Graduate Student Analysis and Visualization

Group 11 : Yoontae Park, Yoobin Cheong, Ilias Arvanitakis, Joseph Schuman

December 6, 2021

1. Abstract

The main goal of this project is to shed light on the key characteristics that admitted Masters students at the CDS possess. Extensive data analysis methods and visualization tools will be utilized, producing a report that will be useful for the faculty during the admissions process. Of equal importance are the career paths that students follow during their summer internships and after their graduation. We are aiming to delve deeper in this topic by analyzing the available career data and presenting our results in a way that future CDS Masters applicants can discern the comparative advantage of the program compared to alternative ones.

2. Introduction of the dataset

Communicating with the administration of the CDS and after the approval of the department we got access to individual student data, as well as the internship and career outcomes. The data were provided in such a way that anonymity was ensured, by removing information such as the names of the students and their date of birth. Because the country of origin is another factor that could undermine the anonymity of the data, the dataset would only mention the US, China, India, which have the highest representation in the student population, whereas the remaining students are classified as 'Other'. Given sensitivities regarding gender and ethnicity, our team did not think it was appropriate to use this dataset to compare academic performance of students based on these factors. Comparisons of this nature can easily be taken out of context and interpreted in ways that may perpetuate gender or racial bias. As a result we excluded this kind of analysis from our findings.

The dataset includes qualitative and quantitative information for each student. On the qualitative side the dataset includes information about each student's country of origin; their US citizenship status (i.e., US citizens, US permanent residents, international); and their gender. On the quantitative side, the dataset includes undergraduate GPA and GRE scores. The GPA is usually on a scale of 4.0. For the GRE, we have each student's scores for the Quantitative, Verbal, and Writing as well as their corresponding percentiles. Additionally, TOEFL scores are included where relevant.

Our key objective for the data analysis is to get some metrics about the dispersion of the GRE and GPA scores for the admitted CDS masters students. We also deem that it would be highly interesting to dig into the correlations between the GPA, GRE and TOEFL scores.

Finally, we will also make visualizations, utilizing all the above data, trying to make them easy to follow and understood by the potential reader of the report. For the part of the data that is focused on the career outcomes, our capabilities for data analysis are narrower. We have data for the summer internships of the past three years. These data come from DS-GA 1009: Practical training course and contain information about the company, the location and the job title of the internship. We have a similar dataset over the past 3 years for the career outcomes after graduation. These data come from the exit survey that each student completes after graduation. Our main goal with these data is to create visualizations that will be easy for the reader to follow and clearly portray the outstanding career paths that students follow both as interns and as graduates and also point out the professional advantage that CDS students have compared to other programs. Although not taught in the lectures, we came to the conclusion that we will also use the Tableau as a complement to the visualizations that python produces. Ultimately, we hope that

our report will help the department with the admission process and also with the improvement of the website by integrating the visualized information.

3. Data preprocessing

3.1 Current Student Profile

As we have already discussed, the scores of GRE, GPA, and TOEFL were provided by the CDS, for three consecutive school years. They include the classes of 2019-2020, 2020-2021 and 2021-2022. For each of the above years we have 284, 272, and 308 observations respectively with 11 columns including Legal Sex, US Citizenship Status, Foreign Citizenship, Undergraduate GPA, TOEFL, GRE Quantitative, Verbal and Writing scores, each with their corresponding percentile. With the scores provided in separate excel sheets per academic year, each individual sheet needs to be read separately and combined as a dataframe to make the analysis. For loading the files and computing basic statistics, Numpy and Pandas packages were used. For plotting and visualizations, Matplotlib.pyplot and Seaborn packages were also employed. For some of the aforementioned columns there were either missing values or values that were in the wrong scale. Below we will delve into more detail about the preprocessing of our data.

TOEFL: Based on the initial observation of the dataset, the TOEFL column has many missing values (i.e. there are only 55 TOEFL scores out of 308 students for 2021-2022), so it cannot effectively represent the corresponding class or provide accurate measure of the analysis. Therefore, the analysis does not include the TOEFL data.

Undergraduate GPA: The data cleaning process includes updating null values to mean values and converting GPAs that are not in the scale of 4.0. To effectively calculate the mean that represents all the cohorts, we averaged all the GPAs that were on the scale of 4.0. To convert GPAs that were not in the 4.0 scale, the following rules were applied. If the GPA was greater than 4 but less than 5, we would scale this GPA so it is within the scale of 4. If the GPA is greater than or equal to 5 but less than 10, the same scaling would take place. The same method would be employed if the GPA is greater than or equal to 65 and below 100. If the GPA would not fall into any of the categories above, we would just use the average GPA as a way to fill the null values.

GRE (Quantitative, Verbal, Writing) Scores: Based on the initial observations, there are 4 missing values for each class so data cleaning is required to update these null values. All the GRE scores are in the same scale without any outliers, so the null values are converted into a mean of Quantitative, Verbal, and Writing scores of the corresponding class.

CDS student composition based on gender, country, resident status: There are only two missing values in the entire data of all three classes, and there is no standard of estimating the gender based on the data given, so these two null values would be removed from the dataset when the gender based statistics are calculated. Note that they can be used for the statistics that are not gender-based. For Foreign Citizenship Status, we were advised by the department that all missing values should be assumed as US Citizenship. For US Citizenship Status, ‘1’ represents citizens, ‘3’ represents permanent residents, and ‘4’ represents international students. To get a better understanding, a separate column indicating the status in words was added. There are 5 missing values out of all three-year data, and there is no standard of estimating the citizenship from the data so the null values would be dropped from the data.

