Final Report

Investigating High School Location's Impact on Enrollment Rate and the Correlation Between Acceptance and Enrollment Rates to UC San Diego

Problem Statement

The UC System, a network of ten campuses across California (nine of which offer undergraduate degrees), is one of the largest and most prestigious public university systems globally. Each year, the UC system receives hundreds of thousands of applications from a diverse pool of prospective scholars, including in-state, out of state, and international students. Although UC San Diego's reputation for academic excellence and accessibility is undeniable, in order to determine the factors influencing the acceptance of admission offers, there is a need to investigate whether the admission and yield rates for students of a certain demographic may differ from those of another relatively homogenous group. For example, where an applicant grows up has been said to be a factor in admission due to differing yield rates based on proximity to the university's campus. Therefore, applicants to UCSD can be categorized based on this variable by splitting them into in-state and out-of-state applicants. This project aims to assess the relationship between a student's location and their probability of enrollment conditional on acceptance within the vast pool of applicants to UCSD, the second-most applied to university in both the system and nation as a whole.

Motivations and Significance

Due to admissions departments not informing applicants of the exact reasons as to why they were accepted or not, prospective college freshmen are left in the dark as to the importance of various characteristics used in the decision-making process happening behind closed doors.

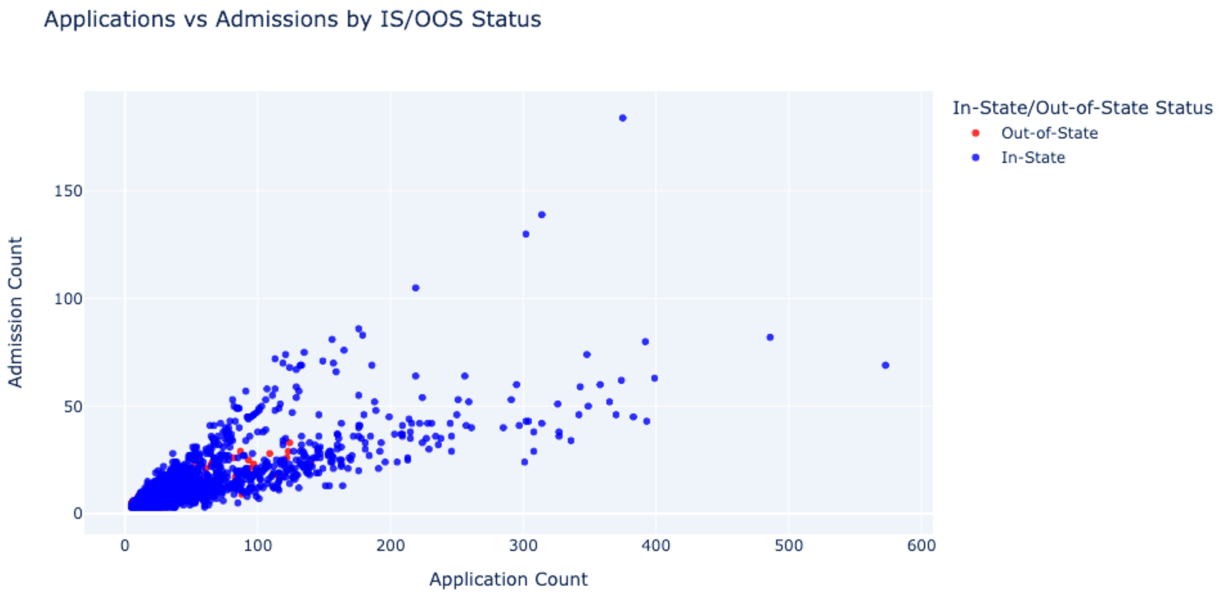
Data Sourcing

The UC system's official website offers all of the necessary data and much more. UCSD-specific data from 2023 can be pulled exclusively by utilizing the filtering tools offered on their page.

