UNIVERSITY CASE STUDY

Analyzing Institutional and Demographic Factors Influencing College Applications Rates

Qin Liu, Marisol Arita
Stat 311 – Regression Analysis

Introduction

The objective of this study is to explore the factors influencing college graduation rates across various institutions in the United States. By examining a dataset of 777 universities and colleges, we aim to uncover relationships between acceptance rates and institutional, demographic, and financial characteristics. These findings will provide insights into how different attributes contribute to application rates.

We will also explore how variations in college affordability, represented by tuition and associated costs, influence application rates. Furthermore, we will assess the role of faculty qualifications—measured by the percentage of faculty with Ph.D.'s—in shaping number of applications.

This research leverages statistical methods to build models addressing these questions, enabling the identification of key predictors and their significance. Through this investigation, we aim to provide actionable insights for institutions seeking to enhance application rates and improve student experiences.

Data and Methods

DATA

We used a dataset of demographic characteristics and tuition information from 777 universities and colleges across the United States. The entries were provided in a csv file. We performed all statistical analysis using the regression analysis software JMP.

These are the first 24 data entries that show an example of what the csv file contains:

	UniversityName	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
1	Abilene Christi…	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
2	Adelphi Univer…	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
3	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
4	Agnes Scott C…	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
5	Alaska Pacific …	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15
6	Albertson Colle…	Yes	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55
7	Albertus Magn…	Yes	353	340	103	17	45	416	230	13290	5720	500	1500	90	93	11.5	26	8861	63
8	Albion College	Yes	1899	1720	489	37	68	1594	32	13868	4826	450	850	89	100	13.7	37	11487	73
9	Albright College	Yes	1038	839	227	30	63	973	306	15595	4400	300	500	79	84	11.3	23	11644	80
10	Alderson-Broa…	Yes	582	498	172	21	44	799	78	10468	3380	660	1800	40	41	11.5	15	8991	52
11	Alfred University	Yes	1732	1425	472	37	75	1830	110	16548	5406	500	600	82	88	11.3	31	10932	73
12	Allegheny Coll…	Yes	2652	1900	484	44	77	1707	44	17080	4440	400	600	73	91	9.9	41	11711	76
13	Allentown Coll.…	Yes	1179	780	290	38	64	1130	638	9690	4785	600	1000	60	84	13.3	21	7940	74
14	Alma College	Yes	1267	1080	385	44	73	1306	28	12572	4552	400	400	79	87	15.3	32	9305	68
15	Alverno College	Yes	494	313	157	23	46	1317	1235	8352	3640	650	2449	36	69	11.1	26	8127	55
16	American Inter…	Yes	1420	1093	220	9	22	1018	287	8700	4780	450	1400	78	84	14.7	19	7355	69
17	Amherst College	Yes	4302	992	418	83	96	1593	5	19760	5300	660	1598	93	98	8.4	63	21424	100
18	Anderson Univ…	Yes	1216	908	423	19	40	1819	281	10100	3520	550	1100	48	61	12.1	14	7994	59
19	Andrews Unive…	Yes	1130	704	322	14	23	1586	326	9996	3090	900	1320	62	66	11.5	18	10908	46
20	Angelo State U…	No	3540	2001	1016	24	54	4190	1512	5130	3592	500	2000	60	62	23.1	5	4010	34
21	Antioch Univer…	Yes	713	661	252	25	44	712	23	15476	3336	400	1100	69	82	11.3	35	42926	48
22	Appalachian St…	No	7313	4664	1910	20	63	9940	1035	6806	2540	96	2000	83	96	18.3	14	5854	70
23	Aquinas College	Yes	619	516	219	20	51	1251	767	11208	4124	350	1615	55	65	12.7	25	6584	65
24	Arizona State …	No	12809	10308	3761	24	49	22593	7585	7434	4850	700	2100	88	93	18.9	5	4602	48

VARIABLES

Variables in the dataset:
Apps = Number of applications received
UniversityName= Name of the university
Private = Public/private indicator
Accept = Number of applicants accepted
Enroll = Number of new students enrolled
Top10perc = New students from top 10 % of high school class
Top25perc = New students from top 25 % of high school class
F.Undergrad = Number of full-time undergraduates
P.Undergrad = Number of part-time undergraduates
Outstate = Out-of-state tuition
Room.Board = Room and board costs
Books = Estimated book costs
Personal = Estimated personal spending
PhD = Percent of faculty with Ph.D.'s
Terminal = Percent of faculty with terminal degree
S.F.Ratio = Student/faculty ratio
perc.alumni = Percent of alumni who donate
Expend = Instructional expenditure per student
Grad.Rate = Graduation rate

We used **apps** as our response variable (y) and the remaining variables as independent variables to be analyzed and filtered to find the best fit for our model and the appropriate predictor variables (x). Since the **UniversityName** column contains all unique row entries, we concluded that having so many unique levels adds unnecessary complexity, and thus it will not be included as an independent variable in the model. We also performed some data transformation to convert the Top10perc, TOP 25perc, PhD, Terminal, perc.alumni, Grad.Rate numbers to a decimal (50% -> .50 i.e.) and give new variables names as follows:

