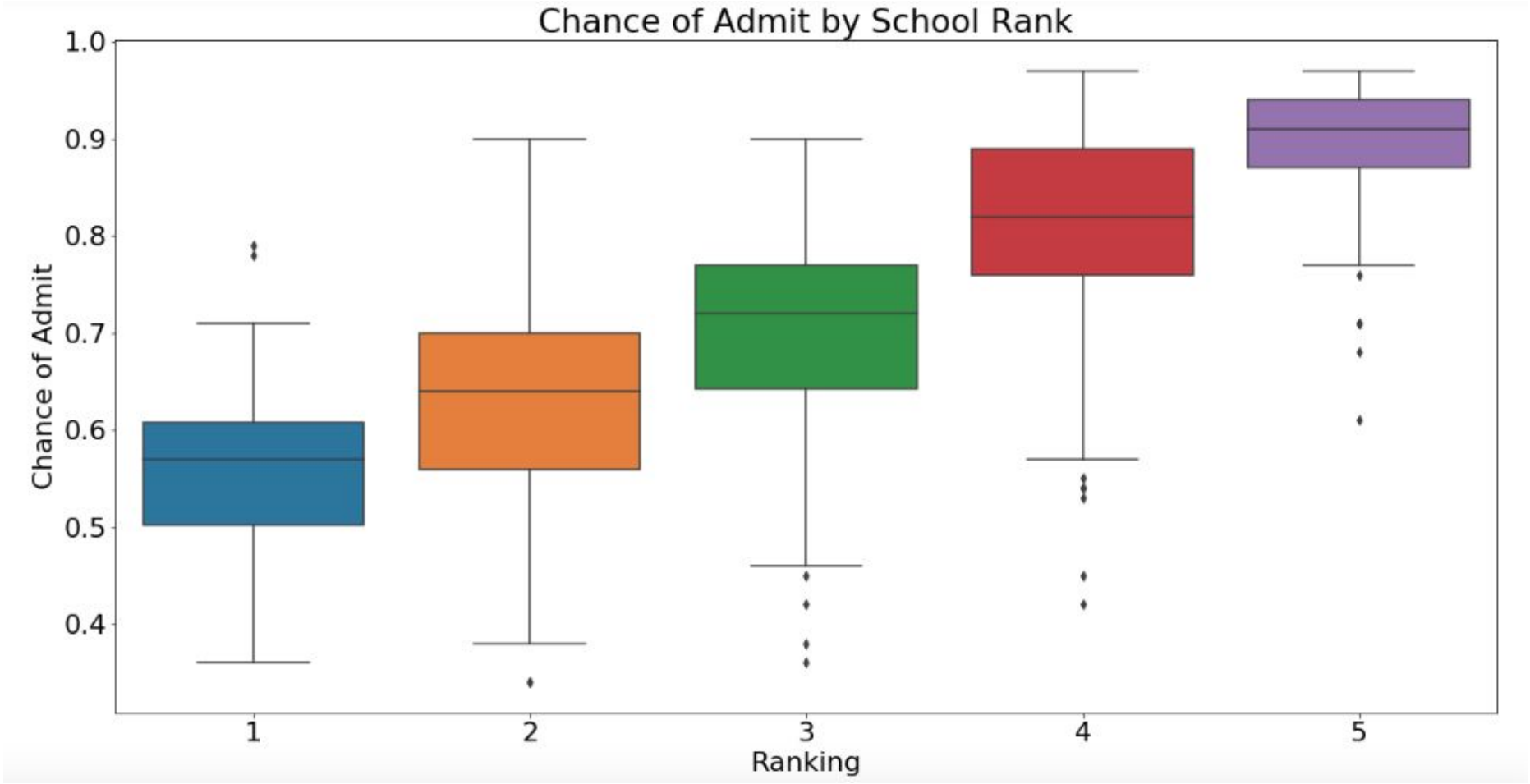
Analyzing Graduate School Admission

Using Supervised Learning Algorithms to Predict Graduate School Admission

The Problem

Every year nearly 2 million students apply to graduate school. A growing number of applicants and shrinking admissions rate among elite programs poses challenges for prospective students.

In this dataset, the overall mean chance of admission is 72%. Yet for students from the lowest ranked schools the mean chance for admission is 56%.


Chance of Admit by School Rank
Chance of Admit
1.0
0.9
0.8
0.7
0.6
0.5
0.4
1
2
3
4
5
Ranking

Research Question

Many academically talented students attend lower ranked schools. Often these students face greater financial challenges and balance work and family responsibilities that impact CGPA.

What are factors that can help undergraduate students from lower ranked schools improve their chances of acceptance into graduate school?

Goal: Increase number of students who see graduate school as a possibility.

It is generally understood that the most **critical factors to get into graduate school** are:

- **Research Experience** Professors love to see research experience, especially with colleagues they are acquainted with and in the field of the candidate's interest. Published papers at important conferences and being a first or second author on an academic publication adds tremendous value.
- **Recommendations** Widely regarded as the most important factor in admissions, good recommendations, especially from Professors who are highly regarded or well known in their field, is instrumental to admission.
- **Undergraduate GPA** Many PhD programs have strict GPA cutoffs of 3.5 or 3.6. Admissions likes to not only see a strong GPA, but advanced graduate level classes that show your expertise.
- **GRE** While GREs are typically dismissed as merely a sanity check or a screening filter that may raise a red flag or two, we investigate how true this is later in the article.
- **Repute of your undergraduate institution** While many like to downplay this factor, it anecdotally seems to make a difference. Perhaps this is because it correlates with not only how smart you may have been to be admitted there, but the strength of the curriculum, professors more likely to be known at other prestigious universities and a better calibration of your GPA and coursework.

http://debarghyadas.com/writes/the-grad-school-statistics-we-never-had/

Influential Data

Outside the scope of the current study

Legacy

A 2009 paper by Dr. Michael Hurwitz, *Harvard Graduate School of Education* concluded:

- The **odds of admission** for application with **legacy status** are **3.3 times** the odds for those without legacy status
- The magnitude of this legacy admission advantage depends greatly on the **nature of familial ties** between the outcome college and the **selectivity of the outcome college**

Potential Impact

- Contribute to improving application process for students in lower ranked schools
- Provide valuable information to applicants concerning factors they should emphasize on their application
- Dispel myths about what factors contribute to graduate school acceptance

The Data

GRE Scores (290 to 340)

TOEFL Scores (92 to 120

University Rating (1 to 5)

Statement of Purpose (1 to 5)

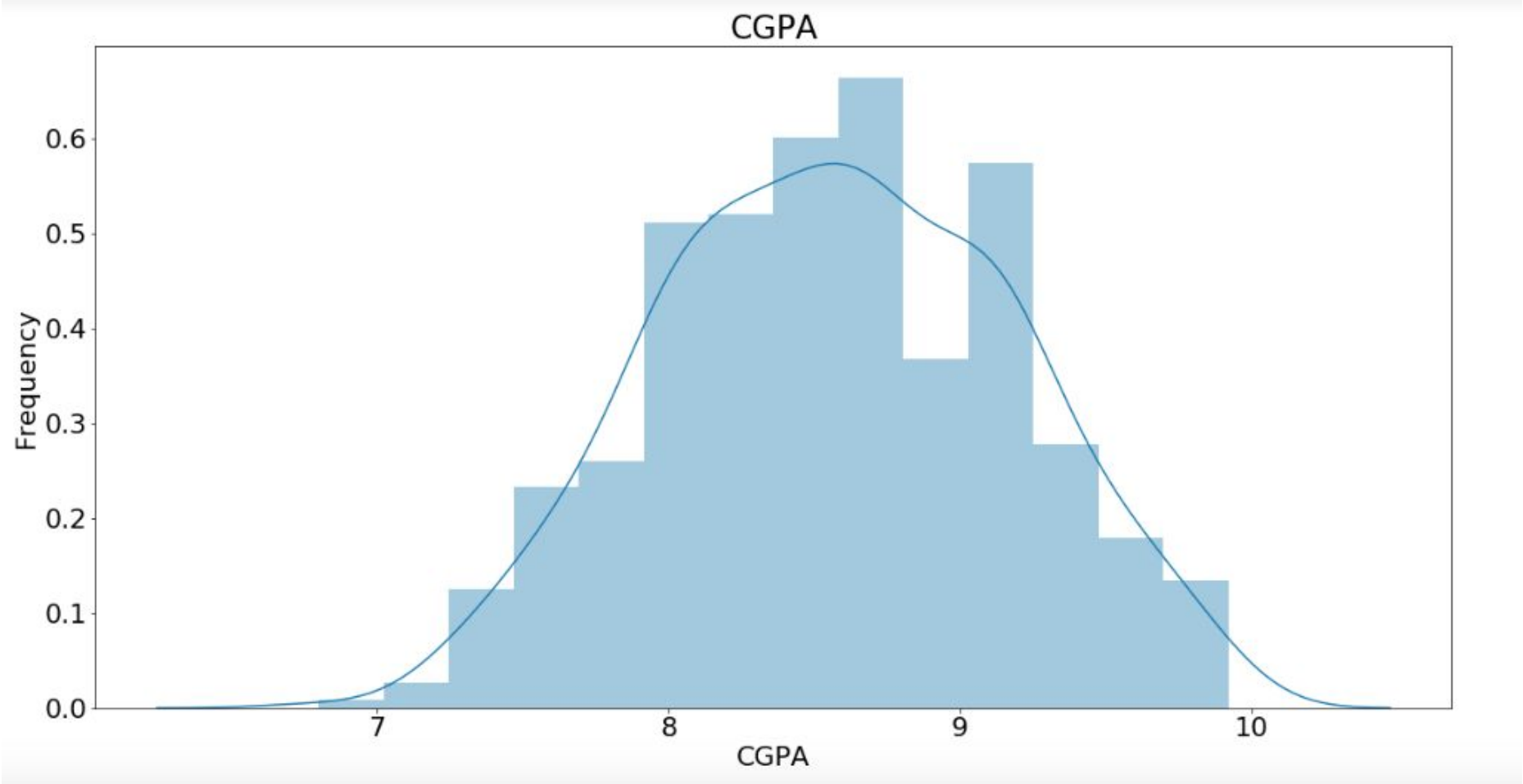
Letter of Recommendation Strength (1 to 5)

Undergraduate CGPA (6.8 to 9.92)