3.2 Career Outcomes / Summer Internship Outcomes

Initially we had to merge the *Summer Internship Outcomes* and *Career Outcomes* datasets. *Summer Internship Outcomes* had 4 columns, 'Semester', 'Organization', 'Job title' and 'Job Location' whereas *Career Outcomes* had 3 columns 'Organization', 'Job title' and 'Job Location'. By creating the 'Semester' column in *Career Outcomes* dataset and adding each year as a value (i.e. 'Career(20-21)'), we were able to merge those datasets row-wise, resulting in a 222 row dataset with 4 columns. After that we checked for duplicated rows. There were quite a few, which makes sense since many students had the same job at the same company at the same location. Finally, we checked for null values and noise data for each column. For the 'Semester' column we found that the 'Summer 2019' category included an extra spacing at the end which we had to remove through string manipulation. Thus, we converted 'Summer 2019 ' into 'Summer 2019'. Column wise we made several changes that can be explained below.

Organization: there was one null value which could not be identified and as a result we removed this row. Also, in this column there were typos. The same company name was written in many different ways. By merging some noise values into actual values, unique values in the 'Organization' column decreased from 137 to 113.

Job Location: there were 44 null values. We were not able to decide how to replace null values for job location. Deleting rows would significantly decrease our sample. Therefore, we simply converted all the null values into 'unidentified'. For the cleaning part, we decided to convert all the information into the state level.

Job Title: only 6 values were missing which were eventually marked as 'unidentified'. We also decided to create the following categories: 'Data Engineer', 'Data Scientist', 'Data Analyst', 'Software Engineer', 'Machine Learning Engineer', 'Product Manager', 'Quantitative Analyst', 'Research Scientist', and 'Others'.

Since for both the career outcomes and the internships we only had three columns of information, we decided initially to merge them and add more columns based on the characteristics of the company. Utilizing Linkedin data, we created two extra categories, location of the HQ, and employee size. In terms of the missing locations, we converted them into 'unidentified'. Using the 'Global Industry Classification Standard', we merged several industry categories and by creating a function in python we were able to convert the already existing industries to the new standard. As a result of this the final dataset had the following columns 'Semester', 'Organization_cleaned', 'Industry_category', 'Company_size', 'Final_loc', 'Job Category', and a total of 221 rows. Concluding the preprocessing of the entire career dataset, we eventually split this into internships and career outcomes, which we used to do our research and visualizations on.

4. Student academic performance

Class Statistics for 2019-2020, 2020-2021, 2021-2022: The mean scores of GRE Quantitative, Verbal, Writing for each class are as follows: the mean GRE Quantitative score is 166.95 (90.03 Percentile), Verbal is 159.08 (79.94 Percentile), Writing is 4.12 (61.70 Percentile), GPA 3.67 with standard deviation of 0.23. As for the resident status composition of the CDS program, 63.5% of the students are international students. The percentage of students with foreign country citizenships in the program is 68.3% which does not correspond to the percentage of non-US citizenships which is 66.3%. The difference of 2.0% can be interpreted as possible dual citizenship.

Gender composition per class: The proportion of men in the CDS program is 54.6% and the proportion of women is 45.4%. Although it is contrary to the national statistics of US college students

(source 1) that the ratio of women students (59.5% in 2021) enrolled in US colleges exceeded the ratio of men (41.5% in 2021) enrolled in US colleges, it can be interpreted as a common characteristic of STEM/Engineering colleges. As a part of STEM fields of study, considering that the male ratio in the industry is over 70%, it can be said that the male to female ratio in the CDS program is very evenly distributed.

Advanced Approaches (Regression, Correlation): GRE scores are based on two different scales, a scale of 170 for Quantitative and Verbal and a scale of 6 for Writing, and each GRE test is not standardized but the percentiles are. Therefore, the percentile of GRE scores are used in the analysis instead of the actual scores. Based on the correlation coefficient plots, the GRE Quantitative and Verbal scores are positively correlated with a coefficient of 0.27 whereas the GRE Quantitative and Writing scores are negatively correlated with a coefficient of -0.16. The absolute values of the correlation coefficients are 0.15 or 0.20, and it indicates that they are not strongly correlated. The GRE Verbal and Writing scores have a relatively high correlation (0.40), which is plausible as both are a subject in the same liberal arts. The GPA (converted into 4.0 scale) and GRE Quantitative scores have a positive correlation with a coefficient of 0.28. However, GPA is negatively correlated with the verbal score at R of -0.16. Using a heatmap provides general intuition of correlations between several categories at once, which helps to compute correlation comparisons across data.

5. Student career outcome

Organization: What we can conclude is that many students did an internship or found a full-time job in big tech companies such as Amazon, Apple, Google, IBM, and Facebook.

Industry: Summer internships were mostly in Tech and Finance with shares of 50% and 25% respectively. Yearly trends were also similar, as the Tech and Finance industries were the two most popular for internships. The same is true for career outcomes. Tech stands at 53% and Finance at 20%. The similar trend holds, as the Tech and Finance industry were the two industries that students would get a job. In 2019-2020 however, careers in Finance were relatively low compared to the rest of the years (Only 9%). We would guess that this is due to the fact that we had very little career outcome data given that the cohort of students is about 150, and what we have is just about 30 for every year.

