Higher Education Institutions & Student Enrollment

An Introductory Analysis on IPDES 2012-2020 Data

BY CARLOTTA AMADUZZI
CAMADUZ1@UMBC.EDU

DATA 606 ~ CAPSTONE PROJECT
UNIVERSITY OF MARYLAND BALTIMORE COUNTY
DR. CHAOJIE WANG

Project Overview

- Data collection & cleaning
- Initial exploratory analysis
- Application of Machine Learning Algorithms
- Conclusions & Further Analysis

Rationale

Access to Higher Education is essential

to maintain a more equitable society	to inspire individual professional and financial growth	to ensure social mobility

Understanding the factors that influence undergraduate enrollment in Higher Education Institutions (HEIs) can have a double positive effect:

on the HEIs themselves which are for all practical purposes fairly large (business) organizations	on the impact increased levels of education can have on society as a whole

Research Focus & Results

Based on publicly reported information regarding HEIs, **do the selected and reported features affect student enrollment choice**?

With an eye to I.D.E.A. (Inclusion, Diversity, Equity, and Access), **is there evidence of changes in enrollment in HEIs over time**?

In particular, with an eye towards Standardized Tests and Blind Admission Policies, are there changes emerging over time?

The Data

IES NCES National Center for Education Statistics MENU

IPEDS

Publicly available information

Integrated Postsecondary Education Data System (IPEDS) web site (https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx)

D**ifferent data files** reported by Federally Funded HEIs

Institutional ID data
Institutions Offerings
Students' Enrollment data
Admissions' Policies

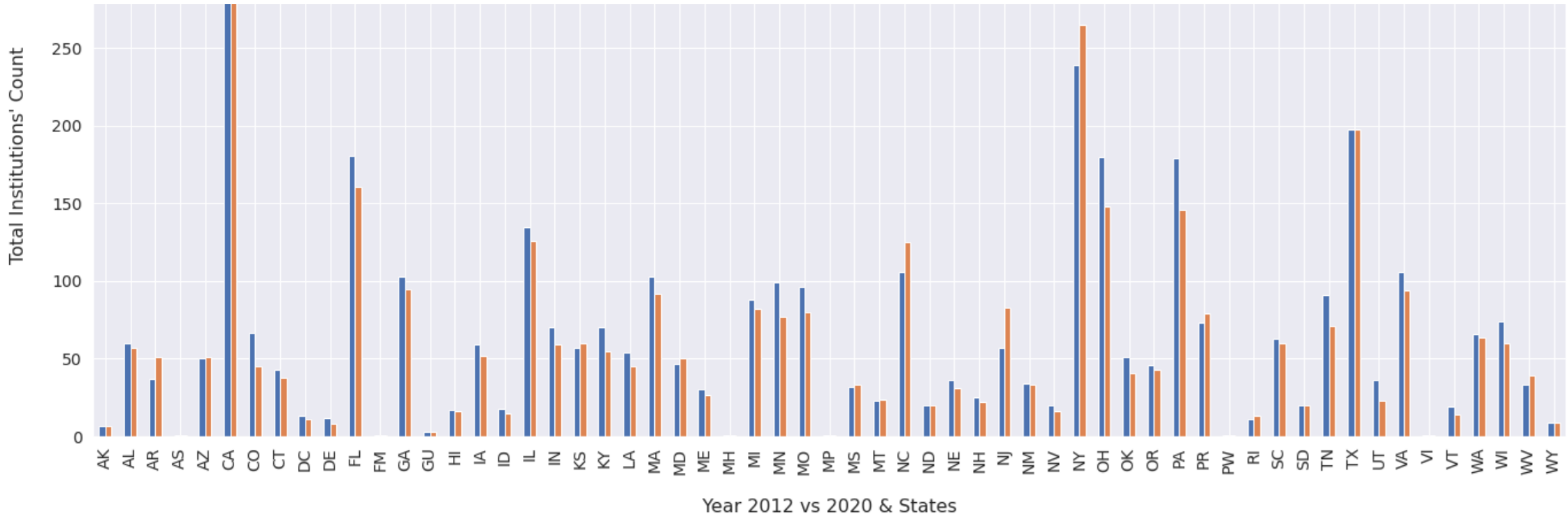
Only **reported** data

Status of Education

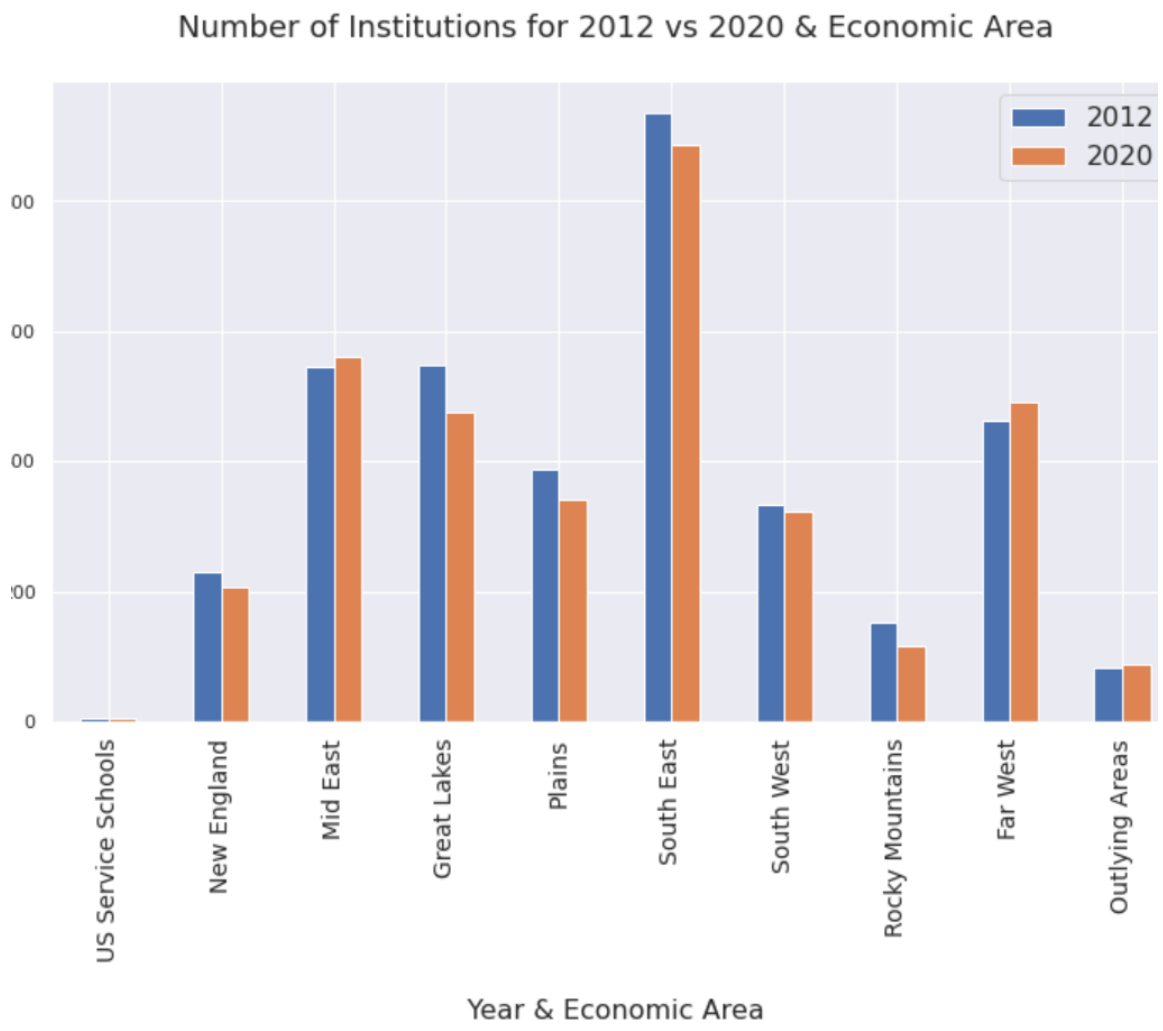
Higher Education Institutions

Enrollment

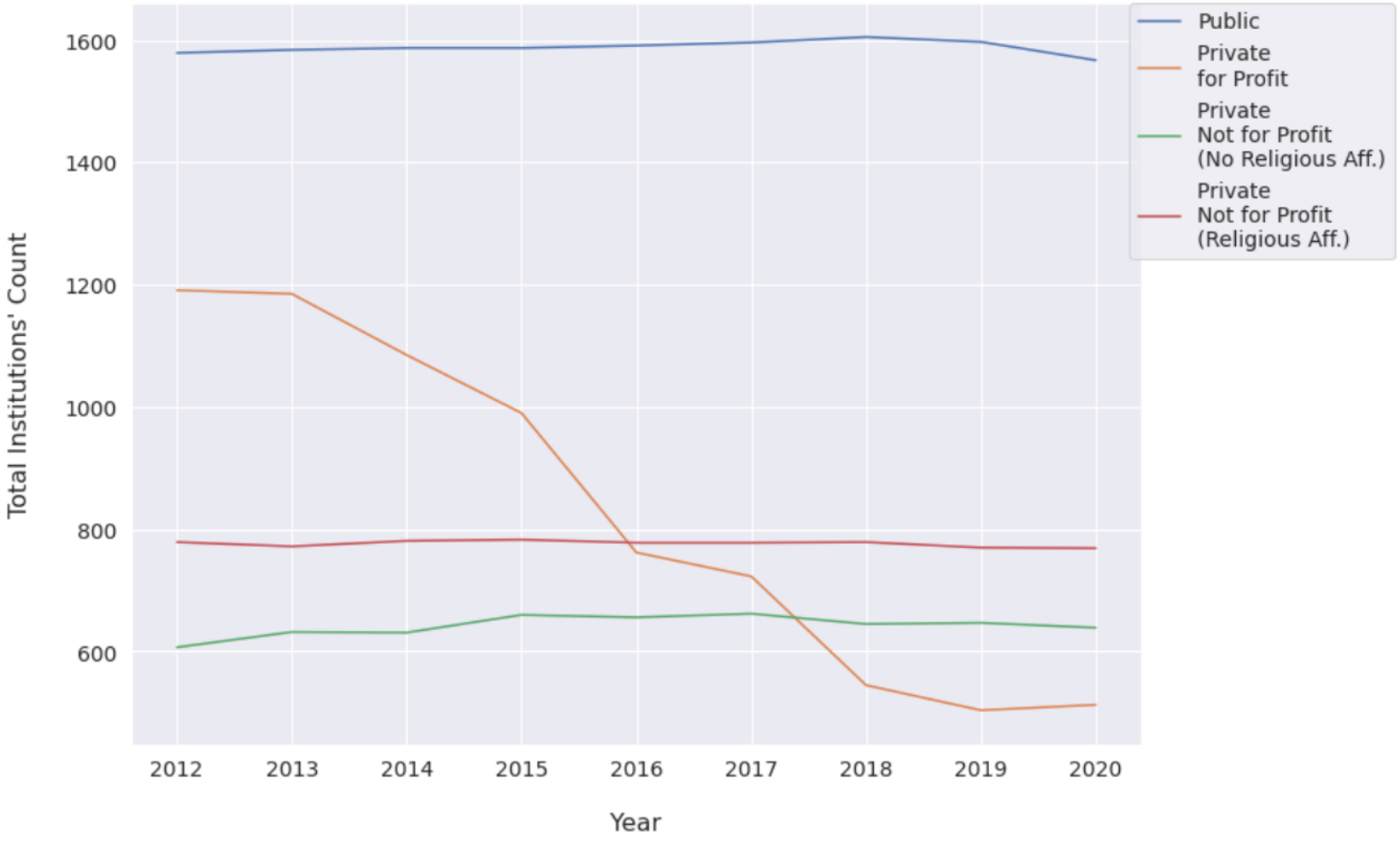
Selection Processes



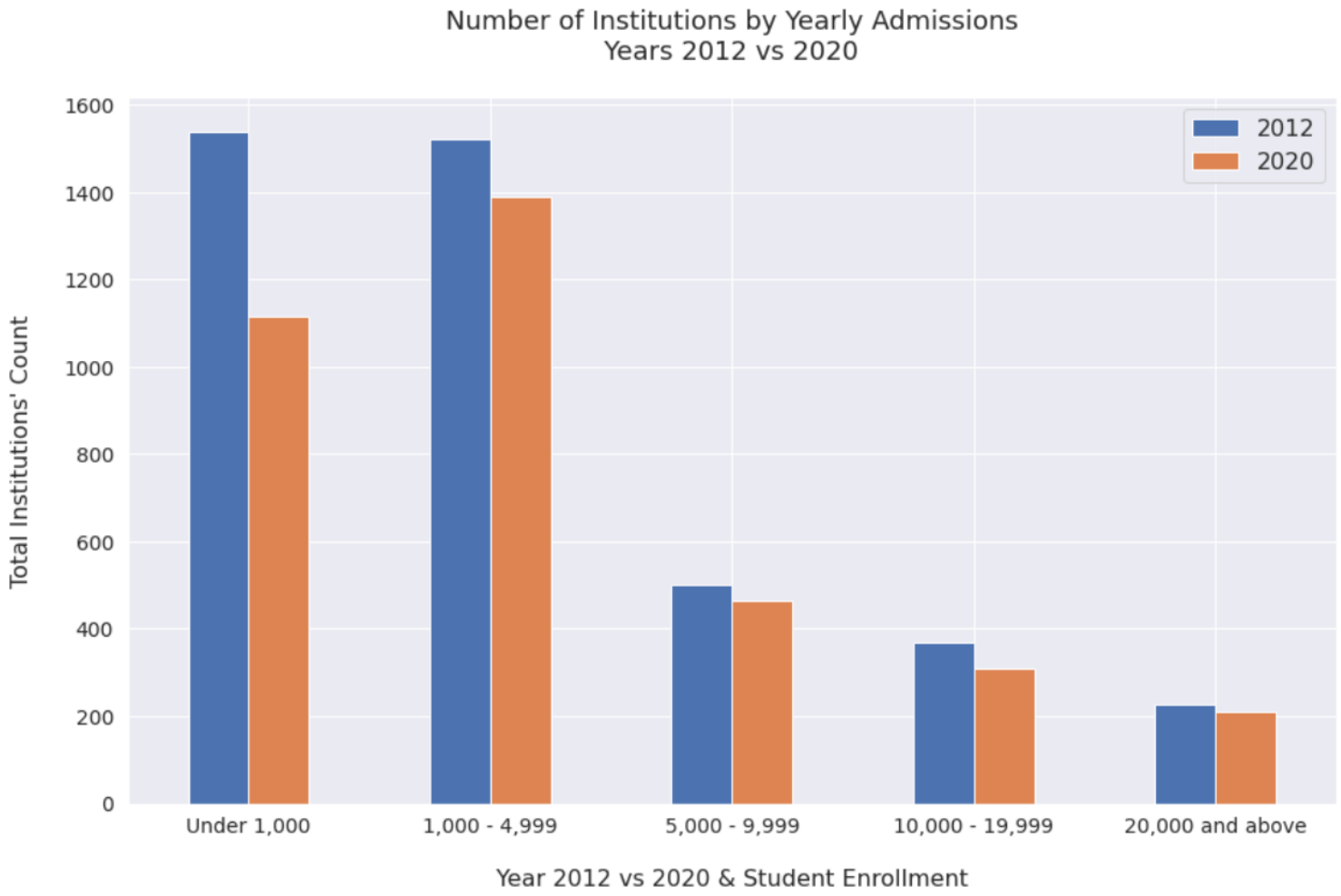
HEIs ~ Contraction across almost all States



