DATA 603 Final Project

Higher Education Enrollment

by Carlotta Amaduzzi
18th August 2021

The Data

- The data was downloaded from the **Integrated Postsecondary Education Data System** (IPEDS) web site (https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx)
- It refers to **three different files** of data relative to 2020 Higher Education Institutions data that were merged together after cleaning
  - Institutional ID data
  - Students' Cost data
  - Students' Enrollment data
- Focus was placed exclusively on **data reported** by the Institutions themselves

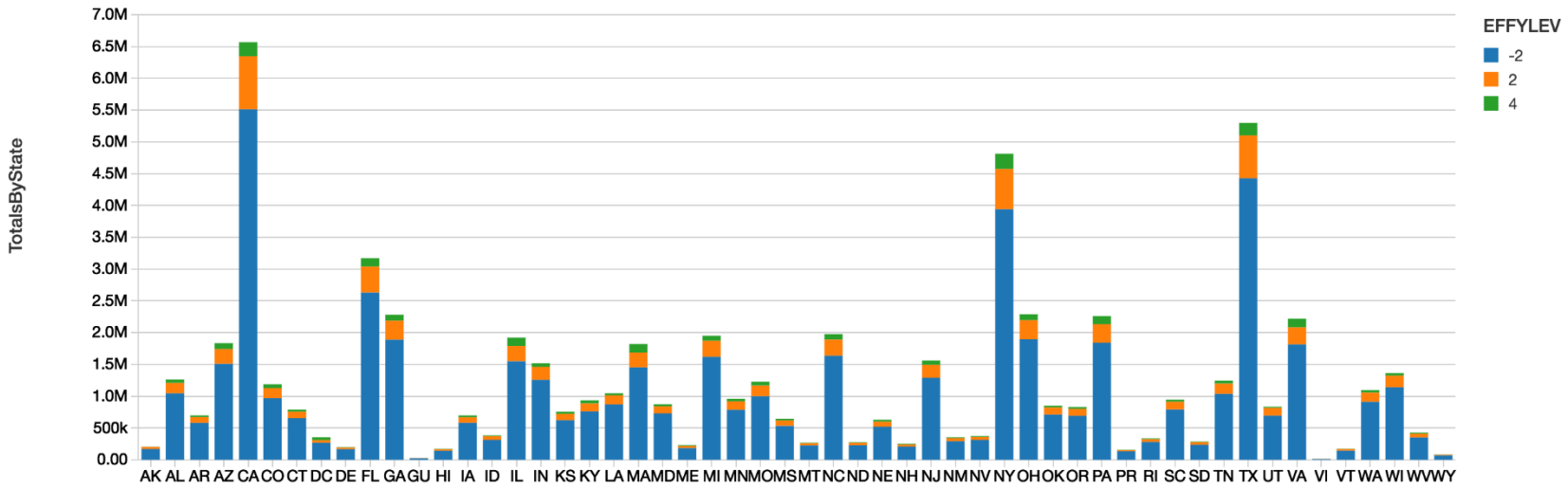
The Cleaning Process

- Institutions' Identification information
  - By State & US Territories and By Group (EIN Number)
- Tuition & Other Cost Information
  - Only Reported data
  - Current data vs. Historical Data
  - Separating Institutions focused on Undergraduate Programs only
- Students' Enrollment
  - Only Reported

