Linear Regression Final Project
US College Tuition Prediction

Miho Hunter
4/20/2022

Introduction

20-Year Average Tuition Growth Among National Universities, 2002-2022

— Private — Out-of-state — In-state

Question:

What makes college so expensive?

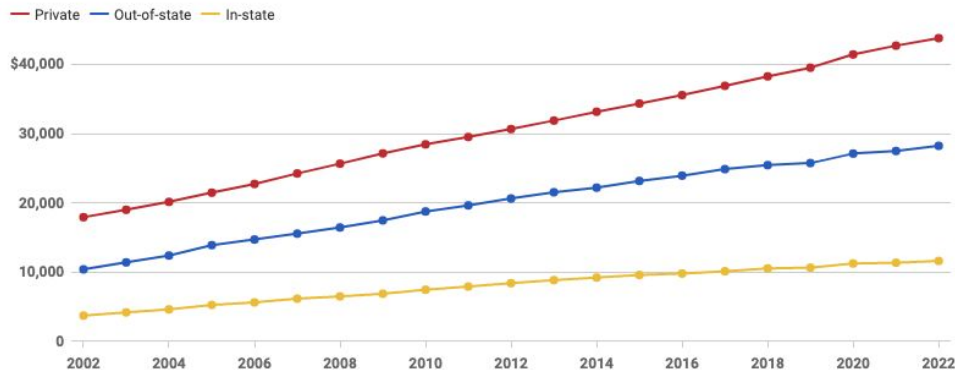Can we predict tuition?

Source: U.S. News & World Report • Get the data

Data Used:

★ America's Top Colleges by Forbes
★ Cost of College List by the Stephen Phillips Memorial Scholarship Fund Inc.
★ US College data
★ Global 2000 List By the Center for World University Rankings

Tools:

★ Selenium, Beautiful Soup

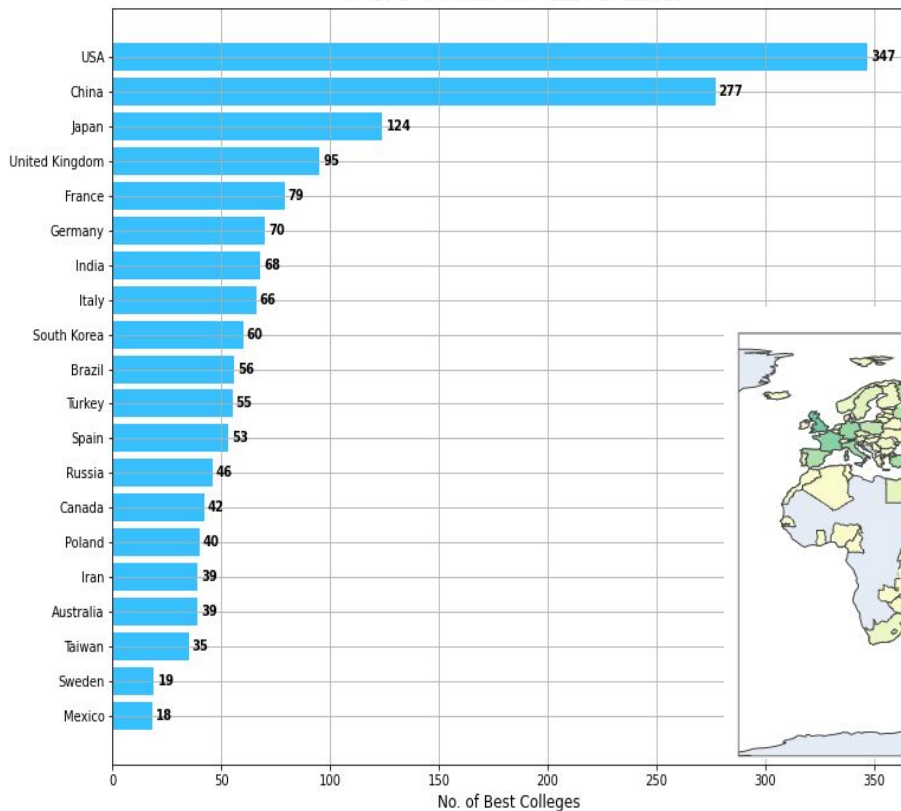★ Pandas, Matplotlib, Plotly Express

★ SKLearn, Statsmodels

★ Fuzzywuzzy

Top 20 Countries with Best Colleges

TOP 20 COUNTRIES WITH BEST COLLEGES

Country	No. of Best Colleges
USA	347
China	277
Japan	124
United Kingdom	95
France	79
Germany	70
India	68
Italy	66
South Korea	60
Brazil	56
Turkey	55
Spain	53
Russia	46
Canada	42
Poland	40
Iran	39
Australia	39
Taiwan	35
Sweden	19
Mexico	18

No. of Best Colleges

GDP by Country

#	Country	GDP (abbrev.)
1	United States	$19.485 trillion
2	China	$12.238 trillion
3	Japan	$4.872 trillion
4	Germany	$3.693 trillion

56 more rows

https://www.worldometers.info › gdp › gdp-by-country
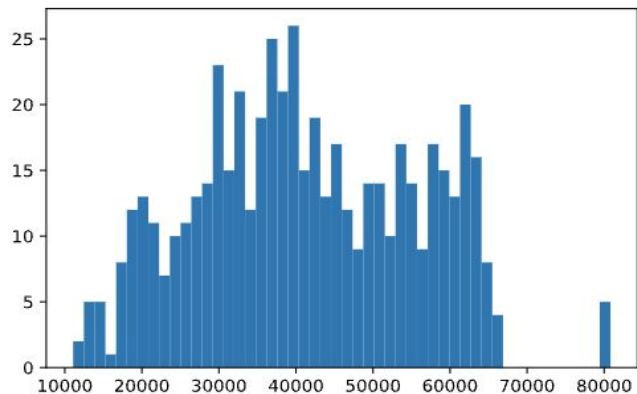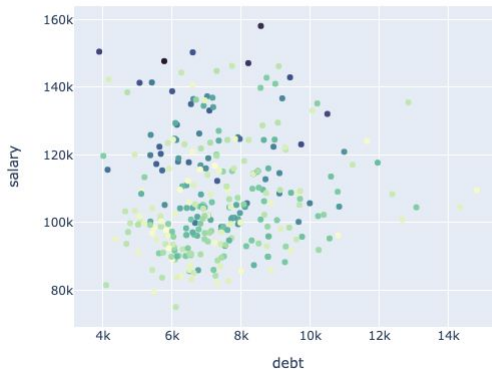
GDP by Country - Worldometer
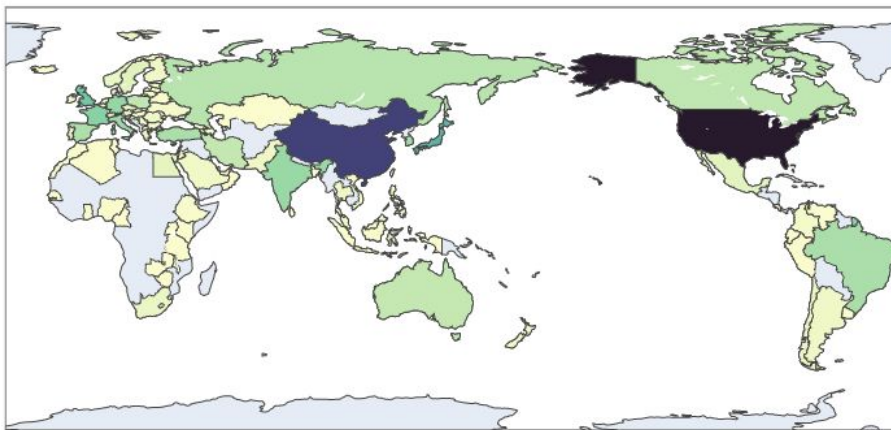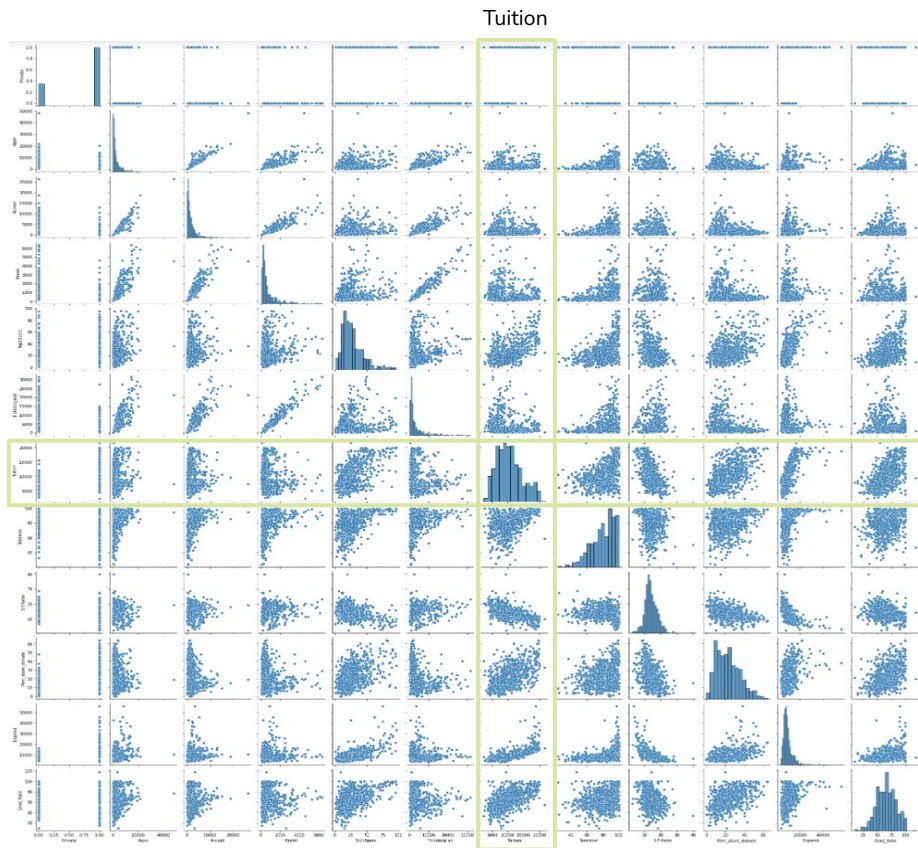
Tuition and Salary