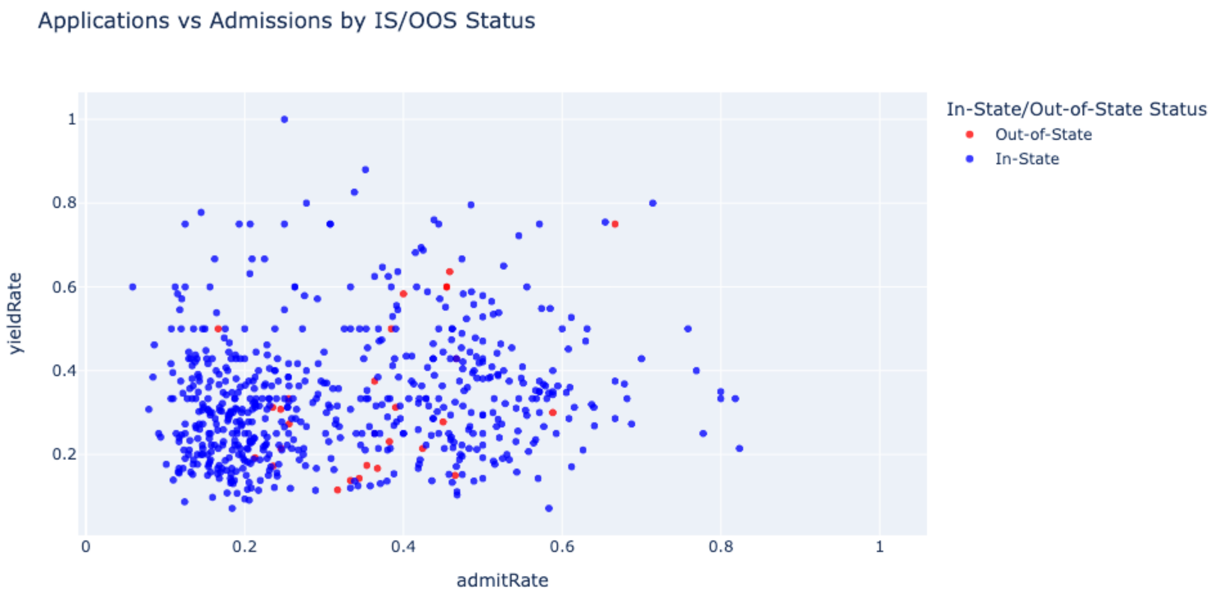
Description of the Data

We curated two separate DataFrames with all the data needed by utilizing the filtering tools available on the UC system's site and deriving some new columns ourselves with Python and Pandas. The first DataFrame has each American high school (2759 of them) as the unique identifier with columns including "City", "County/State/Country" (California high schools list their respective county, but others list their state), "Enr GPA" (average GPA of admitted students who enrolled), "Adm GPA" (average GPA of all admitted students), and "GPA Diff" (a column we derived by subtracting "Enr GPA" from "Adm GPA"). The second DataFrame (after also having been cleaned) uses each American high school (6916 of them) as the unique identifier as well, followed by columns including "City", "County/State/Country" (following the same rules as that of the previous DataFrame), "App Ct" (number of applicants, but is missing if the true value is less than five for the sake of privacy, "Adm Ct" (number of admitted students), "Enr Ct" (number of admitted students who enrolled), and "IS/OOS" (a binary variable column we derived that tells if the high school is within California or not). The second DataFrame contains nearly three times as many high schools because the first excludes those with less than five applicants.

