DATA 606 Capstone Project

Part 1 ~ by Carlotta Amaduzzi
camaduz1@umbc.edu

as of February 27, 2022

Overview

- 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

The questions I am interested in looking into are the following:

1) Based on publicly reported information regarding HEIs, what features seem to affect student enrollment choice the most?

2) With an eye to I.D.E.A. (Inclusion, Diversity, Equity, and Access), do HEIs with different structural characteristics, fare differently across the US?

3) Are new policies adopted by HEIs, such as standardized-tests-blind admission policies, having an effect on students' enrollment?

The Data

- The data initially used for this project is all publicly available information
- 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 **different files** of data relative to HEIs
  - Institutional ID data
  - Institutions Offerings
  - Students' Enrollment data
- Focus was placed exclusively on **data directly reported** by the HEIs themselves

Data Cleaning

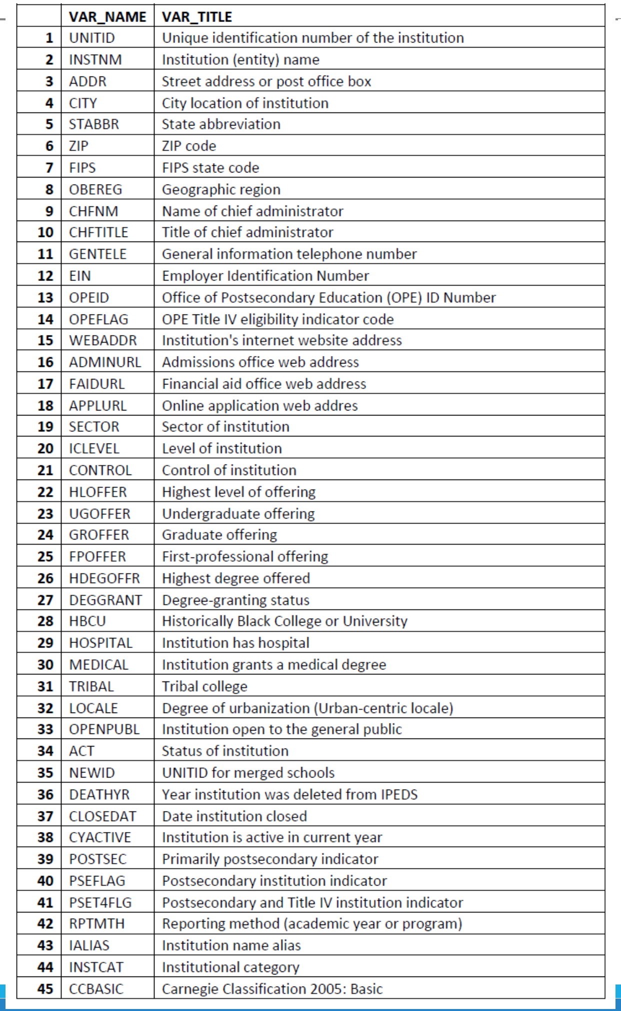
- During this initial phase of data cleaning primary focus has been placed on acquiring a sound sub-set of HEIs for which up-to-date data was available for the 2019-2020 academic year.

- Dictionary files accompanying each datafile have been extensively used to identify the salient features retained for the analysis and to understand the reported data

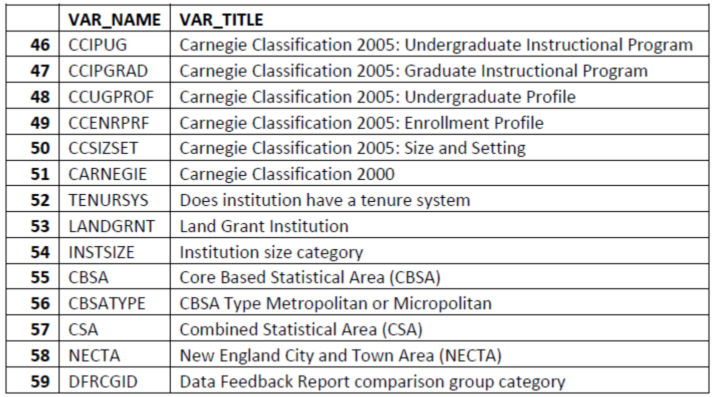
- The short-list of institutions selected will be used to complete a comparative analysis across time, if possible