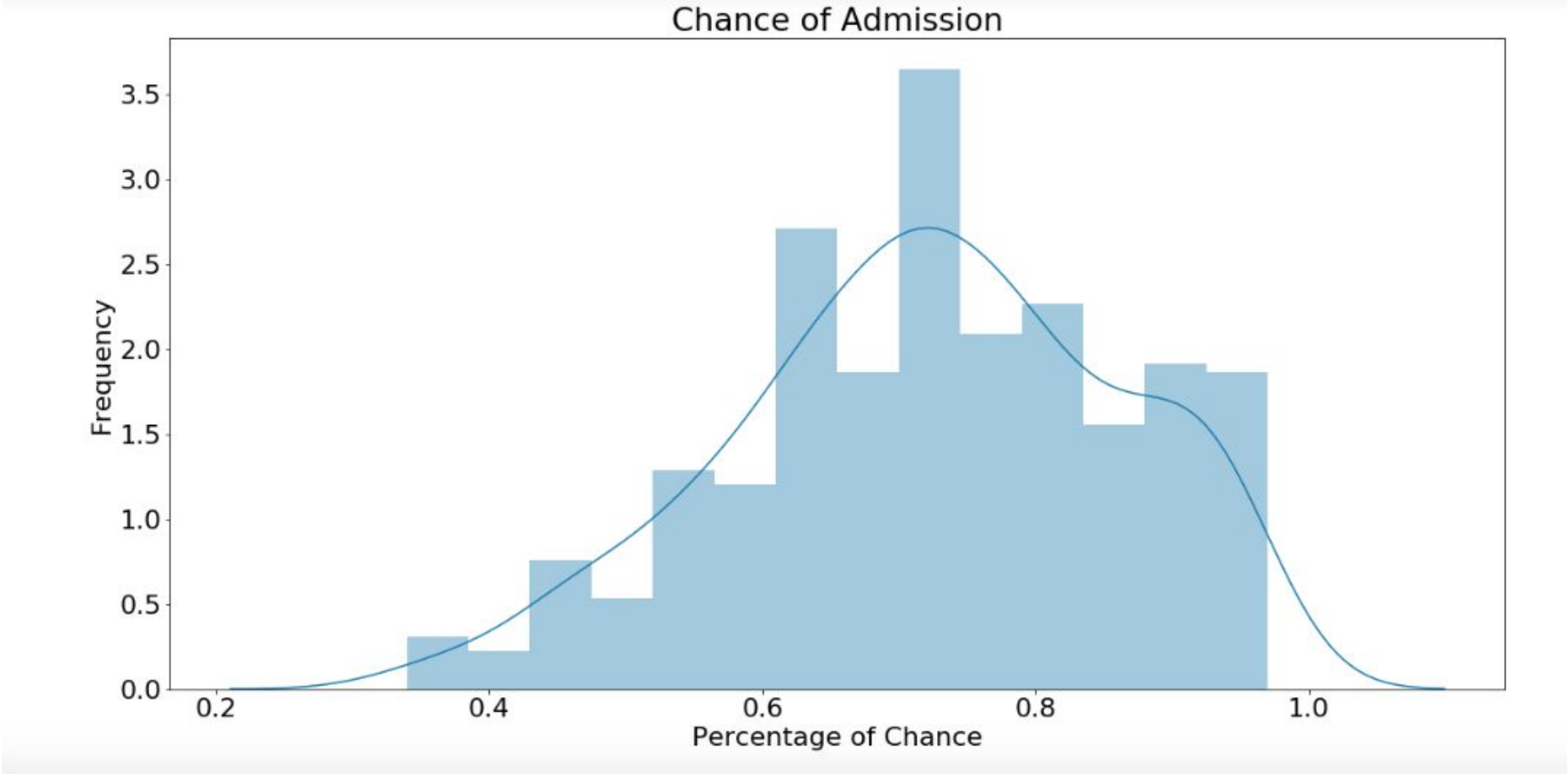
Research Experience (0 or 1)

Predict: Chance of Admit (0.34 to 0.97)