Company size: The majority of the internships were in large companies with more than 10000 employees. This would make sense since the majority of the students are international and big companies are more capable of sponsoring visas. In terms of career outcomes, even if we don't know about domestic vs international student distribution, we might assume big companies have more openings for a data scientist job.

Location: Most of the internships are located in NY (41%), CA (24%), WA (11%), and MA (7%). This year, likely due to covid, we see a shift to other places as well, with companies allowing remote work students may have chosen to apply to internships with greater geographical diversity. Career Outcome follows a similar path as the summer internships. This year was the most diverse in terms of location distribution.

Job Category: The majority of students are usually employed as data scientists or data analysts as interns. For full time offers, most students got a position as data scientists, followed by software engineers and machine learning engineers.

Appendix:

- Sources:
  Lecture notes - numpy, pandas, plots and graphs: https://brightspace.nyu.edu/

  Seaborn: https://seaborn.pydata.org/

  Matplotlib: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.plot.html

  Tableau official website: https://www.tableau.com/

  'Global Industry Classification Standard':
  https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard

  Gender composition in STEM:
  https://www.aauw.org/resources/research/the-stem-gap/

  Linkedin Company information: https://www.linkedin.com/jobs/

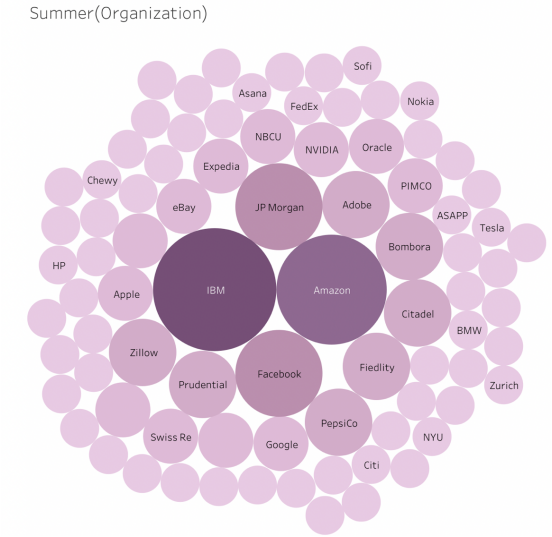
  Current CDS webpage: https://cds.nyu.edu/careers-ds/

  CMU career outcome statistics:
  https://www.cmu.edu/career/outcomes/post-grad-dashboard.html

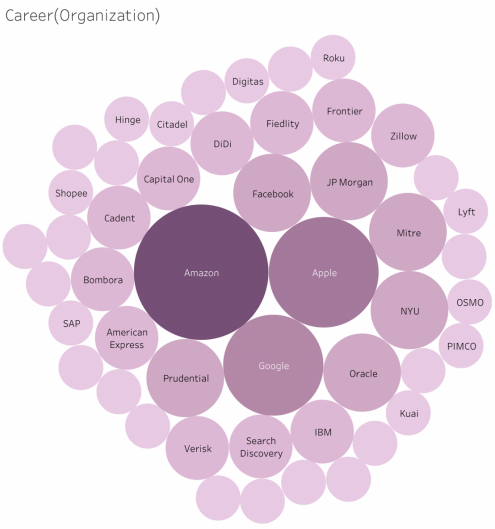
  MIT class profile:
  https://mitsloan.mit.edu/master-of-business-analytics/admissions/class-2022-profile

- Data visualization: Create plots and graphs using Pandas (including what we've learned in the class) and Tableau

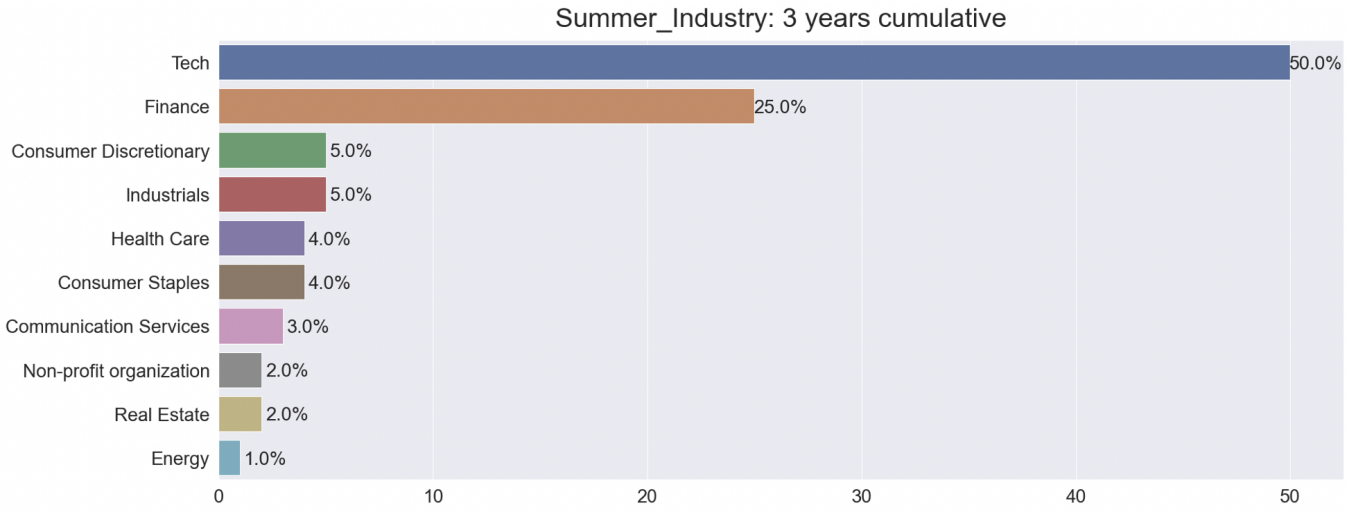
[Summer Internship: Organization - Tableau]