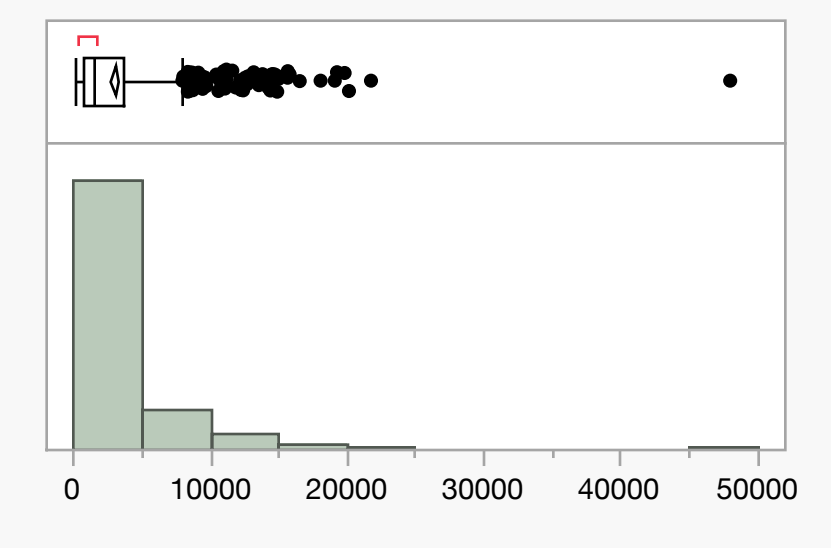
TOP10%, TOP 25%, PHD%, Terminal%, Alum%, Grad Rate% and saved in a new file titled 'University Data_updated.csv'.

Exploratory Data Analysis

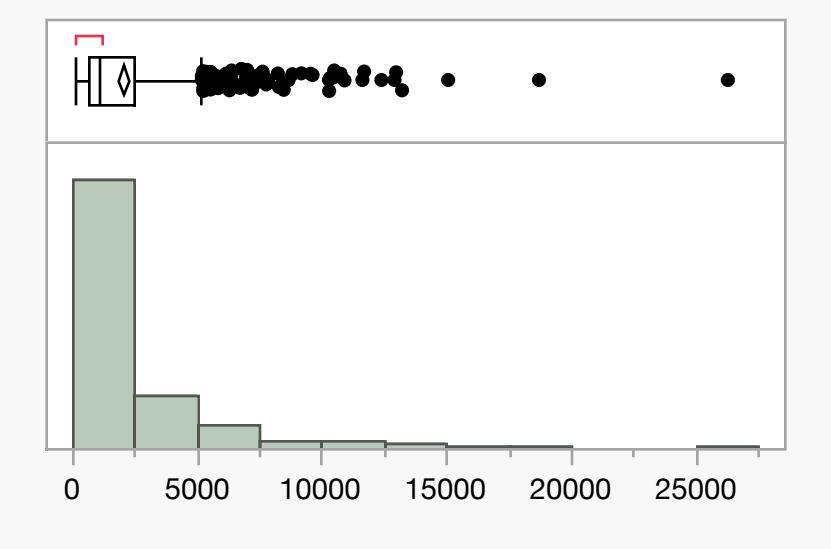
We performed exploratory analysis to see different patterns within the data.

Examples:

Apps



Accept



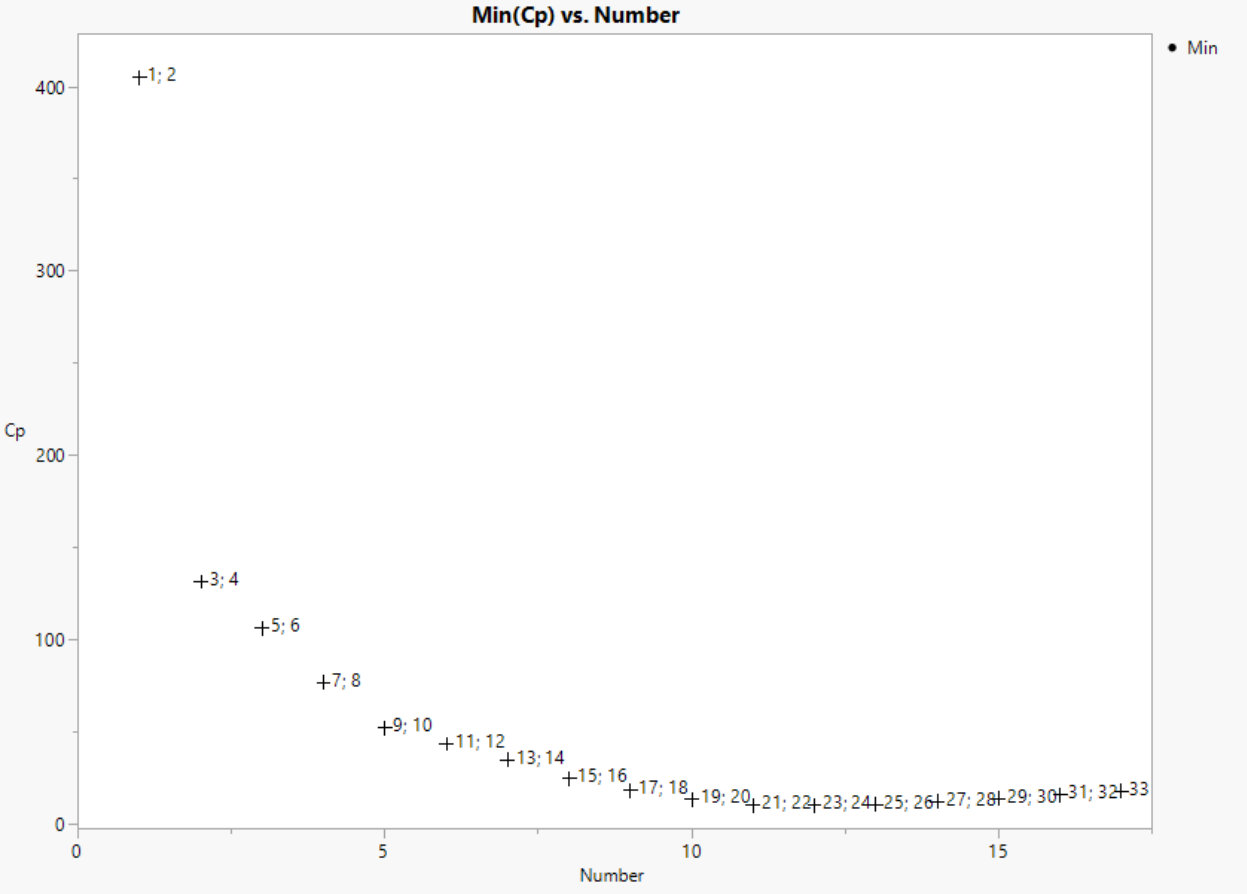
There is one outlier for number of applications and acceptances on row 484 – Rutgers at New Brunswick, which has 48094 applications, and 26330 applicants accepted. These numbers seem high considering the upper end of applications received lie between 10,000 – 20,000 applications for Ivy League Universities such as Yale and University of Pennsylvania, and significantly less acceptances. It seems unlikely 26,000 applicants were admitted as freshman class. We foresee this data entry will be an influential data point but will perform some tests to confirm.

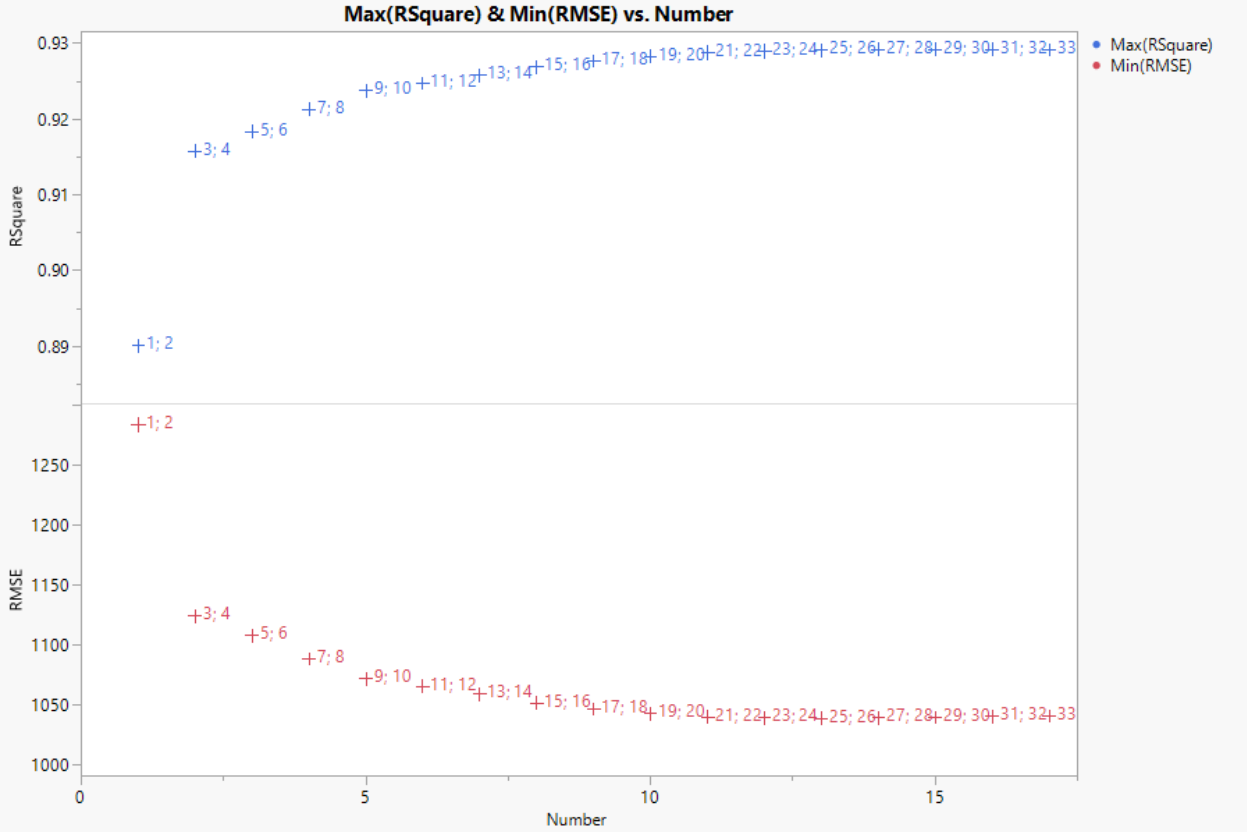
Initial Model:

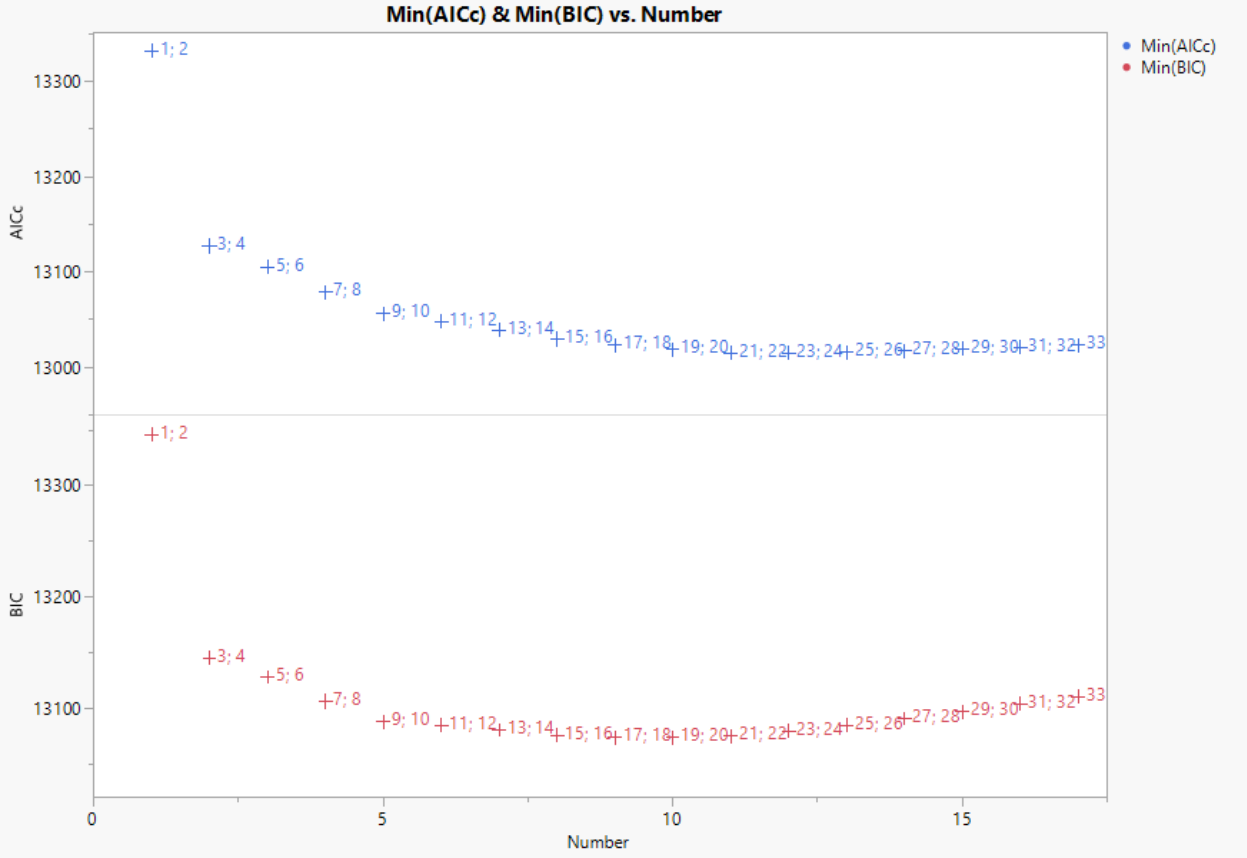
We employed the All-Possible-Regression Selection Method to evaluate the 17 viable variables. This allows us to screen for the important independent variables.

Ordered up to best 2 models up to 17 terms per model.

