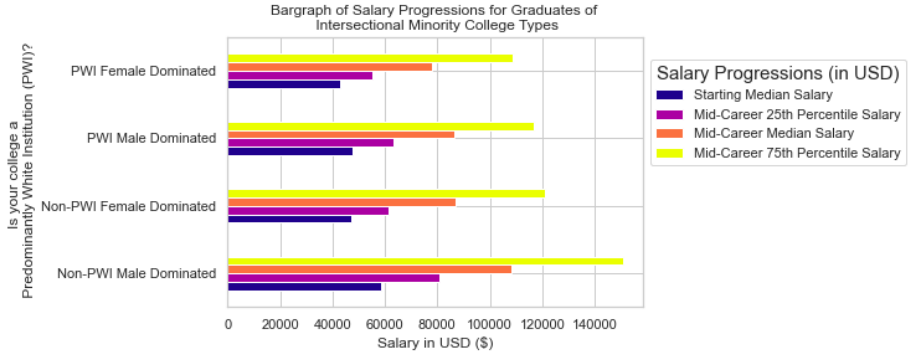
**What were some struggles encountering this question?**

My biggest struggle includes thinking up a graph for the question. Since the question needs data from 2 different categorical variables and their corresponding qualitative assessments, I wanted a graph that has 'Mostly Women' in 1 axes, 'Mostly White' in another axes, and have the results mapped out in quantitative salaries. This would make visualization of data easier to read since I don't want to risk people being confused by the writing of my 4 categories.

Solution: I decided to map out the categories in a bar graph from the previous question after trying out several ideas like a crossbar, categorical scatterplot, distribution plot etc. I mapped it as a horizontal bar graph so you can see the differences more. I tried to highlight the differences more clearly by detailing the .annotate function to resemble a bracket but unfortunately I wasn't able to figure that out for multiple difference coordinates.

**[#3] Result and Interesting Analysis**:
Interestingly, graduates from the non-PWI Male Dominated colleges earn the most salary throughout their career. Non-PWI Female Dominated earns the second highest salary, PWI Male Dominated college graduates follow, and PWI Female Dominated college graduates earn the lowest. The ANOVA test shows a p-value 0.00 which means that the difference between means is statistically significant. Also coincidentally, there are a bigger number of PWIs than non-PWIs.

Non-PWIs tend to have higher ranking programs because of an increased importance of diversity in the image of education while a handful of the PWIs might be from smaller colleges in lower earning states. Another thing to notice is that PWI Male Dominated college graduates initially earn the 2nd highest starting salary, but went down to 3rd place when looking at Mid-Career Salary results. These results are super interesting to me because I would think that the Male dominated PWIs would earn the most but maybe if we gained a more equal number of colleges across the same 4 types we could have reached a more credible solution.




**Question #4:**

- How important is socioeconomic class in post-undergraduate salary?

**[#4] Answering the Question:**
The steps I used to answer this question begins with:

- Grouping average values of different Salary progressions (Starting vs. Mid-Career) against the different School Types through .groupby, sorting values in ascending order for quicker analysis. This shows which type of college earns the most.
- Grouping the average values of freshmen who receive any form of financial aid against the different School Types. This shows which type of colleges admit students who need financial assistance.
- Plotting a line graph with different y axes values on either side, allowing me to map the Starting and Mid-Career Salary data against the Financial Aid data and let viewers see the contrast.
- I used subplots to allow for multiple y axes, mapping 2 twin graphs that records different Starting & Mid-Career salaries for each School Type alongside the original graph that records % of financial aid students with each School Type.
- Using .twinx stacks the graphs on top of each other, allowing for 3 different lines with different axes ranges to coexist. To highlight it, they are labeled with different colors.

**What were some of the struggles?**

My biggest struggle was to figure out how to fit all 3 axes together. Mid-Career Salary will have a higher amount base than Starting because of the time progression, while % of financial aid students have double digits while salaries are usually at a 6 digit range. Thankfully, I found this documentation:https://matplotlib.org/stable/gallery/spines/multiple_yaxis_with_spines.html#sphx-glr-gallery-spines-multiple-yaxis-with-spines-py to help me create the final line graph.

**[#4] Result and Interesting Analysis**:

As expected, Ivy League college graduates earn more salary throughout their career. Engineering and Liberal Arts comes second then third highest, then Party, then State having lowest average median Starting and Mid-Career Salary. I was pretty surprised about the order of the last 2, but my guess is that Party schools are generally categorized to be larger state/private schools while State colleges include those smaller less known or ranked universities that can contribute to the lower earning value. Interestingly enough, Ivy League colleges have the least amount of freshmen certifying for any financial aid while Engineering takes in the most % of freshmen with financial aid.

This highlights very important things about higher education. First, the income loop that wealthy families have to easily accumulate wealth since richer people can afford more resources like tutoring, test prep etc. to gain better stats and get admitted to better ranked schools. Second, the data behind Engineering type schools show society's push for STEM. Higher education allocating more funds to admit low-income students if they are pursuing STEM can be a way to get out of the poverty cycle for most students.