# STA 141A Project Proposal

Thy Pham (thypham@ucdavis.edu), Andrew Chen (ayhchen@ucdavis.edu), Tyler Lee (Tyylee@ucdavis.edu), Chaewon Park (cwpark@ucdavis.edu), Christina Deng (ccdeng@ucdavis.edu)

**Expected Contributions**: Tyler Lee (yellow), Thy Pham (purple), Christina Deng (purple), Chaewon Park (blue), Andrew Chen (blue)

Written Report

1. Executive summary
2. Project description, background of problem, overall goal(s)
3. Describe datasets, datasources, using words, summary statistics and 2d scatter plots with pairs() or ggpairs()
4. Describe your methodology and why you chose these methods (for each question)
5. Describe your analysis and findings (for each question)
6. Conclusion

R Code and Analysis

1. Exploratory Data analysis: Summary statistics
2. Linear regression for question 2
3. ANOVA test for question 1

**Brief Description of Project**: To explore real data science experience by posting questions, finding data sources, exploring/visualizing data, and deploying/validating statistical models, we chose the Education & Career Success dataset on Kaggle to work on it. Using this dataset, we will focus on investigating the relationship between educational background, personal attributes, and starting salary. Specifically, we will explore whether there are differences in average starting salary based on gender and field of study, and examine how the factors(student's university ranking, university GPA, networking score, and soft skill score) affect/enable us to predict the starting salary. Through data analysis/statistical modeling, we will uncover the potential relationship between personal background variables and career outcomes to see how personal/academic factors shape early professional success.

**Overall Goal:**

Our project aims to provide both statistical and social insights. Statistically, we will demonstrate the use of regression analysis, ANOVA testing, and model selection to draw insightful conclusions based on the educational dataset from Kaggle. Socially, our findings may help educators and policy makers understand trends across demographics. And these insights can further help promote career readiness, particularly for underrepresented groups.

**Brief description of dataset:** The dataset we are using for this project, titled "Education & Career Success," is from Kaggle, where the values of this dataset were generated using observed trends of the real-world rather than sampling and collecting information from a population. This data contains information of 5000 students and 20 variables pertaining to various measures of academic and professional success as well as the educational background of the students. There are 4 categorical variables such as gender, field of study, current job level, and whether or not they were an entrepreneur, with the remaining 16 being numerical variables. This dataset will be used to address our driving questions of:

1. *Is there a difference in average starting salary based on gender and field of study?*
2. *Does a student's university ranking, university GPA, networking score and soft skill score accurately predict their starting salary?*

**Background of Problem**: We see a lot of variance in the starting salary of an undergraduate, but how much of that is due to industry demands and how much of that is due to career readiness and educational background. This is what we aim to get more insight on. Determining if there is a difference in starting salary between genders, and or industries can help with policy making. For example, if we see there is a large difference in means for a certain gender, we can help promote diversity, and equity, and include programs to help promote equity in the workforce. We will also analyze the variables that correlate with career salary, such as university prestige and academic performance. Through data visualization, analytic summary, and statistical modeling, we are aiming not only to find out the underlying relationship between the biggest factors on salary but also to provide insight to promote career equity. The information that we gain from this project can be used to design targeted programs and experiences that expand access to educational and professional opportunities, particularly for underrepresented or disadvantaged groups.

**Methodologies (what and why)**

We plan to use multiple linear regression, possibly considering interactions in the model as well, to determine which factors are relevant to predicting the starting salary. To discern the necessary factors from all the options and reach a final model, we will follow the AIC model criteria which is good for a prediction model.

To test if there is a difference in average starting salary based on gender and field of study, we will use the ANOVA test to see if mean starting salaries are different across groups; we may also consider the interaction effect between gender and field of study.

I expect these methodologies to be useful because I believe the variables will have a linear pattern (e.g. better university ranking or better university GPA will result in higher starting salary) which will make multiple linear regression useful for predicting salary. I also think the ANOVA test will help us determine if there are truly different starting salaries between different gender and field of study groups because the ANOVA test checks if the means between groups are statistically different.

Data Source:
Education and Career Readiness
https://www.kaggle.com/datasets/adilshamim8/education-and-career-success