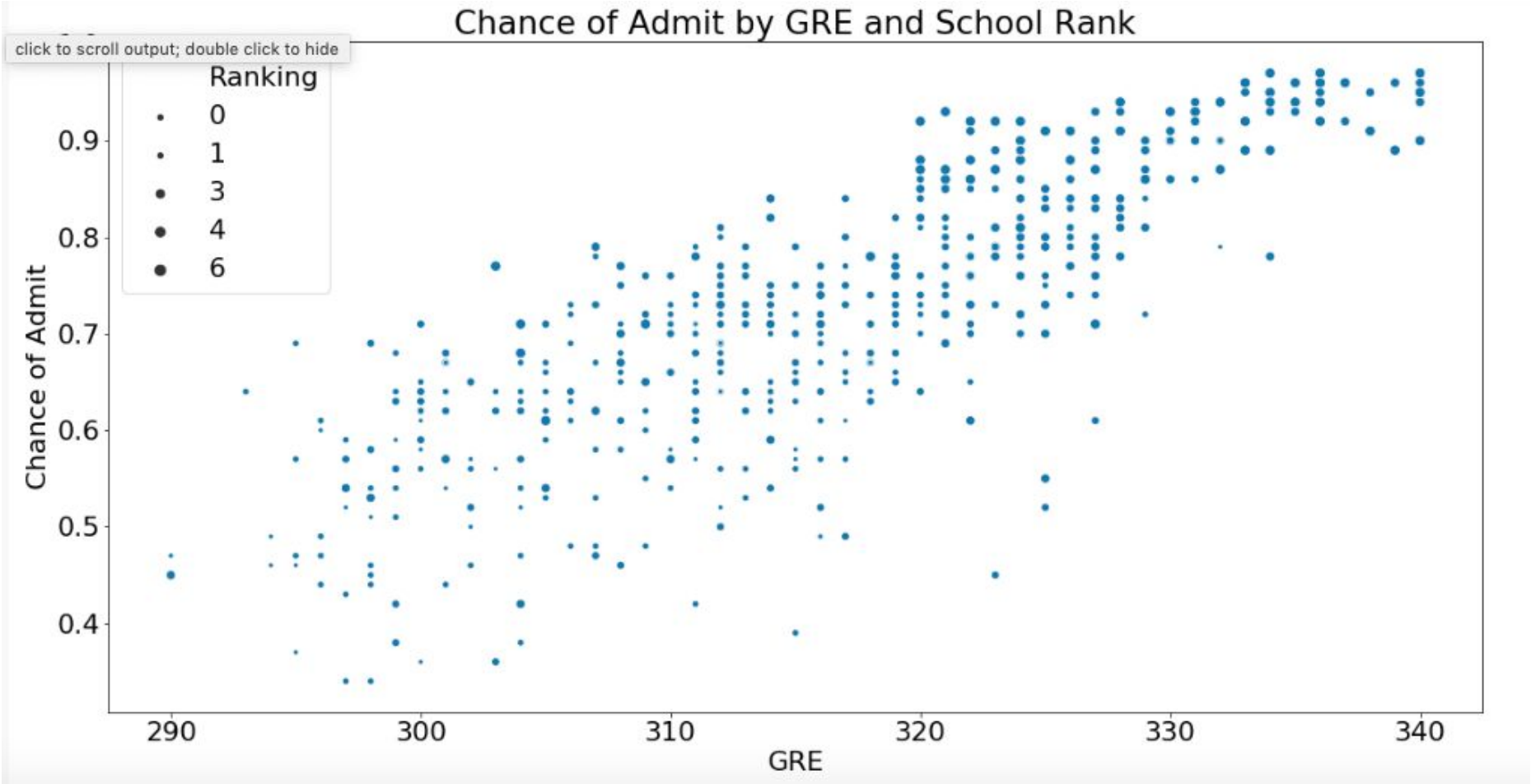
Exploratory Data Analysis



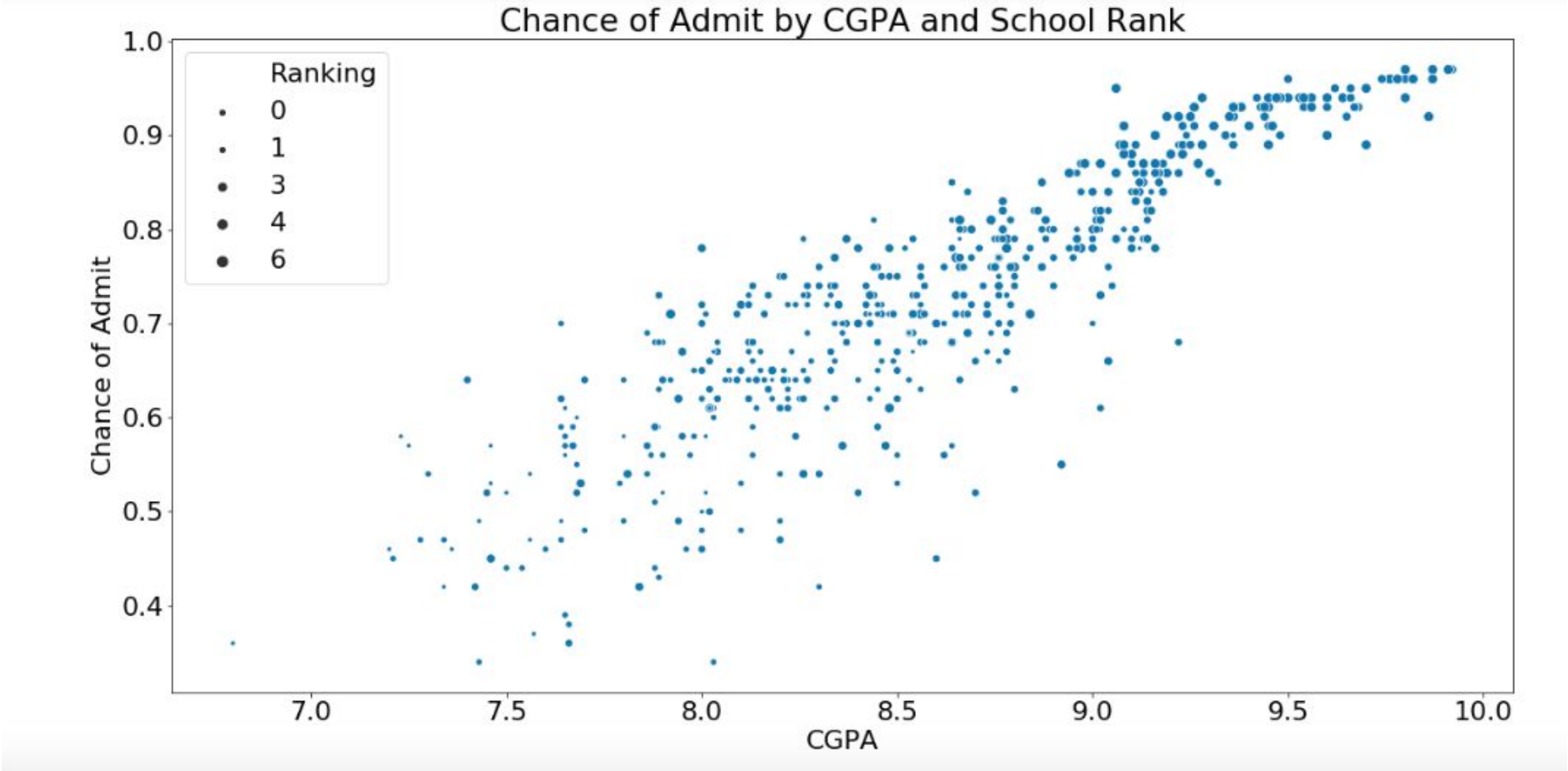
Normal distribution of cumulative GPA



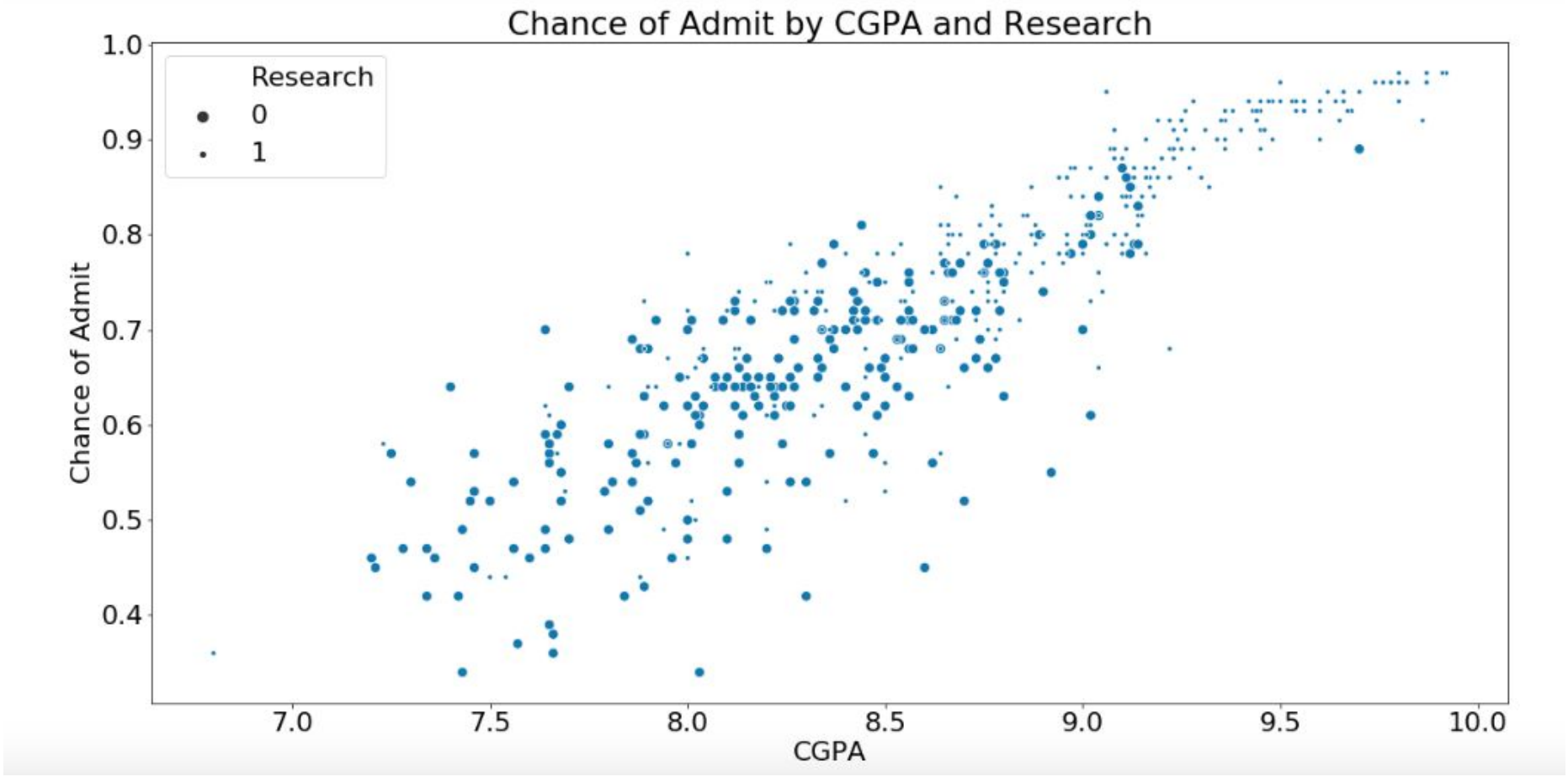
Chance of Admission slightly skewed to the