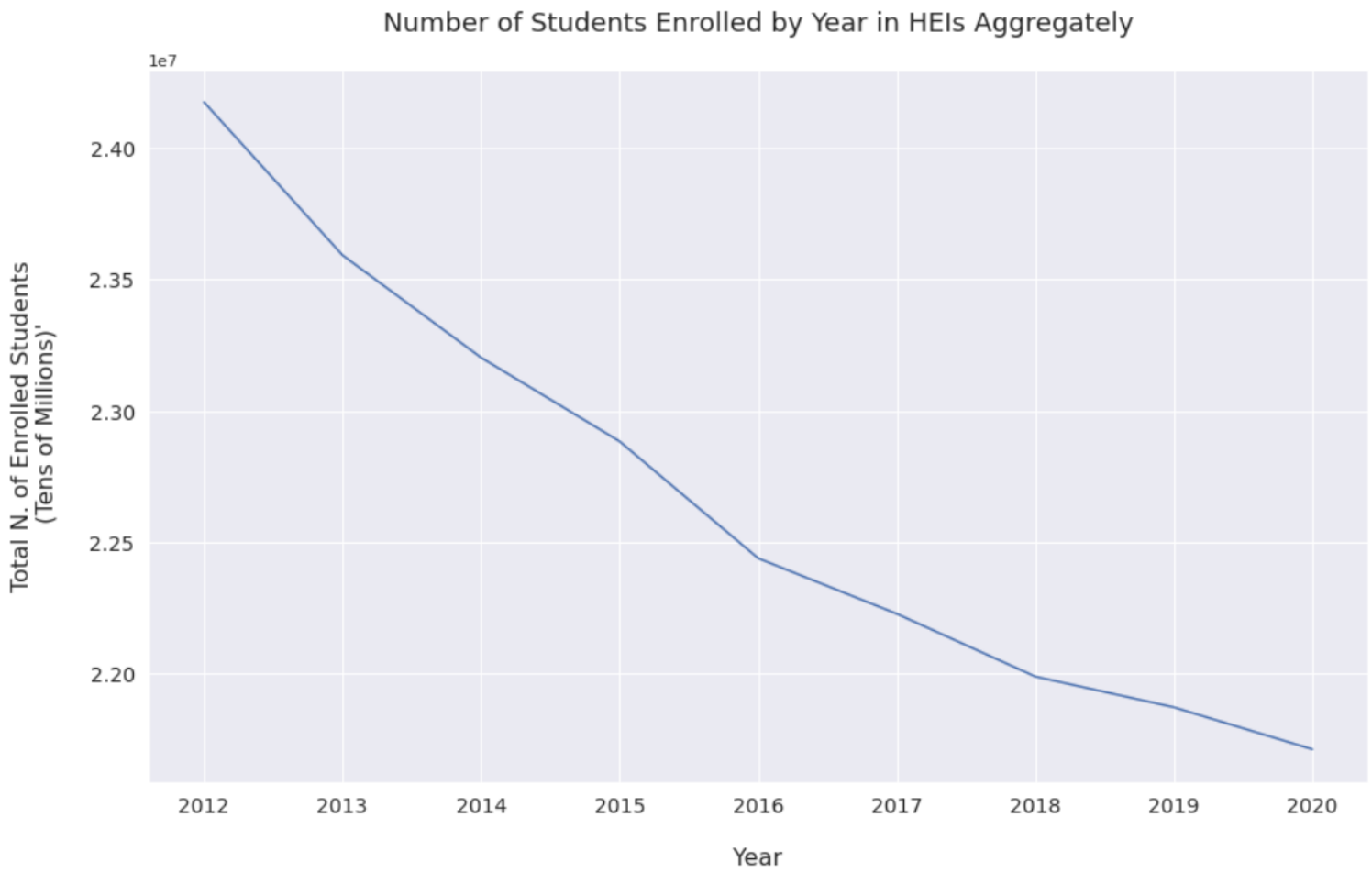
HEIs ~ Contraction across almost all Economic Areas



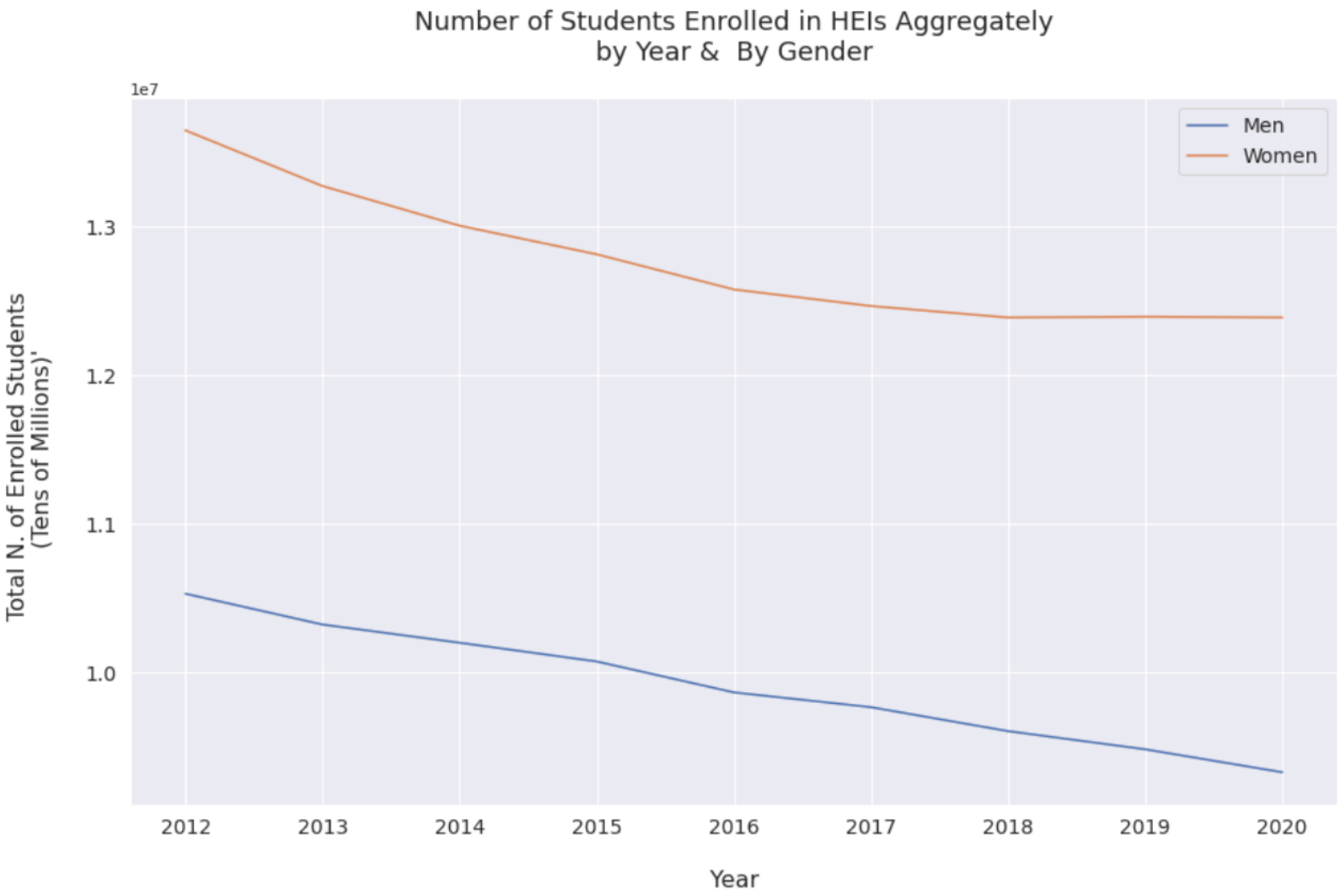
HEIs ~ Significant Contraction in Private For-Profit Institutions


Number of Institutions by Yearly Admissions
Years 2012 vs 2020
Total Institutions' Count
1600
1400
1200
1000
800
600
400
200
0
2012
2020
Under 1,000
1,000 - 4,999
5,000 - 9,999
10,000 - 19,999
20,000 and above
Year 2012 vs 2020 & Student Enrollment

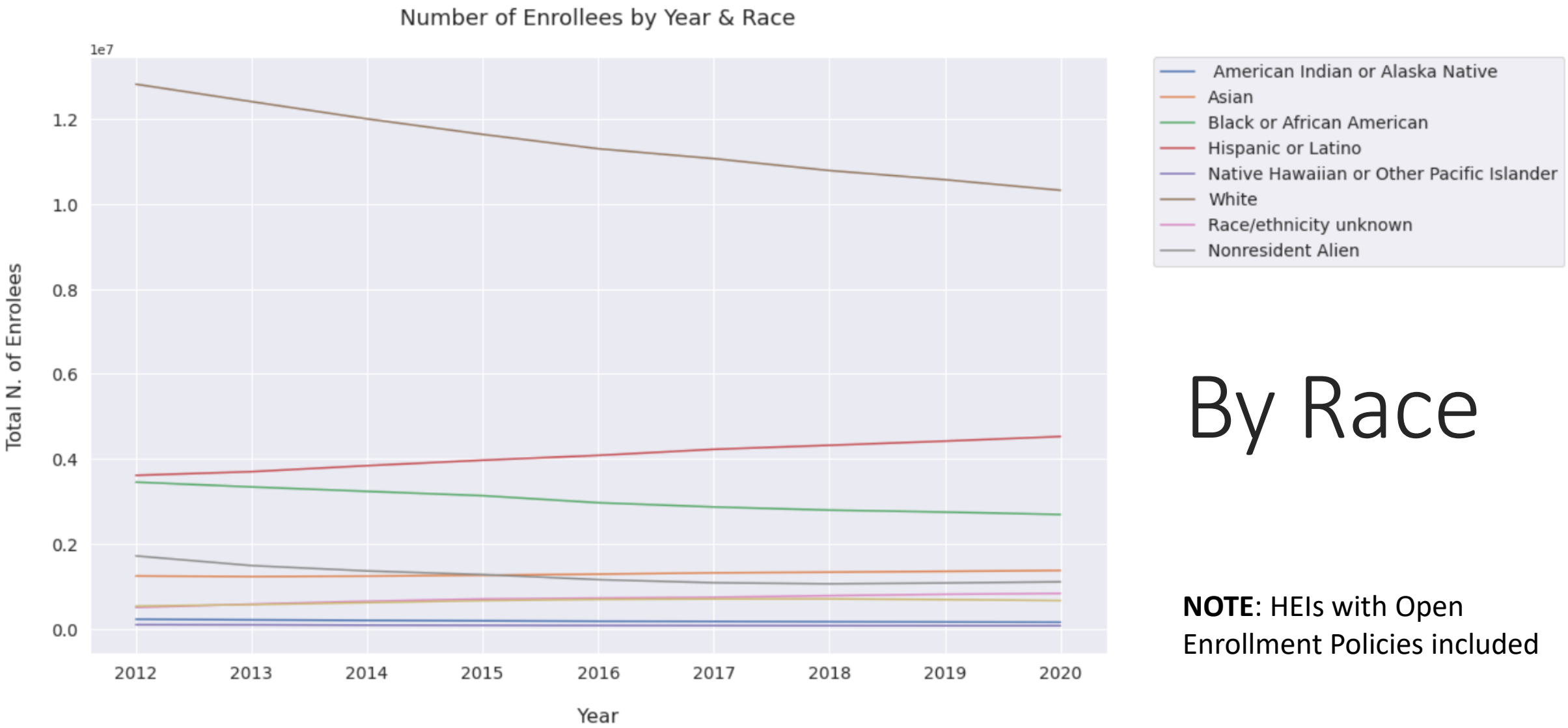
By Size


Number of Students Enrolled by Year in HEIs Aggregately
1e7
2.40
2.35
2.30
2.25
2.20
Total N. of Enrolled Students
(Tens of Millions)'
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year

Enrollment

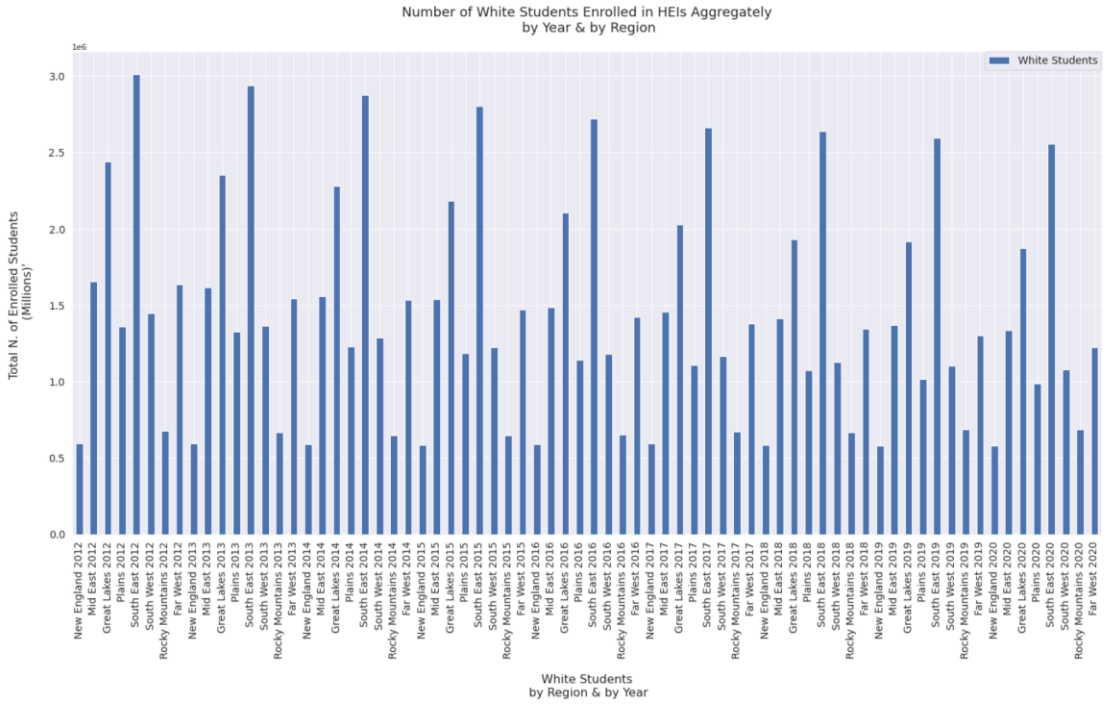

Number of Students Enrolled in HEIs Aggregately
by Year & By Gender
1e7
1.3
1.2
1.1
1.0
Total N. of Enrolled Students
(Tens of Millions)'
Men
Women
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year

