Factors Influencing Academic Performance: A Data-Driven Analysis of Student Grades at UCSD

Spring 2022 Enrollment Data

Mert Ozer
mozer@ucsd.edu

Kang Gun Ham
kham@ucsd.edu

Jason Gu
jig036@ucsd.edu

Yanhao Guo
yag007@ucsd.edu

1 Introduction

1.1 Statement of the Problem

This analysis investigates the various factors that influence the average grades received by students at the University of California, San Diego (UCSD). Specifically, we aim to identify and quantify the impact of variables such as class size, enrollment numbers, evaluation rates, recommendation rates, study hours, and departmental differences on the average grades. By constructing and analyzing multiple regression models, we seek to uncover the relationships between these predictors and the students' academic performance as measured by their grades.

1.2 Motivation for the Investigation

The motivation for this investigation stems from the growing emphasis on data-driven decision-making in education. As UCSD increasingly collects and analyzes data related to student performance, there is an opportunity to leverage this information to improve educational practices and policies. Moreover, with the rising competition in the academic environment, understanding the factors that contribute to better grades can help UCSD differentiate itself by providing a higher quality of education and a more supportive learning environment. Being students ourselves, we are particularly motivated to undertake this analysis as it directly affects my peers and me. This analysis aims to provide actionable insights based on empirical data, ultimately benefiting both students and faculty at UCSD.

1.3 Relevance of the Problem

Understanding the determinants of student performance is crucial for UCSD as it strives to improve academic outcomes and overall student satisfaction. Grades are a key indicator of student success and can influence future educational and career opportunities. By identifying the factors that significantly impact grades, educators and administrators can implement targeted strategies to enhance teaching effectiveness, optimize class sizes, and allocate resources more efficiently. As a student at UCSD, this investigation also holds personal relevance, as the insights derived from this analysis can directly impact the quality of education and academic support provided to fellow students.

1.4 Relation to Project Proposal

Our final report is significantly different from our project proposal. Initially, our project aimed to analyze UC San Diego's Winter 2024 enrollment data to optimize resource allocation, focusing on metrics like the ratio of waitlisted to enrolled students and departmental enrollment trends. However, we encountered limitations with the initial dataset, which only included class size and department as possible features to experiment. This narrow scope hindered our ability to apply advanced feature selection techniques, such as stepwise selection or LASSO regression, due to the lack of sufficient covariates

Given these constraints, we decided to change our dataset and shift our focus to incorporate additional features that could provide a more comprehensive analysis. We sourced new data that included variables such as class evaluations, recommendation rates, study hours, and other course-related metrics. This enrichment allowed us to explore a broader range of factors influencing student performance, particularly average grades, which are more relevant and insightful from a student's perspective.

In short, our shift from analyzing enrollment metrics to investigating average grades was driven by both data limitations and the goal of enhancing the relevance and impact of our study. This approach not only addresses the initial objectives of understanding course dynamics but also provides more actionable insights for improving student outcomes and academic experiences at UCSD.

2 Dataset

2.1 Dataset Overview

In this project, we sourced our data from the UCSD-Historical-Enrollment-Data GitHub repository and CAPE. However, UCSD has discontinued CAPE and is moving forward with a similar but new form called SET. With limited access to data, we decided to focus on the Spring 2022 Quarter for this analysis.

The Spring 2022 dataset from UCSD-Historical-Enrollment-Data GitHub consists of 8,533 rows [Link]. Each row represents an individual class enrollment section. It also contains 6 columns, each representing a different attribute of the course sections. The **subj_course_id** column is a nominal data type that contains identifiers for each subject course. Similarly, **sec_code** and **sec_id** are also identifiers for the section of those same classes. For example, the class DSC 180A may have a section A01 and a section B02. The **instructor** column lists the name of the professor teaching the course. The **total_seats** columns are discrete data that indicate the total number of seats offered during the respective quarter. Lastly, the **meetings** column has information on lecture and discussion times for that section such as class duration, schedule, and location.

The Spring 2022 CAPES dataset consists of 1267 rows and 11 columns. Each row records the course and the respective instructors. The **instructor** column records the instructor of the course. The **sub_course** column contains an identifier for each course and the **course** column includes the course title. The **term** includes the quarter in which the evaluation belongs, in our case, it will be Spring 2022. The **evals_made** records the number of students who made the evaluation. The percentage of students who recommend the course and professor is also recorded in **rcmd_class** and **rcmd_instr** respectively. Finally, the **study_hr_wk**, **avg_grade_exp**, and **avg_grade_rec** record the numerical data of hours spent in class and student's grades in the course.

2.2 Data preparations

2.2.1 Merge two dataframe

At the beginning of our analysis, we merged two datasets to create comprehensive data that could support a robust analysis. Each dataset was enhanced by adding a term column to indicate the specific academic term, such as 'SP22' for Spring 2022. This term column was essential for distinguishing data from different terms and ensuring accurate analysis across the datasets. The merge operation was performed on the common columns term and sub_course, allowing us to consolidate course-specific data efficiently. (We choose to focus on the SP22 quarter only by the end, but one can easily expand the analysis scope to the three quarters we have processed and uploaded)

2.2.2 Data Cleaning

When doing data cleaning, we identified and handled missing values by examining the non-null counts across the columns (we also need to detect some negative values like '-1' that also indicates a null value). Besides, we also took care of outliers using 3 standard deviations as the threshold. Additionally, we standardized the formats of categorical variables and ensured text fields were clean by checking and removing any existing whitespace.

2.2.3 Feature Engineering

Feature engineering involved creating new variables that could provide additional insights into the factors affecting student performance. Key features included:

- **Class Schedule and Length**: We parsed the meetings column to extract the class schedule (e.g., 'TuTh' for Tuesday and Thursday) and the duration of each class session. This information was stored in new columns class_schedule and class_length.
- **Enrollment Rate**: We calculated the enrollment rate as the ratio of enrolled students to total seats available, which was stored in a new column capacity_utilization. This metric helped in understanding the demand and capacity utilization of each course.
- **Categorical Variables**: We converted categorical variables such as department_name, class_schedule, and class_length into categorical data types to facilitate statistical analysis and regression modeling. This involved creating dummy variables for these categorical features, allowing us to include them in our regression models effectively.