Model	Number	RSquare	RMSE	AICc	BIC	Cp
Accept	1	0.8901	1283.85	13332.0	13345.9	404.9881
Enroll	1	0.7171	2059.80	14066.6	14080.6	2259.2184
Accept,TOP10%	2	0.9158	1124.58	13127.2	13145.8	131.6805
Accept,Expend	2	0.9106	1158.93	13173.9	13192.5	187.6665
Accept,TOP10%,Expend	3	0.9183	1108.13	13105.3	13128.5	106.3306
Accept,TOP10%,TOP 25%	3	0.9177	1112.11	13110.9	13134.1	112.6218
Accept,TOP10%,Outstate,Expend	4	0.9213	1088.79	13079.0	13106.8	76.9421
Private{Yes-No},Accept,TOP10%,Expend	4	0.9202	1095.97	13089.2	13117.0	88.1154
Accept,Enroll,TOP10%,Outstate,Expend	5	0.9238	1072.08	13056.0	13088.4	52.1894
Accept,TOP10%,Outstate,Room.Board,Expend	5	0.9227	1079.76	13067.1	13099.5	63.9343
Accept,Enroll,TOP10%,Outstate,Room.Board,Expend	6	0.9247	1065.82	13047.9	13084.9	43.6154
Accept,Enroll,TOP10%,TOP 25%,Outstate,Expend	6	0.9247	1065.88	13048.0	13085.0	43.7141
Accept,Enroll,TOP10%,TOP 25%,Outstate,Room.Board,Expend	7	0.9258	1059.27	13039.4	13081.0	34.6983
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,Outstate,Expend	7	0.9257	1060.05	13040.5	13082.2	35.8694
Private{Yes-No},Accept,Enroll,TOP10%,Outstate,Room.Board,PHD%,Expend	8	0.9269	1052.02	13029.7	13076.0	24.8261
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,Outstate,Room.Board,Expend	8	0.9268	1052.88	13031.0	13077.3	26.1057
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,Outstate,Room.Board,PHD%,Expend	9	0.9277	1046.89	13023.2	13074.0	18.1927
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,Outstate,Room.Board,Terminal%,Expend	9	0.9275	1048.44	13025.5	13076.3	20.4840
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,Outstate,Room.Board,PHD%,Expend,Grad Rate%	10	0.9283	1042.99	13018.4	13073.9	13.4156
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,Outstate,Room.Board,PHD%,Expend	10	0.9281	1044.83	13021.2	13076.6	16.1304
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,Outstate,Room.Board,PHD%,Expend,Grad Rate%	11	0.9288	1040.09	13015.1	13075.2	10.1548
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,P.Undergrad,Outstate,Room.Board,PHD%,Expend,Grad Rate%	11	0.9287	1040.96	13016.4	13076.5	11.4288
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,PHD%,Expend,Grad Rate%	12	0.9290	1039.35	13015.1	13079.7	10.0816
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,Outstate,Room.Board,PHD%,S.F.Ratio,Expend,Grad Rate%	12	0.9289	1039.86	13015.9	13080.5	10.8288
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,PHD%,S.F.Ratio,Expend,Grad Rate%	13	0.9291	1039.10	13015.8	13085.0	10.7113
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,PHD%,Terminal%,Expend,Grad Rate%	13	0.9290	1039.71	13016.7	13085.9	11.6078
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,PHD%,Terminal%,S.F.Ratio,Expend,Grad Rate%	14	0.9292	1039.48	13017.4	13091.2	12.2735
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,Personal,PHD%,S.F.Ratio,Expend,Grad Rate%	14	0.9291	1039.59	13017.6	13091.4	12.4275
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,Personal,PHD%,Terminal%,S.F.Ratio,Expend,Grad Rate%	15	0.9292	1039.98	13019.2	13097.6	14.0093
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,Books,PHD%,Terminal%,S.F.Ratio,Expend,Grad Rate%	15	0.9292	1040.14	13019.5	13097.8	14.2432
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,Books,Personal,PHD%,Terminal%,S.F.Ratio,Expend,Grad Rate%	16	0.9292	1040.66	13021.3	13104.2	16.0019
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,Personal,PHD%,Terminal%,S.F.Ratio,Alum%,Expend,Grad Rate%	16	0.9292	1040.67	13021.3	13104.2	16.0077
Private{Yes-No},Accept,Enroll,TOP10%,TOP 25%,F.Undergrad,P.Undergrad,Outstate,Room.Board,Books,Personal,PHD%,Terminal%,S.F.Ratio,Alum%,Expend,Grad Rate%	17	0.9292	1041.35	13023.4	13110.9	18.0000




Max(RSquare) & Min(RMSE) vs. Number
Max(RSquare)
Min(RMSE)
RSquare
RMSE
Number


Min(AICc) & Min(BIC) vs. Number
Min(AICc)
Min(BIC)
AICc
BIC
Number

We picked the model with one of the highest R^2 values, with one of the lowest Cp values that we believe would create an uncomplicated, yet powerful model. We can see that 11 independent variables should be included in the group as most important variables. By analyzing metrics such as R^2, RMSE, C_p, AIC and BIC, we identified a set of potentially significant predictors prioritizing a balance of strong predictive power without adding unnecessary complexity or the risk of overfitting and developed the initial model as shown below:

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11}$
- $y - Apps$
- $x_1 - \begin{Bmatrix} 1\ if\ private \\ 0\ otherwise \end{Bmatrix}\ base\ is\ not\ private$
- $x_2 - Accept$
- $x_3 - Enroll$
- $x_4 - Top\ 10\%$
- $x_5 - top\ 25\%$
- $x_6 - F.Undergrad$
- $x_7 - Outstate$
- $x_8 - \ Room.Board$
- $x_9 - PHD\ \%$
- $x_{10} - Expend$
- $x_{11} - \ Grad\ Rate\%$

We performed tests to assess the initial fit of the model. F-test to assess whether the regression model explains a significant portion of the variability in the response variable compared to the variability due to error, as well as individual T-tests to assess the validity of each predictor.

F-test

$$H_0: \beta_1 \ldots\ldots \beta_{11} = 0\ vs$$
$$H_a: Some\ \beta_i \neq 0$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	1.0796e+10	981429011	907.2312
Error	765	827565469	1081784.9	Prob > F
C. Total	776	1.1623e+10		<.0001*

A low p-value of < .0001 which is less than a .05 significance level indicates a strong and statistically significant model. A high F ratio also denotes that the model explains more than random error.