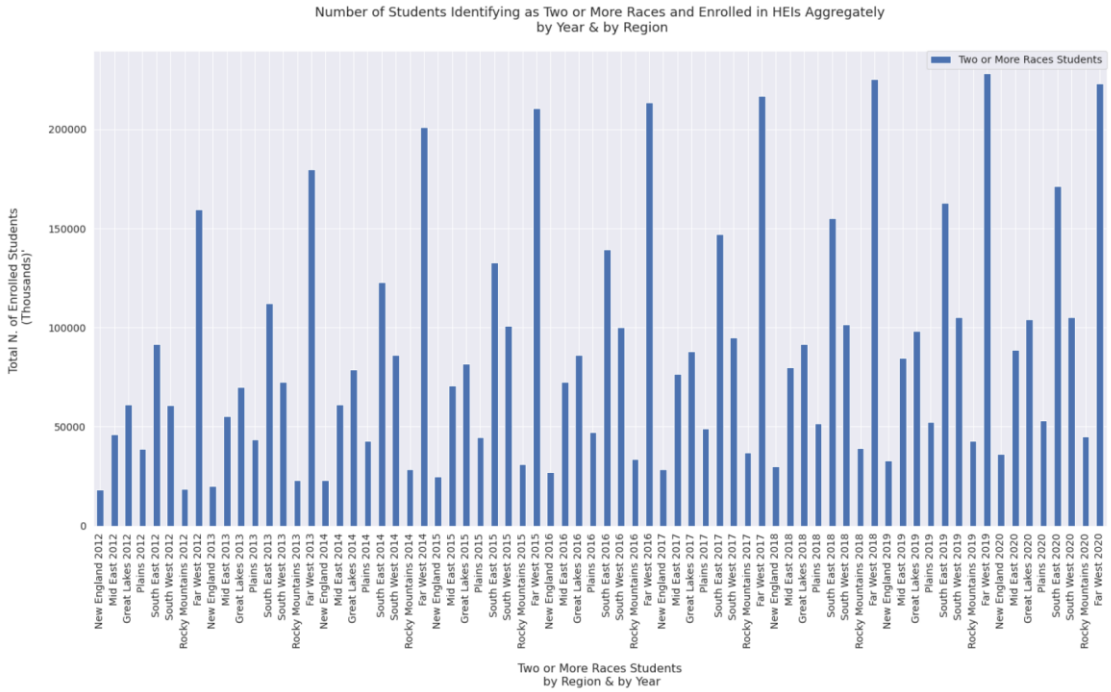
By Gender



By Race

NOTE: HEIs with Open Enrollment Policies included





Whites vs Two or More Races

Step I

Focus on 2020 Data to select best approach

Step II

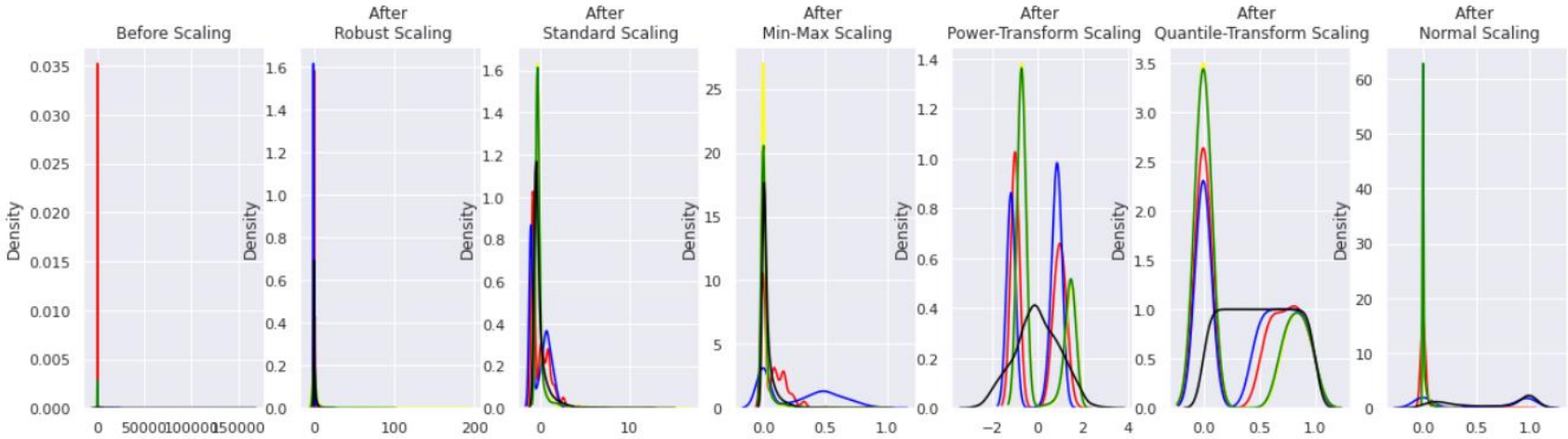
Apply Strategy on 2012-2020 Data

2020 Data

- 45 Variables (5 Numerical Variables vs the rest Categorical
  Application Fees; Room and Board fees (aggregately);
  Total n. SAT scores ;Total n. ACT Scores ; Total n. of Enrollments (Target)

- Unbalanced Data (Preprocessing required)
  - OneHotEncoder vs. – RobustScaler

- Linear Regression



2020 Data

- Linear Regression

- 67% accuracy on Training data vs. 60% accuracy on Testing data

- Explained Variance score (also known as the Coefficient of Determinatior R Squadre) is only 60%

- Predictions consistently underestimate Actual data

- Causes:
  - Outliers
  - Unbalanced data
  - Simple test case no feature pre-selection

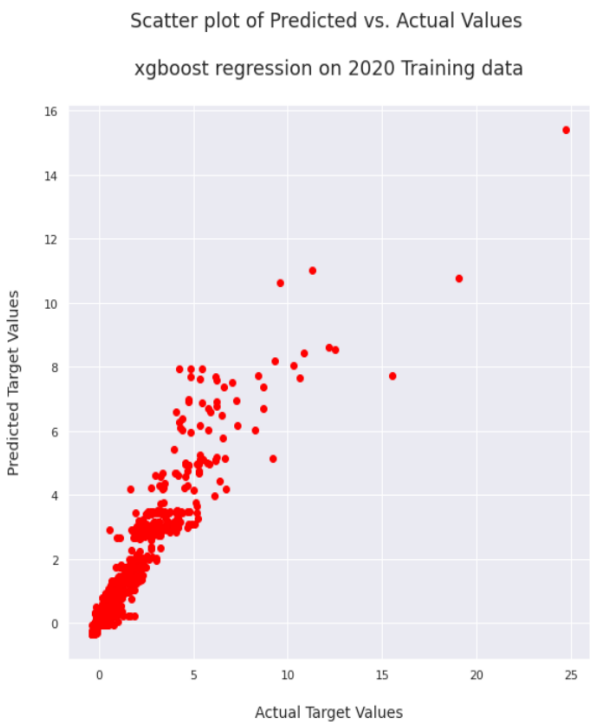
2020 Data

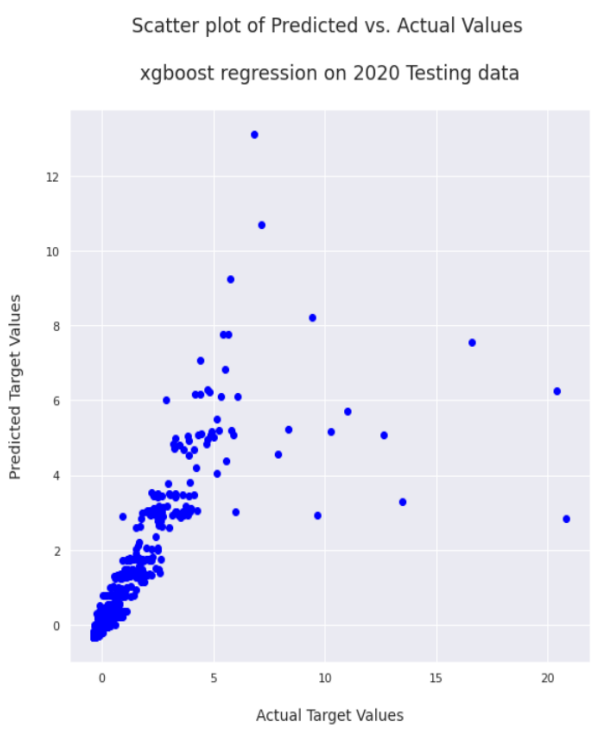
- XGBoost Regression–
  gradient descent paired with random tree selection to handle large datasets and improve predictions

- Training data accuracy of approximately 90%

- Testing data accuracy of about 68%

- Overall improvement of about 13%.





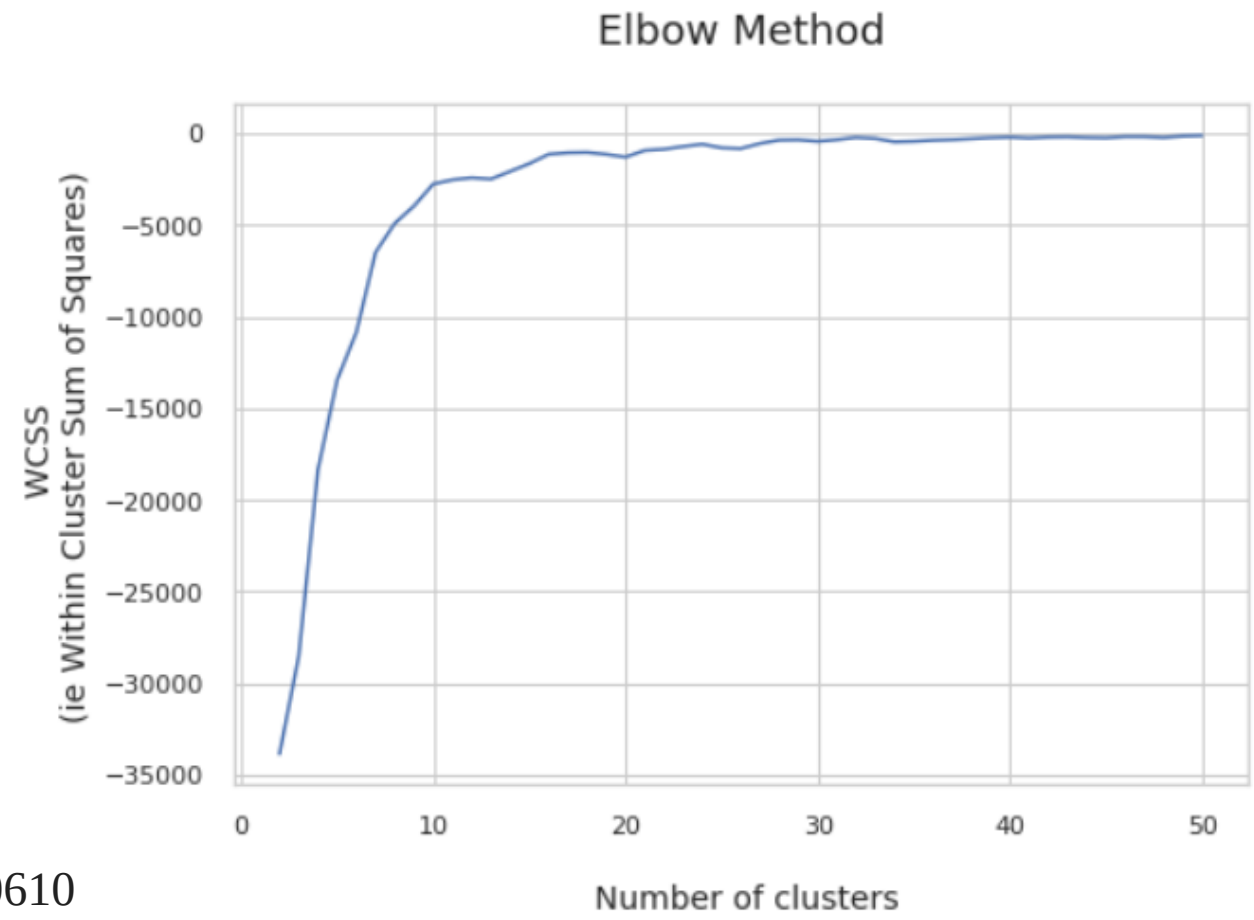
2020 Data ~ K Means

- **Clustering by State**

- Tested k value between 2 and 50

  Best 9 ~ Economic areas

- Gradient of the inertia is negative =
  = function is decreasing
  (Distance between points and centroids)



Homogeneity score for 9 number of clusters is: 0.0610
Completeness score for 9 number of clusters is: 0.1645

2020 Data ~ K Means

- **Clustering by Institutional Size**

Homogeneity score for 2 number of clusters is: 0.08
Completeness score for 2 number of clusters is: 0.62