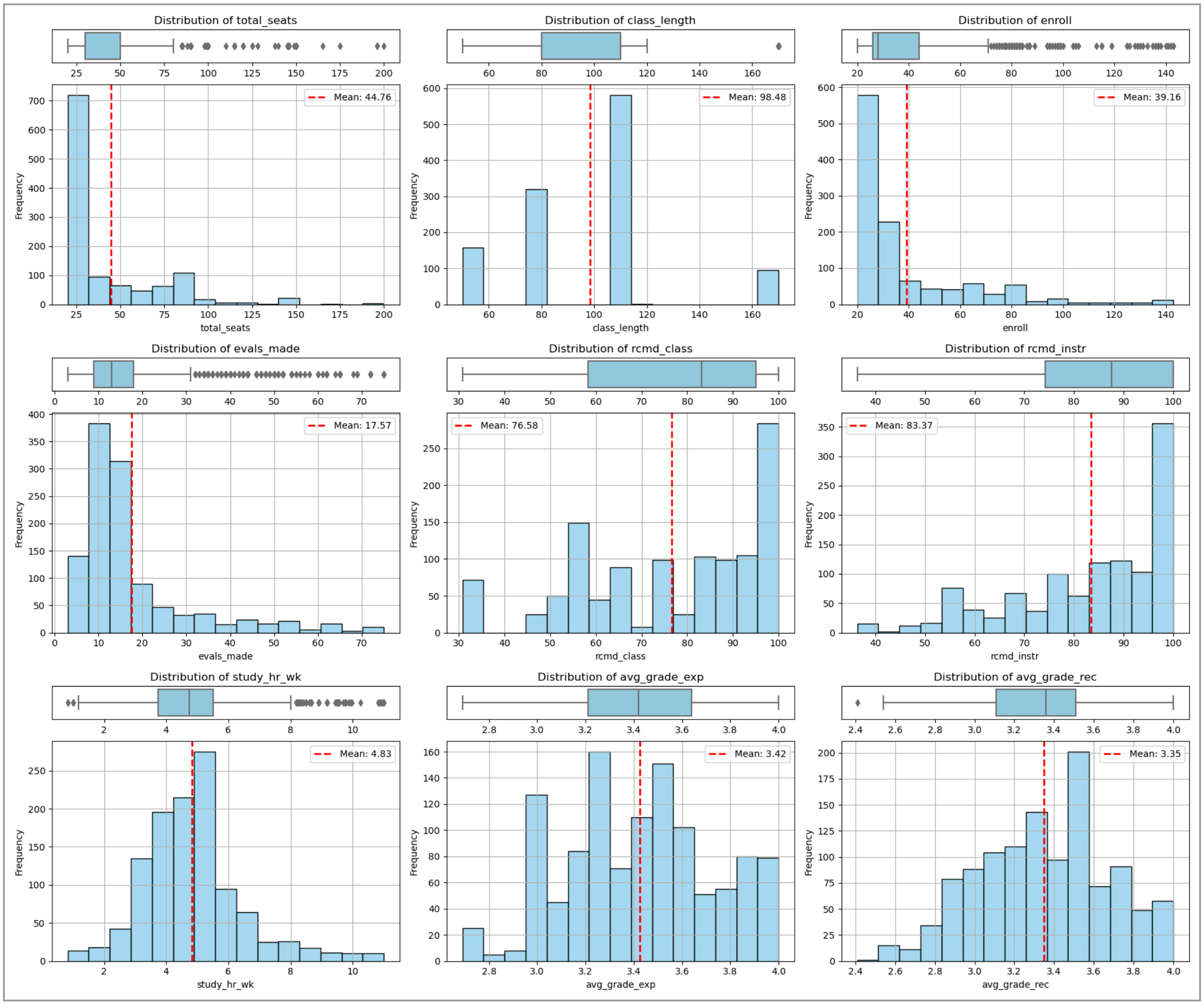
2.3 Descriptive Statistics

In this section, we want to provide some descriptive statistics to provide an overview of the cleaned dataset. First, we calculated the central tendency for each numerical variable. The average number of seats offered per course section is 44.76 seats, with a standard deviation of 27.96. This suggests that there is a moderate variation in class sizes. The median of 30 indicates that half of the sections had less than 30 seats.

Now, let us look at the class duration. On average, each class was 98.48 minutes long. The median class length was 110 minutes, with 25% of classes being 80 minutes or shorter, and 75% being 110 minutes or shorter. These statistics reflect the typical scheduling pattern at UCSD.

For the CAPE dataset, we want to highlight the recommendation rates. The mean percentage of students who recommended the course and instructor was 76.58% and 83.37% respectively. These two metrics have a standard deviation of 20.55% and 16.11% respectively, indicating quite a bit of variation. It is worth noting that the median of these two metrics is 83.10% and 87.50% This suggests that most courses and instructors are well-regarded by the students.

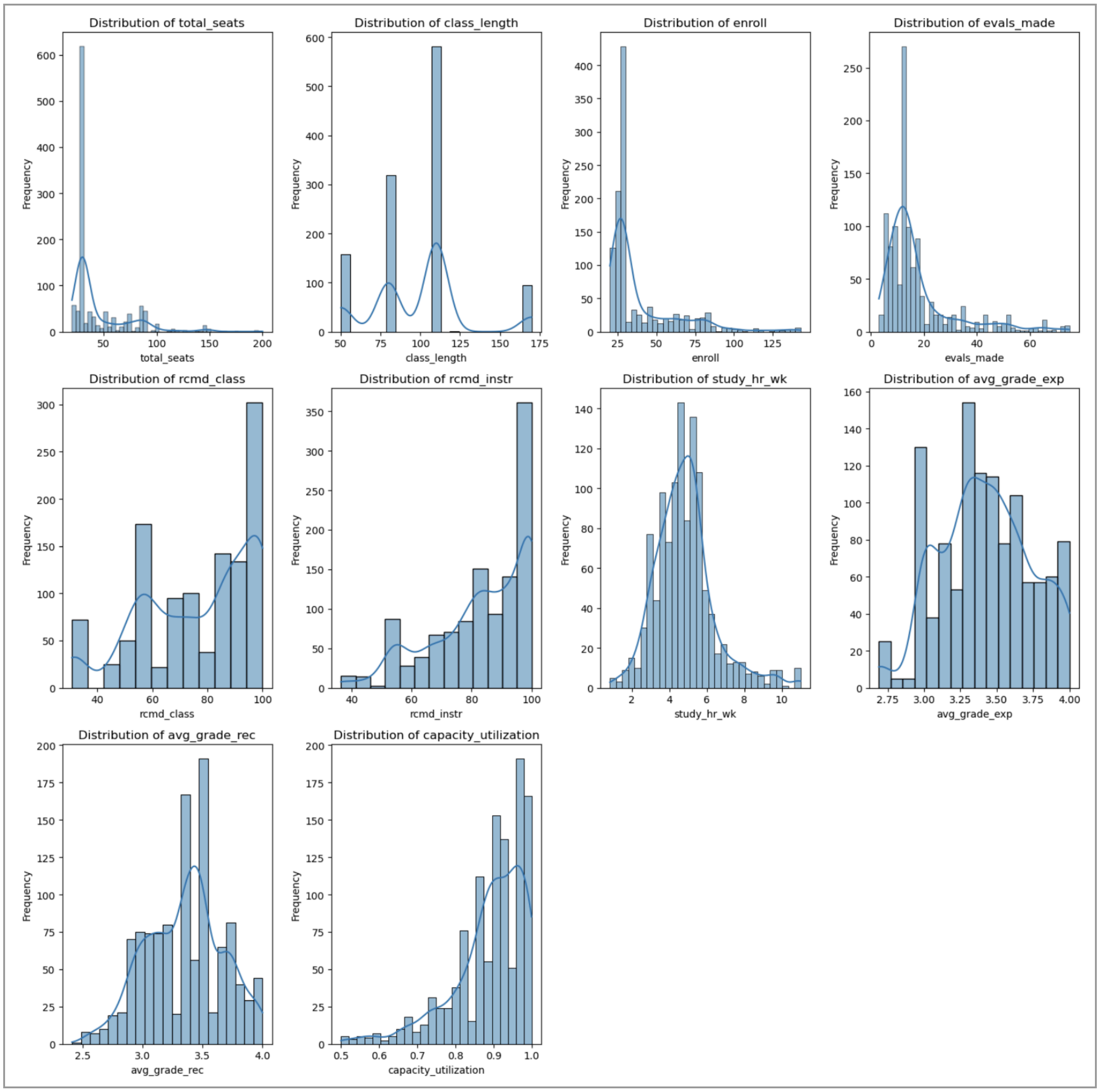
Below is a visual representation to complement some of the statistics mentioned above. Each numerical variable is presented with a histogram with a red line representing the mean. We've also included a box plot to provide a more nuanced view of the spread of data.



3 Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

Here we take a look at the distribution of our features from our data individually on their histograms. These histograms would help us understand the spread of each feature and better equip us to make predictions based on this data.



Distribution of Class Lengths

The histogram for class lengths reveals a distribution with peaks around specific duration. This suggests that class sessions are largely standardized, falling mainly into a few specific lengths. The noticeable peaks could represent the typical class periods (possibly 50 and 80 minutes) structured around UCSD's standard academic scheduling blocks. Classes longer than the most common durations are less frequent, possibly reflecting specialized or extended sessions for labs or seminars. Understanding these patterns can provide suggestions in optimizing classroom allocations and scheduling practices.