Individual T-tests

$$H_0: \beta_i = 0 \; vs$$
$$H_a: \beta_i \neq 0$$

When performing a t-test on each of the variables we find that when using a significance level of .05, all predictor variables we chose have p-values less than a .05 significance level, so we reject the null hypothesis and thus we know they are statistically significant in the model and predicting y.

Indicator Function Parameterization

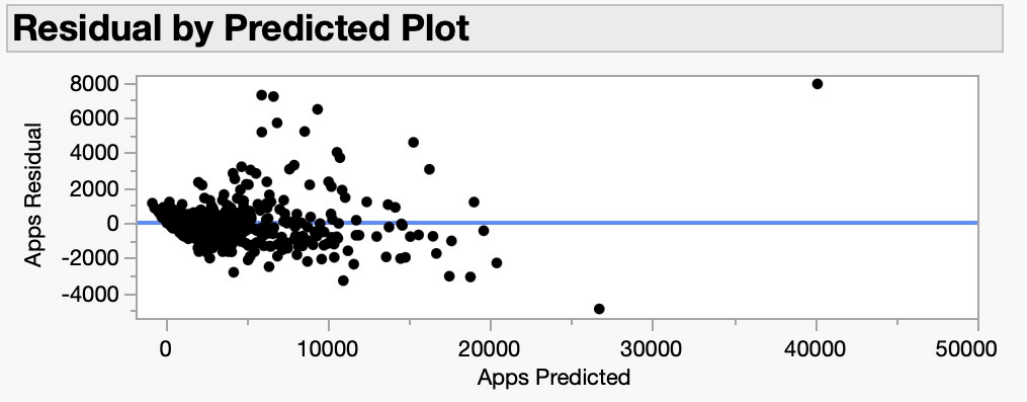
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-134.7296	265.1574	-0.51	0.6115
Private[Yes]	-521.3161	134.2171	-3.88	0.0001*
Accept	1.5834576	0.039997	39.59	<.0001*
Enroll	-0.900946	0.18437	-4.89	<.0001*
TOP10%	4968.8361	550.6257	9.02	<.0001*
TOP 25%	-1471.781	442.0853	-3.33	0.0009*
F.Undergrad	0.0718664	0.031295	2.30	0.0219*
Outstate	-0.090667	0.01832	-4.95	<.0001*
Room.Board	0.1553891	0.046688	3.33	0.0009*
PHD%	-1036.195	311.4193	-3.33	0.0009*
Expend	0.0731212	0.011388	6.42	<.0001*
Grad Rate%	799.6556	282.8119	2.83	0.0048*

Summary of Fit

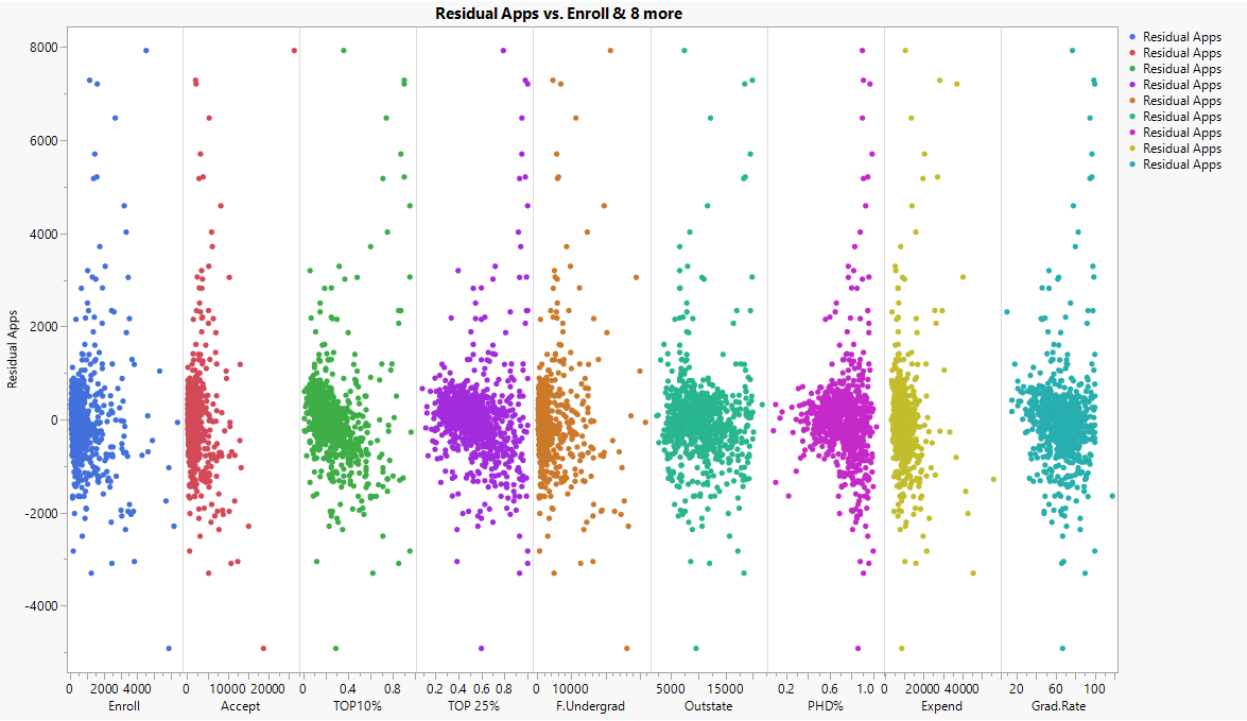
RSquare	0.928801
RSquare Adj	0.927777
Root Mean Square Error	1040.089
Mean of Response	3001.638
Observations (or Sum Wgts)	777