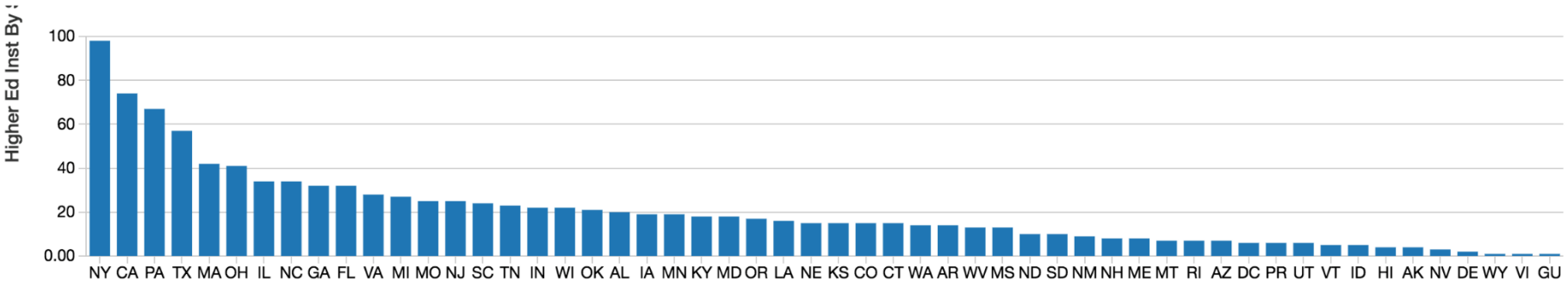
Merging the Data

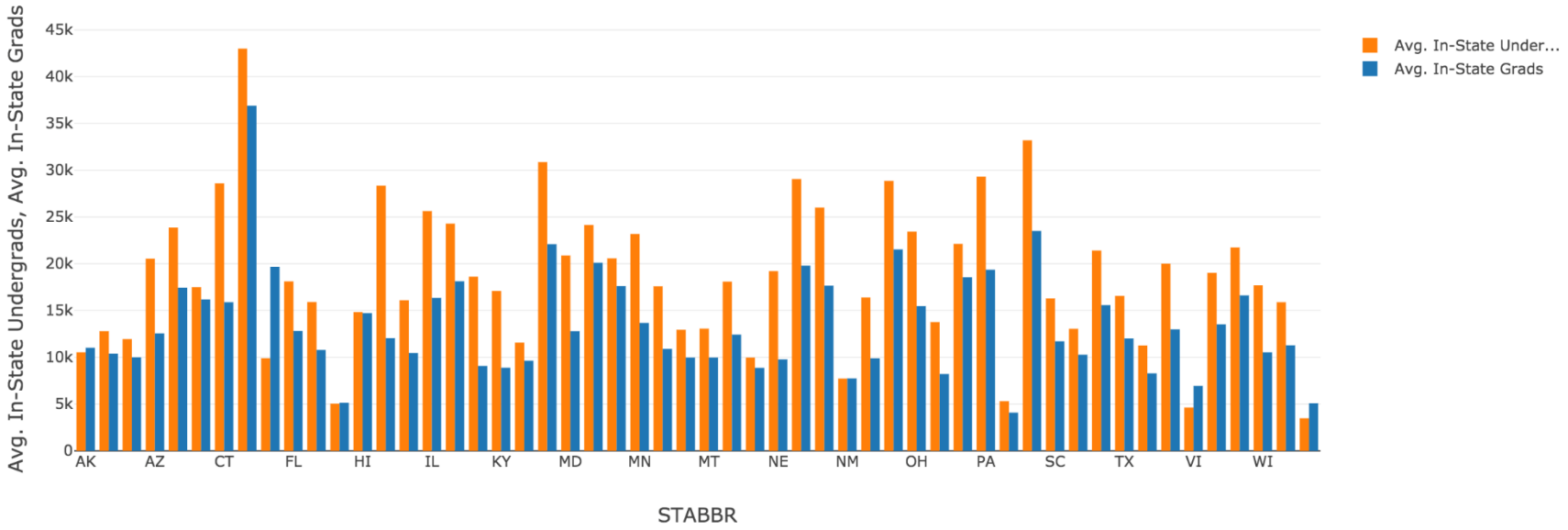
STABBR	Higher Ed Inst By State
NY	98
CA	74
PA	67
TX	57
MA	42
OH	41
IL	34
NC	34
GA	32
FL	32
VA	28

- Students' reported Status – problematic
- Taking a closer look at the Reported Data
  - Overall by State
  - For Maryland Reporting Institutions
- Students' Reported Enrollment Status is problematic