Distribution of Enrollment From Dataset

The distribution of enrollment shows a strong correlation with the total seats within those classes, suggesting that class sizes are a strong predictor of enrollment numbers. However, the distribution here is slightly skewed towards lower enrollments relative to total seats, with peaks less sharp and more spread out. This dispersion highlights the variability in how full classes get, possibly indicating courses that are either highly popular or under-attended. The analysis of these patterns can be critical for understanding student preferences and could guide course offerings (balancing popular courses) and identifying underutilized ones that might need reevaluation.

Distribution of Capacity Utilization From Dataset

The capacity utilization graph shows a significant skew toward higher utilization rates, with a massive spike from 90 - 100% utilization. This indicates that a large number of classes are operating at full or close to capacity, which could imply a well-matched offering to demand or a potential strain on resources where demand exceeds supply. The lower utilization rates, visible from 0 to around 0.5, are less common but suggest that some classes significantly underutilize the available space. These cases warrant further investigation to determine if they result from an overestimation of demand, niche subject matter, or suboptimal scheduling or marketing of these courses.

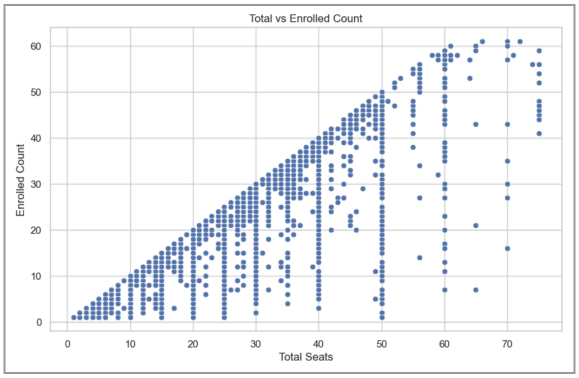
Distribution of Evaluation Participation From Dataset

The evaluations made from students shown within its histogram have peaks that likely align with the class sizes but show some variability as well. Lower peaks may indicate that classes where there are fewer students who feel compelled or able to provide feedback due to the class engagement levels. Analyzing this could help improve the feedback system at UCSD and encourage higher participation rates for all classes offered. This could ensure comprehensive and constructive input for course adjustments for the benefits of professor evaluation.

3.2 Bivariate Analysis

Here we look at a scatter plot of total seats versus the enrollment count to visualize their mutual relationship with each other and a series of box plots of departments versus capacity utilization to understand which departments are highly demanded and which are less demanded.

Scatter Plot of Total Seats and Enrollment Count

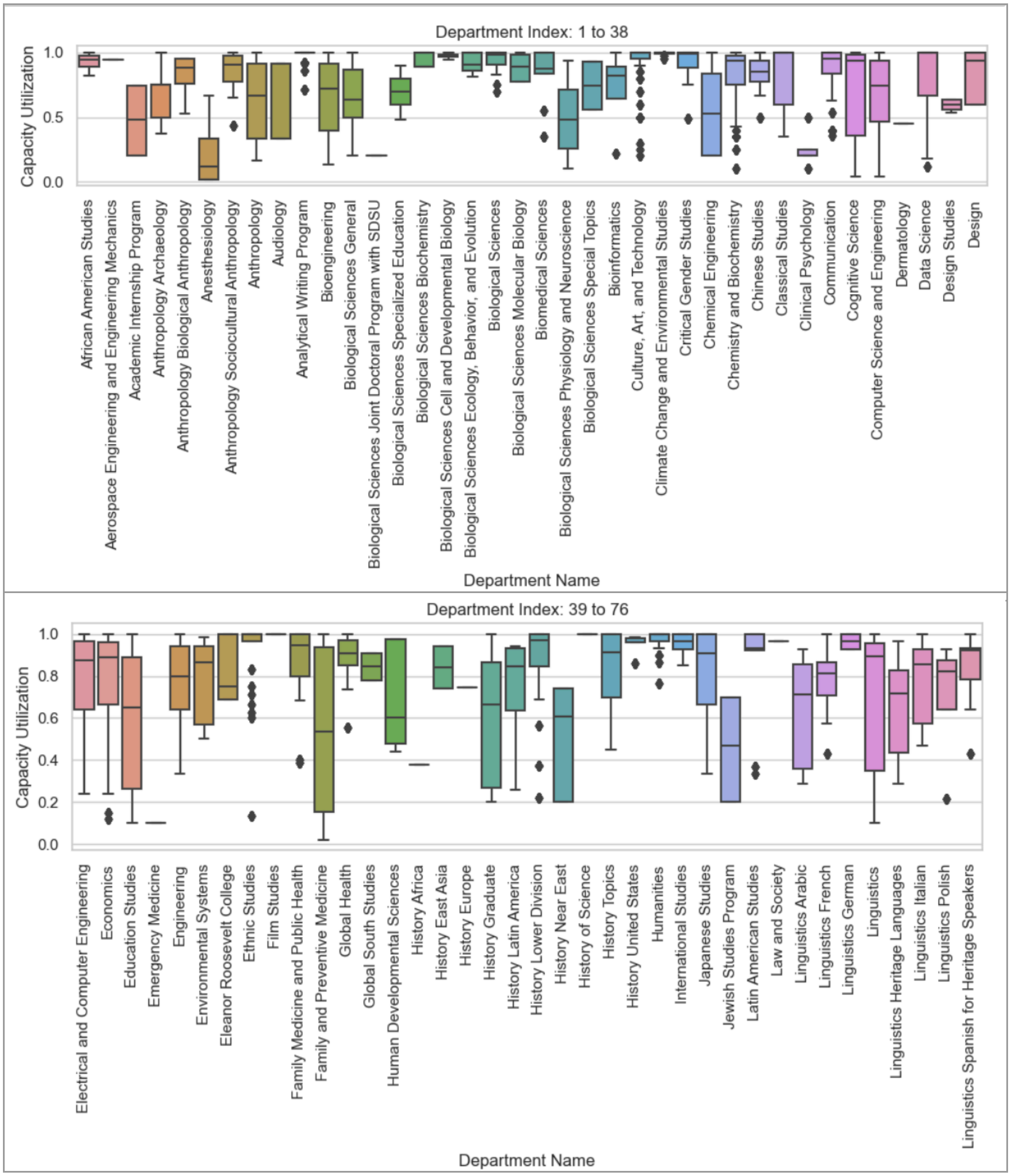


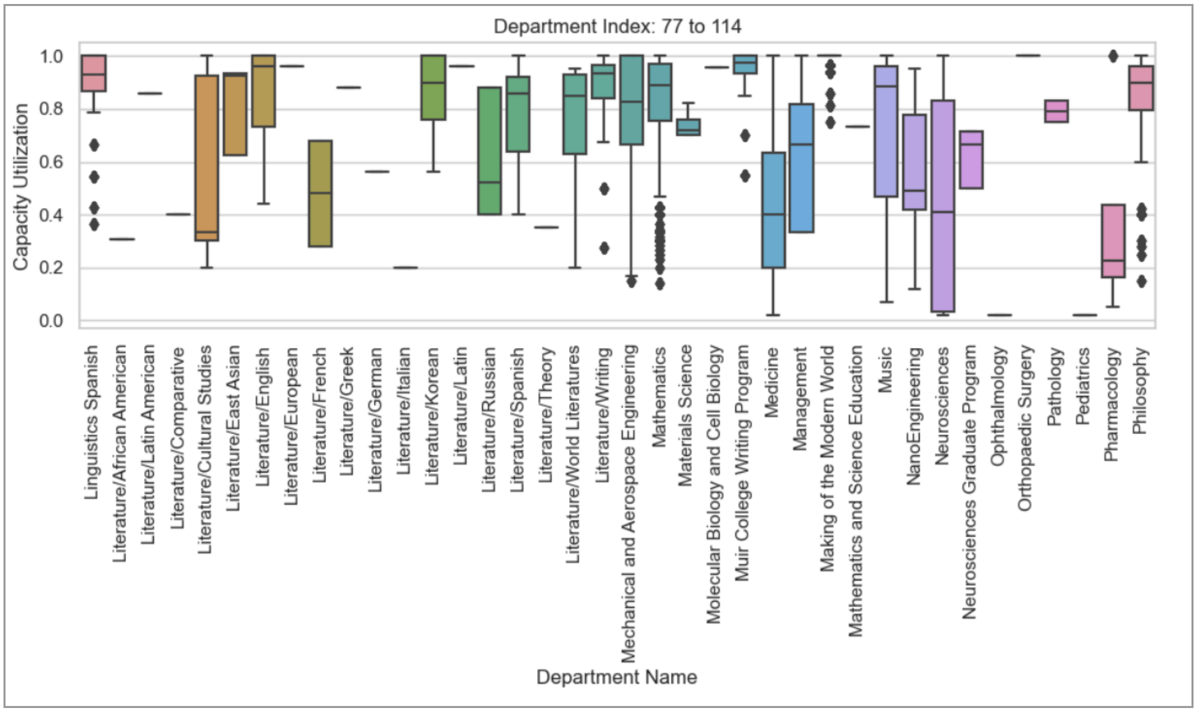
1. Positive Linear Relationship: The plot shows a clear positive linear trend indicating that, generally, classes with more available seats tend to have higher enrollment. This suggests that course capacity is a