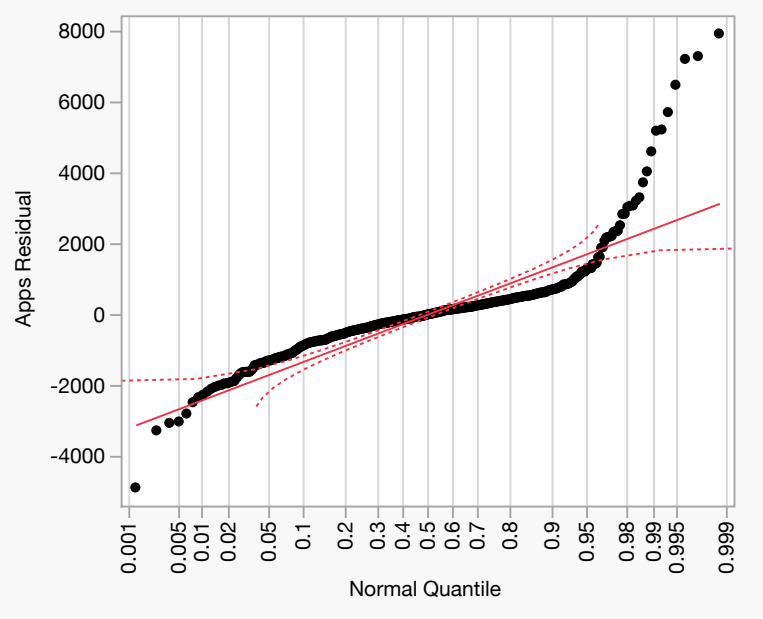
R^2 and R^2 adjusted are 92.88% and 92.78% respectively, meaning 92% of the variability is explained by the model which suggests a strong fit.

RESIDUAL ANALYSIS



We check to make sure the assumptions are met. The residual plot however showed residuals aren't fully random in pattern. Residual variability seems to increase as the predicted values increase. This suggests heteroscedasticity which violates one of the assumptions of linear regression. We can see some outliers that may also be influential in regards to the model and skewing results.





We can see in the last two graphs that there are potentially influential outliers. The residual **apps** vs. x's also suggests there may be some interactions or non-linear relationships we didn't capture with the initial model. In this normal quantile plot the residuals deviate from the line especially at the ends (or tails). This deviation indicates that the residuals are not evenly distributed, which violates an assumption in linear regression.

To fix the problem we excluded two data entries (row 484 and row 462) which are outliers in the model skewing results as can be observed in the residual by predicted plot, as well as row 484 being > 1 and having a very high number when looking at Cook's D influence. We also performed a square root transformation on the response variable (y), to stabilize the model and variance.

Σ	minal%	S.F.Ratio	perc.alumni	Alum%	Expend	Grad.Rate	Grad Rate%	Cook's D Influence Apps
484	0.95	19.5	19	0.19	10474	77	0.77	4.3826097405

Final model:

Our final model is a multiple linear regression model with a square root transformation applied to the response variable (y) which is **apps** or number of applications received by a respective university. We employ a combination of continuous and categorical variables to achieve the best

fit model. We also employed interaction terms and quadratic terms to encapsulate the full complexity of the research question as a simpler regression model would not achieve on its own.

The square root transformation was chosen as a way to stabilize the variance and therefore improve the normality of the residuals as seen below:

$$\sqrt{y} = 11.53 - 6.1x_1 + 0.02x_2 + 0.003x_3 + 24x_4 - 4.5x_5 - 0.00013x_6 - 0.0002x_7 + 0.001x_8 - 2.28x_9 + 0.0002x_{10} + 7.8x_{11} - 0.0000006\,{x_2}^2 - 0.0000005{x_3}^2 - 0.000065x_1x_7 + 0.0002x_1x_{10}$$

- $y - Apps$
- $x_1 - \begin{Bmatrix} 1\ if\ private \\ 0\ otherwise \end{Bmatrix}\ base\ is\ not\ private$
- $x_2 - Accept$
- $x_3 - Enroll$
- $x_4 - Top\ 10\%$
- $x_5 - top\ 25\%$
- $x_6 - F.Undergrad$
- $x_7 - Outstate$
- $x_8 - \ Room.Board$
- $x_9 - PHD\ \%$
- $x_{10} - Expend$
- $x_{11} - \ Grad\ Rate\%$