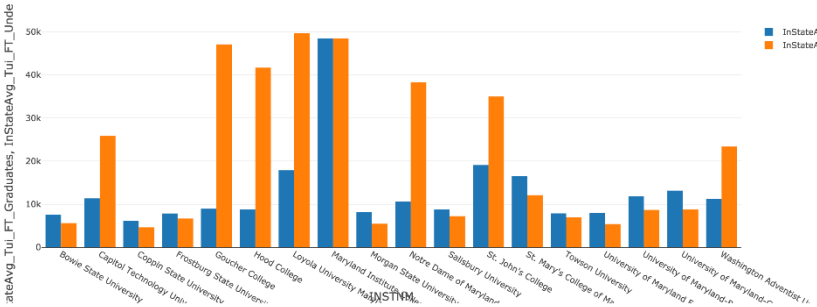


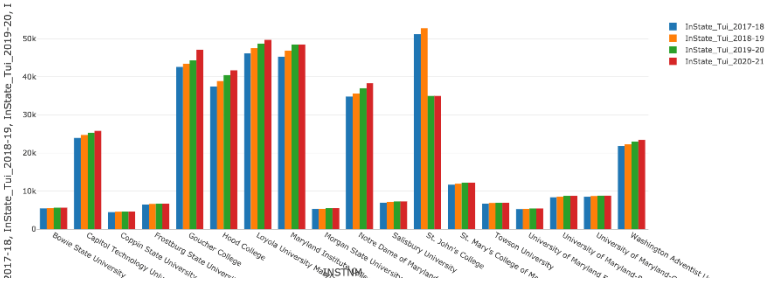
The Overall Picture

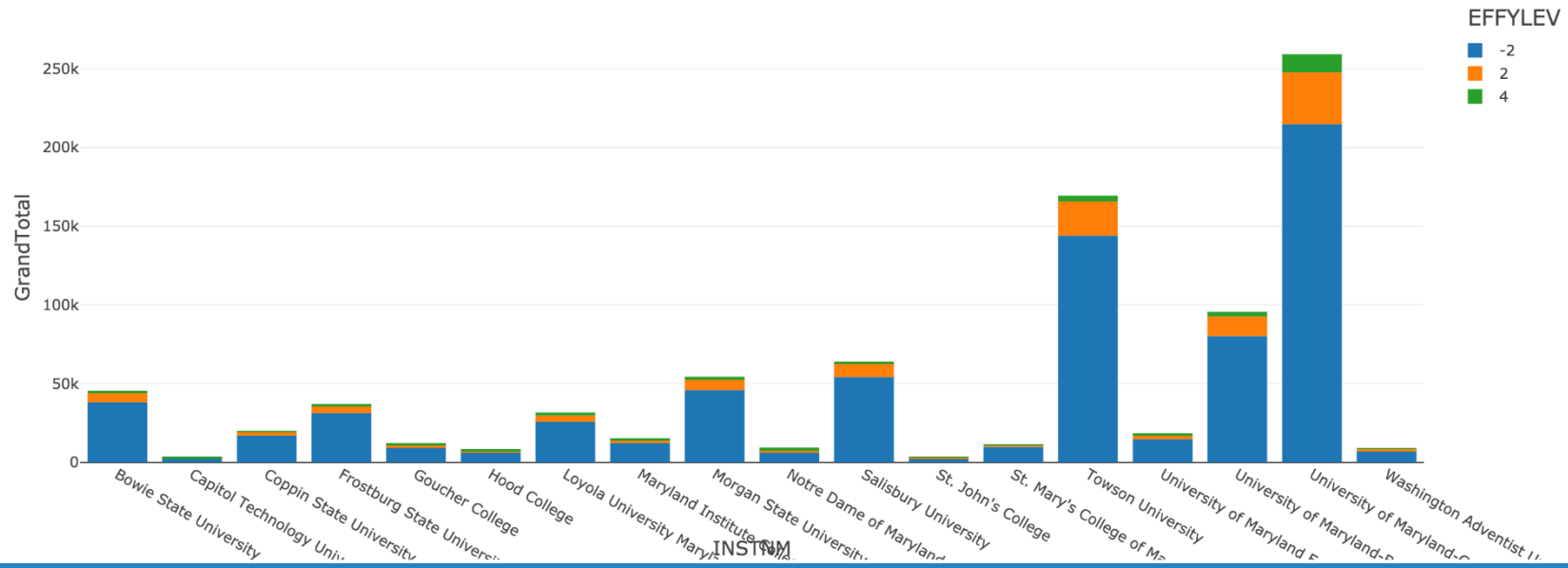




The Overall Picture ~ Maryland

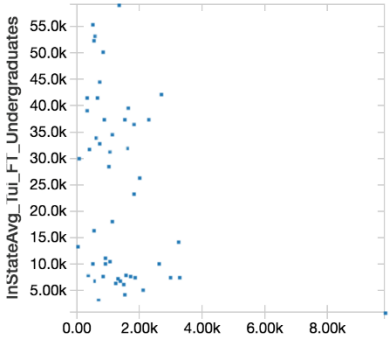


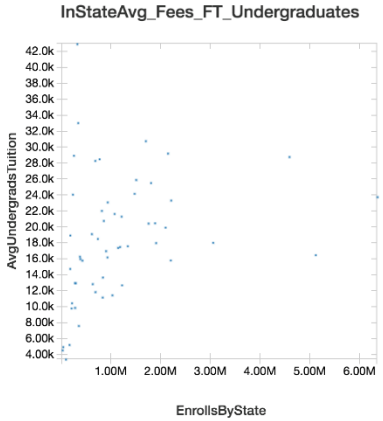


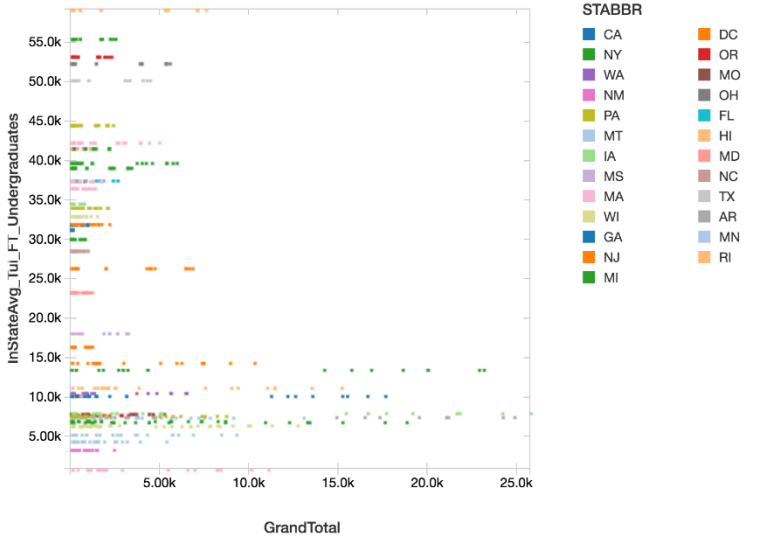


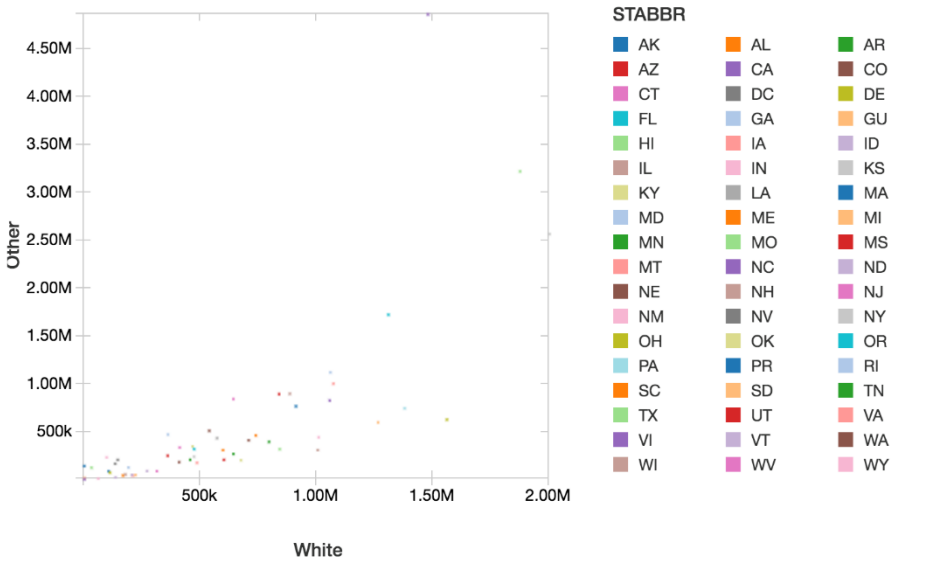
The Analysis Process

- Taking a Peek
- Correlation
- Deep dive on Enrollment









There are 17542 rows in the training set, and 4313 in the test set

	summary	White	Other
1	count	17542	17542
2	mean	1474.4468133622163	1332.4873446585339
3	stddev	3084.0036360212935	3147.7407683693336
4	min	0.0	0.0
5	25%	54.0	66.0
6	50%	349.0	314.0
7	75%	1348.0	1126.0
8	max	39575.0	56217.0