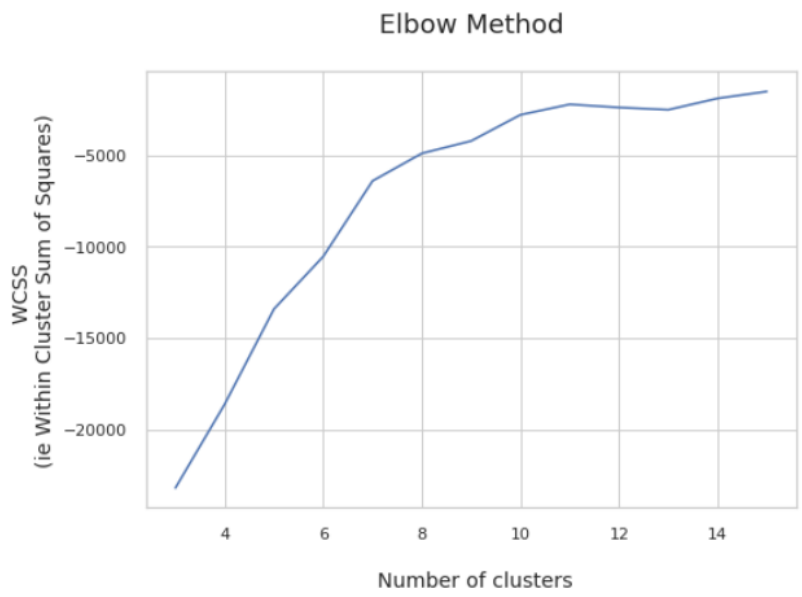
Homogeneity score for 3 number of clusters is: 0.13
Completeness score for 3 number of clusters is: 0.51

Homogeneity score for 4 number of clusters is: 0.14
Completeness score for 4 number of clusters is: 0.42

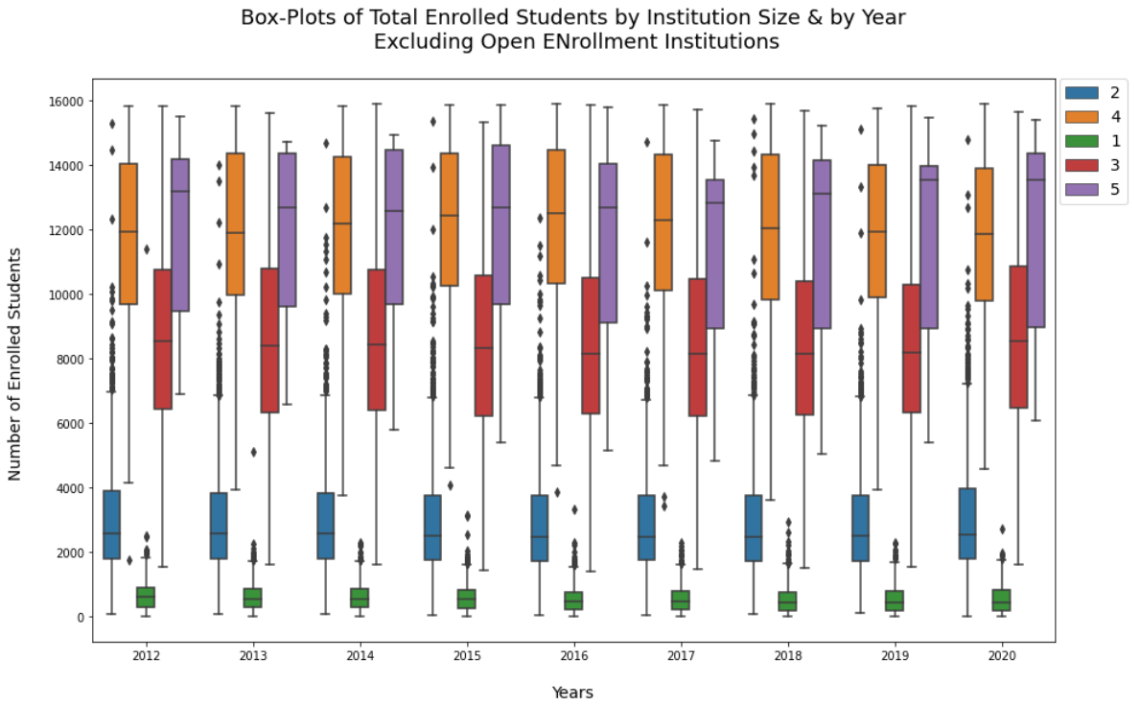
Homogeneity score for **5** number of clusters is: 0.15
Completeness score for **5** number of clusters is: 0.34

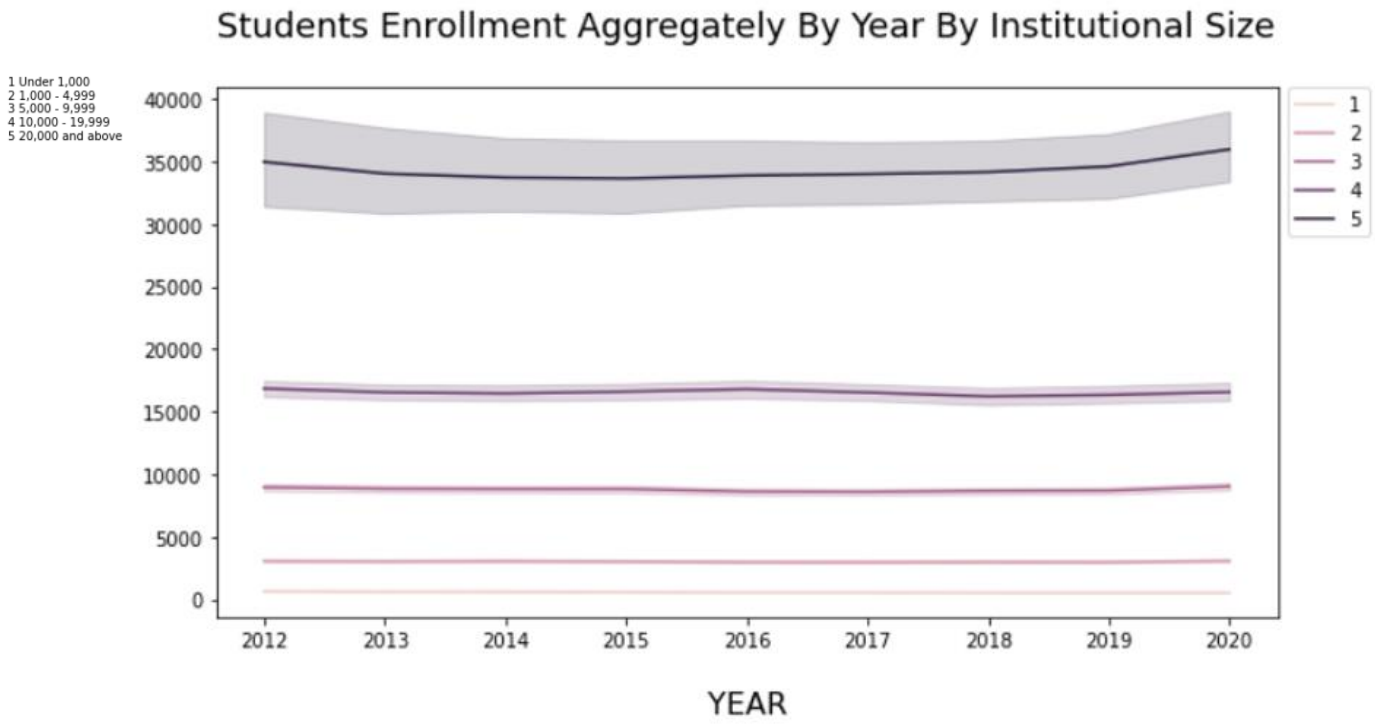
Homogeneity & **Completeness**
Both Higher than for k clusters of States!

```
Classification Report:
                    precision    recall  f1-score   support

1       << 1,000         0.37      1.00      0.54       292
2       1,000 - 4,999    0.00      0.00      0.00       377
3       5,000 - 9,999    0.00      0.00      0.00       134
4       10,000 - 19,999  0.28      0.34      0.31        98
5       20,000 <<        0.64      0.15      0.24        62

           accuracy                           0.35       963
          macro avg      0.26      0.30      0.22       963
       weighted avg      0.18      0.35      0.21       963

Accuracy:  0.34683281412253375
[-23211.20706272 -18626.45440892 -13418.61458282 -10543.8090176
  -6408.39391476  -4902.07411114  -4222.99300745  -2792.44541542
  -2219.945871    -2391.54757732  -2515.55385751  -1892.9195458
  -1516.2906226 ]
```



2012 – 2020 Data





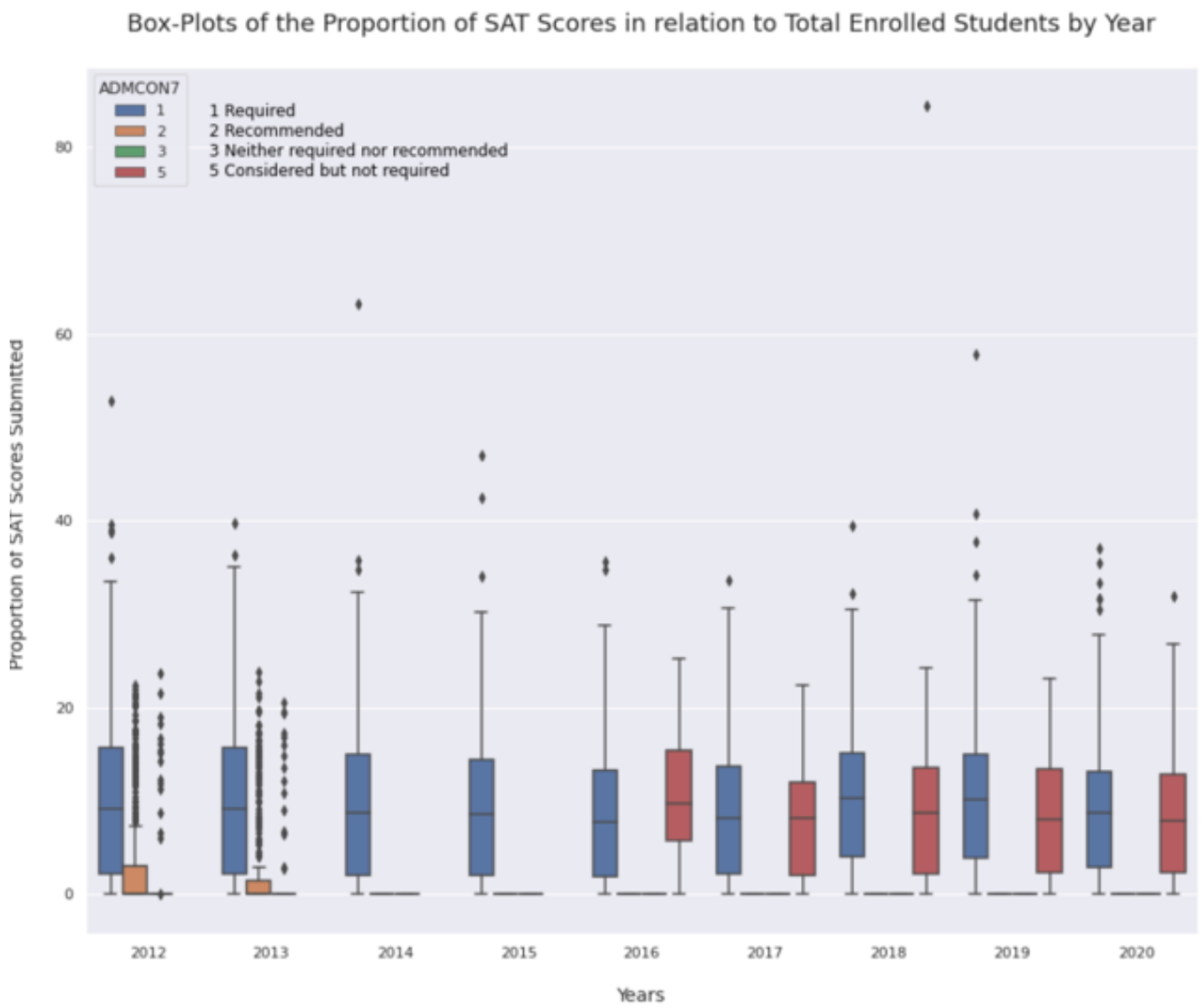
Students' Enrollment

- Student enrollment by Institutional Size
- Not fluctuating as much as we might expect
- Larger HEIs experienced the largest fluctuations over time

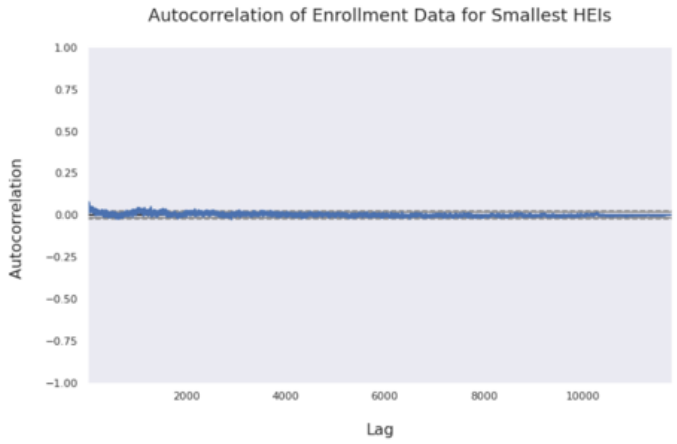
2012-2020 SAT Scores' Submissions

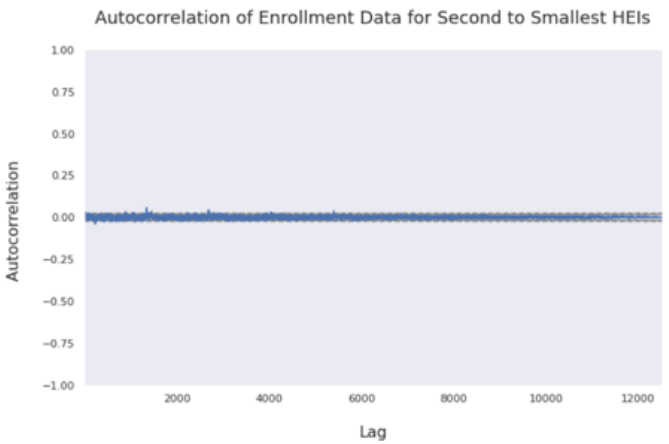
- Number of Institutions Recommending SAT Scores
  vs
  HEIs Considering but not requiring SAT Scores