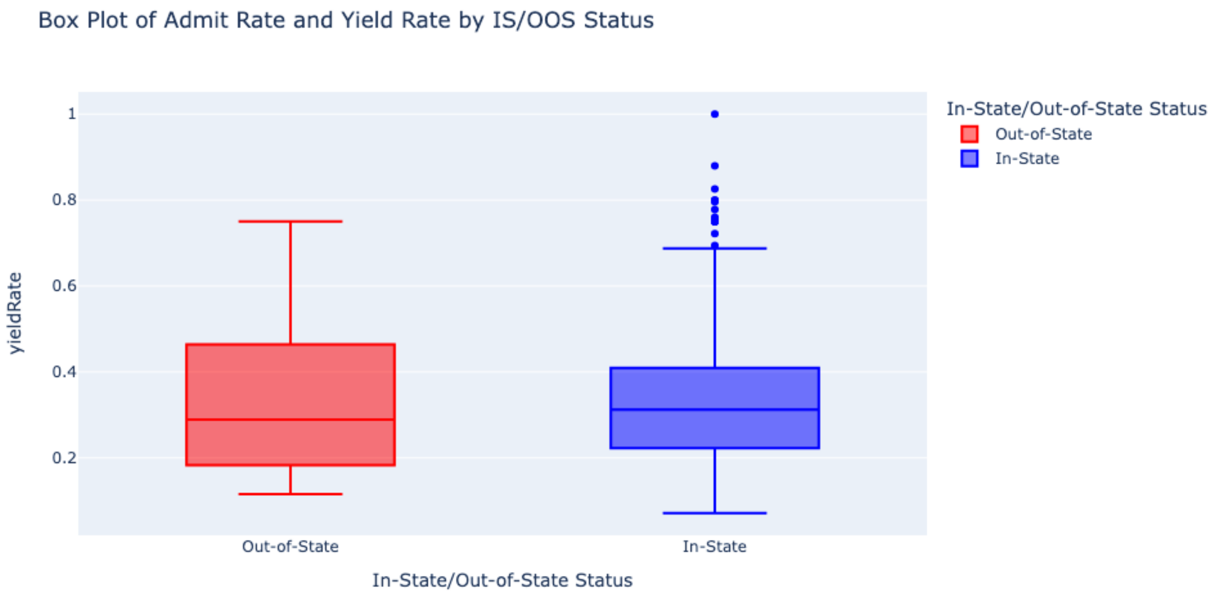
Exploratory Data Analysis



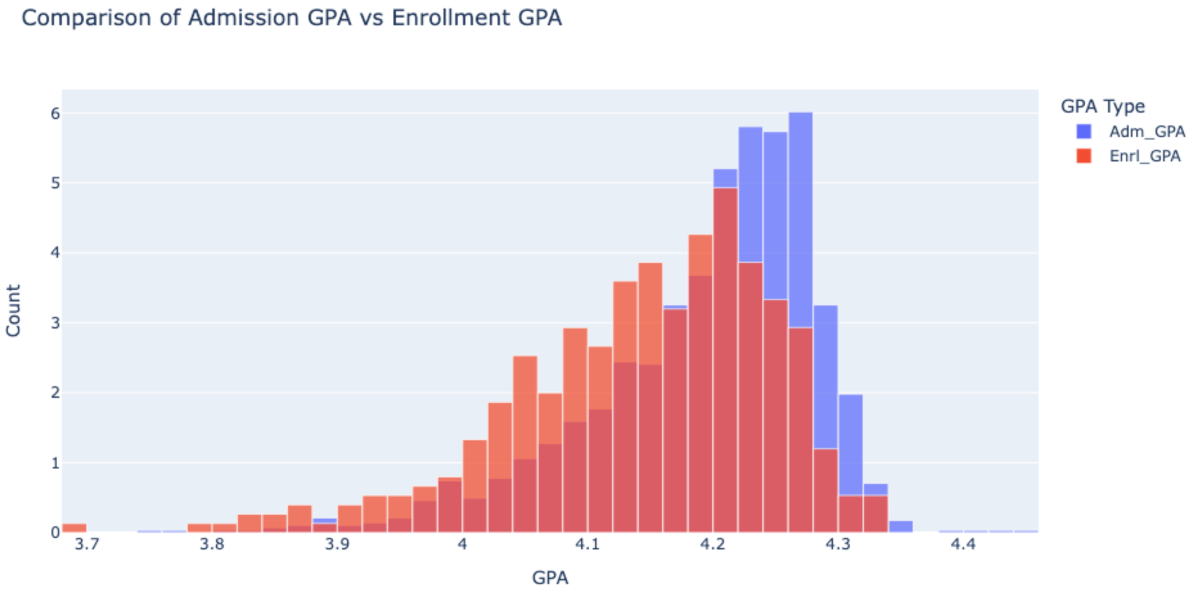
The above scatterplot shows there is a roughly linear relationship between the number of students applying to UCSD and the number of those who are admitted for a given high school, regardless of if it is in-state or not. However, it also reveals that there are a lot more data points for a particular group (In-State), which makes it difficult to discern that pattern. This visualization motivates the need to compare the proportions or rates rather than the count, as it may not be the best parameter for the yield rate.



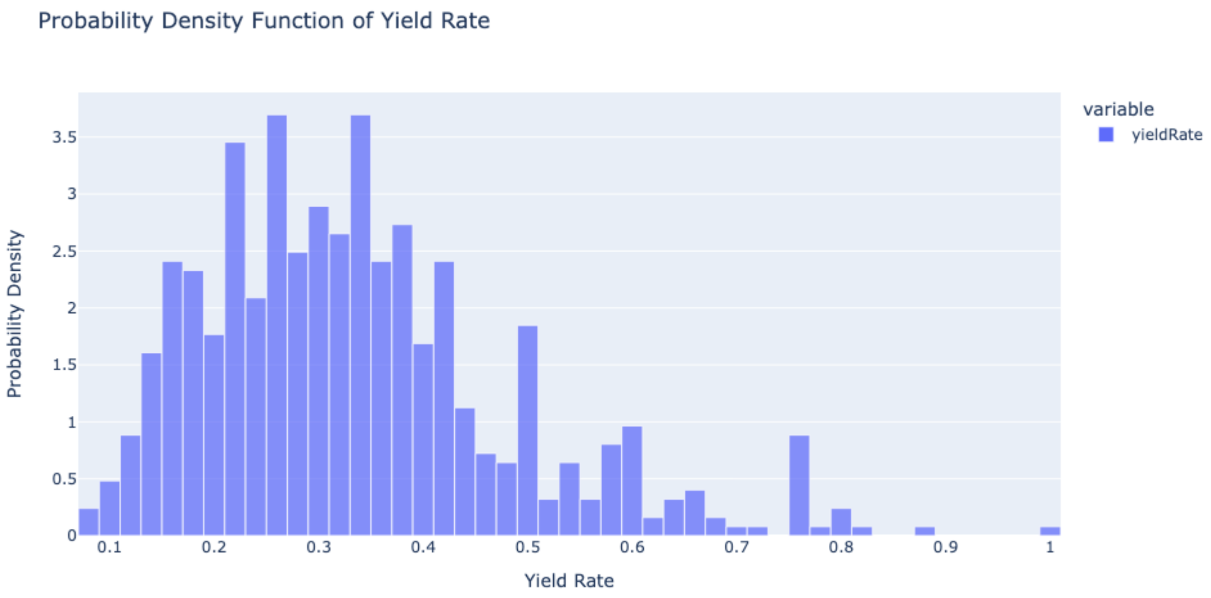
The above scatterplot shows no obvious relationship between a high school's admit rate and yield rate, regardless of if it's in-state or out-of-state. Looking at the proportions reveals a relationship substantially different than the one between Admission and Application Counts.



The above box plot shows In-State high schools have a slightly higher median yield rate than that of Out-of-State high schools, but they have a smaller IQR and more significant extrema. Both distributions seem to be right-skewed.



The above histogram shows the distributions of admitted GPA and enrolled GPA have a similar shape (both being left-skewed), but the enrolled GPA has a noticeably lower mean GPA.