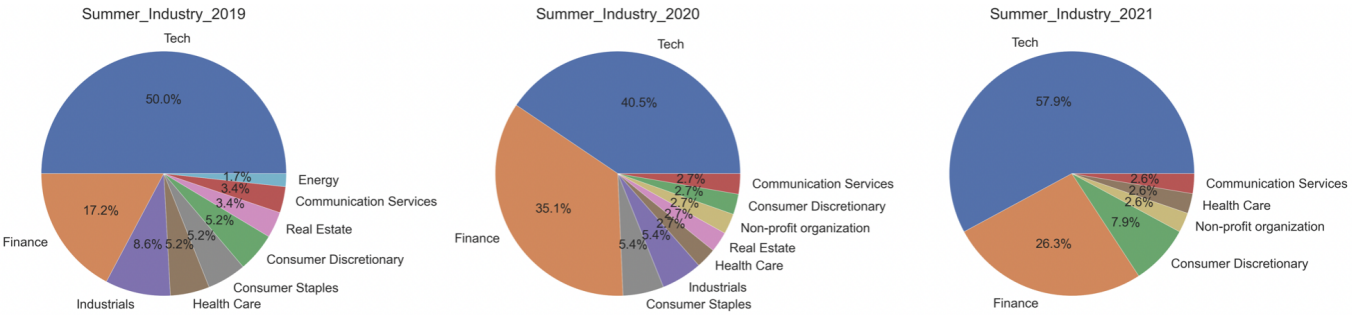
[Career Outcome: Organization - Tableau]



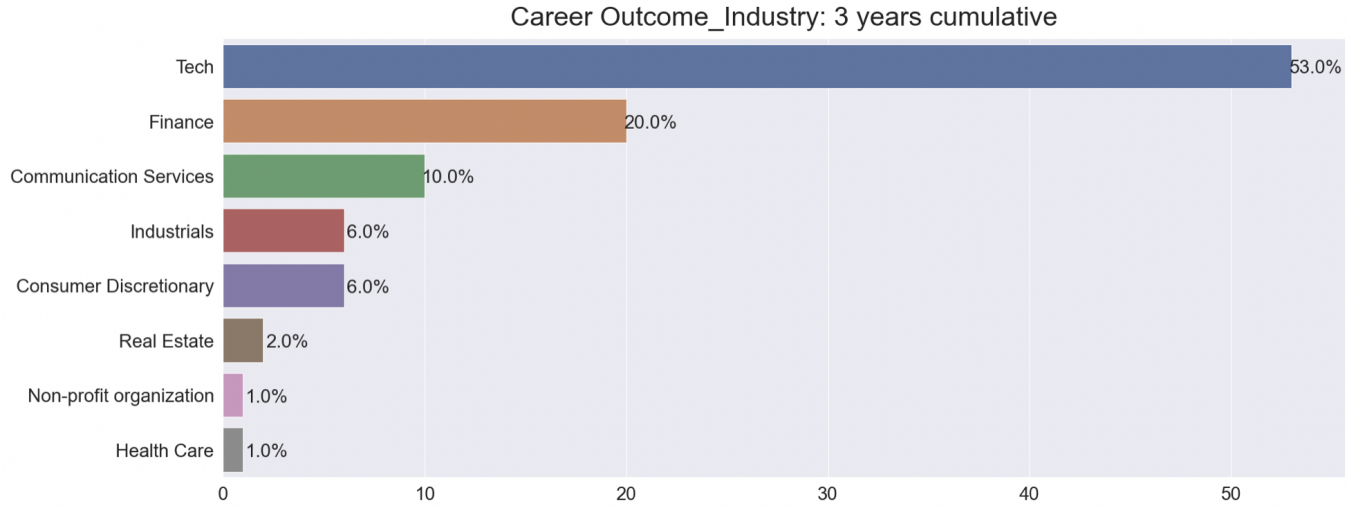
[Summer Internship: Industry 3 years distribution]



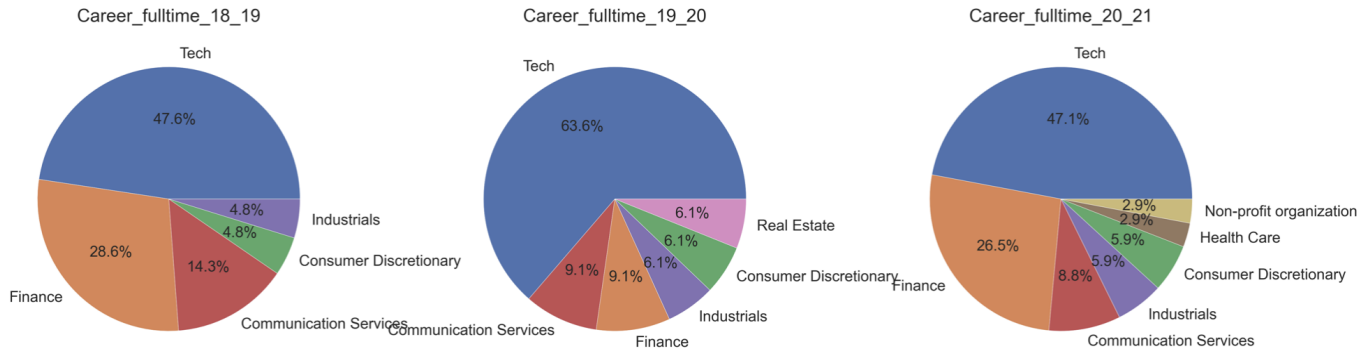
[Summer Internship: Industry 3 years trend]