Salary vs Debt

Salary vs Score
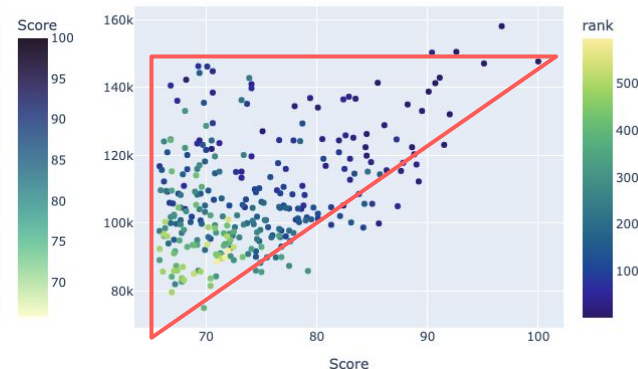
Name	Tuition	Is_Private	rank	state	salary
Bates College	80825.0	1	226.0	ME	124500.0
Connecticut College	80678.0	1	154.0	CT	103400.0
Amherst College	80350.0	1	16.0	MA	127100.0
Colby College	80044.0	1	92.0	ME	105700.0
Union College	79931.0	1	NaN	NaN	NaN
...	...	...	...	...	...
Eastern Connecticut State University	12796.0	0	404.0	CT	87400.0
Brigham Young University	12730.0	1	53.0	UT	113400.0
Massachusetts College of Liberal Arts	12579.0	0	NaN	NaN	NaN
Midwestern State University	12380.0	0	NaN	NaN	NaN
Fitchburg State University	11081.0	0	NaN	NaN	NaN

Bates College in ME had the highest tuition.
$80825 per year!