right - indicating that prospective students have a strong chance of being admitted to a graduate program



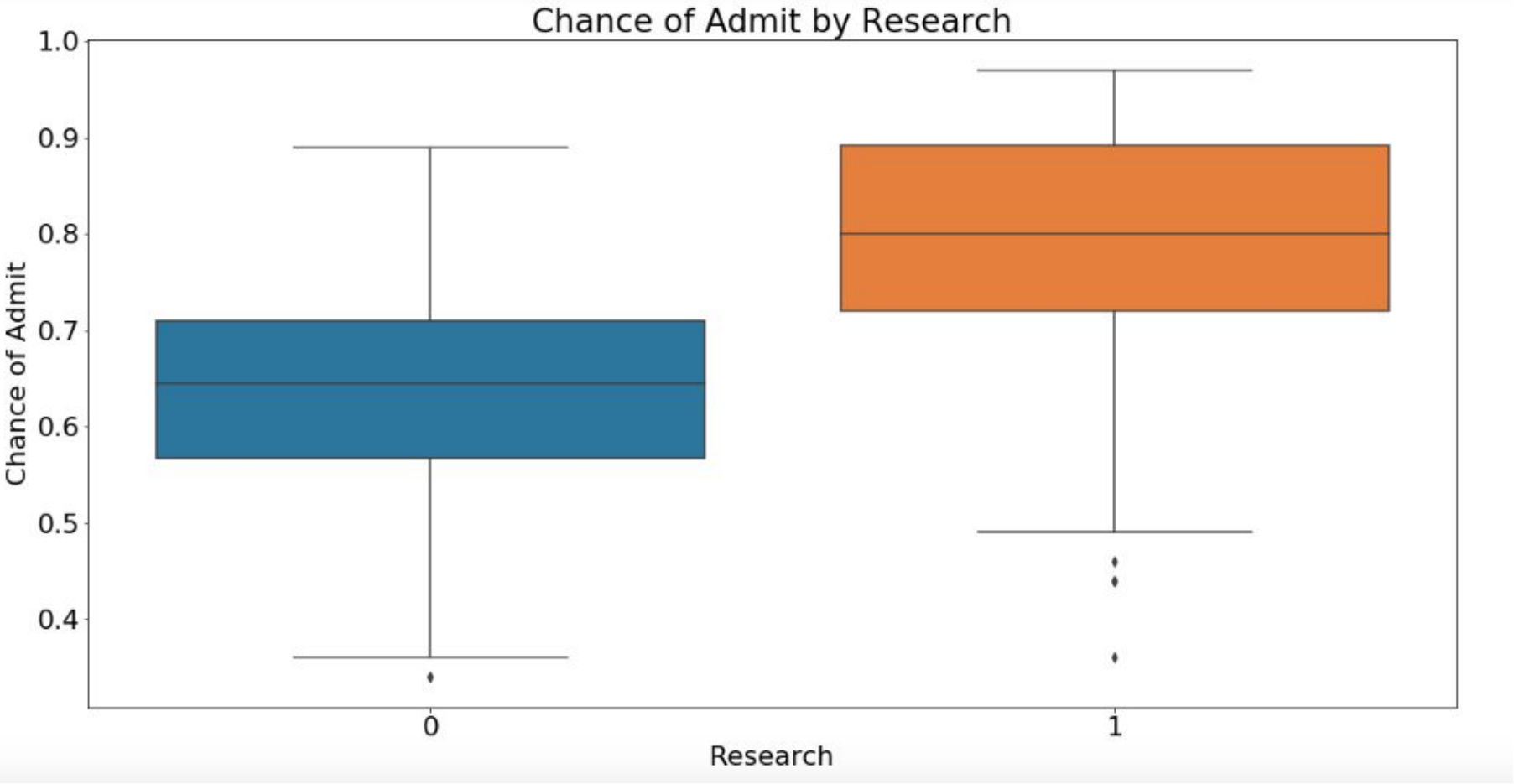
Higher GRE scores and higher ranked school leads to greater chance of admission