Summary of Fit

RSquare	0.943075
RSquare Adj	0.94195
Root Mean Square Error	6.376345
Mean of Response	47.11381
Observations (or Sum Wgts)	775

Analysis of Variance

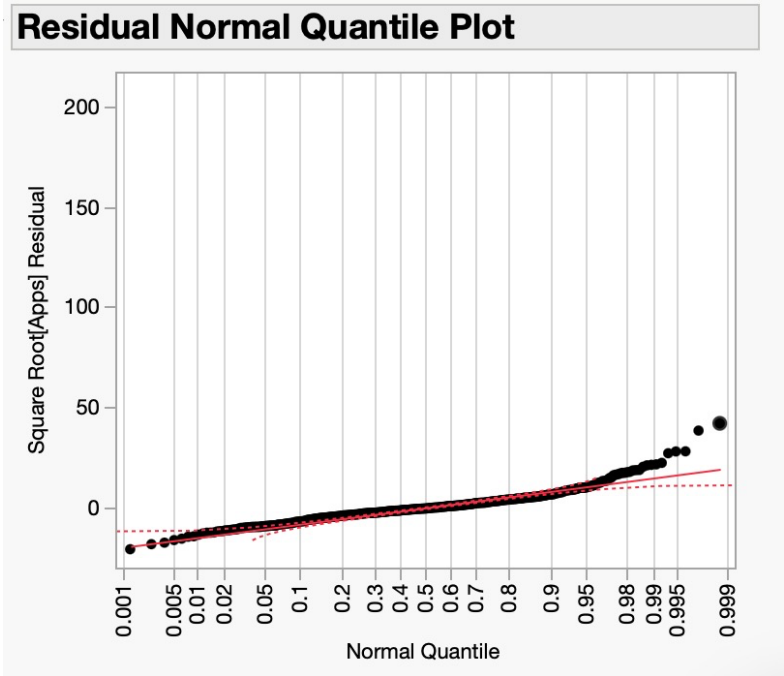
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	15	511239.88	34082.7	838.2815
Error	759	30859.25	40.7	Prob > F
C. Total	774	542099.13		<.0001*

Indicator Function Parameterization				
Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	11.532139	2.279442	5.06	<.0001*
Private[Yes]	-6.110039	1.94222	-3.15	0.0017*
Accept	0.0154622	0.000668	23.14	<.0001*
Enroll	0.0031411	0.001871	1.68	0.0937
TOP10%	24.285941	3.40009	7.14	<.0001*
TOP 25%	-4.487567	2.715212	-1.65	0.0988
F.Undergrad	-0.00013	0.000196	-0.66	0.5063
Outstate	-0.000184	0.000258	-0.71	0.4760
Room.Board	0.0011821	0.00029	4.07	<.0001*
PHD%	-2.279341	1.934173	-1.18	0.2390
Expend	0.0002043	0.000207	0.99	0.3241
Grad Rate%	7.8279204	1.741353	4.50	<.0001*
Accept*Accept	-6.126e-7	4.86e-8	-12.60	<.0001*
Enroll*Enroll	-4.707e-7	2.753e-7	-1.71	0.0877
Private[Yes]*Outstate	-6.535e-5	0.000271	-0.24	0.8093
Private[Yes]*Expend	0.0001871	0.000211	0.89	0.3750

In the summary of fit table, we observe R^2 and R^2 adjusted are both high and have improved since the initial model. At 94.3% and 94.2% when adjusting for number of predictors, they indicate that the variability is mostly explained by the predictors chosen for the model and have a better fit with the standardizing of response variable y. The model in this case has a strong fit. The root mean square error (RMSE) is 6.37 which is low and indicates strong accuracy when it comes to predicting.

When performing individual t-tests we see that Private, Accept, top 10%, Room and Board, Grad Rate are very impactful on the model.

When reading the ANOVA table we observe a low p-value which confirms that the model is statistically significant when predicting the number of applications (y). A higher F ratio also suggests that the predictors (x) have a strong relationship to the response variable y.



The updated residual normal quantile plot has also improved, and the data points are closer to the line and normal distribution.

Conclusion

This study investigated the institutional and demographic factors influencing the number of applications received by U.S. universities. Using a comprehensive dataset and advanced regression techniques, we identified key predictors of application rates, including institutional type, affordability measures (tuition and associated costs), and student-to-faculty ratios. Our final model demonstrated strong predictive power with an adjusted R^2 of 94.2%, indicating that the variability in applications is largely explained by the selected predictors.

The analysis highlights key factors influencing college application trends, such as academic prestige (top 10% of high school class) and affordability (room and board costs). Surprisingly, a higher percentage of faculty with Ph.D.s appears to negatively impact applications, potentially reflecting applicant preferences or perceptions. These insights can help institutions strategically enhance enrollment by balancing costs, emphasizing academic achievements, and addressing perceptions of faculty composition. However, we noted that outliers, such as Rutgers University, significantly impacted the initial analysis, necessitating their exclusion and the use of a transformation to stabilize variance.

In summary, universities aiming to increase application numbers should prioritize affordability, target students comprising of the top 10% of graduating classes, and emphasize university graduation rates. By addressing these factors strategically, institutions can enhance their appeal to prospective students, ultimately driving growth and diversity in their applicant pools.