Good school ≠ High Tuition

Salary vs Tuition

Features

Tuition
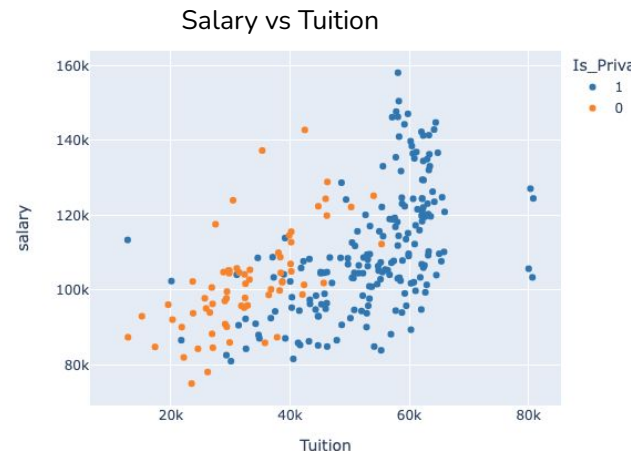
Showing high correlation, but number of applications, number of applicants accepted, and number of students enrolled do not determine each other.
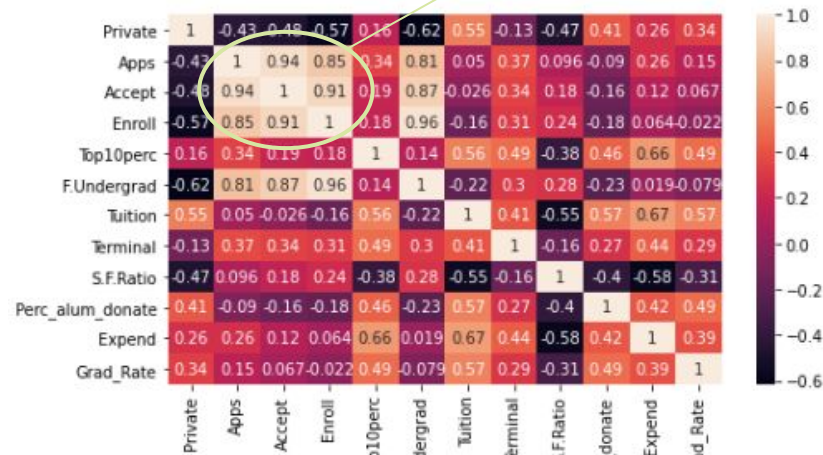
Linear Regression Model