Institutional Descriptive data

- **Data File**: hd2020.csv
- **Dictionary**: hd2020.xls
- The data file contains 59 different variables that help unequivocally identify each HEI
- Final Selected features: (after data cleaning)
  - 'UNITID’,
  - 'INSTNM’
  - 'STABBR’
  - 'OBEREG’
  - 'HLOFFER’
  - 'GROFFER’
  - 'HDEGOFFR’
  - 'HBCU’
  - 'LOCALE’
  - 'POSTSEC’
  - 'INSTCAT’
  - 'INSTSIZE'

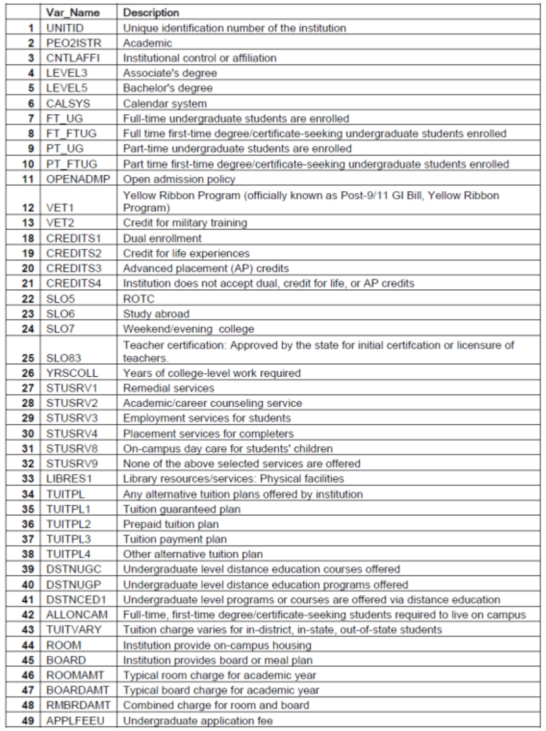


	VAR_NAME	VAR_TITLE
1	UNITID	Unique identification number of the institution
2	INSTNM	Institution (entity) name
3	ADDR	Street address or post office box
4	CITY	City location of institution
5	STABBR	State abbreviation
6	ZIP	ZIP code
7	FIPS	FIPS state code
8	OBEREG	Geographic region
9	CHFNM	Name of chief administrator
10	CHFTITLE	Title of chief administrator
11	GENTELE	General information telephone number
12	EIN	Employer Identification Number
13	OPEID	Office of Postsecondary Education (OPE) ID Number
14	OPEFLAG	OPE Title IV eligibility indicator code
15	WEBADDR	Institution's internet website address
16	ADMINURL	Admissions office web address
17	FAIDURL	Financial aid office web address
18	APPLURL	Online application web addres
19	SECTOR	Sector of institution
20	ICLEVEL	Level of institution
21	CONTROL	Control of institution
22	HLOFFER	Highest level of offering
23	UGOFFER	Undergraduate offering
24	GROFFER	Graduate offering
25	FPOFFER	First-professional offering
26	HDEGOFFR	Highest degree offered
27	DEGGRANT	Degree-granting status
28	HBCU	Historically Black College or University
29	HOSPITAL	Institution has hospital
30	MEDICAL	Institution grants a medical degree
31	TRIBAL	Tribal college
32	LOCALE	Degree of urbanization (Urban-centric locale)
33	OPENPUBL	Institution open to the general public
34	ACT	Status of institution
35	NEWID	UNITID for merged schools
36	DEATHYR	Year institution was deleted from IPEDS
37	CLOSEDAT	Date institution closed
38	CYACTIVE	Institution is active in current year
39	POSTSEC	Primarily postsecondary indicator
40	PSEFLAG	Postsecondary institution indicator
41	PSET4FLG	Postsecondary and Title IV institution indicator
42	RPTMTH	Reporting method (academic year or program)
43	IALIAS	Institution name alias
44	INSTCAT	Institutional category
45	CCBASIC	Carnegie Classification 2005: Basic



	VAR_NAME	VAR_TITLE
46	CCIPUG	Carnegie Classification 2005: Undergraduate Instructional Program
47	CCIPGRAD	Carnegie Classification 2005: Graduate Instructional Program
48	CCUGPROF	Carnegie Classification 2005: Undergraduate Profile
49	CCENRPRF	Carnegie Classification 2005: Enrollment Profile
50	CCSIZSET	Carnegie Classification 2005: Size and Setting
51	CARNEGIE	Carnegie Classification 2000
52	TENURSYS	Does institution have a tenure system
53	LANDGRNT	Land Grant Institution
54	INSTSIZE	Institution size category
55	CBSA	Core Based Statistical Area (CBSA)
56	CBSATYPE	CBSA Type Metropolitan or Micropolitan
57	CSA	Combined Statistical Area (CSA)
58	NECTA	New England City and Town Area (NECTA)
59	DFRCGID	Data Feedback Report comparison group category