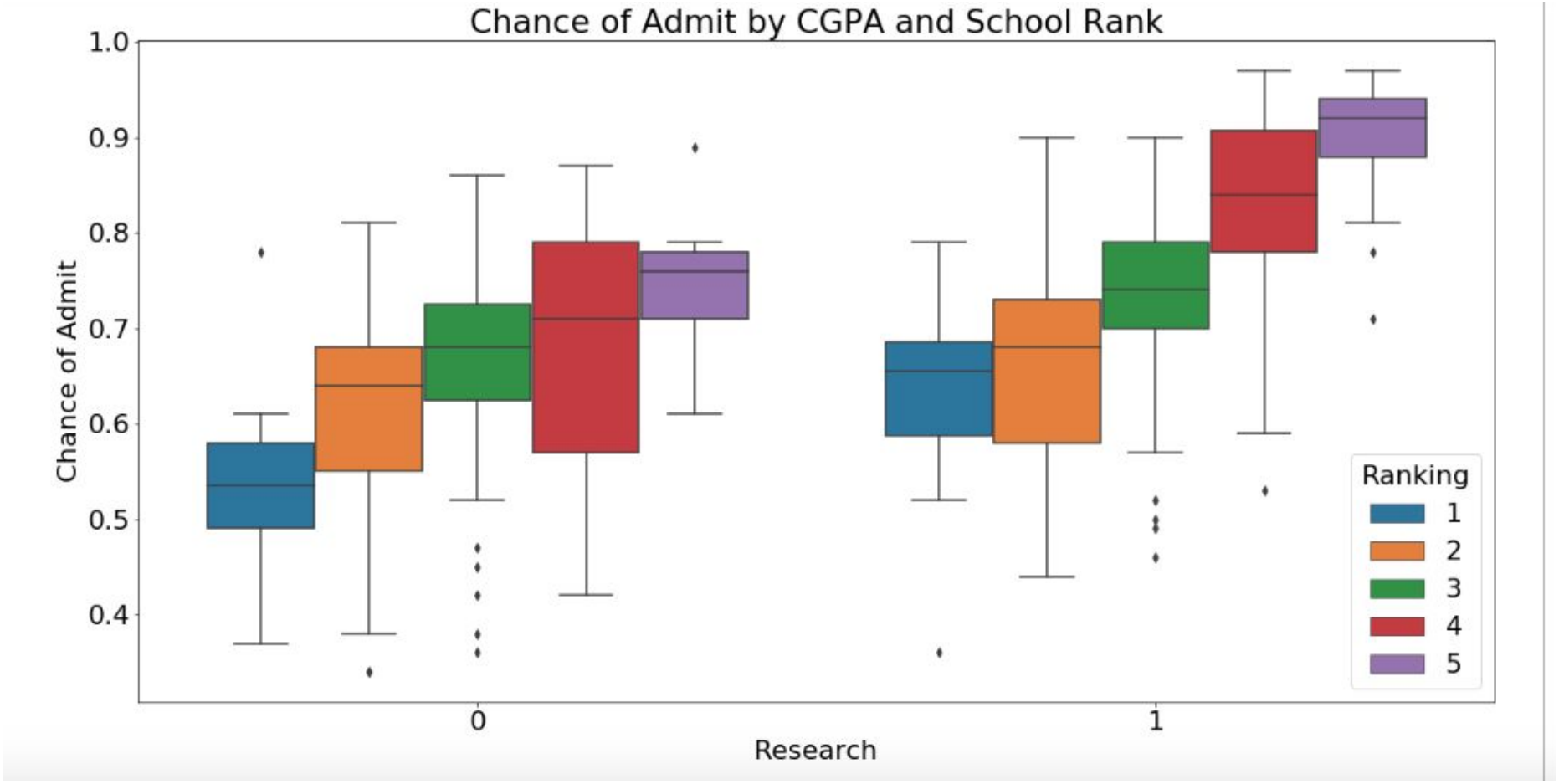
High CGPA and strong school ranking increases chance of admission