BASELINE	MODEL IMPROVEMENT	FINAL

R^2 = 0.719

★ High P value features
★ Used 80/20 test/train split to validate

★ Dropped features with high P value ('S.F.Ratio', 'F.Undergrad', 'Top10perc', 'Apps')
★ Scaled features to standard normal distribution
★ Used train/test split and K-fold cross validation and checked scores for each model
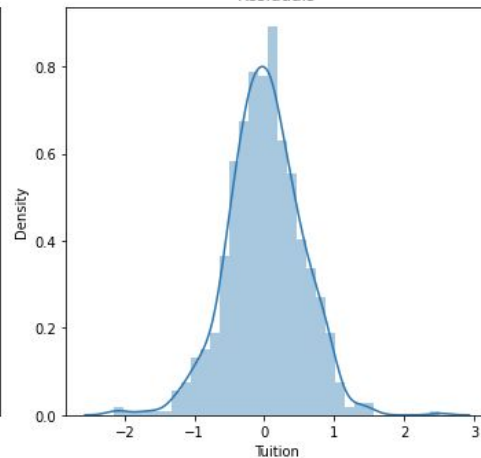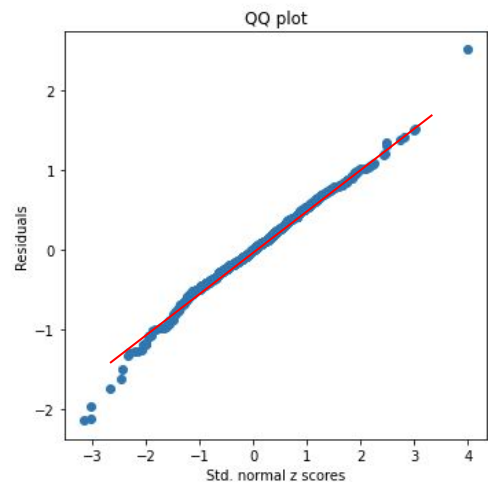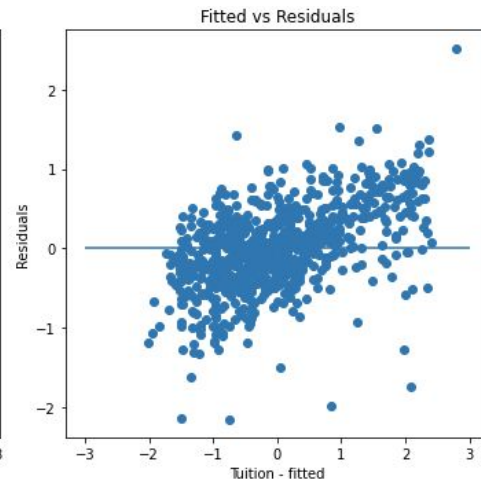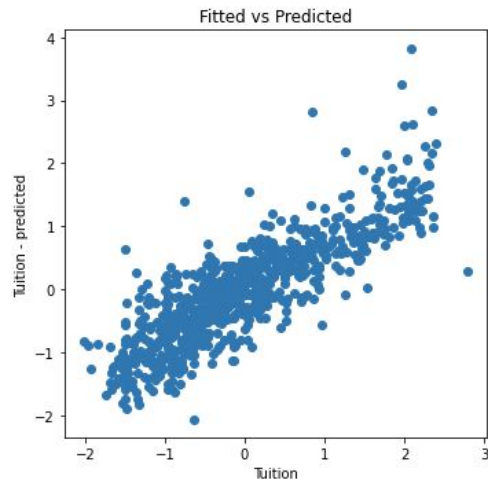
R^2 = 0.730
MAE 0.444
RMSE 0.559

★ R^2 train = 0.7303
★ R^2 test = 0.7210 (slightly overfit, but not bad)
★ Assumptions checked!

Assumptions

❏ Linearity
❏ Homoscedasticity
❏ Independence
❏ Normality

Assumptions hold good!

Conclusions

★ The baseline model with all features had R^2 = 0.719.
★ R^2 was improved to 0.730 for the final model with the selected and scaled features.

The Estimated Regression Equation:

Tuition = -0.0062 + 0.3305*private - 0.3262* Apps + 0.6109*Accept -0.3093*Enroll + 0.1086* Top10perc +0.1569 *Terminal +0.1557*Perc_alum_donate +0.3454 * Expend + 0.1338 *Grad_Rate

Appendix

- ❏ America's Top Colleges by Forbes

  https://www.forbes.com/top-colleges/

- ❏ Cost of College List by the Stephen Phillips Memorial Scholarship Fund Inc.

  https://phillips-scholarship.org/new-applicants/cost-of-college-list/

- ❏ US College data

  https://www.kaggle.com/datasets/yashgpt/us-college-data

- ❏ Global 2000 List By the Center for World University Rankings

  https://cwur.org/2021-22.php