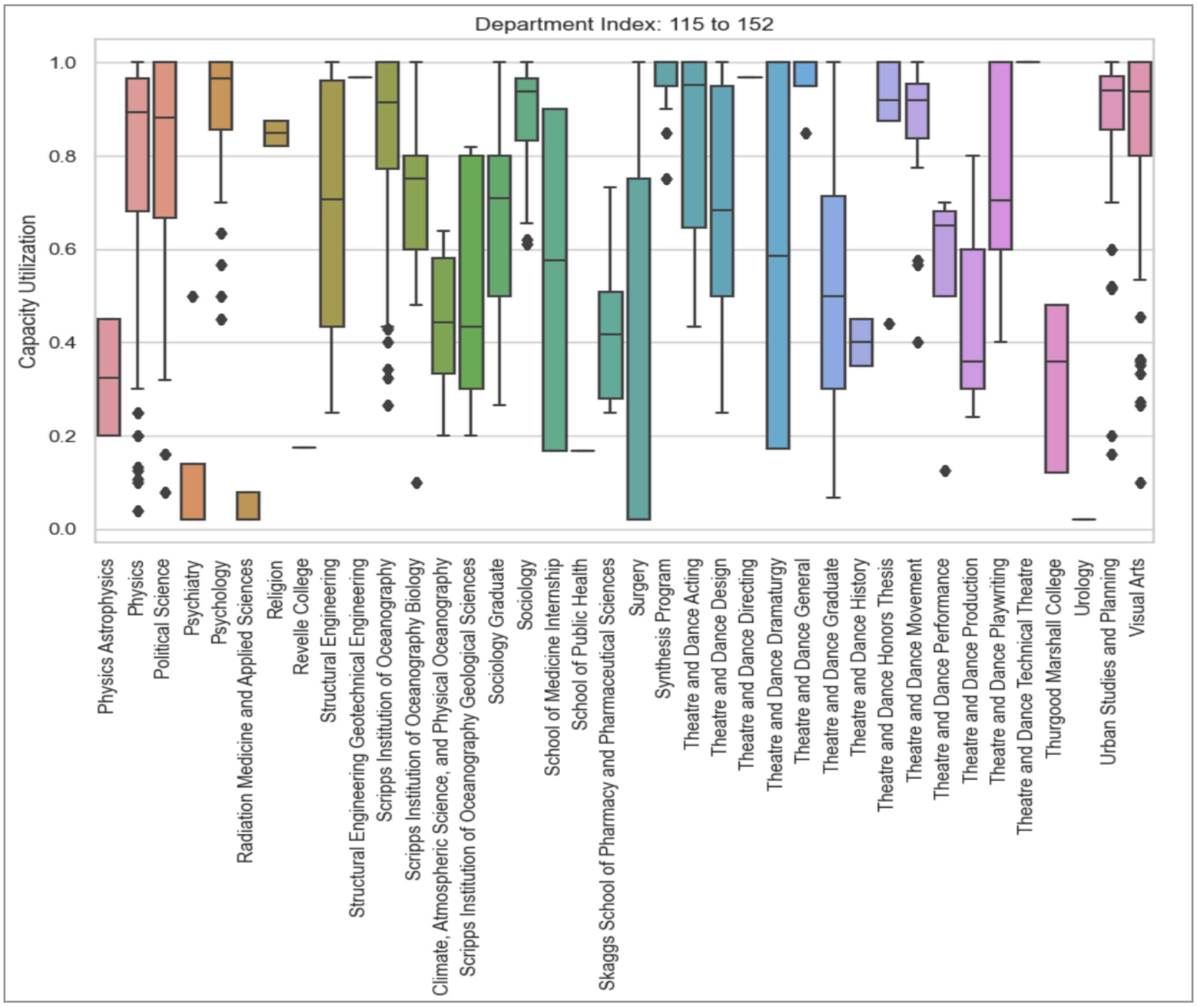
significant factor influencing how many students enroll.

2. Maximum Capacity Enrollment: Many points are clustered along the line where enrolled_ct equals total, especially as class sizes increase. This pattern indicates that larger classes are often filled, which might be a sign of high demand for these courses or a limited offering that compels students to fill available spots.
3. Variability in Smaller Classes: For smaller class sizes (less than 30 seats), there is noticeable variability in how full these classes are. Some small classes are at or near capacity, while others have significantly fewer students than the total available seats. This variability could be due to several factors, such as the niche nature of the course, scheduling conflicts with other popular classes, or less interest in the specific subject matter.

Barplot of Capacity Utilization By Department




Department Index: 77 to 114
Capacity Utilization
1.0
0.8
0.6
0.4
0.2
0.0
Linguistics Spanish
Literature/African American
Literature/Latin American
Literature/Comparative
Literature/Cultural Studies
Literature/East Asian
Literature/English
Literature/European
Literature/French
Literature/Greek
Literature/German
Literature/Italian
Literature/Korean
Literature/Latin
Literature/Russian
Literature/Spanish
Literature/Theory
Literature/World Literatures
Literature/Writing
Mechanical and Aerospace Engineering
Mathematics
Materials Science
Molecular Biology and Cell Biology
Muir College Writing Program
Medicine
Management
Making of the Modern World
Mathematics and Science Education
Music
NanoEngineering
Neurosciences
Neurosciences Graduate Program
Ophthalmology
Orthopaedic Surgery
Pathology
Pediatrics
Pharmacology
Philosophy
Department Name


Department Index: 115 to 152
Capacity Utilization
1.0
0.8
0.6
0.4
0.2
0.0
Physics Astrophysics
Physics
Political Science
Psychiatry
Psychology
Radiation Medicine and Applied Sciences
Religion
Revelle College
Structural Engineering
Structural Engineering Geotechnical Engineering
Scripps Institution of Oceanography
Scripps Institution of Oceanography Biology
Climate, Atmospheric Science, and Physical Oceanography
Scripps Institution of Oceanography Geological Sciences
Sociology Graduate
Sociology
School of Medicine Internship
School of Public Health
Skaggs School of Pharmacy and Pharmaceutical Sciences
Surgery
Synthesis Program
Theatre and Dance Acting
Theatre and Dance Design
Theatre and Dance Directing
Theatre and Dance Dramaturgy
Theatre and Dance General
Theatre and Dance Graduate
Theatre and Dance History
Theatre and Dance Honors Thesis
Theatre and Dance Movement
Theatre and Dance Performance
Theatre and Dance Production
Theatre and Dance Playwriting
Theatre and Dance Technical Theatre
Thurgood Marshall College
Urology
Urban Studies and Planning
Visual Arts

Observations Across the Graphs

High Utilization in Science and Technology Departments:

- Departments like Computer Science, Engineering, and Biological Sciences consistently show high capacity utilization. This trend is noticeable across the first and third graphs, which include many of the STEM fields. High utilization in these departments may be driven by a strong demand due to the growing emphasis on STEM careers and the high value placed on these degrees in the job market.
- Another factor could be that these departments often have well-defined lab components or hands-on sessions that require specific student-to-equipment ratios, thus capping the class sizes more stringently and leading to higher utilization.