The above probability density function shows a roughly normal distribution with a slight right skew. Yield rates can be seen to range from 0% to 100%, but a majority of the area under the curve lies between 15% and 45%. The function resembles a Chi-squared distribution very closely

References and Analyses

The UC admissions public dataset is vast, covering numerous features for each institution (as it relates to both universities and high schools) to focus on. Other researchers have used the publicly available admissions data to make different conclusions about admission rate. One example of this is “Were Minority Students Discouraged from Applying to University of California Campuses after the Affirmative Action Ban?” (Antonovics and Backes 2013), which focuses on applications of minority students to UCLA and Berkeley. Although many other studies use this publicly available data, we could not find any focusing purely on high school locations and yield rates as it specifically applies for UCSD, meaning no prior studies are relevant for our statistical analysis.

Analyses

We performed three kinds of statistical analyses on the datasets, each with a different approach to analyzing if there is a statistically significant difference between the yield rate of in-state and out-of-state high schools. Our first analysis is a hypothesis test that analyzes the difference in means of the yield rates between high schools within California and those that are not. Our second analysis is linear regression, comparing two models, one of which with a set of covariates that includes the school's location, and one that does not. Our third analysis is a logistic regression model, harnessing the predictive power of yield rate, along with other variables, in regards to a school being in-state or out-of-state. All of these analyses point towards how a high school being in-state or out-of-state influences its yield rate for UC San Diego.

Interpretations

Test 1 : Difference in Means

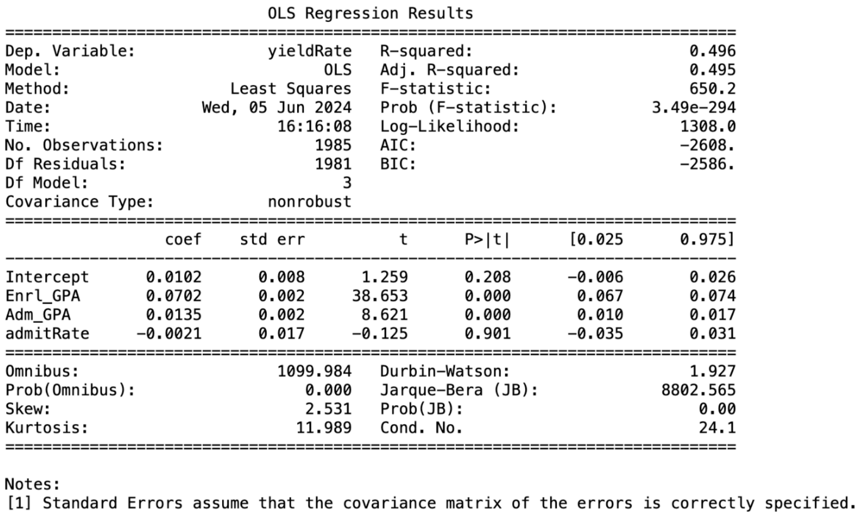
For our first difference in means test, the p-value (of about $1.1 \cdot 10^{-110}$) is so many times smaller than 0.05 that the null hypothesis stating the in-state and out-of-state yield rates are the same can be easily rejected.

Part of why this p-value could be so drastically low is likely attributable to a combination of how the null values are imputed and the specific value we imputed both of them with, possibly artificially expanding the gap between the average in-state high school yield rate and average out-of-state high school yield rate. For high schools with enrollment counts less than three, the admissions website has their value listed as null (for the sake of privacy). This means we are assuming these schools have an enrollment rate of zero percent by using 0 for the imputation even though they could have an enrollment rate as high as 100% (for example, two students could have applied and be granted admission, but the school's admit count would be considered null and then imputed as zero). This would reduce the average for both in-state and out-of-state schools, but would have a much greater impact on the average out-of-state school yield rate because a greater proportion of them have null for their yield count (about 99% of out-of state high schools have less than three students enrolling as compared to only about 71% of in-state schools).

One way of correcting this issue is going about the difference in means test in the same manner, but using a more meaningful metric for the imputed value that also adheres to the null hypothesis. An applicable value for this task could be the overall yield rate across all American high schools students. This can be done by dividing the total number of enrolled students by the total number of admitted students (both values can also be found at this UC website page). This overall yield rate (which ends up being about 22%) can be used instead of 0 for the imputed value and aligns with the null hypothesis that assumes the mean yield rate for in-state and out-of-state high schools are equal. This process results in a higher p-value about $4.6 \cdot 10^{-40}$, but one that still allows for apparent rejection of the null hypothesis at a 5% significance level, indicating that there is a significant difference between the mean yield rate of in-state high schools and out-of-state high schools. Due to being higher, this p-value does not reject the null with the same room to spare as the one from the initial test, but was generated by a process that can be deemed more reliable.