Services Offered Descriptive data



	Var_Name	Description
1	UNITID	Unique identification number of the institution
2	PEO2ISTR	Academic
3	CNTLAFFI	Institutional control or affiliation
4	LEVEL3	Associate's degree
5	LEVEL5	Bachelor's degree
6	CALSYS	Calendar system
7	FT_UG	Full-time undergraduate students are enrolled
8	FT_FTUG	Full time first-time degree/certificate-seeking undergraduate students enrolled
9	PT_UG	Part-time undergraduate students are enrolled
10	PT_FTUG	Part time first-time degree/certificate-seeking undergraduate students enrolled
11	OPENADMP	Open admission policy
12	VET1	Yellow Ribbon Program (officially known as Post-9/11 GI Bill, Yellow Ribbon Program)
13	VET2	Credit for military training
18	CREDITS1	Dual enrollment
19	CREDITS2	Credit for life experiences
20	CREDITS3	Advanced placement (AP) credits
21	CREDITS4	Institution does not accept dual, credit for life, or AP credits
22	SLO5	ROTC
23	SLO6	Study abroad
24	SLO7	Weekend/evening college
25	SLO83	Teacher certification: Approved by the state for initial certifcation or licensure of teachers.
26	YRSCOLL	Years of college-level work required
27	STUSRV1	Remedial services
28	STUSRV2	Academic/career counseling service
29	STUSRV3	Employment services for students
30	STUSRV4	Placement services for completers
31	STUSRV8	On-campus day care for students' children
32	STUSRV9	None of the above selected services are offered
33	LIBRES1	Library resources/services: Physical facilities
34	TUITPL	Any alternative tuition plans offered by institution
35	TUITPL1	Tuition guaranteed plan
36	TUITPL2	Prepaid tuition plan
37	TUITPL3	Tuition payment plan
38	TUITPL4	Other alternative tuition plan
39	DSTNUGC	Undergraduate level distance education courses offered
40	DSTNUGP	Undergraduate level distance education programs offered
41	DSTNCED1	Undergraduate level programs or courses are offered via distance education
42	ALLONCAM	Full-time, first-time degree/certificate-seeking students required to live on campus
43	TUITVARY	Tuition charge varies for in-district, in-state, out-of-state students
44	ROOM	Institution provide on-campus housing
45	BOARD	Institution provides board or meal plan
46	ROOMAMT	Typical room charge for academic year
47	BOARDAMT	Typical board charge for academic year
48	RMBRDAMT	Combined charge for room and board
49	APPLFEEU	Undergraduate application fee

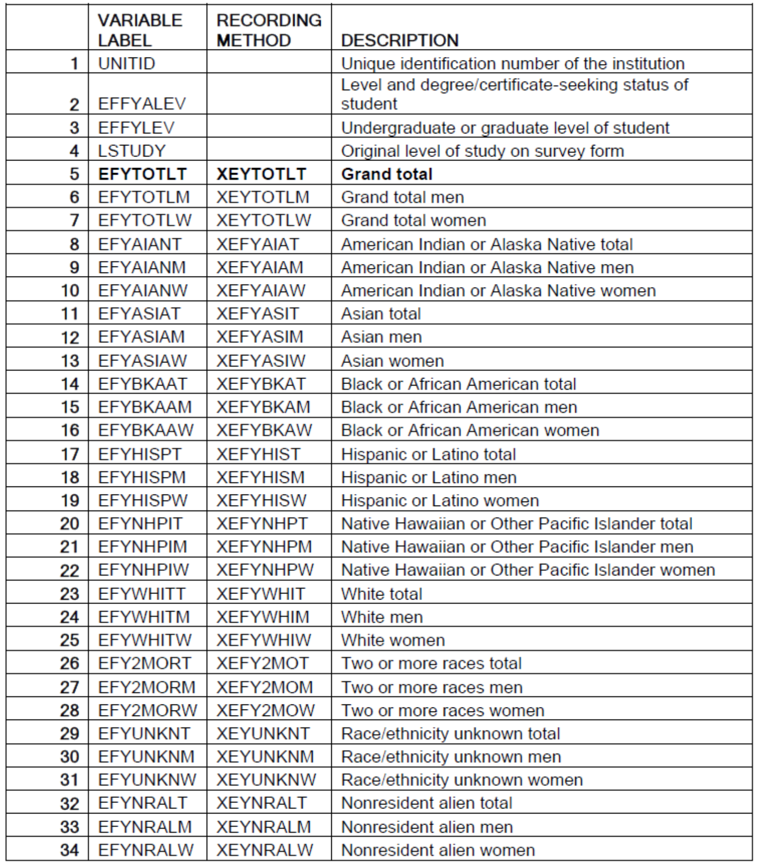
- **Data File**: ic2020.csv
- **Dictionary**: ic2020.xls
- The data file contains 49 different variables that offer information regarding HEIs services offerings
- Final Selected features: (after data cleaning)