Significant Utilization in Cultural Studies Departments:

- Departments focusing on cultural studies, such as Ethnic Studies and Gender Studies, also display considerable capacity utilization. These departments, seen prominently in the second graph, might experience fluctuating enrollments influenced by social trends and growing awareness and interest in social justice and cultural discourse.

Speculations on Capacity Utilization Trends

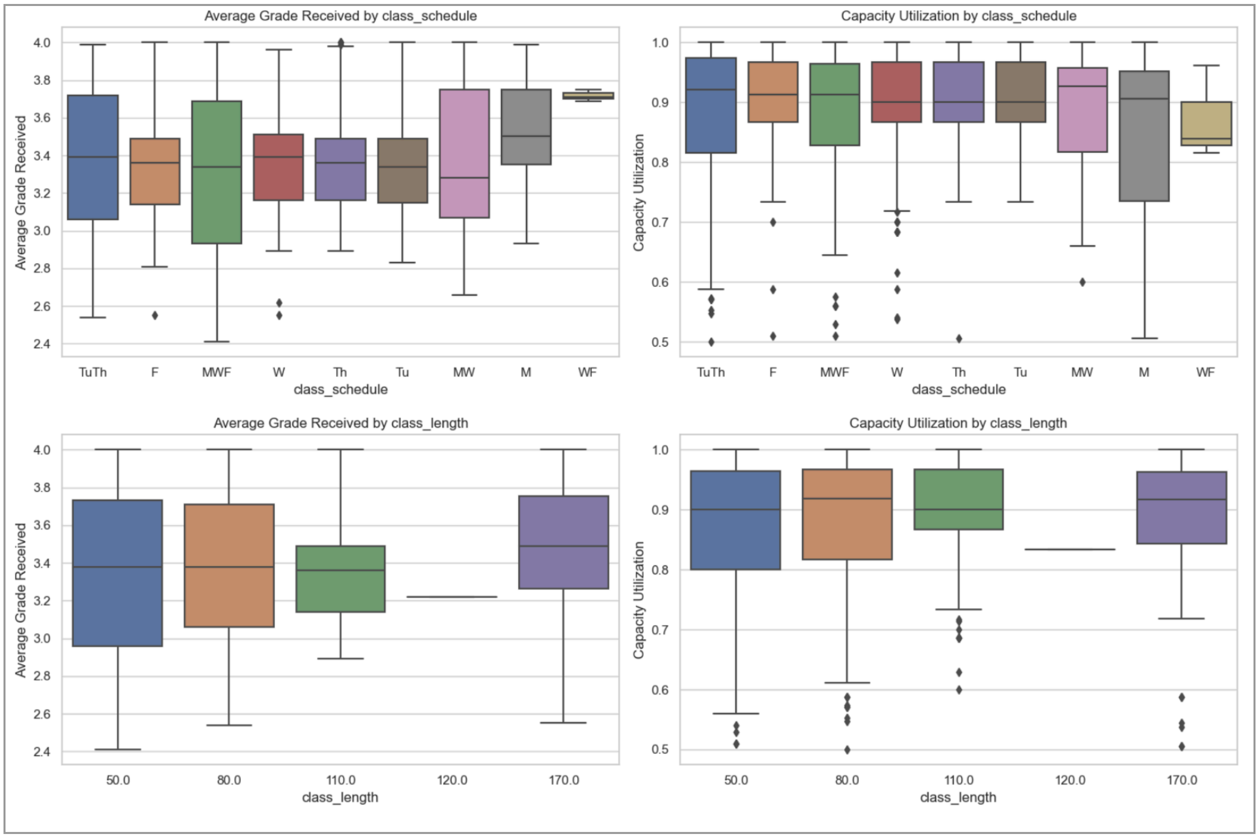
- Curriculum Influence: For the science and technology departments, the curricular requirements that necessitate sequential course completion could be ensuring steady enrollment, thereby maintaining high capacity utilization.
- Cultural Relevance: The departments involved in cultural studies may be benefiting from the increasing societal focus on diversity and inclusion, which could encourage more students to enroll in these courses.

Concerns with Lower Capacity Utilization

Some departments, particularly Emergency Medicine, Radiation Medicine and Applied Sciences, Ophthalmology, Pediatrics, and Urology, show notably lower capacity utilization. Several factors could explain these observations:

- Seasonal Demand: It's plausible that these departments face lower demand during the winter quarter. Certain courses, especially those involving clinical practice or outdoor activity, might be less feasible or popular during colder months.
- Curricular Structure: Students in these fields may often follow strict curricular paths that do not offer much flexibility for taking courses out of sequence, which could lead to underutilization in specific terms if the core courses are scheduled in other quarters.
- Specialized Nature: These fields are highly specialized; thus, they might naturally have smaller student bodies, leading to lower absolute numbers even if the program is appropriately sized for the interest and career prospects in these areas.

Average Grade and Capacity Utilization by Class Schedule and Class Length



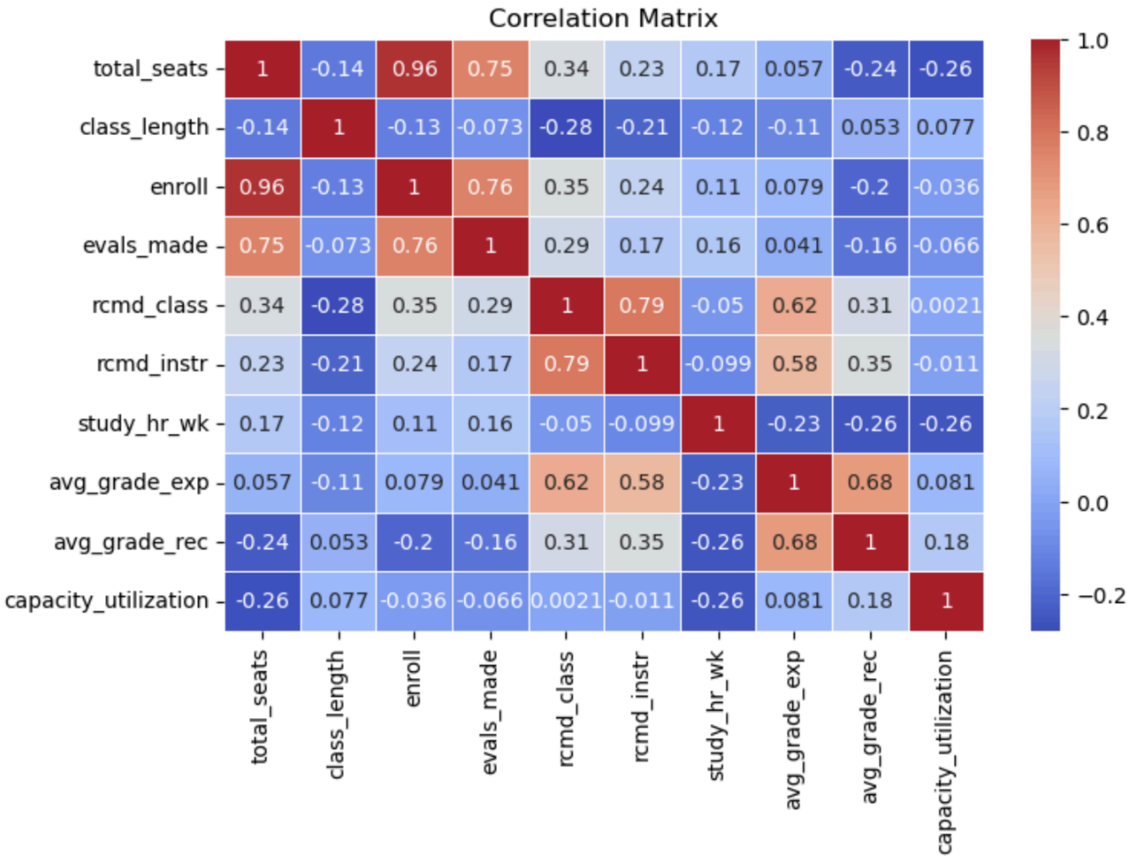
Analysis of Average Grade Received by Class Schedule and Length