Conducting research and high CGPA leads to greater chance of admission



Increased chance of admission if applicant has conducted research

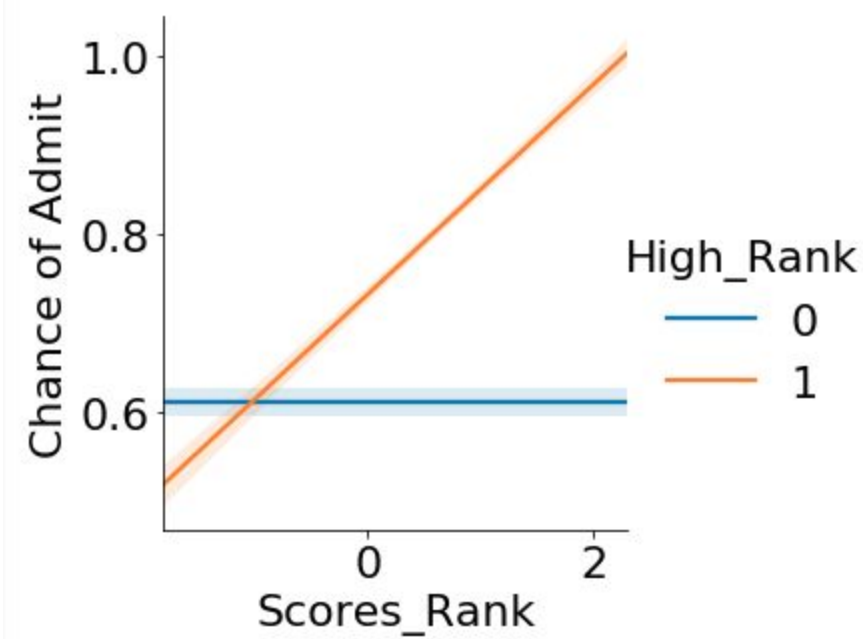


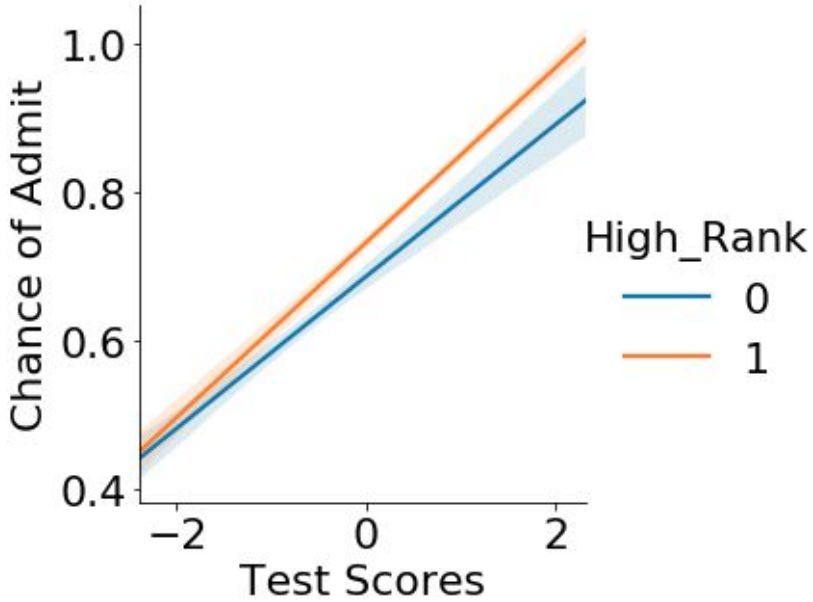
There is an interaction between research and rank on chance of admission

Feature Engineering

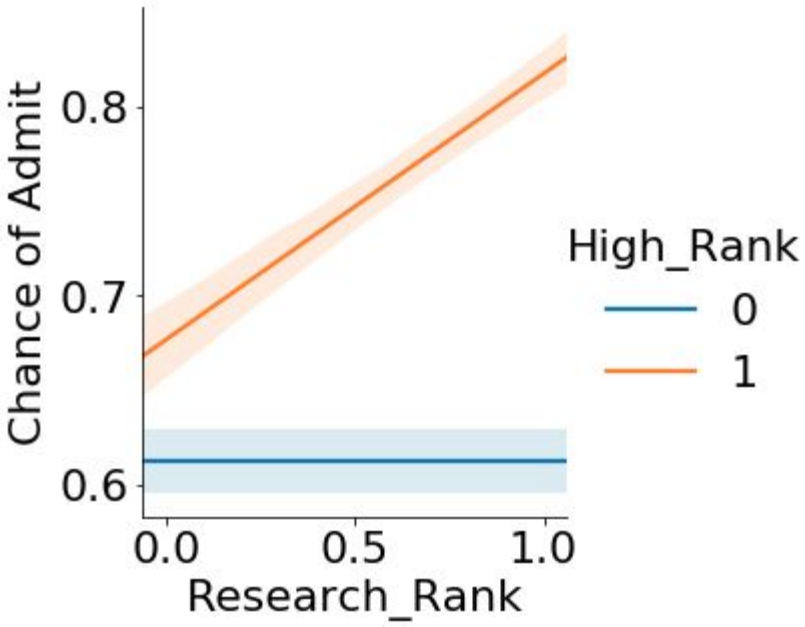
I created new features from the data:

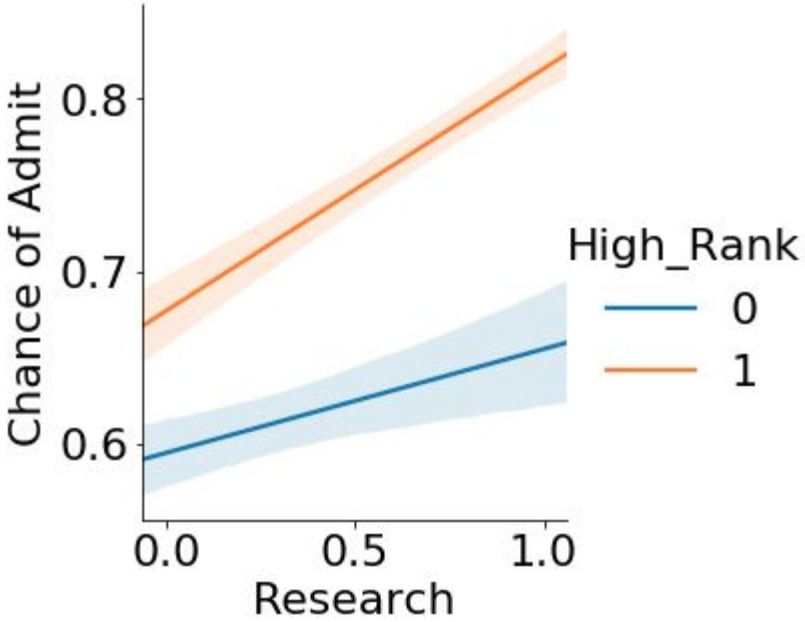
- Test Scores (TOEFL and GRE)
- Written Docs (SOP and LOR)
- High Rank (Higher ranked school)
- Scores Rank (Test scores & High Rank)
- Research Rank (Research & High Rank)





When test scores are evaluated alone (without considering the interaction between test scores and rank) there is a fairly similar chance of admission for students from high and low ranking schools.