[Career Outcome: Industry 3 years distribution]



[Career Outcome: Industry 3 years trend]



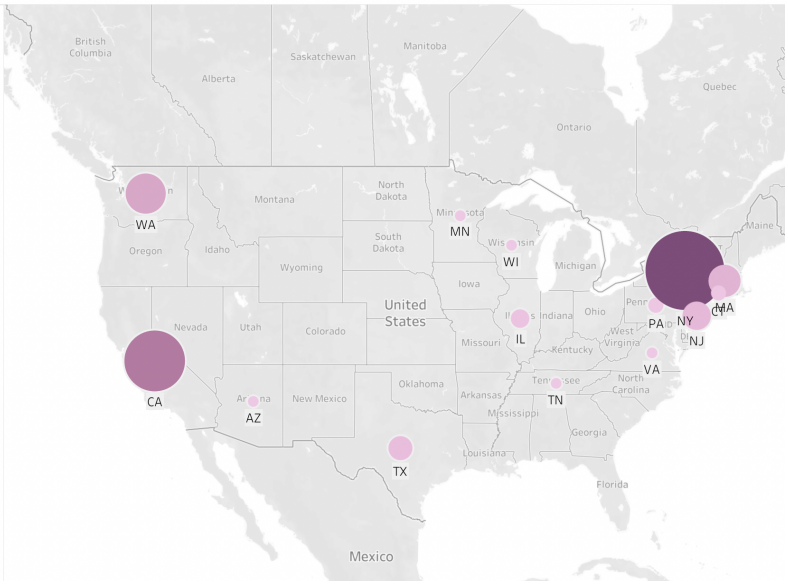
[Summer Internship: Company Size]

	Company_size
10,001+	0.59
1,001-5,000	0.14
51-200	0.08
11-50	0.06
5,001-10,000	0.04
2-10	0.04
501-1,000	0.04
201-500	0.02

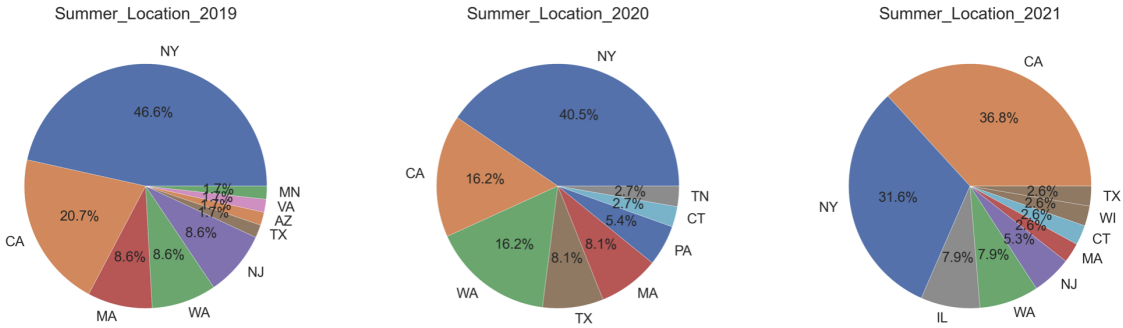
[Career Outcome: Company Size]

	Company_size
10,001+	0.65
5,001-10,000	0.10
51-200	0.09
1,001-5,000	0.08
11-50	0.03
201-500	0.03
501-1,000	0.01

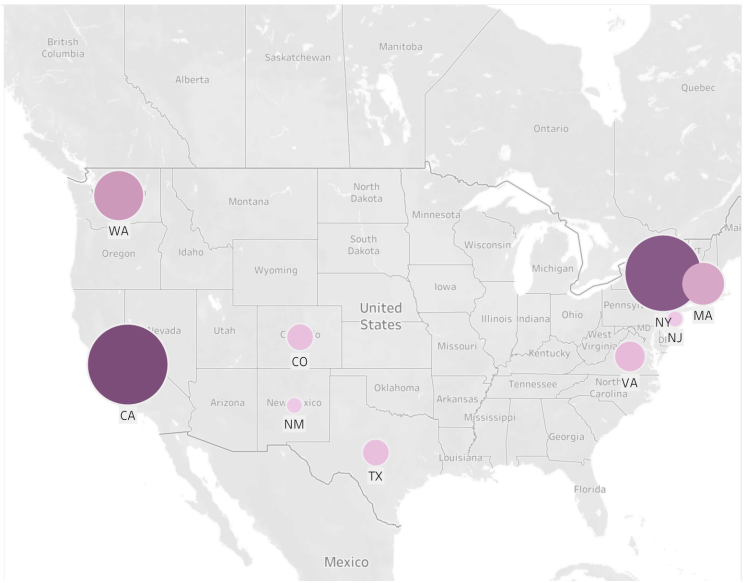
[Summer Internship: Job Location 3 years distribution - Tableau]