The boxplots depicting average grades by class schedule suggest varying academic outcomes across different scheduling formats. Courses held on Tuesdays and Thursdays (TuTh) appear to have a slightly higher median grade, possibly indicating that the spread of classes over fewer, but longer, days per week may facilitate better student performance or engagement. On the contrary, classes scheduled for single days, such as Monday (M), Wednesday (W), and Friday (F), display a wider range of grades with a lower median, which might suggest challenges in maintaining consistency or depth in course delivery.

Analysis of Capacity Utilization by Class Schedule and Length

Capacity utilization across different class schedules shows a broad uniformity in median values, with slight variations that might reflect the preferences or availability of the student body concerning their schedules. Classes scheduled early in the week (Monday and Wednesday) show a denser concentration near full capacity, which might indicate a student preference for beginning the week with more intensive schedules. Interestingly, classes that meet less frequently during the week, such as those only on Tuesday (Tu) or Thursday (Th), have a broader range, suggesting that these might either be specialized courses with limited demand or alternatively, highly popular courses that reach full capacity quickly.

3.3 Correlation Matrix Analysis



	total_seats	class_length	enroll	evals_made	rcmd_class	rcmd_instr	study_hr_wk	avg_grade_exp	avg_grade_rec	capacity_utilization
total_seats	1	-0.14	0.96	0.75	0.34	0.23	0.17	0.057	-0.24	-0.26
class_length	-0.14	1	-0.13	-0.073	-0.28	-0.21	-0.12	-0.11	0.053	0.077
enroll	0.96	-0.13	1	0.76	0.35	0.24	0.11	0.079	-0.2	-0.036
evals_made	0.75	-0.073	0.76	1	0.29	0.17	0.16	0.041	-0.16	-0.066
rcmd_class	0.34	-0.28	0.35	0.29	1	0.79	-0.05	0.62	0.31	0.0021
rcmd_instr	0.23	-0.21	0.24	0.17	0.79	1	-0.099	0.58	0.35	-0.011
study_hr_wk	0.17	-0.12	0.11	0.16	-0.05	-0.099	1	-0.23	-0.26	-0.26
avg_grade_exp	0.057	-0.11	0.079	0.041	0.62	0.58	-0.23	1	0.68	0.081
avg_grade_rec	-0.24	0.053	-0.2	-0.16	0.31	0.35	-0.26	0.68	1	0.18
capacity_utilization	-0.26	0.077	-0.036	-0.066	0.0021	-0.011	-0.26	0.081	0.18	1

The correlation matrix visualizes the relationships between various variables related to class structure and academic outcomes. There is a strong positive correlation (0.96) between **'total_seats'** and **'enroll'**, indicating that classes designed to accommodate more students generally have higher enrollments. This relationship suggests that course capacity is typically well-utilized at UCSD.

Another significant correlation (0.75) exists between 'total_seats' and 'evals_made', suggesting that larger classes tend to have more evaluations completed, likely due to the higher number of students. On the other hand, **'capacity_utilization'** shows a negative correlation with **'total_seats'** (-0.26) and 'enroll' (-0.20), indicating that larger classes, despite their higher enrollment numbers, may not always be proportionately filled to their designated capacity. Furthermore, 'avg_grade_rec' (average grade received) has a slightly negative correlation with **'total_seats'** (-0.24) and **'enroll'** (-0.16), saying that larger classes could be associated with slightly lower average grades.

4 Statistical Analysis

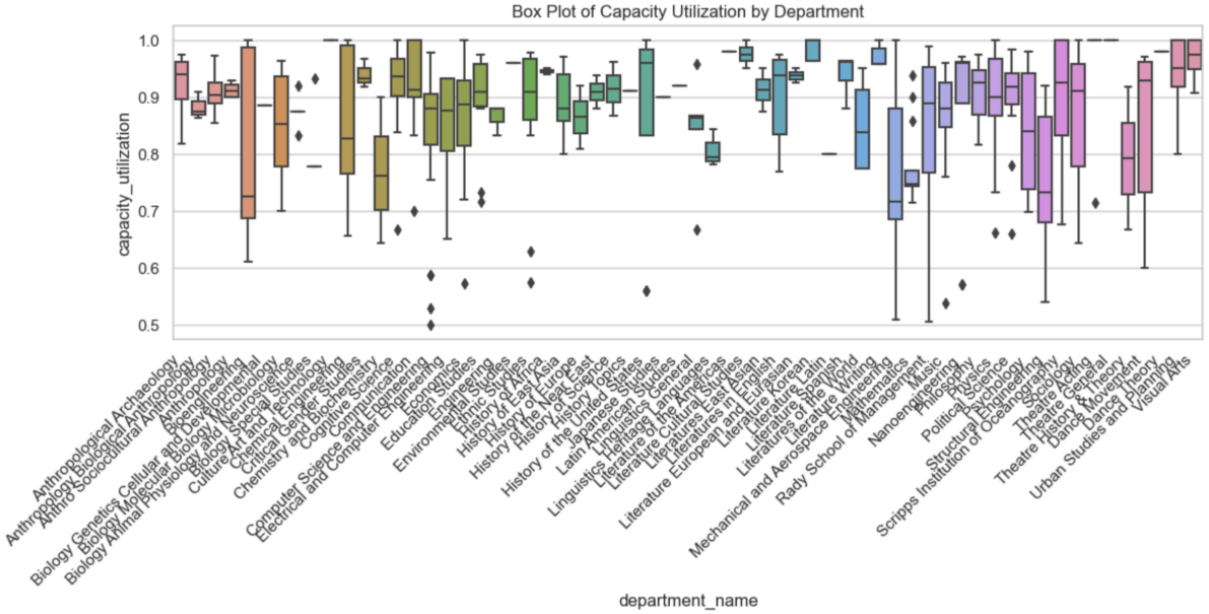
ANOVA Test for Differences in Capacity Utilization Across Departments:
To assess whether there is a statistically significant variation in capacity utilization across different departments at UC San Diego, we employ ANOVA (Analysis of Variance):

- Null Hypothesis (H0): There is no significant difference in capacity utilization across departments.
- Alternative Hypothesis (Ha): There is a significant difference in capacity utilization across departments.

Test Statistics: We use the ANOVA F-statistic as our test statistic for this analysis. The ANOVA F-statistic is particularly suited for comparing the means of more than two groups simultaneously by analyzing the variance within each group compared to the variance between the groups.

Significance Level: We set a significance level at 5% for this test. If the p-value is less than 0.05, we reject the null hypothesis, indicating that at least one department has a significantly different capacity utilization rate compared to others.

ANOVA Test Execution: We will compute the F-statistic and p-value using the 'capacity_utilization' across different 'department_name' groups in our data.



Based on the box plot above, the barplot of capacity utilization by the department from our EDA, as well as the ANOVA results from our code, we noted that F-statistic = 4.65, p-value = 0.000000. The ANOVA (Analysis of Variance) test was conducted to determine if there are statistically significant differences in capacity utilization among the various departments at UCSD. The F-statistic from the ANOVA results is approximately 4.65, and the p-value is exceedingly small indicating strong statistical evidence.