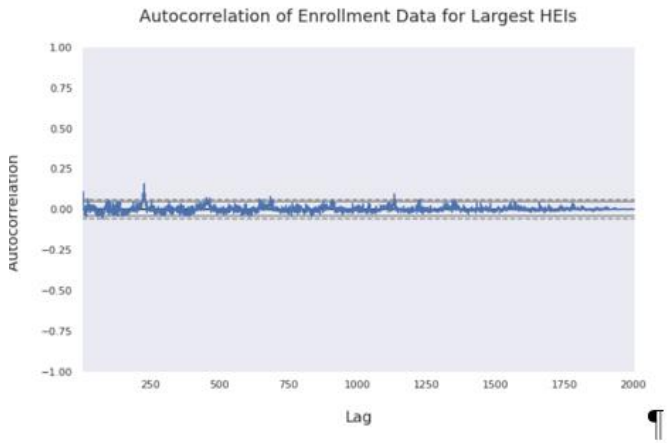
- Change in policy or change in reporting?
- (Same is true for ACT Scores)

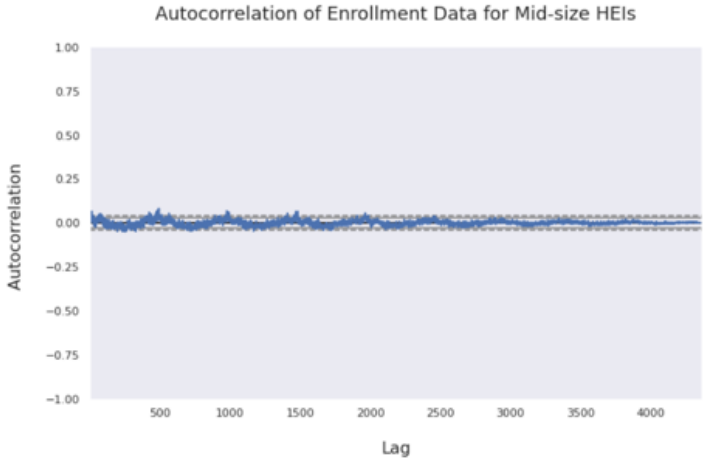


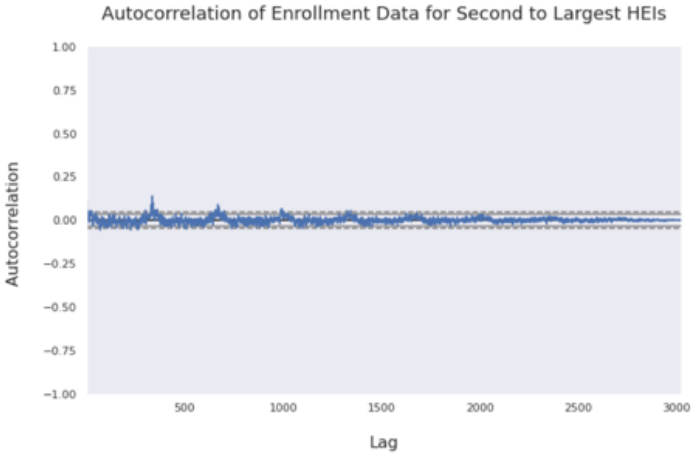
2012 – 2020 Enrollment By HEIs Size & Lag





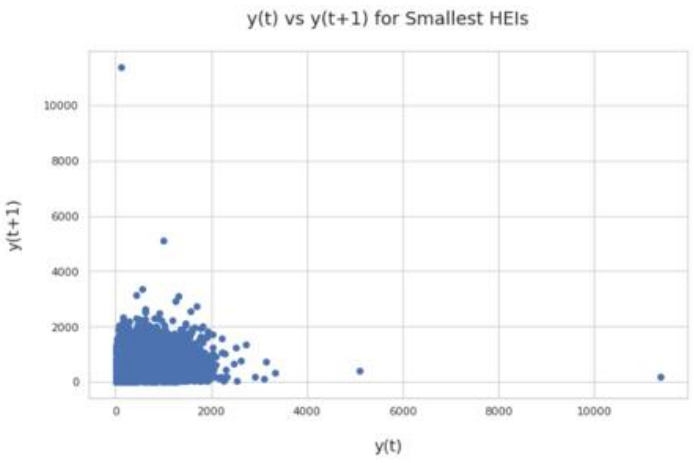


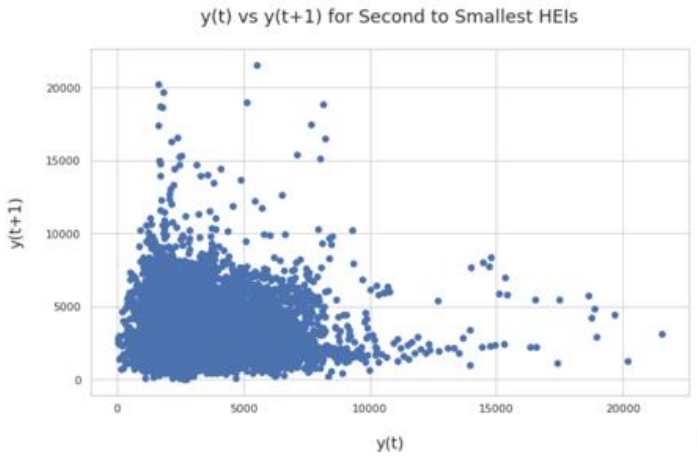


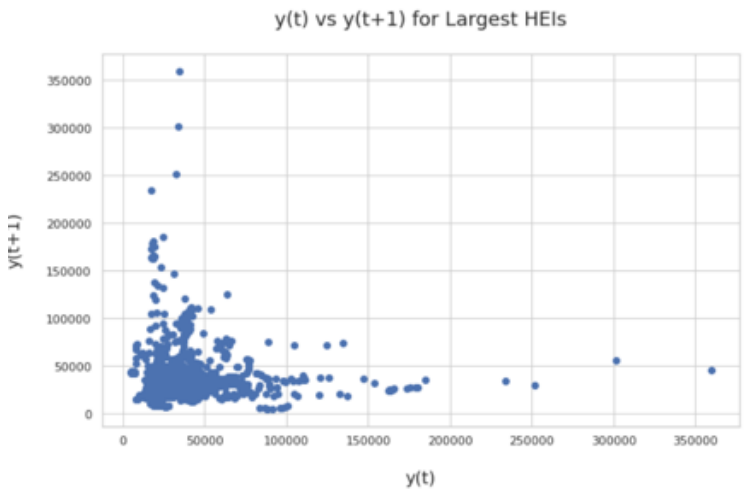


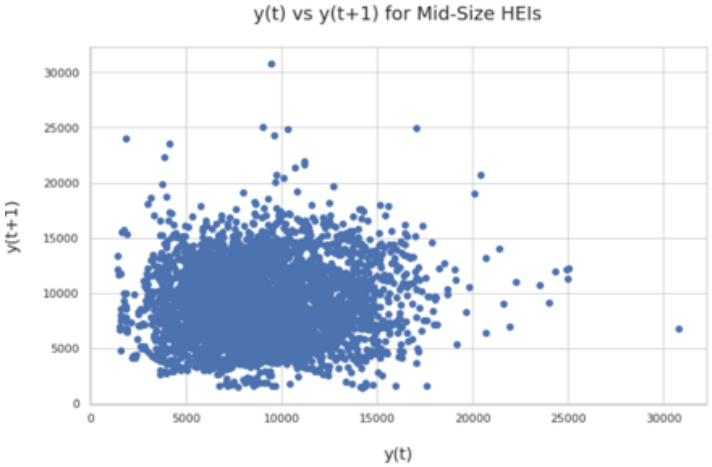
- No particularly significant autocorrelation for the data
- Larger HEIs may be the most autocorrelated with previous levels of enrollment