The Analysis Process

- White Students vs. All Other Groups
  - Simple Regression Model

```
Coefficient: [0.6671596257120478]
Intercept: 348.7959605234763
The equation for the Linear Regression Line is (Predicted Other Students' Enrollment) = [0.6671596257120478] *(White Students Enrollment) + (348.7959605234763)
```

```
Root Mean Squared Error: 2382.131133
R2: 0.427259
```

  - Does NOT improve much without Outliers

```
Root Mean Squared Error: 1708.767035
R2: 0.564757
```

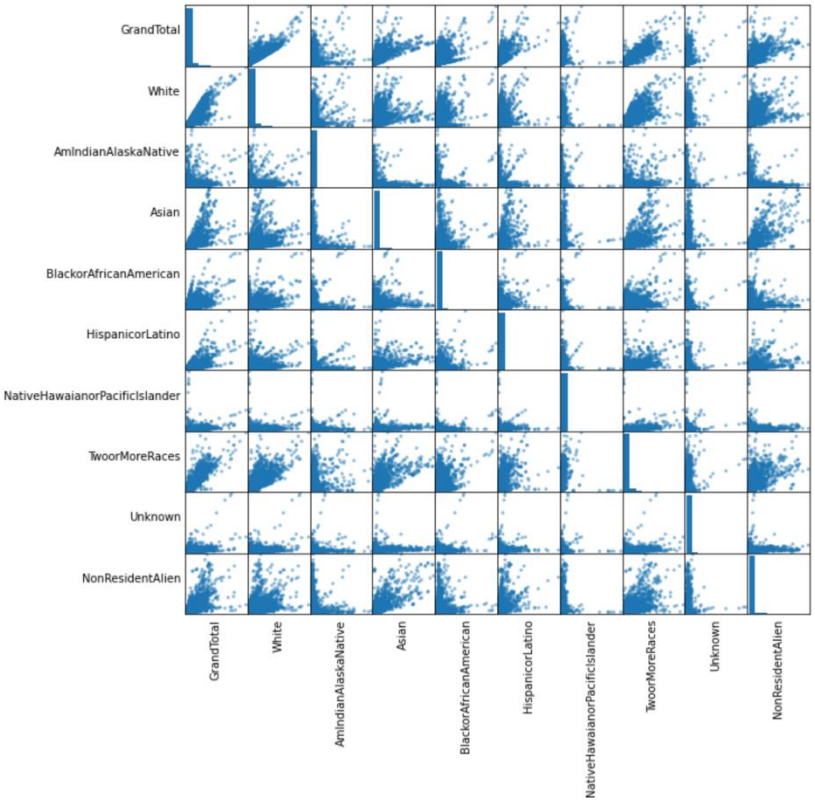
  - Better Using Decision Tree Model & Eliminating Grand Total Column (Double Count)

```
RMSE is 2071.2361498805044
R2 is 0.6355405742755694
```

The Analysis Process

- Overall Correlations by Race
  - Multiple Regression Model
  - Still Not Great

	feature	importance
18	White	0.999648
17	NativeHawaianorPacificIslander	0.000352
0	UNITIDindex	0.000000
30	InState_Tui&Fee_2019-20	0.000000
32	InState_Fees_2020-21	0.000000
33	InState_Tui&Fee_2020-21	0.000000
34	OutOfState_Tui_2017-18	0.000000
35	OutOfState_Fees_2017-18	0.000000




```
Coefficient: [-1.7800547481586966,-1.247238563483539,0.2427181347677768,-0.10131106116475028,-3.3005800695925567,10.794583836908682,0.939291278580745,1.251591322763446]
Intercept: 259.92769758062485
Root Mean Squared Error: 1586.727272
R2: 0.735272
```

Conclusions

- More analysis is necessary
- Institutions' Recognition seems to play a role in enrollment (more than cost)
- Expanding the data to include unreported or historical data
- Expanding the data to include Students' Completions NOT possible (data corrupted)
- K-Means might be an avenue to pursue