Interpretation:

- F-statistic: The F-statistic value of 4.65 suggests that the between-group variability (differences in capacity utilization among departments) is larger than the within-group variability (variations within each department). This difference in variances is statistically significant.

- P-value: A p-value of 0.000000 signifies that the observed data is highly unlikely under the null hypothesis. In the context of our analysis, this implies a strong rejection of the null hypothesis that capacity utilization is consistent across all departments.

Conclusion:
Given the very low p-value we found, we reject the null hypothesis and accept the alternative hypothesis that there are significant differences in capacity utilization across departments. This result prompts further investigation into specific departmental differences and suggests that resource allocation strategies may need to be department-specific to be effective.

5 Linear Regression Model

5.1 Baseline Full Linear Regression Model

The baseline linear regression model serves as the initial step in our analysis to understand the relationship between the dependent variable `avg_grade_rec` (average grade received) and covariates, including both continuous and categorical predictors. **The formula is as below:**

formula = 'avg_grade_rec ~ enroll + evals_made + rcmd_class + rcmd_instr + capacity_utilization + study_hr_wk + avg_grade_exp + C(department_name) + C(class_schedule) + C(class_length)'

This model provides an overview of how covariates contribute to the average grade received in the courses. **The Mean Squared Error (MSE) of this model is 0.03 and R-squared (R²) is 0.702.**

5.2 Model with statistically significant features (P-value threshold: 0.05)

Based on the model summary above, we pitched statistically significant features and built a new model. The coefficient p-value of `rcmd_class` is 0.387 so we delete this covariate and proceed to the test model. However, this did not result in any significant difference in MSE and R-square.

Formula:

formula_pvalue = 'avg_grade_rec ~ enroll + evals_made + rcmd_instr + capacity_utilization + study_hr_wk + avg_grade_exp + C(class_schedule) + C(department_name) + C(class_length)'

The Mean Squared Error (MSE) of this model is 0.03 and R-squared (R²) is 0.702.

5.3 Model Diagnostics

We proceed to perform a quick assumption check over the regression model we have built above, and based on the diagnostics result, we will then perform some extra feature engineering or adjustments.

5.3.1 Check for Multicollinearity using VIF

The model appears to have significant multicollinearity issues, as indicated by the high VIF values for several features. Notably, `total_seats`, `class_length`, `enroll`, and `rcmd_class` exhibit exceptionally high VIF values, which suggest that these features are highly correlated with each other and/or with other features in the model.

This multicollinearity can lead to unreliable coefficient estimates, making it difficult to determine the individual effect of each predictor on the response variable `avg_grade_rec`. While some categorical variables (e.g., `department names` have lower VIF values, the presence of such high multicollinearity among key predictors warrants consideration of remedial measures such as PCA or regularization methods to stabilize the model and improve interpretability, which we will cover later in this report.

VIF Table for the features:

Features	VIF (Variance Inflation Factor)
total_seats	151.410281
class_length	53.911198
enroll	161.099779
evals_made	8.943361
….	….

5.3.2 Re-generate model after using VIF to filter multicollinearity (Threshold: VIF < 10)

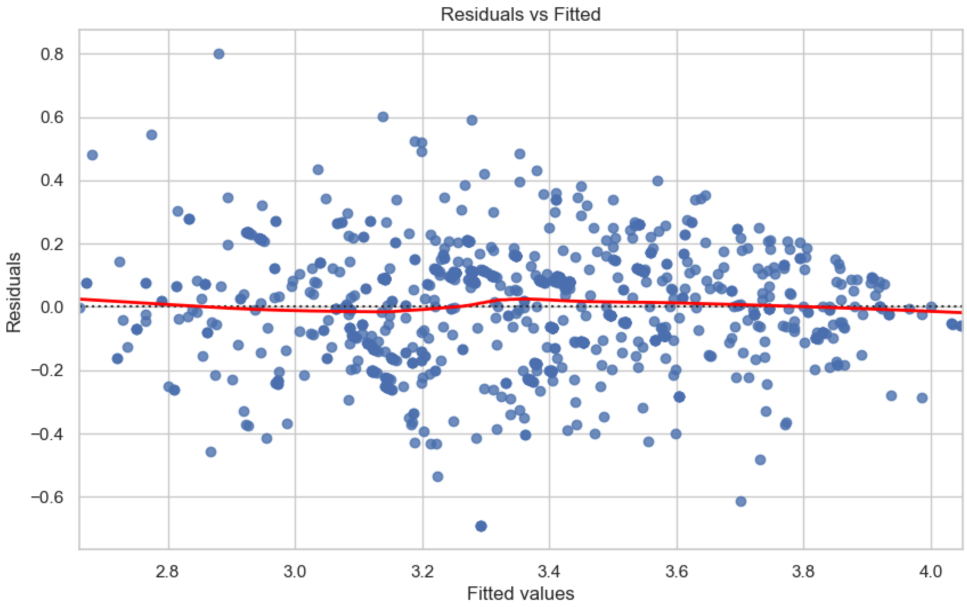
Mean Squared Error (MSE) of the model is 0.05, and R-squared (R^2) is 0.506

After filtering out features with high Variance Inflation Factor (VIF) values, the model's performance has declined significantly. This suggests that while multicollinearity was reduced, the predictive power and accuracy of the model were adversely affected, indicating that some of the removed features were important for explaining the variability in the response variable.

Nonetheless, it's important to note that models with high multicollinearity might show good prediction power on the training data due to overfitting, but they often perform poorly on unseen data. **Therefore,**

balancing multicollinearity reduction and maintaining sufficient predictive power is crucial for developing a robust model that generalizes well to new data.

5.3.3 Check Linearity/homoscedasticity with residual plot



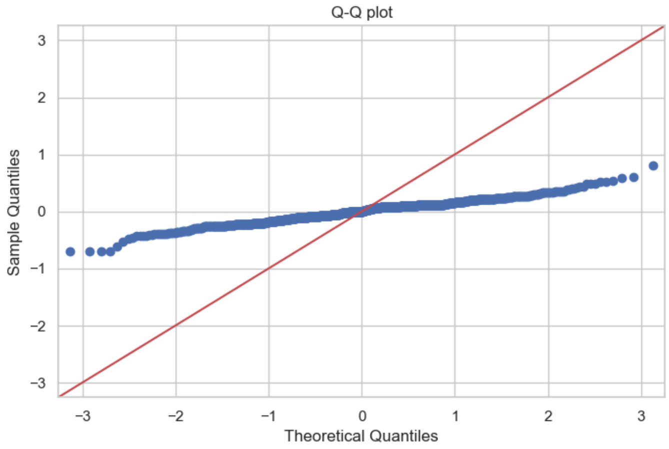
- **Linearity assumption analysis:**

The residual plot shows residuals scattered around the horizontal axis (zero line), which generally suggests that the model's assumption of linearity is reasonably met. **However, there appears to be a slight curve in the red Lowess line, particularly around the fitted values of 3.2 to 3.4.** This suggests a potential mild non-linearity in the relationship between the predictors and the response variable.

- **Homoscedasticity assumption analysis:**

Homoscedasticity means that the residuals should have constant variance across all levels of fitted values. **In this plot, the spread of the residuals appears to be somewhat consistent across the range of fitted values.**

5.3.4 Check the normality of residuals with QQ plot and Shapiro-wilk test



The Q-Q plot reveals deviations from the normal distribution, particularly at the tails, where the left tail shows lighter tails and the right tail shows heavier tails than expected under normality. These deviations suggest that the residuals are not normally distributed, which can affect the validity of statistical inferences, such as confidence intervals and p-values for the coefficients. Addressing this non-normality, possibly through data transformations or robust regression methods, would improve the model's reliability.