- 'UNITID’
- 'CNTLAFFI’
- 'LEVEL3’
- 'LEVEL5’
- 'CALSYS’
- 'FT_UG’
- 'FT_FTUG’
- 'PT_UG’
- 'PT_FTUG’
- 'OPENADMP’
- 'VET1’
- 'VET2’
- 'CREDITS1’
- 'CREDITS2’
- 'CREDITS3’
- 'CREDITS4’
- 'SLO5‘
- ’SLO6’
- 'SLO7’
- 'SLO83’
- 'YRSCOLL’
- 'STUSRV1’
- 'STUSRV2’
- 'STUSRV3’
- 'STUSRV4’
- 'STUSRV8’
- 'STUSRV9’
- 'LIBRES1’
- 'TUITPL’
- 'TUITPL1’
- 'TUITPL2’
- 'TUITPL3’
- 'TUITPL4’
- 'DSTNUGC’
- 'DSTNUGP’
- 'DSTNCED1’
- 'ALLONCAM’
- 'TUITVARY’
- 'ROOM’
- 'BOARD’
- 'APPLFEEU’
- 'RMANDBRDAMT'

Student Enrollment data

- **Data File**: effy2020.csv
- **Dictionary**: effy2020.xls
- The data file contains 34 different variables that offer information regarding HEIs students
(not including 30 Imputation Variables)

- 'UNITID’
- 'EFYTOTLT’
- 'EFYTOTLM’
- 'EFYTOTLW’
- 'EFYAIANT’
- 'EFYAIANM’
- 'EFYAIANW’
- 'EFYASIAT’
- 'EFYASIAM’
- 'EFYASIAW’
- 'EFYBKAAT’
- 'EFYBKAAM',
- 'EFYBKAAW’
- 'EFYHISPT’
- 'EFYHISPM’
- 'EFYHISPW’
- 'EFYNHPIT’
- 'EFYNHPIM',
- 'EFYNHPIW’
- 'EFYWHITT’
- 'EFYWHITM’
- 'EFYWHITW’
- 'EFY2MORT’
- 'EFY2MORM’
- 'EFY2MORW’
- 'EFYUNKNT’
- 'EFYUNKNM’
- 'EFYUNKNW’
- 'EFYNRALT’
- 'EFYNRALM’
- 'EFYNRALW'



	VARIABLE LABEL	RECORDING METHOD	DESCRIPTION
1	UNITID		Unique identification number of the institution
2	EFFYALEV		Level and degree/certificate-seeking status of student
3	EFFYLEV		Undergraduate or graduate level of student
4	LSTUDY		Original level of study on survey form
5	**EFYTOTLT**	**XEYTOTLT**	**Grand total**
6	EFYTOTLM	XEYTOTLM	Grand total men
7	EFYTOTLW	XEYTOTLW	Grand total women
8	EFYAIANT	XEFYAIAT	American Indian or Alaska Native total
9	EFYAIANM	XEFYAIAM	American Indian or Alaska Native men
10	EFYAIANW	XEFYAIAW	American Indian or Alaska Native women
11	EFYASIAT	XEFYASIT	Asian total
12	EFYASIAM	XEFYASIM	Asian men
13	EFYASIAW	XEFYASIW	Asian women
14	EFYBKAAT	XEFYBKAT	Black or African American total
15	EFYBKAAM	XEFYBKAM	Black or African American men
16	EFYBKAAW	XEFYBKAW	Black or African American women
17	EFYHISPT	XEFYHIST	Hispanic or Latino total
18	EFYHISPM	XEFYHISM	Hispanic or Latino men
19	EFYHISPW	XEFYHISW	Hispanic or Latino women
20	EFYNHPIT	XEFYNHPT	Native Hawaiian or Other Pacific Islander total
21	EFYNHPIM	XEFYNHPM	Native Hawaiian or Other Pacific Islander men
22	EFYNHPIW	XEFYNHPW	Native Hawaiian or Other Pacific Islander women
23	EFYWHITT	XEFYWHIT	White total
24	EFYWHITM	XEFYWHIM	White men
25	EFYWHITW	XEFYWHIW	White women
26	EFY2MORT	XEFY2MOT	Two or more races total
27	EFY2MORM	XEFY2MOM	Two or more races men
28	EFY2MORW	XEFY2MOW	Two or more races women
29	EFYUNKNT	XEYUNKNT	Race/ethnicity unknown total
30	EFYUNKNM	XEYUNKNM	Race/ethnicity unknown men
31	EFYUNKNW	XEYUNKNW	Race/ethnicity unknown women
32	EFYNRALT	XEYNRALT	Nonresident alien total
33	EFYNRALM	XEYNRALM	Nonresident alien men
34	EFYNRALW	XEYNRALW	Nonresident alien women