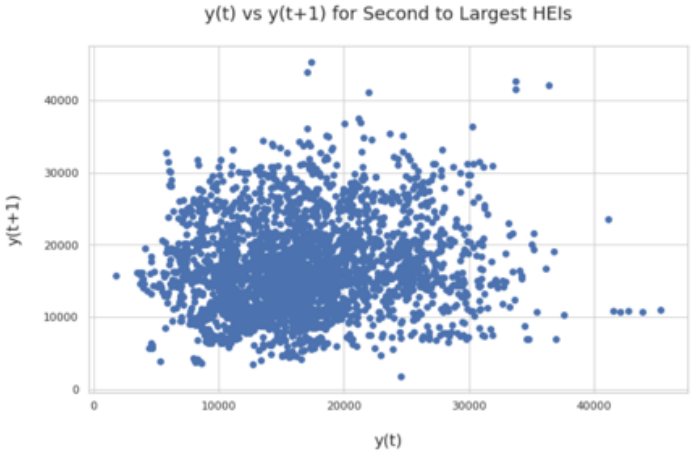
Enrollment By HEIs Size & Lag





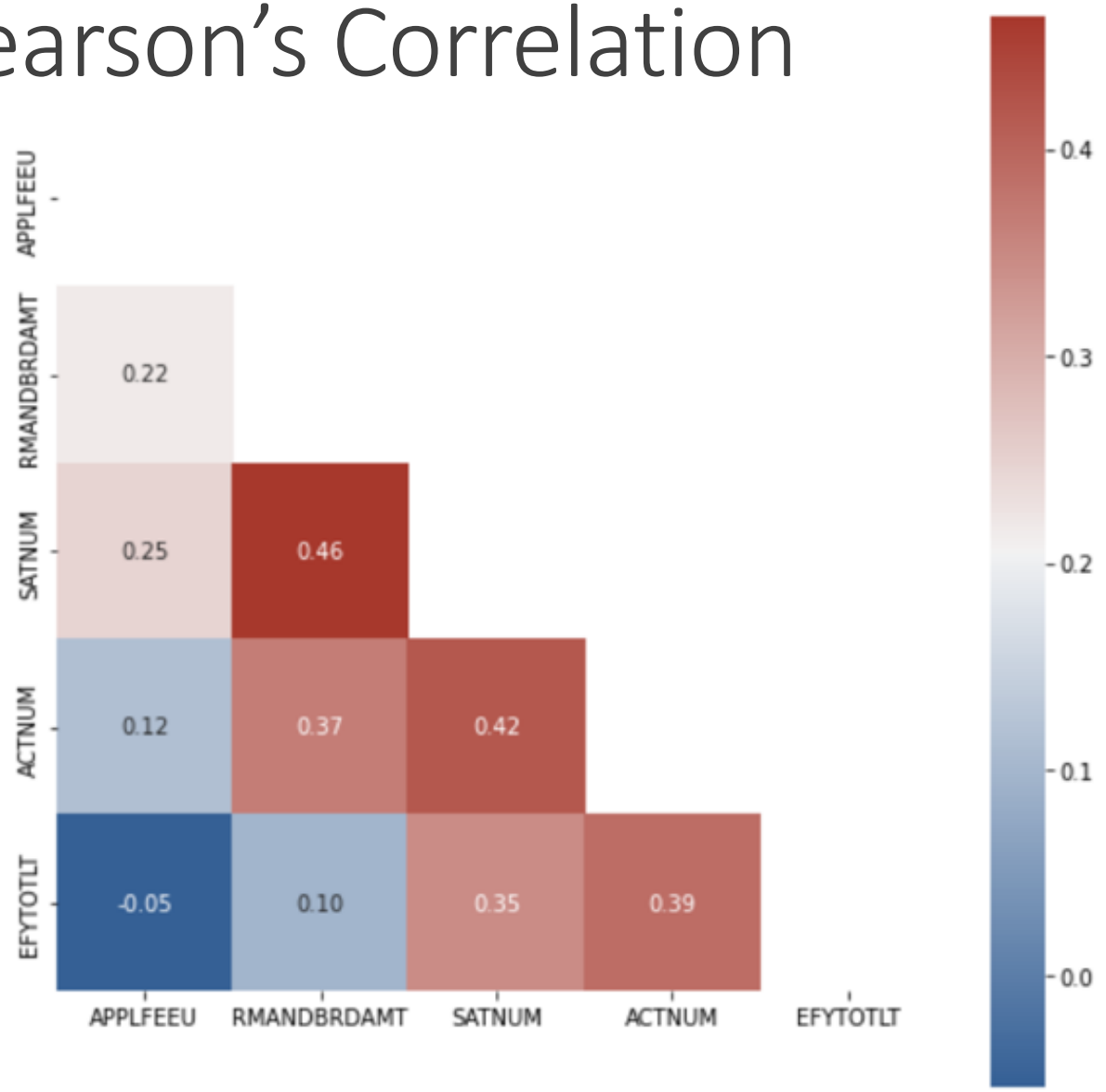


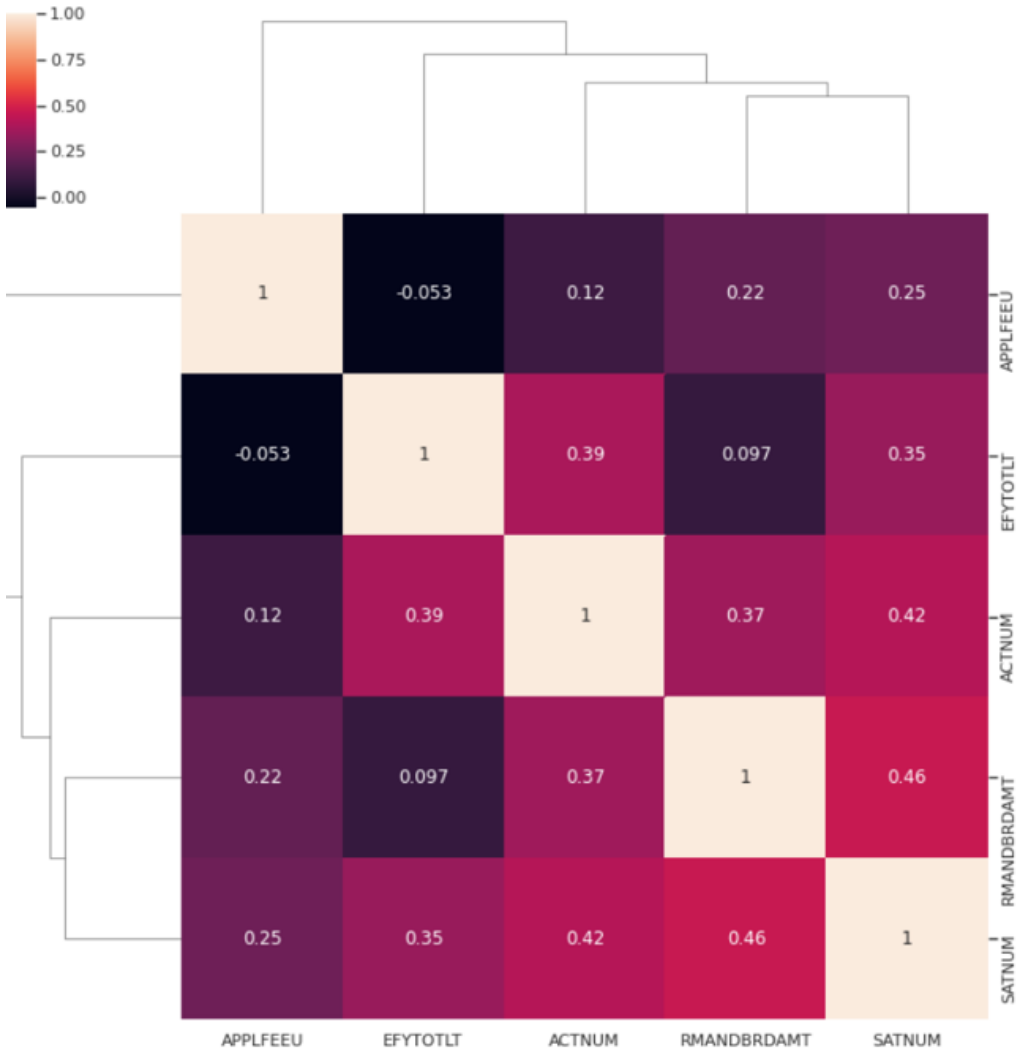




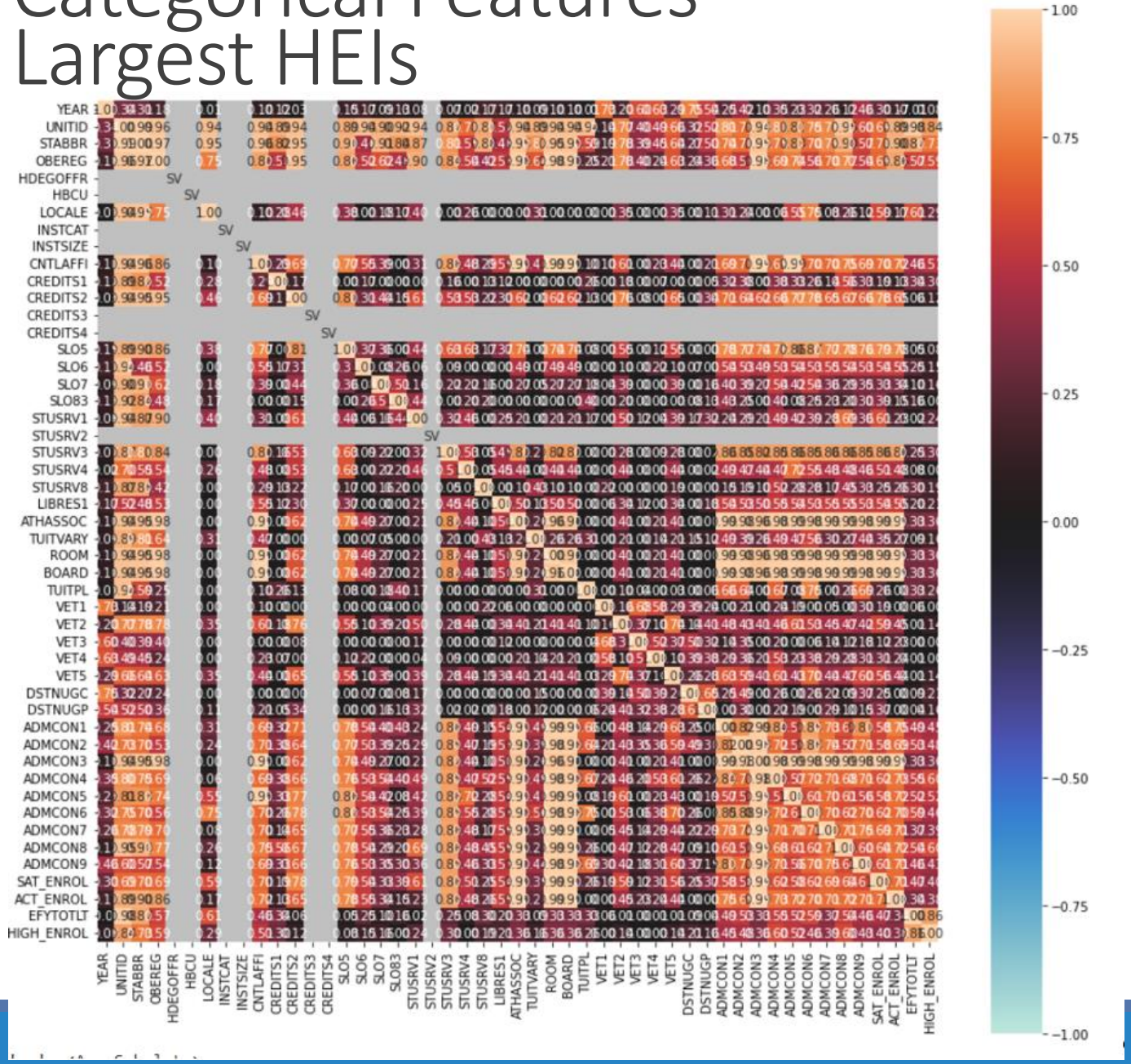
- No particularly significant autocorrelation for the data
- Larger HEIs may be the most autocorrelated

Quantitative Features Pearson's Correlation





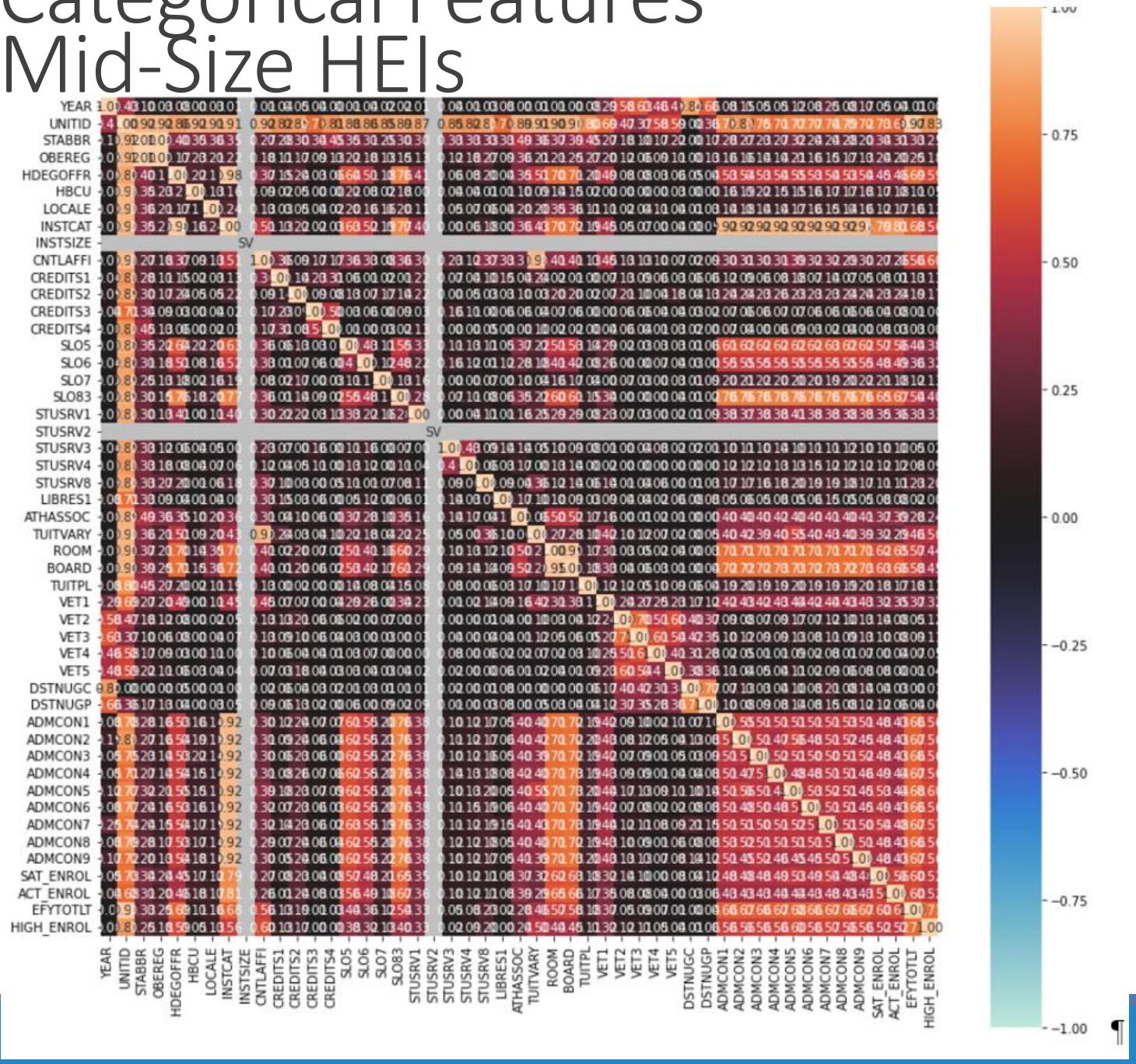
Categorical Features Largest HEIs



Largest HEIs ~ Categorical Variables' Associations

- Location seems to matter
- Traditional High School related variables are associated to enrollment
- Gives us a sense of how uniform this HEIs are
- None are are HBCU
- All offer both Undergraduate and Graduate programs (which makes sense)
- Most are Private organizations (86 vs 8) with the ones Not for profit constituting the larger portion (72 vs 14)
- All accept Advance Placement credits (at least on paper)
- All have academic counseling offices

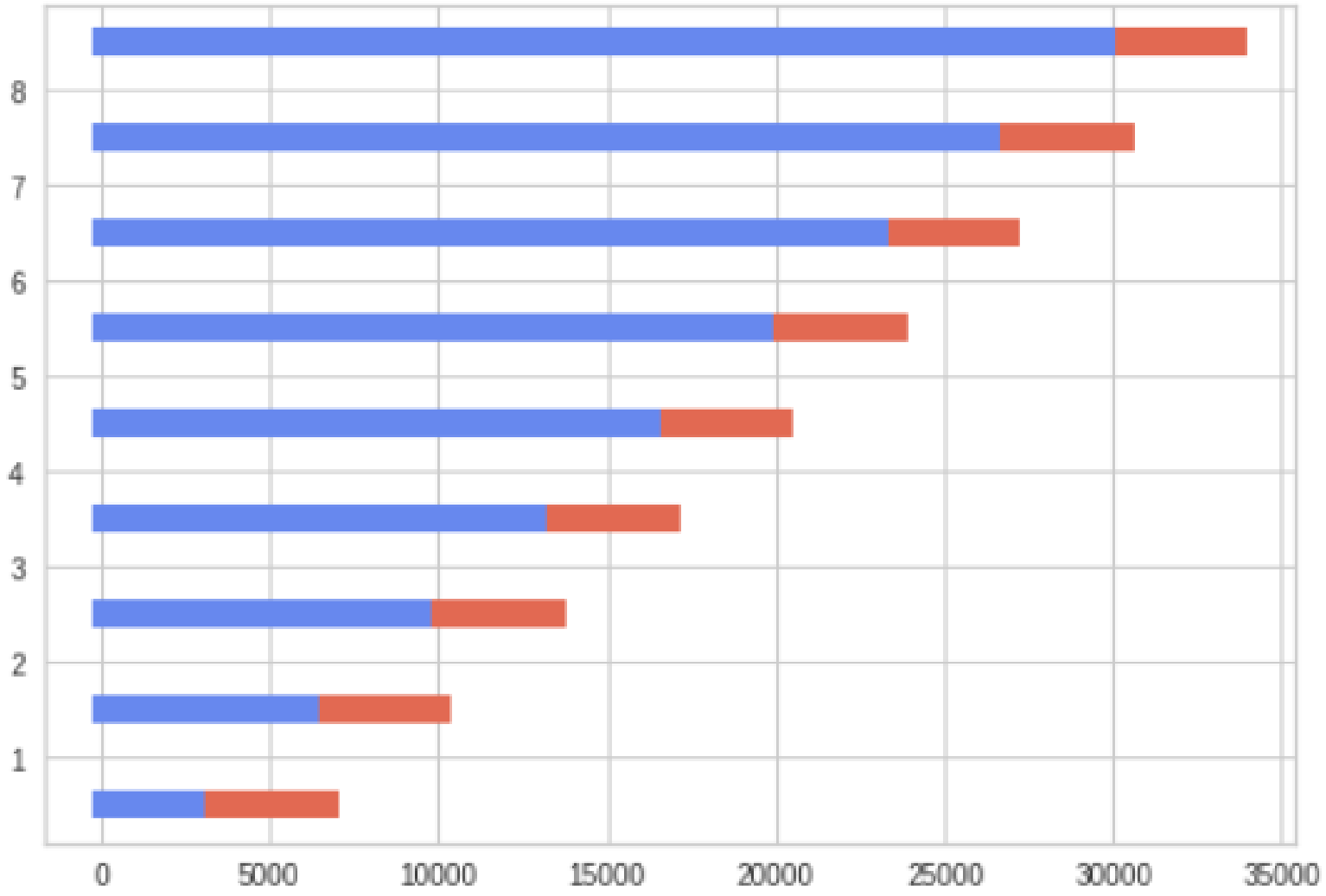
Categorical Features
Mid-Size HEIs



Mid-Sized HEIs ~ Categorical Variables' Associations

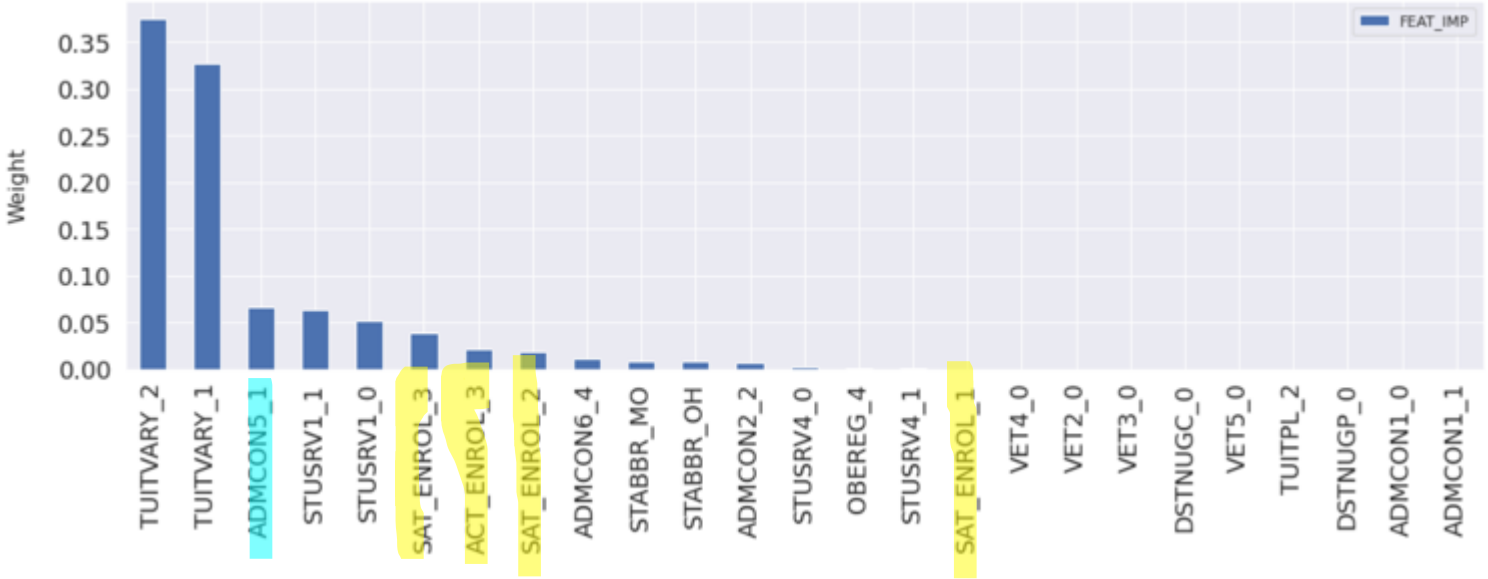
- Greatest number of categorical variables are relevant
- Traditional High School related variables are associated to enrollment
-