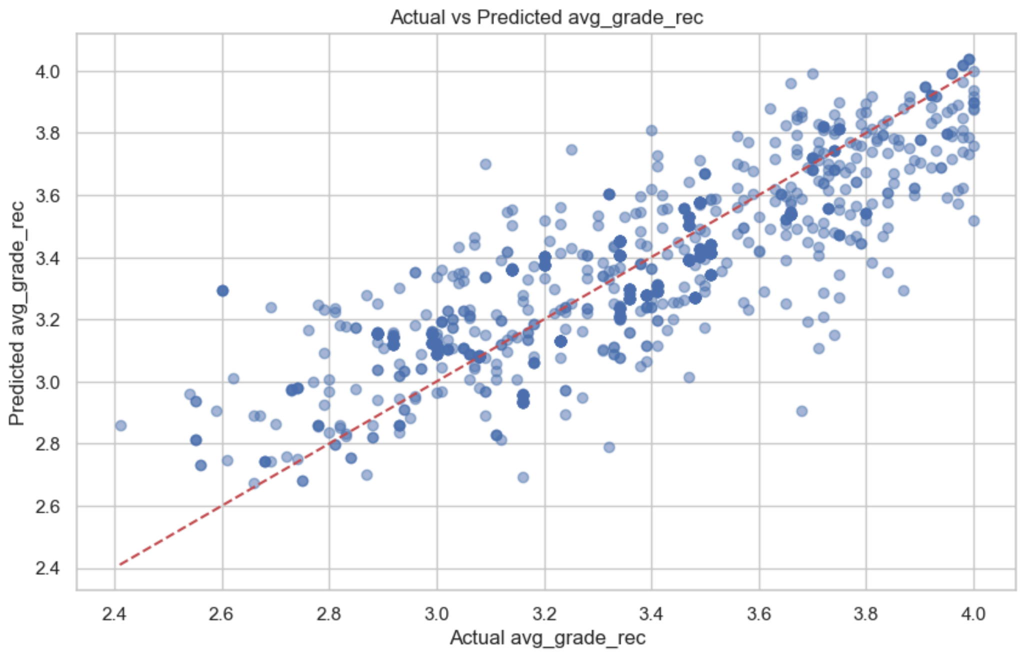
- **Shapiro-Wilk Test Statistic is 0.984, and p-value is 7.22658710650137e-10**

Despite the test statistic being close to 1, the extremely small p-value indicates that the residuals are not normally distributed according to the Shapiro-Wilk test, which aligns with the conclusion we get from the QQ plot above.

5.4 Backward feature selection based on BIC criterion

The backward selection process has yielded a model with a final criterion (BIC) value of -218.60 compared to the initial criterion of -211.99. The selected features include `avg_grade_exp`, `study_hr_wk`, `evals_made`, `class_length`, `department_name`, `capacity_utilization`, `enroll`, and `rcmd_instr`. This set of features suggests that these variables are the most significant predictors for the response variable avg_grade_rec, providing a balance between model complexity and goodness-of-fit.

We then proceed to fit a linear regression model with the feature we selected,
Formula:
formula3 = 'avg_grade_rec ~ evals_made + C(department_name) + C(class_length) + enroll + capacity_utilization + rcmd_instr + avg_grade_exp + study_hr_wk'



The model, utilizing features selected through backward selection based on BIC, demonstrates a reasonable performance with an R-squared (R^2) value of 0.695, indicating that approximately 69.5% of the variance in the response variable is explained by the model. The Mean Squared Error (MSE) of 0.0315 reflects the average squared difference between the observed actual outcomes and the outcomes predicted by the model. The scatter plot of actual versus predicted values shows a strong linear relationship, suggesting that the model predictions are closely aligned with the actual data.

5.5 Perform Ridge/Lasso regression on our model

The **Ridge regression model with α = 0.5 produced a Mean Squared Error (MSE) of 0.0307 and an R-squared (R^2) value of 0.703.** This indicates that the model explains approximately 70.3% of the variance in the response variable `avg_grade_rec`. The relatively low MSE suggests that the predictions are quite accurate, and the higher R^2 value demonstrates a strong fit of the model to the data. The chosen regularization parameter α = 0.5 seems to balance well between fitting the training data and preventing overfitting.

The **Lasso regression model with α = 0.05 yielded a Mean Squared Error (MSE) of 0.0340 and an R-squared (R^2) value of 0.671.** This shows that the model explains approximately 67.1% of the variance in the response variable `avg_grade_rec`. The higher MSE compared to the model with α = 0.5 suggests that the predictions are less accurate. The decrease in R^2 indicates that the model fit is not as strong. The smaller regularization parameter α = 0.05 provides less penalization, which might have led to overfitting or not sufficiently reducing the impact of multicollinearity among predictors.

6. Interpretation of Results (Ridge regression as final model)

The results of the Ridge regression model provide valuable insights into the factors influencing the average grade received in courses. The first seven coefficients, representing **`total_seats`, `enroll`, `evals_made`, `rcmd_class`, `rcmd_instr`**, **`capacity_utilization`**, and **`study_hr_wk`,** reveal both positive and negative relationships with the dependent variable, **`avg_grade_rec`.**

For instance, the negative coefficient for **`total_seats`** (-0.079486) indicates that larger class sizes are associated with slightly lower average grades. This finding suggests that as the number of seats in a course increases, the average grade received tends to decrease, possibly due to less individual attention or more challenging class dynamics. On the other hand, the positive coefficient for **`enroll`** (0.051555) implies that higher enrollment numbers are linked to higher average grades. This could reflect that popular courses with higher enrollment attract more motivated students or better instructors, leading to better overall performance.

The coefficients for **`evals_made`** (0.016275), **`rcmd_class`** (0.012235), and `rcmd_instr` (0.018372) are all positive, suggesting that more evaluations and higher recommendation rates are associated with higher average grades. These relationships highlight the importance of student feedback and satisfaction in achieving better academic outcomes. Additionally, the positive coefficient for **`capacity_utilization`** (0.012607) and **`study_hr_wk`** (0.021235) reinforces the notion that efficient utilization of available seats and the amount of time students spend studying are crucial factors in determining average grades.

Overall, the results are presented and easy to understand, with the interpretations being consistent with the analyses performed. The model highlights key factors that contribute to the average grade received, providing actionable insights for educational institutions to enhance their course offerings and improve student performance.

7 Discussion

7.1 Summary of Finding

Our analysis of UCSD's Spring 2022 enrollment data highlighted several factors that impact student grades. We found that larger class sizes tend to have slightly lower average grades, while courses with higher enrollment numbers and positive evaluations from students often see better grades. Additionally, classes that are nearly full and students who spend more time studying generally perform better. These insights suggest that optimizing these factors could help improve academic performance at UCSD.

7.2 Limitations of Our Model and Future Research

- **Non-Normal Distribution of Variables**

  Firstly, it is crucial to highlight that many variables in our dataset, such as capacity_utilization, exhibit non-normal distributions, such as left-skewed distributions. This skewness can potentially impact the accuracy and reliability of our predictive models. This issue is evident in the QQ plots, which show deviations from the expected normal distribution. To address this in future work, we will need to focus on feature engineering techniques to transform these variables into more normal distributions, such as log transformations or other normalization methods.

- **Potential Multicollinearity**

  Despite efforts to mitigate multicollinearity by calculating Variance Inflation Factors (VIF) and removing highly collinear variables, some multicollinearity may persist among the predictors. This can affect the stability and interpretability of the regression coefficients. Future analyses could explore more advanced regularization techniques, such as LASSO or Ridge regression, to further address this issue.

- **Interaction and Non-Linear Terms**

  Our analysis primarily focused on linear relationships between the predictors and the response variable. However, real-world data often involves interaction effects and non-linear relationships. Future models should consider including interaction terms and non-linear terms, such as quadratic terms, to better capture the complexity of these relationships. For example, the impact of study hours on grades might differ depending on class size, and such interactions should be explored.