Current Status of the data

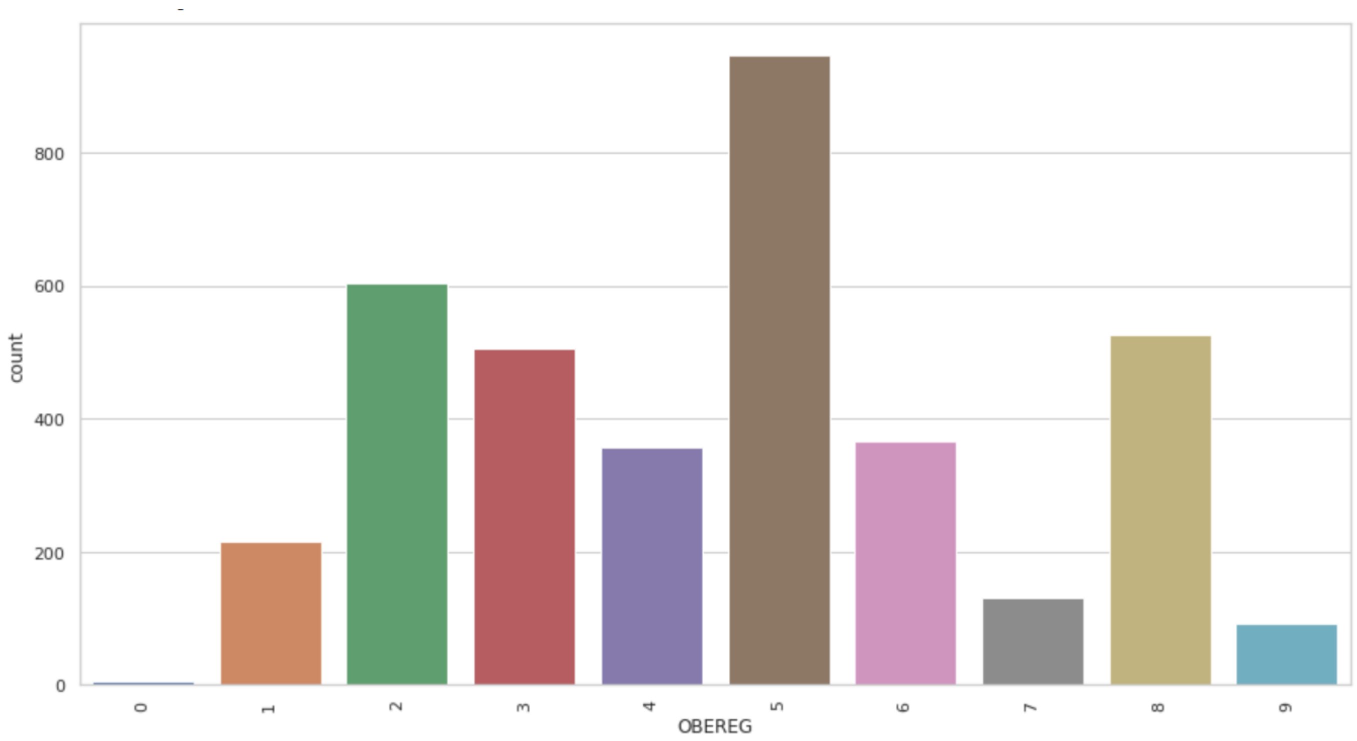
- The files have been merged and the data cleansed
- The institutions shortlisted are 3547
- The data refers to 59 US States & Territories
- The current data refers only to 2019-2020

Initial Exploratory Analysis

While not specifically tied to the Research Questions a couple of initial data visualizations have been created

The graph to the right represents the distributions of HEIs across the Bureau of Economic Analysis' categories.