When test scores are evaluated alone (without considering the interaction between research and rank) chance of admission for students from lower ranking schools increases

Models

Initial Modeling Results - Regression - all variables

Algorithm	**Cross Validation Accuracy**	**Accuracy Mean**
Linear Regression	0.67, 0.79, 0.86, 0.81, 0.89	0.81 (+/- 0.15)
Random Forest	**0.72, 0.55, 0.78, 0.65, 0.87**	**0.84(+/- 0.08)**

Note: When running OLS, GRE is negatively correlated with chance of admit. The feature test scores will be used in future models to account for the negative correlation

Initial Modeling Results - Classification - all variables

Algorithm	**Cross Validation Accuracy**	**Accuracy Mean**
Logistic Regression	**0.84, 0.86, 0.78, 0.88, 0.92**	**0.87 (+/- 0.09)**
Random Forest	0.82, 0.80, 0.78, 0.82, 0.74	0.81 (+/- 0.09)
Support Vector Machine	**0.84, 0.86, 0.82, 0.86, 0.92**	**0.87(+/- 0.07)**
KNN	0.86, 0.74, 0.82, 0.84, 0.84	0.85 (+/- 0.10)
Gradient Boost	0.80, 0.80, 0.74, 0.84, 0.80	0.84 (+/- 0.08)

Combining test scores and accounting for test score relationship with University Rating

Regression Modeling Results [test scores, cgpa, written docs, scores rank]

Algorithm	**Cross Validation Accuracy**	**Accuracy Mean**
Linear Regression	0.67, 0.79, 0.87, 0.81, 0.89	0.81 (+/- 0.16)
Random Forest	**0.70, 0.60, 0.79, 0.67, 0.83**	**0.86 (+/- 0.07)**

Classification Modeling Results [test scores, cgpa, written docs, scores rank

Algorithm	Cross Validation Accuracy	Accuracy Mean
Logistic Regression	**0.86, 0.86, 0.82, 0.84, 0.84**	**0.87 (+/- 0.07)**
Random Forest	0.82, 0.84, 0.76, 0.82, 0.80	0.82 (+/- 0.09)
Support Vector Machine	**0.86, 0.84, 0.82, 0.84, 0.84**	**0.86 (+/- 0.07)**
KNN	0.84, 0.84, 0.80, 0.84, 0.74	0.84 (+/- 0.09)
Gradient Boost	0.80, .084, 0.72, 0.82, 0.74	0.82 (+/- 0.10)

Dropping Scores Rank to evaluate its influence on the model

Regression Modeling Results [test scores, cgpa, written docs]

Algorithm	**Cross Validation Accuracy**	**Accuracy Mean**
Linear Regression	0.67, 0.79, 0.87, 0.81, 0.89	0.81 (+/- 0.16)
Random Forest	**0.72, 0.60, 0.79, 0.67,. 0.82**	**0.82(+/- 0.10)**

Classification Modeling Results [test scores, cgpa, written docs]

Algorithm	Cross Validation Accuracy	Accuracy Mean
Logistic Regression	**0.86, 0.86, 0.82, 0.86, 0.84**	**0.87 (+/- 0.06)**
Random Forest	0.82, 0.84, 0.78, 0.86, 0.80	0.83 (+/- 0.09)
Support Vector Machine	**0.86, 0.84, 0.82, 0.84, 0.82**	**0.86 (+/- 0.07)**
KNN	0.84, 0.84, 0.80, 0.82, 0.82	0.85 (+/- 0.06)
Gradient Boost	0.80, 0.80, 0.68, 0.88, 0.70	0.81 (+/- 0.13)

Features that are considered to be most important for graduate school acceptance

Regression Modeling Results [test scores, cgpa, LOR, Ranking, Research]

Algorithm	**Cross Validation Accuracy**	**Accuracy Mean**
Linear Regression	0.68, 0.79, 0.86, 0.81, 0.89	0.81 (+/- 0.14)
Random Forest	**0.76, 0.60, 0.76, 0.68, 0.80**	**0.84 (+/- 0.08)**

Classification Modeling Results [test scores, cgpa, LOR, Ranking, Research]

Algorithm	Cross Validation Accuracy	Accuracy Mean
Logistic Regression	**0.86, 0.86, 0.82, 0.88, 0.86**	**0.87 (+/- 0.07)**
Random Forest	0.84, 0.80, 0.76, 0.82, 0.84	0.84 (+/- 0.08)
Support Vector Machine	**0.86, 0.84, 0.82, 0.88, 0.86**	**0.87(+/- 0.07)**
KNN	0.80, 0.86, 0.74, 0.78, 0.86	0.85 (+/- 0.10)
Gradient Boost	0.86, 0.84, 0.74, 0.84, 0.70	0.82 (+/- 0.11)

Observations

Strongest regression model is Random Forest - considering interaction between test scores and rank

Strongest classification models are Logistic Regression and SVM - consistent outcomes among all models

Features scores_rank and research rank expose influence of rank on outcome

Key Takeaways

Test scores and CGPA are the most important factors impacting admission to graduate school

According to '*critical factors*' school rank only '*anecdotally makes a difference*' yet according to the data rank does indeed make a difference

Policy Recommendations

Design programs to encourage more undergraduate research for students in lower ranking school

Ensure that students from lower ranked school have equal access to information and support for graduate school admission

Encourage more study on the impact of legacy and familial connections on graduate school admissions.

Further Research

What are the family and financial responsibilities for students by school rank?

Long- term study to evaluate the success of graduate students based on their portfolio for entrance.

How much does SES and gender matter in the application process?