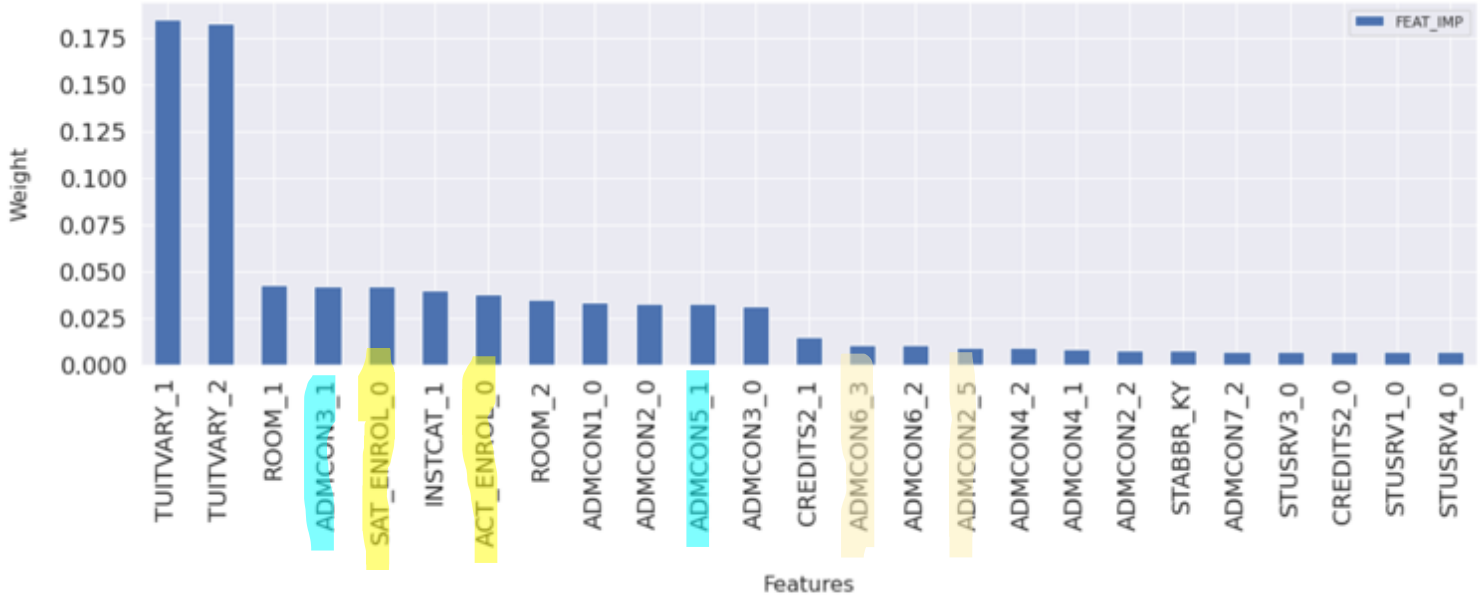
Time Split Method



Walk forward method of selecting Time-Series data

XGBoostRegression ~ Largest HEIs



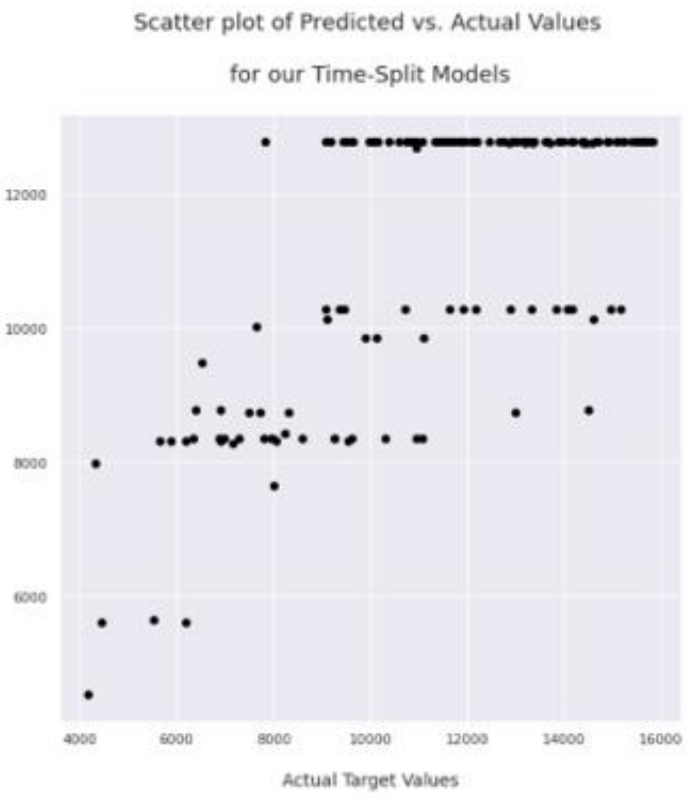


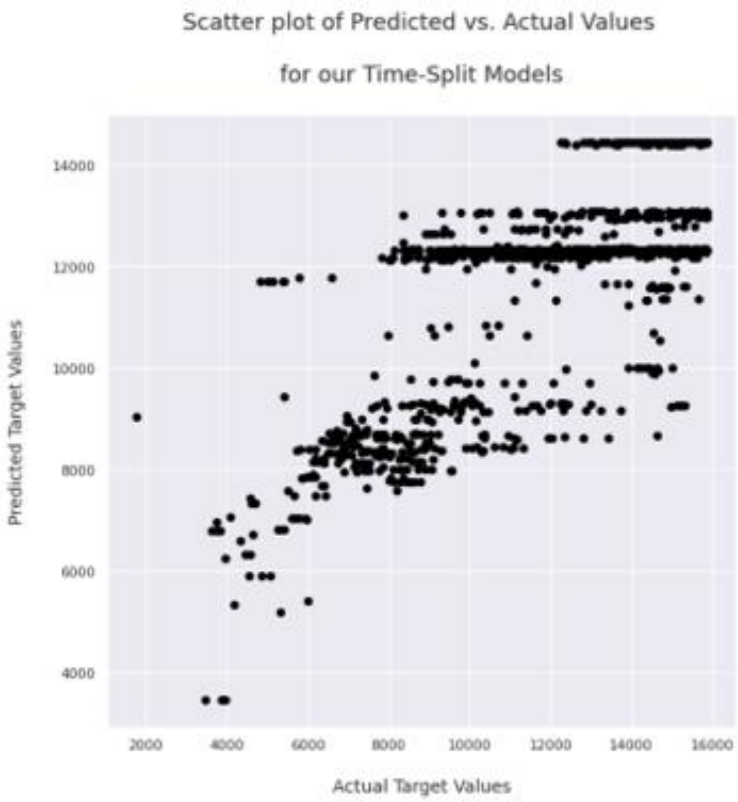
1st Split versus Last Split ~ Largest HEIS

- In State vs Out of State Tuition cancels itself out
- Change in policy and importance with respect to SAT Scores (highlights yellow)
- Recommendations & Secondary School records remain Required for enrollees in many of the Largest schools (highlight Blue)
- Formal demonstration of competencies & School Rank are losing importance (Highlight peach)

1	Required
5	Considered but not required
2	Recommended
3	Neither required nor recommended

XGBoostRegression ~ Largest HEIs





¶

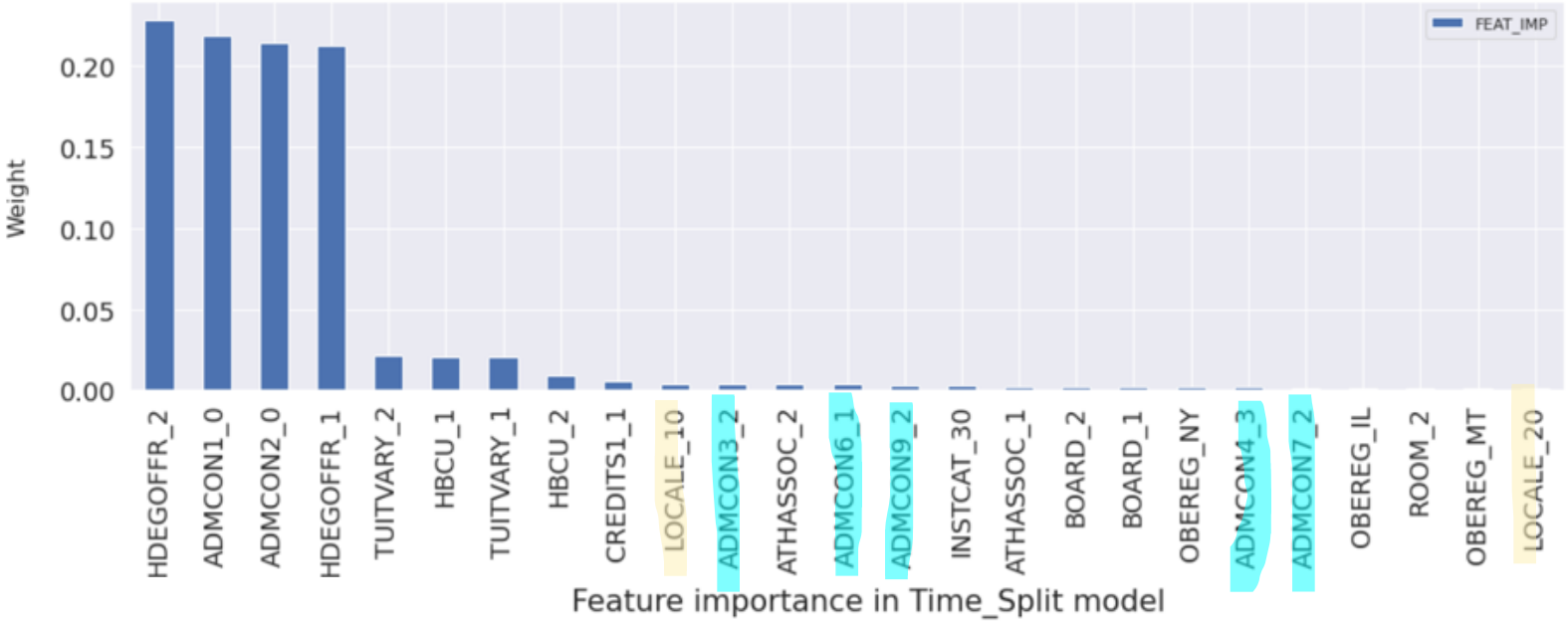
1st Split versus Last Split ~ Largest HEIS

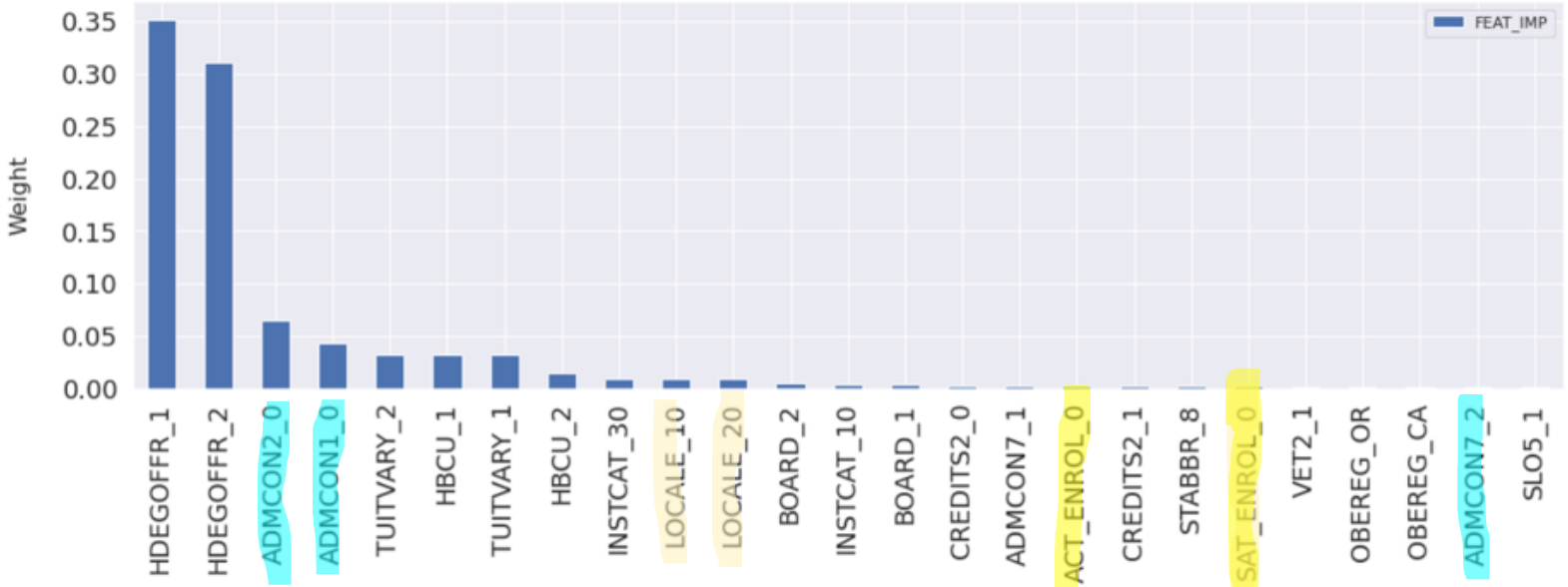
Root Mean Square Error (RMSE) around 2000