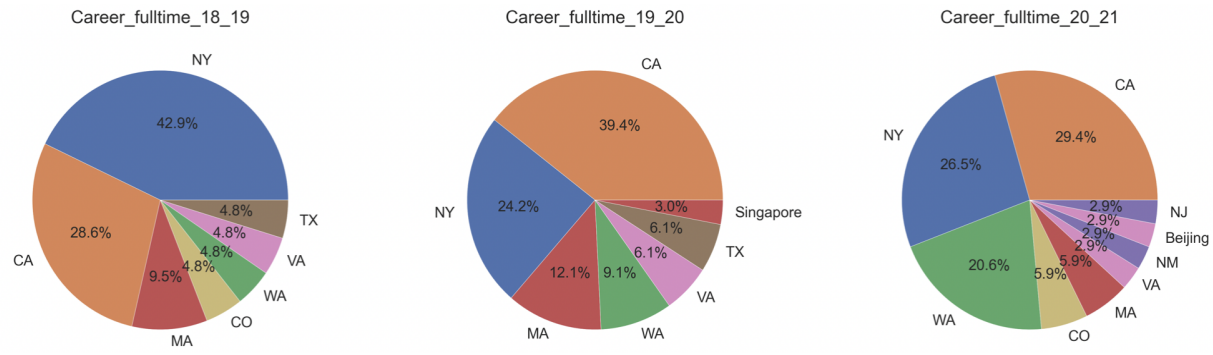
[Summer Internship: Job Location 3 years trend]



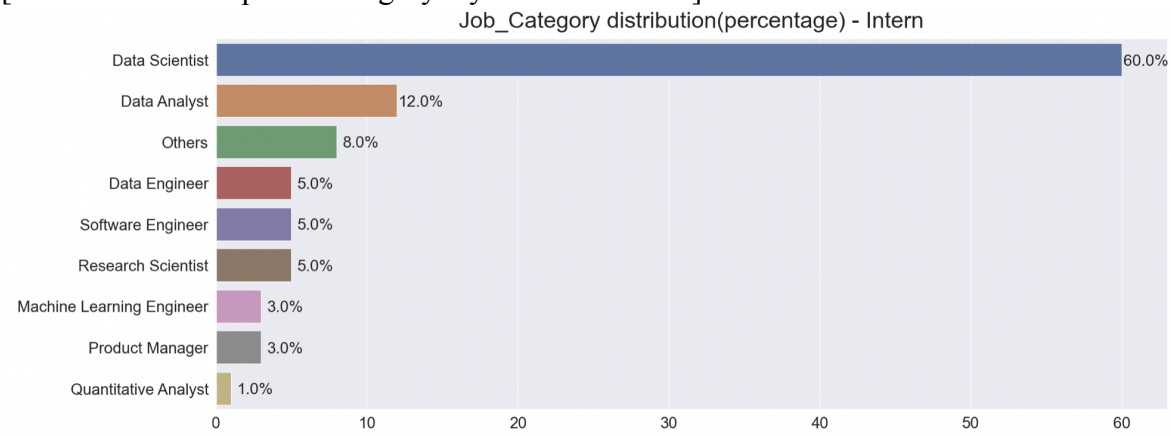
[Career Outcome: Job Location 3 years distribution - Tableau]



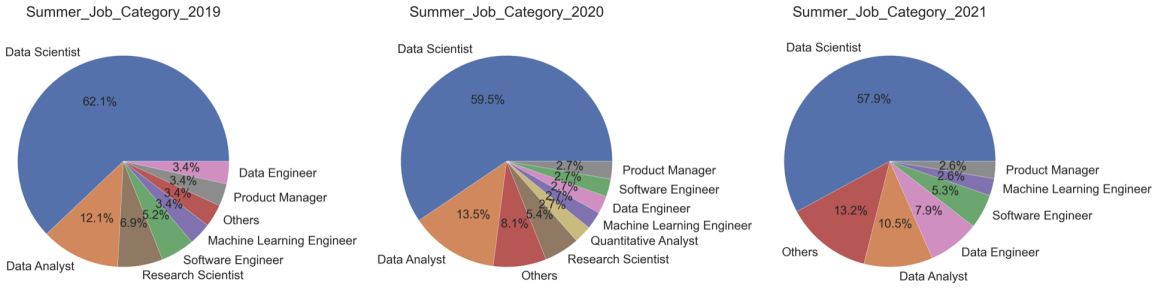
[Career Outcome: Job Location 3 years trend]



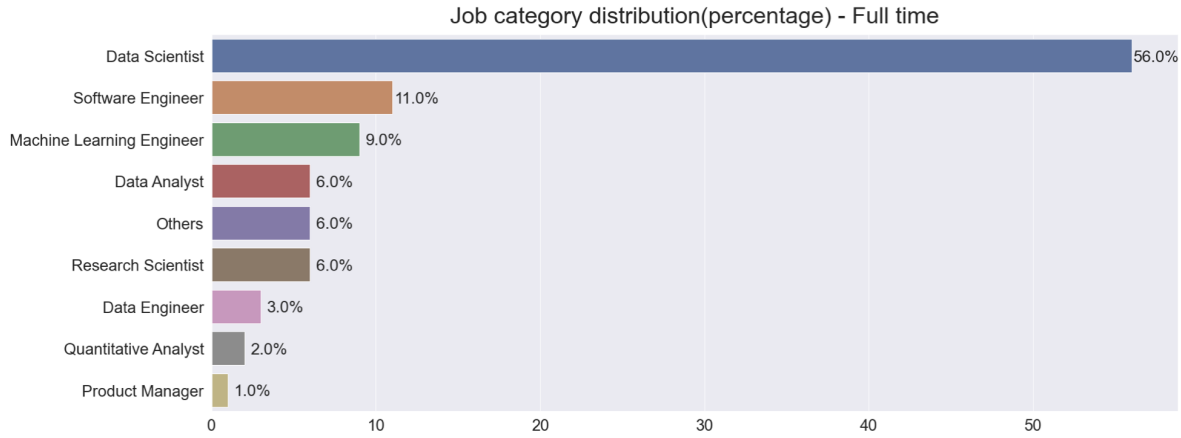
[Summer Internship: Job category 3 years distribution]