Test 2 : Linear Regression

I. Model 1 : yieldRate ~ Enrl_GPA + Adm_GPA + admitRate



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              yieldRate   R-squared:                       0.496
Model:                            OLS   Adj. R-squared:                  0.495
Method:                 Least Squares   F-statistic:                     650.2
Date:                Wed, 05 Jun 2024   Prob (F-statistic):          3.49e-294
Time:                        16:16:08   Log-Likelihood:                 1308.0
No. Observations:                1985   AIC:                            -2608.
Df Residuals:                    1981   BIC:                            -2586.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0102      0.008      1.259      0.208      -0.006       0.026
Enrl_GPA       0.0702      0.002     38.653      0.000       0.067       0.074
Adm_GPA        0.0135      0.002      8.621      0.000       0.010       0.017
admitRate     -0.0021      0.017     -0.125      0.901      -0.035       0.031
==============================================================================
Omnibus:                     1099.984   Durbin-Watson:                   1.927
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8802.565
Skew:                           2.531   Prob(JB):                         0.00
Kurtosis:                      11.989   Cond. No.                         24.1
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Model Fit and Statistical Significance

- **R-squared**: The R-squared value of 0.496 indicates that approximately 49.6% of the variance in the yield rate is explained by our model. This suggests that the model has moderate explanatory power, making it a reasonably good fit for understanding the factors influencing yield rates.
- **Adjusted R-squared**: With a value of 0.495, the adjusted R-squared is slightly lower than the R-squared, reflecting adjustments for the number of predictors used in the model. This slight reduction confirms that the model's robustness is maintained despite the inclusion of multiple predictors.
- **F-statistic**: The F-statistic of 650.2 is a measure of the joint significance of the predictor variables in the model. This high value suggests that the model is statistically significant.

- **Prob (F-statistic)**: The probability associated with the F-statistic is exceedingly low (3.49e-294), indicating a statistically significant overall model fit. This suggests that the observed relationships in the model are highly unlikely to be due to random variation alone.

Coefficients Analysis

- **Intercept**: The model intercept is 0.0102 with a p-value of 0.208. This result suggests that if all predictor variables are zero, the expected yield rate would be approximately 0.0102, though this intercept is not statistically significant.
- **Enrl_GPA**: This coefficient is 0.0702 and is highly significant (p < 0.0001). It indicates a positive and strong relationship between enrolled GPA and yield rate, where each unit increase in enrolled GPA is associated with an increase of 0.0702 in the yield rate.
- **Adm_GPA**: The coefficient for Adm_GPA is 0.0135, also statistically significant (p < 0.0001). This reflects a positive relationship between admitted GPA and yield rate, albeit with a smaller effect compared to enrolled GPA.
- **AdmitRate**: The coefficient for admitRate is -0.0021 with a p-value of 0.901, indicating a non-significant negative relationship with the yield rate. This suggests that higher admission rates may be associated with slightly lower yield rates, although this effect is not statistically significant and should be interpreted with caution.

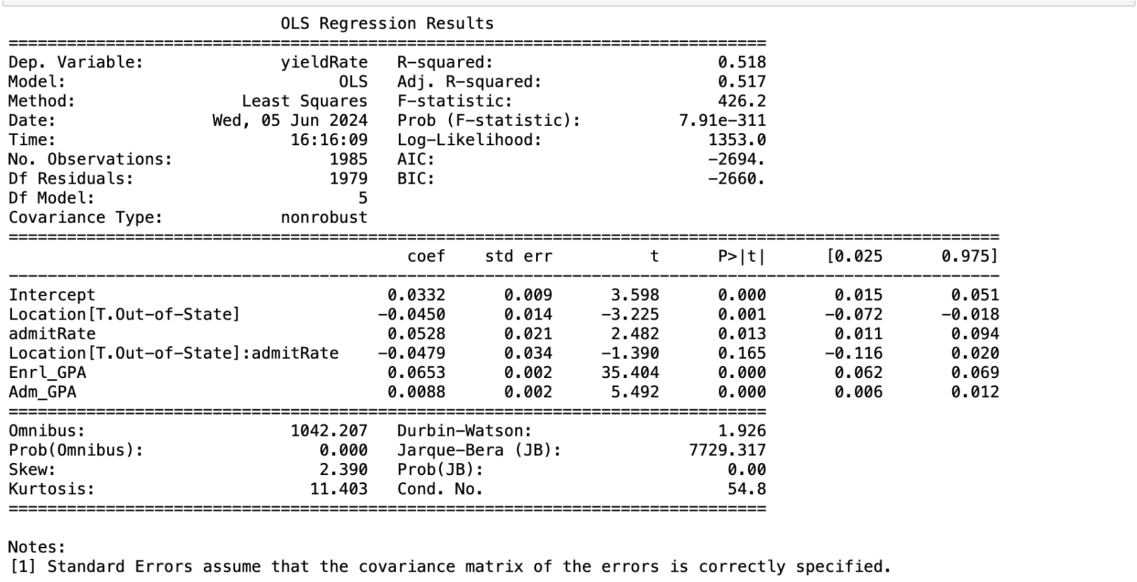
Model Diagnostics

- **Durbin-Watson**: The Durbin-Watson statistic of 1.927 suggests minimal autocorrelation among the residuals, supporting the independence assumption required for linear regression.
- **Omnibus/Prob(Omnibus)**: The significant Omnibus test (p = 0.000) indicates that the residuals of the model are not normally distributed.
- **Jarque-Bera (JB)/Prob(JB)**: Similarly, the Jarque-Bera test is significant (p = 0.00), confirming non-normal distribution of residuals. This highlights potential issues with using linear regression assumptions and may affect the reliability of coefficient estimates and overall model inferences.
- **Skew and Kurtosis**: The skewness of 2.531 and kurtosis of 11.989 further confirm the presence of outliers or extreme values, suggesting a heavy-tailed distribution which deviates from normality.