Training accuracy ranges between (**50.24%, 59.25%)**)
Testing accuracy between (**32%, 62.75%),**

Not a great job at predicting enrollment levels based on the features

XGBoostRegression ~ Medium Sized HEIs



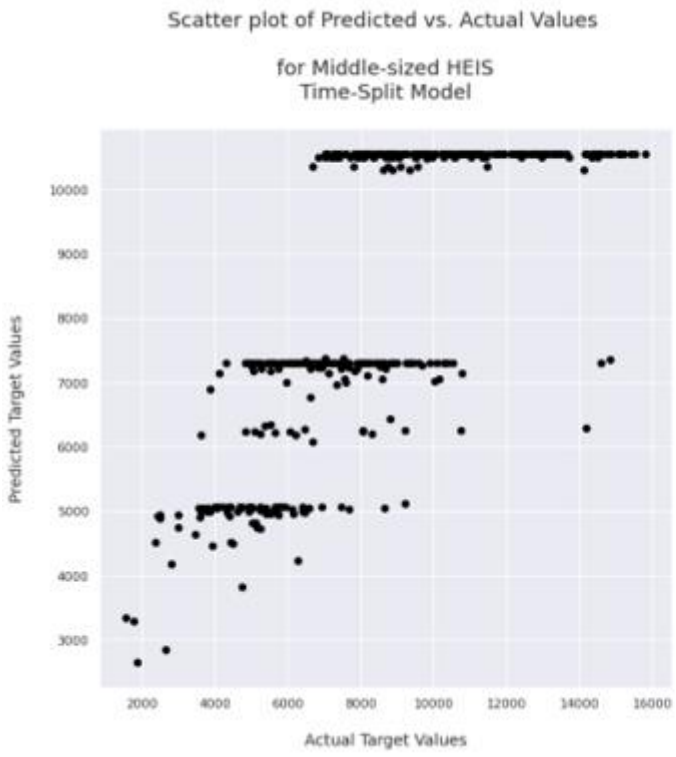


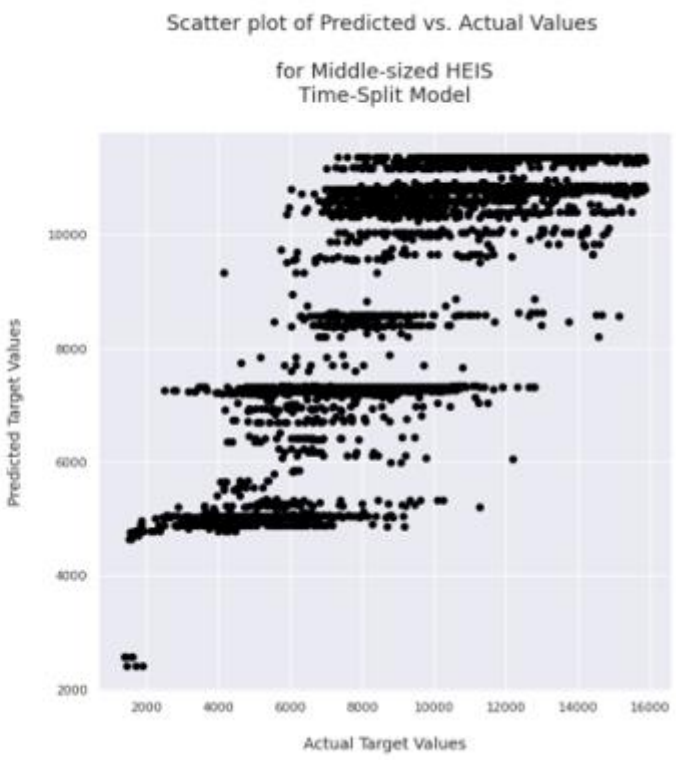
1st Split versus Last Split ~ Mid-Size HEIS

- Graduate Level Courses Cancel each other out (almost)
- HBCU status has lost distinguishing appeal
- SAT Scores are practically not reported Very low proportion & Low importance
- Location matters more
- Shift away from traditional academic reports as a shift in presence of these features appears (blue Highlights)

BUT distinguishing factors between HEIs: Secondary School Rank (2), HS GPA (1) Secondary School Records (3) , Admission Test Scores (7) (Recommended but less influential)

XGBoostRegression ~ Medium Sized HEIs

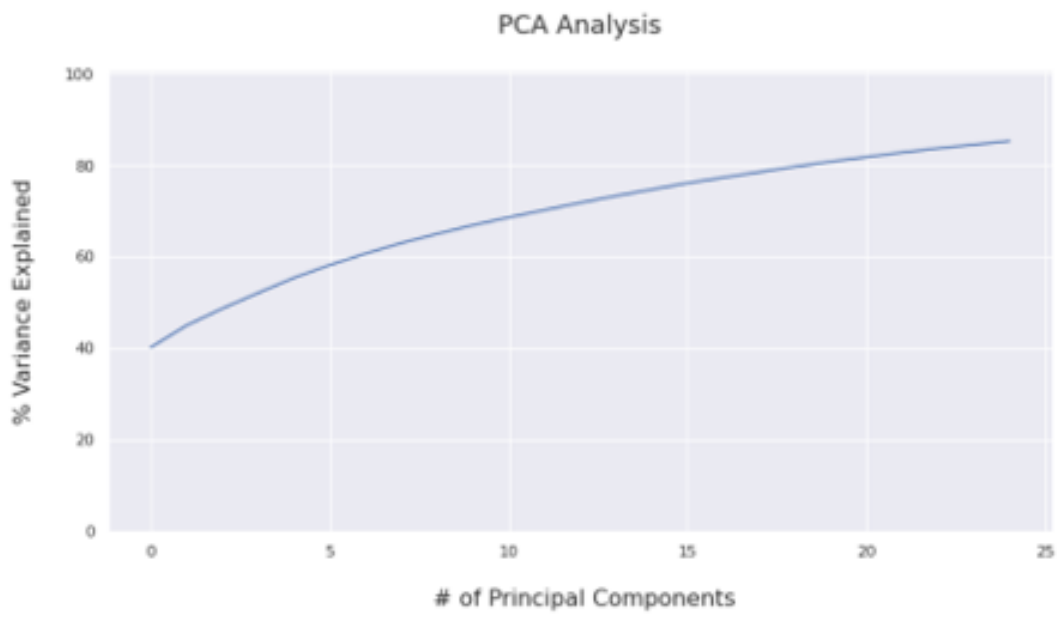


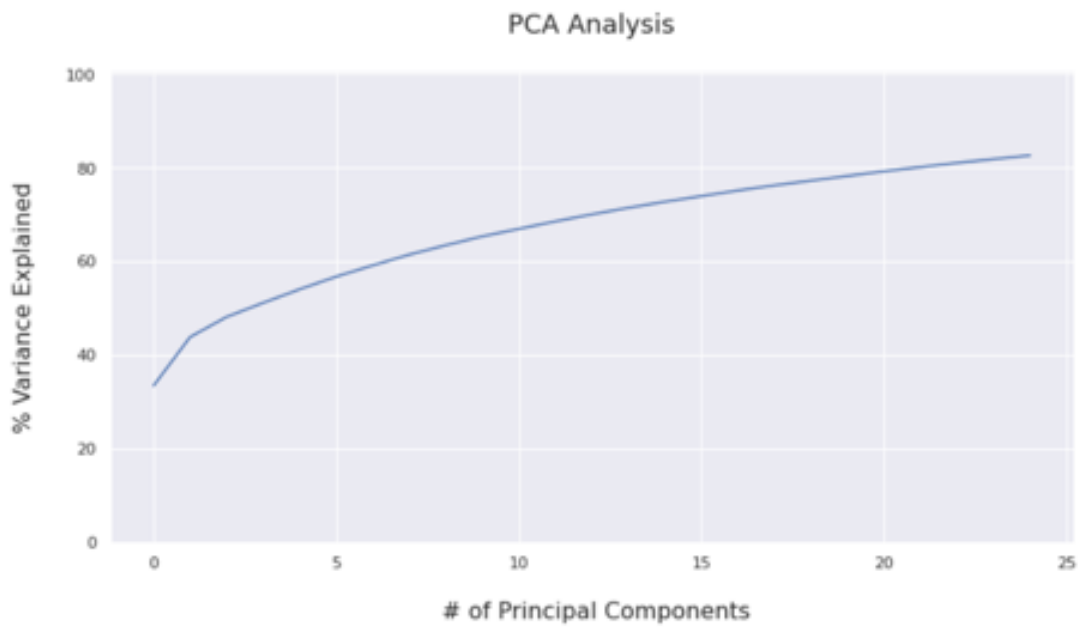


1st Split versus Last Split ~ Mid-Size HEIS

Root Mean Square Error (RMSE) for this size of HEIs ranges between about (1854 and 2074),
Training accuracy ranges between (56.77% and 62.7%)
Testing accuracy between (54.19% and 60.54%)

Better job at predicting enrollment levels based on the features, c
closer around our 45 degree line

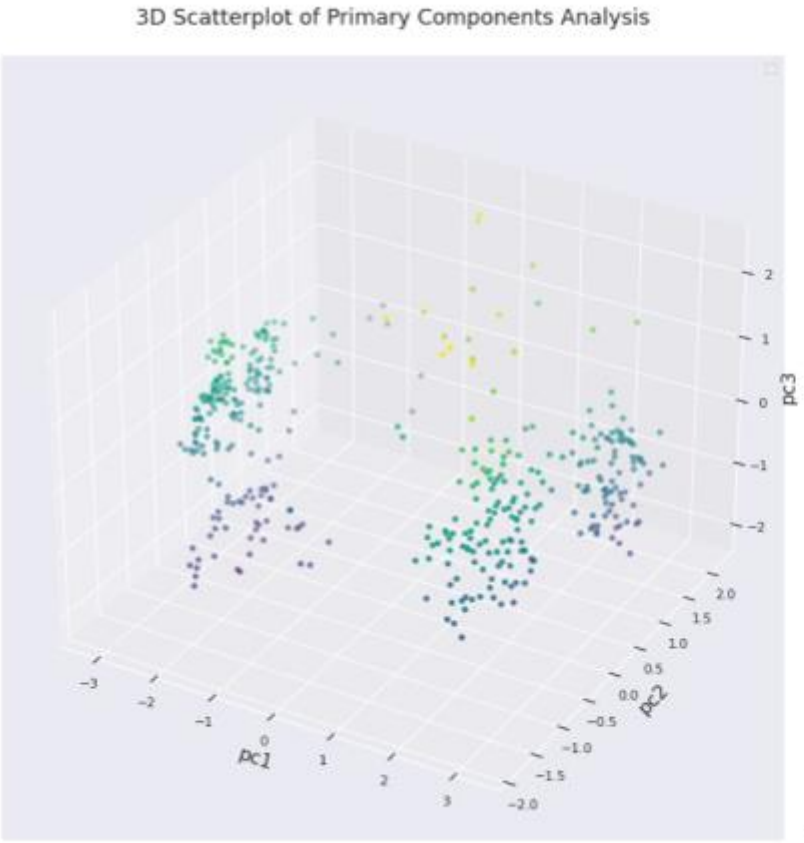


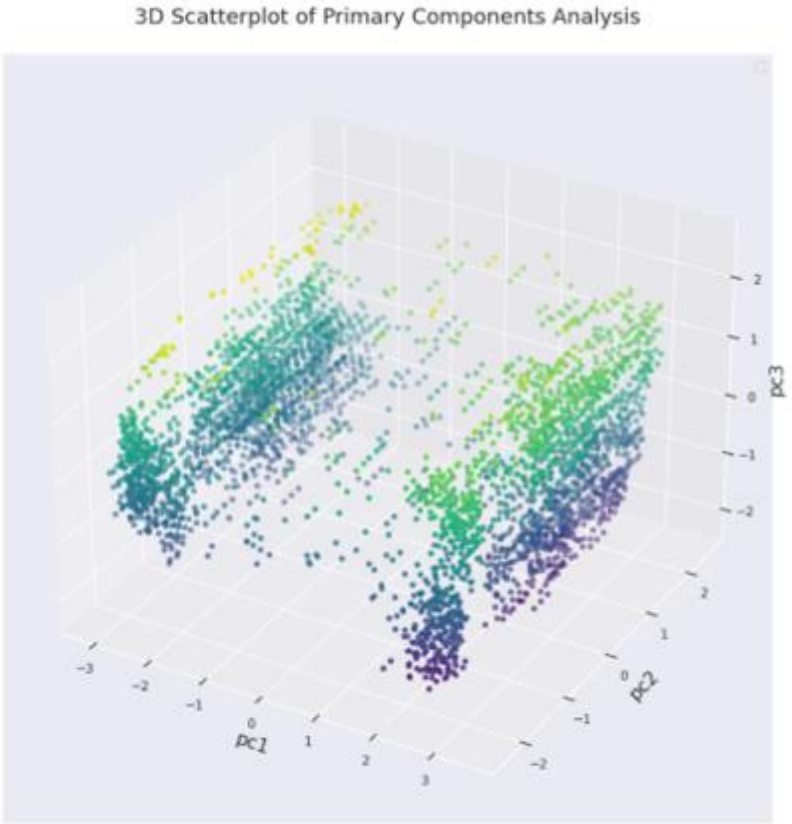


Principal Component Analysis

1st Split versus Last Split ~ Mid-Size HEIS

Cumulative Variance explained by first 3 features 90%





Principal Component Analysis & DBSCAN - Density Based Spatial Clustering Algorithm with Noise

Clustering of data reflects Features importance

Conclusions & Questions

Reported features affect student enrollment choice – BUT we did NOT prove Causal Effects

Changes in enrollment patterns exist – BUT are they reflecting changes in Society or Policy Decisions by HEIs?

Standardized Tests are being used less (differently?) in relation to Students – BUT why? Economics vs Equity?