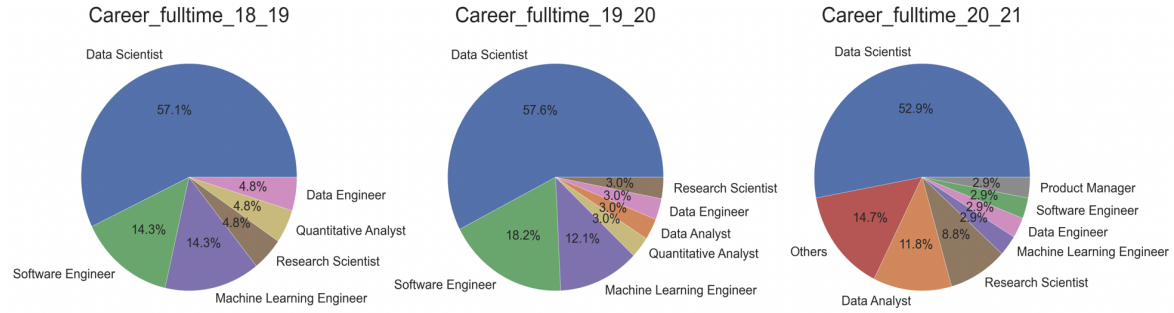
[Summer Internship: Job category 3 years trend]



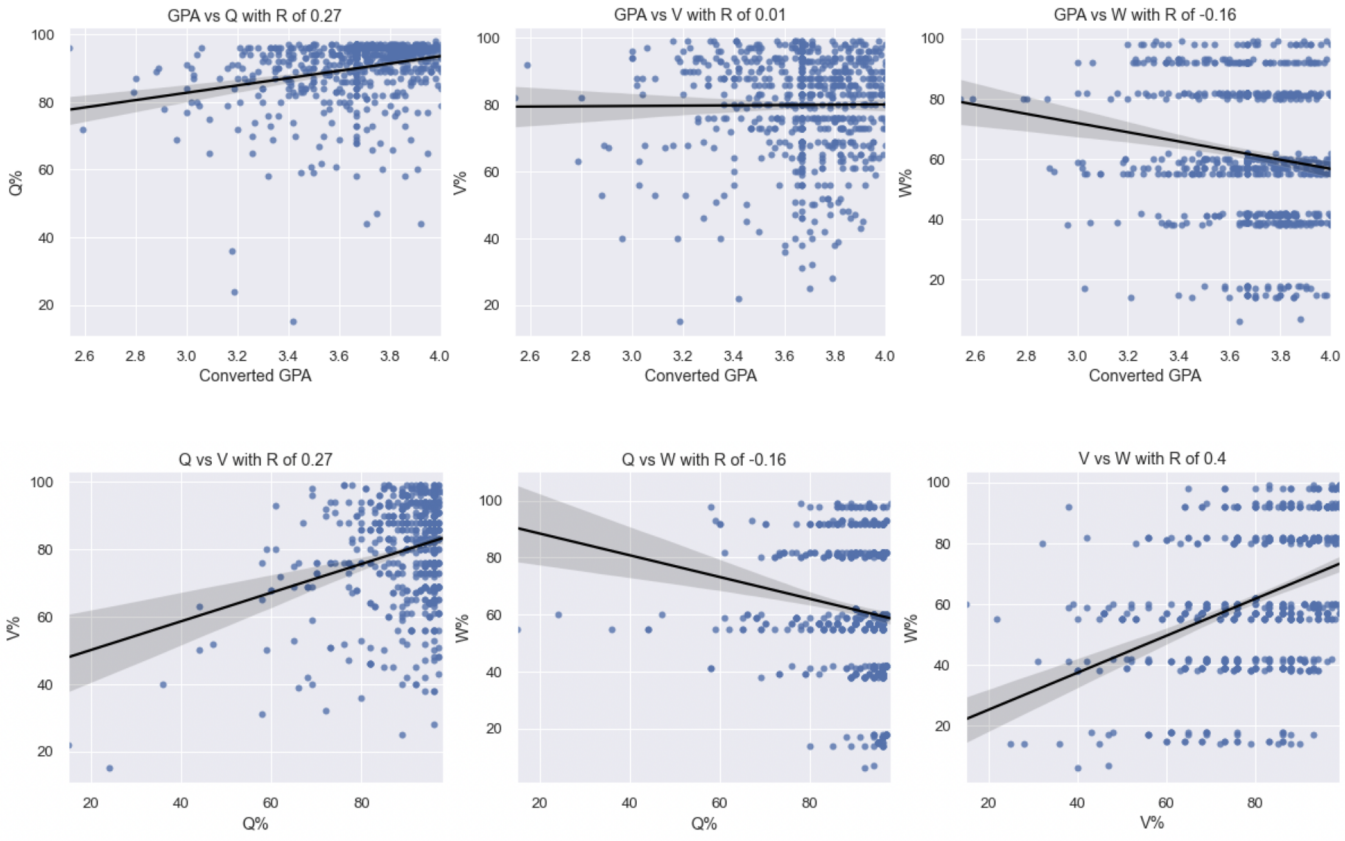
[Career Outcome: Job category 3 years distribution]