Implications and Considerations

The model's ability to explain nearly half of the variance in yield rates primarily through GPA metrics highlights the importance of academic performance as a predictor of yield rate. However, the non-normality of residuals suggests potential violations of OLS assumptions, which might impact the validity of hypothesis testing and confidence intervals. Exploring data transformations or including additional variables could potentially address these issues and improve the model's fit and interpretability.

II. Model 2 : yieldRate ~ Enrl_GPA + Adm_GPA + admitRate + Location*admitRate



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              yieldRate   R-squared:                       0.518
Model:                            OLS   Adj. R-squared:                  0.517
Method:                 Least Squares   F-statistic:                     426.2
Date:                Wed, 05 Jun 2024   Prob (F-statistic):          7.91e-311
Time:                        16:16:09   Log-Likelihood:                 1353.0
No. Observations:                1985   AIC:                            -2694.
Df Residuals:                    1979   BIC:                            -2660.
Df Model:                           5
Covariance Type:            nonrobust
==================================================================================================
                                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                              0.0332      0.009      3.598      0.000       0.015       0.051
Location[T.Out-of-State]              -0.0450      0.014     -3.225      0.001      -0.072      -0.018
admitRate                              0.0528      0.021      2.482      0.013       0.011       0.094
Location[T.Out-of-State]:admitRate    -0.0479      0.034     -1.390      0.165      -0.116       0.020
Enrl_GPA                               0.0653      0.002     35.404      0.000       0.062       0.069
Adm_GPA                                0.0088      0.002      5.492      0.000       0.006       0.012
==============================================================================
Omnibus:                     1042.207   Durbin-Watson:                   1.926
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7729.317
Skew:                           2.390   Prob(JB):                         0.00
Kurtosis:                      11.403   Cond. No.                         54.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Model Fit and Statistical Significance

- **R-squared**: The model has an R-squared of 0.518, which indicates that approximately 51.8% of the variability in the yield rate is explained by the variables included in the model. This represents a slight improvement over the previous model (49.6%).
- **Adjusted R-squared**: With a value of 0.517, the adjusted R-squared is very close to the R-squared, suggesting that there is minimal overfitting despite the addition of more variables into the model.
- **F-statistic**: The F-statistic is 426.2, indicating that the model is statistically significant. This statistic tests whether at least one predictor in the model has a non-zero coefficient, suggesting a robust overall model significance.

- **Prob (F-statistic)**: With a value of 7.91e-311, the probability associated with the F-statistic underscores the statistical significance of the model, confirming that the relationships are not due to chance.

Coefficients Analysis

- **Intercept** : The intercept of the model is 0.0332, statistically significant ($p < 0.001$), setting a baseline yield rate when all predictor variables are at their reference levels.
- **Location[T.Out-of-State]** : The coefficient of -0.0450, significant at the $p = 0.001$ level, suggests that being from out-of-state is associated with a lower yield rate compared to in-state (the likely reference category).
- **admitRate** : The coefficient for admitRate is 0.0528, significant ($p = 0.013$), indicating a positive relationship between the admit rate and the yield rate.
- **Location[T.Out-of-State] :** The interaction term has a coefficient of -0.0479, which is not statistically significant ($p = 0.165$). This suggests that the influence of admit rate on yield rate differs for out-of-state students compared to in-state students, though this difference is not robustly supported by the data.
- **Enrl_GPA** : With a coefficient of 0.0653 and a highly significant p-value ($p < 0.0001$), this indicates that a higher enrolled GPA strongly correlates with an increased yield rate.
- **Adm_GPA** : The coefficient of 0.0088, also statistically significant ($p < 0.0001$), shows a positive but less impactful relationship between admitted GPA and yield rate compared to enrolled GPA.