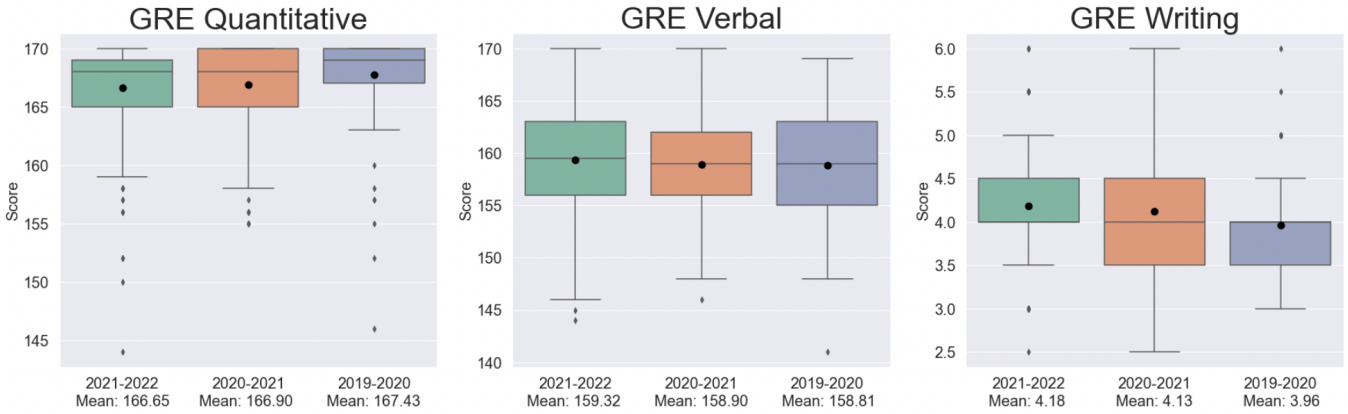
[Career Outcome: Job category 3 years trend]



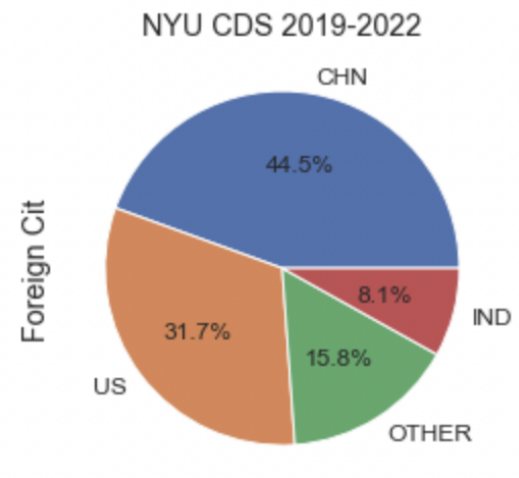
[Regression between student profile features]

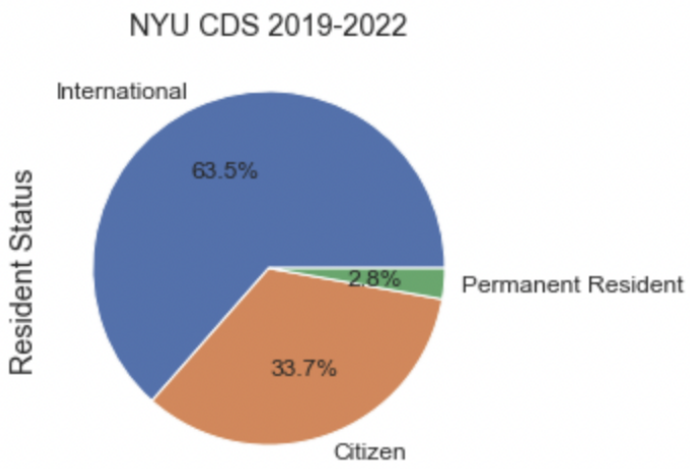


[Box-plot on GRE scores over individual class from 2019 to 2022]

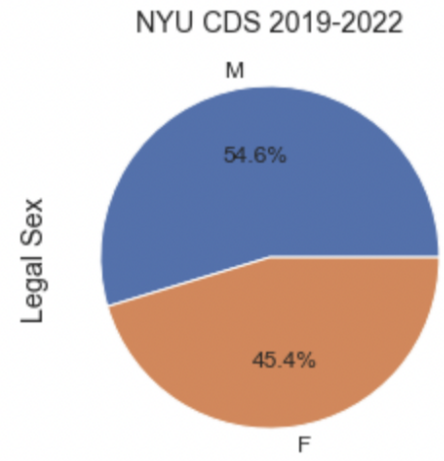


[Pie-chart of CDS Composition based on Diversity]





[Pie-chart of CDS Composition based on Gender]



[Heatmap of student profile]