Model Diagnostics

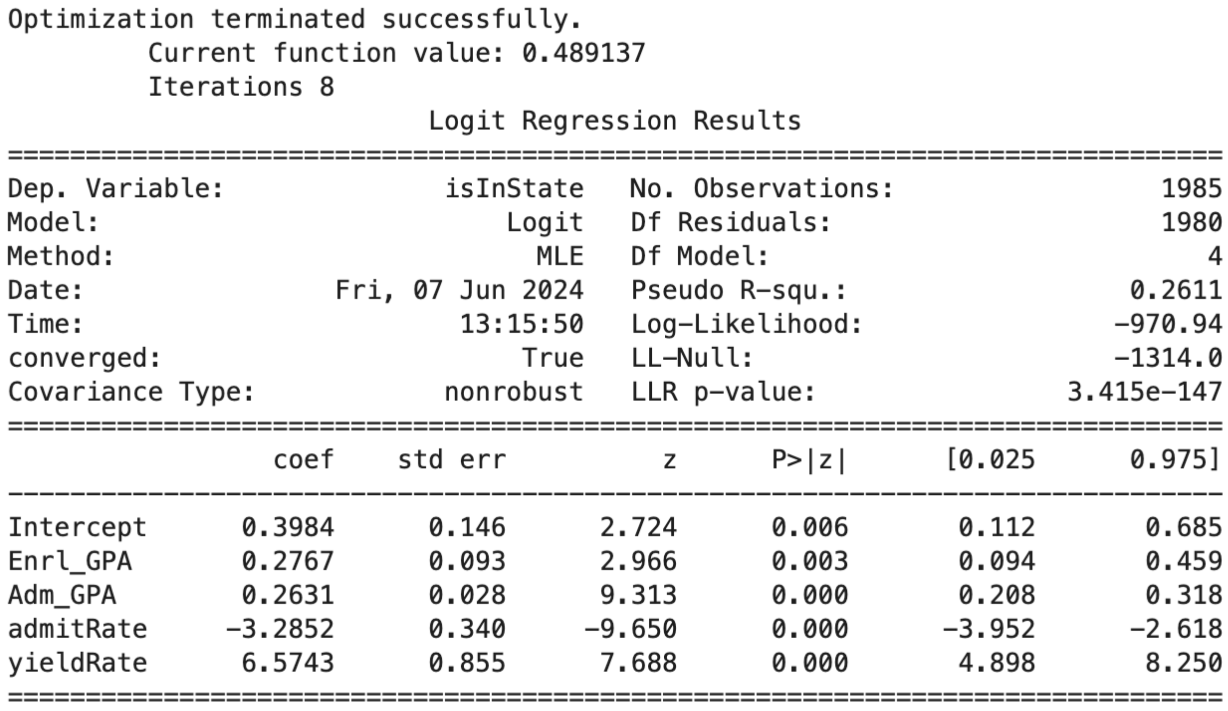
- **Durbin-Watson**: A statistic of 1.926 suggests there is little to no autocorrelation among the residuals, validating the independence of observations, which is crucial for the reliability of the regression analysis.
- **Omnibus/Prob(Omnibus)** and **Jarque-Bera (JB)/Prob(JB)**: Both tests indicate significant non-normality in the residuals (p = 0.000 for Omnibus and p = 0.00 for JB). These results suggest that the assumptions necessary for OLS may not be fully met.
- **Skew and Kurtosis**: Skewness of 2.390 and kurtosis of 11.403 reveal that the residuals are not symmetrically distributed and exhibit heavy tails, indicating the presence of outliers or extreme values.

Implications and Considerations

The introduction of location factors and interaction terms has slightly enhanced the explanatory power of the model but also added complexity, particularly in interpreting the effects where interaction terms are not significant. The significant coefficients for GPA metrics reaffirm their strong influence on yield rate predictions. However, the evident non-normality of residuals raises concerns about the underlying assumptions of OLS, suggesting that further analysis using alternative modeling techniques or transformations might be necessary to provide more reliable predictions and inferences.

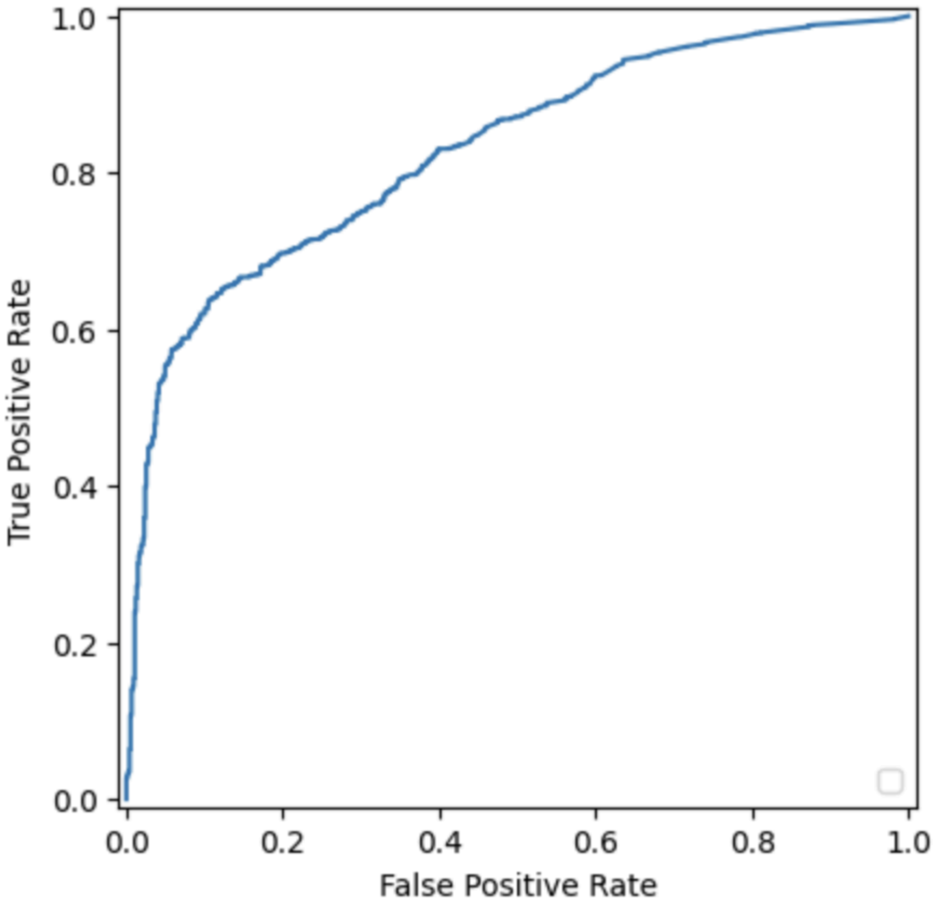
Test 3 : Logistic Regression

For our logistic regression model, the results show that admit GPA, enrollment GPA, admit rate and yield rate are all statistically significant factors for predicting whether a high school was in-state or out-of-state to an arbitrary alpha of 0.05. This is evident by observing the p-values of the model. Looking at the coefficients, we can analyze how each variable affects what the model predicts as in-state. The coefficients show how much more likely the high school identified is in-state, for an infinitesimal change in each value. This is measured in an expected increase / decrease in the log-odds for this outcome. For example, log-odds of being in-state goes up similarly for each change in GPA value. The big conclusion that can be drawn from this model lies in the yield rate coefficient, which is the value that has been most of interest in this analysis. This value is so much higher than the other coefficients, showing that there is an extreme increase in the log-odds of a school being in-state for a slight increase in yield rate. This illustrates a high dependence on yield rate in this logistic regression model predicting in-state schools compared to out-of-state schools.



```
Optimization terminated successfully.
         Current function value: 0.489137
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:              isInState   No. Observations:                 1985
Model:                          Logit   Df Residuals:                     1980
Method:                           MLE   Df Model:                            4
Date:                Fri, 07 Jun 2024   Pseudo R-squ.:                  0.2611
Time:                        13:15:50   Log-Likelihood:                -970.94
converged:                       True   LL-Null:                       -1314.0
Covariance Type:            nonrobust   LLR p-value:                3.415e-147
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.3984      0.146      2.724      0.006       0.112       0.685
Enrl_GPA       0.2767      0.093      2.966      0.003       0.094       0.459
Adm_GPA        0.2631      0.028      9.313      0.000       0.208       0.318
admitRate     -3.2852      0.340     -9.650      0.000      -3.952      -2.618
yieldRate      6.5743      0.855      7.688      0.000       4.898       8.250
==============================================================================
```

The predictive power of the logistic regression model also shows useful results. With an accuracy of 0.7365, this model predicts whether a school is in California far better than chance, with only four variables. Looking at the confusion matrix, this accuracy value is helped a lot by its high number of True Negatives, which in this case is when the model can determine a school is not in California, mostly due to the yield rate. The Receiver Operating Characteristic (ROC) curve shows similar results, mostly zooming in on the model's effectiveness when naming a school as 'in-state' throughout different cut-off values. With an area under the curve value of 0.827, which is far higher than a random chance value of 0.5, this model illustrates that there is evidence towards California high schools showing statistically significantly higher yield rates than schools out of state, where a high school's yield rate to UCSD can reliably predict whether that school is within the borders of California.



Limitations

There are a couple possible limitations with our analysis. One of which is rooted in the source that we gathered our data from. According to the admissions website that our data derived from, 'the number of admits and number of enrollees for any category with fewer than three students are shown as a blank' (University of California Admissions). This means that a good number of high schools with minimal information are missing for privacy reasons. This causes a lot of gaps in the data that might disproportionately affect one category of high schools. This was partially addressed in our analyses for our hypothesis test, but this imputation is not a foolproof solution. Another shortcoming of our analysis is the lack of conclusion for causation, as we cannot determine causation with observational statistical tests. For example, in the case of our difference in means analysis specifically, we know there is a difference in yield rates between in-state high schools and out-of-state ones, but we do not know the reasons leading to said difference.

Conclusion

This study sought to determine the relationship between the geographical location of high schools and their yield rates for UC San Diego. Through various statistical analyses—namely, hypothesis testing, linear regression, and logistic regression—our research highlights a pronounced disparity in the yield rates between in-state and out-of-state high schools, underscoring the strong influence geographic location has on enrollment decisions. The results of the hypothesis tests unequivocally reject the null hypothesis, suggesting a substantial difference in yield rates, while our regression analyses strengthen these findings by quantifying the impact of various covariates including GPA and admit rate. Particularly, the logistic regression model successfully illustrated how yield rates can predict the likelihood of a high school being in-state with notable accuracy.

Despite robust findings, our study encounters limitations inherent to the data's granularity and the absence of causal inference due to its observational nature. These restrictions suggest the necessity for incorporating more granular data in further research and potentially experimental designs to better understand what underpins these geographical disparities such as differences in tuition and financial aid, proximity to one’s hometowns, interest in attending UCSD in particular (as compared to other UCs for example), and other factors.

In conclusion, the insights derived from this study provide valuable implications for admissions strategies at UC San Diego and potentially other universities within the UC system. By recognizing the influence of geographical factors on enrollment, it is conceivable that universities tailor their outreach and admission strategies to better accommodate and possibly predict applicant behaviors based on their origin. Thus, over the course of this comprehensive inferential analysis, recognizing these disparities allows for a more transparent enrollment process for aspiring high school students, enabling them to navigate through the seemingly ambiguous process.