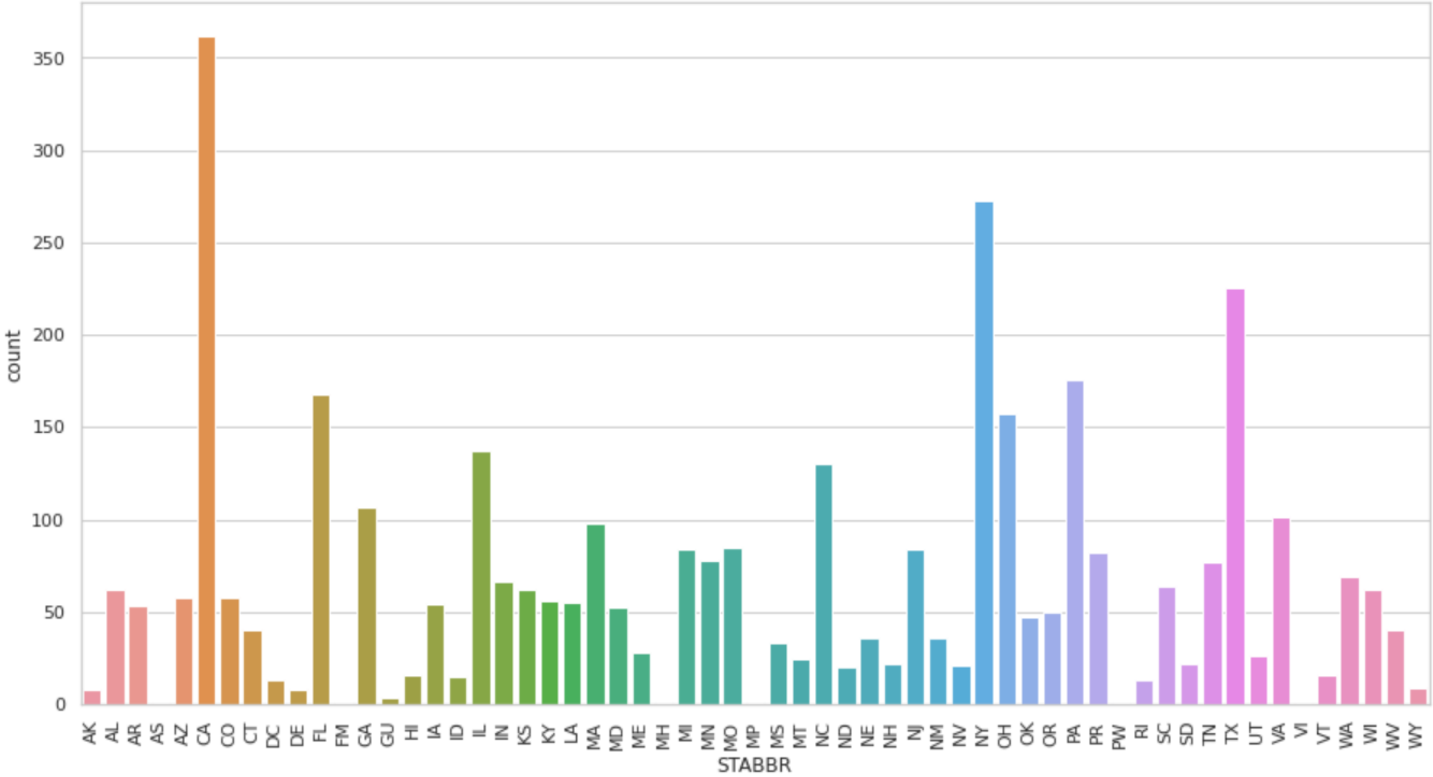
# 0 - US Service schools
# 1 - New England CT ME MA NH RI VT
# 2 - Mid East DE DC MD NJ NY PA
# 3 - Great Lakes IL IN MI OH WI
# 4 - Plains IA KS MN MO NE ND SD
# 5 - Southeast AL AR FL GA KY LA MS NC SC TN VA WV
# 6 - Southwest AZ NM OK TX
# 7 - Rocky Mountains CO ID MT UT WY
# 8 - Far West AK CA HI NV OR WA
# 9 - Outlying areas AS FM GU MH MP PR PW VI -
(Not in the Continental USA)



Initial Exploratory Analysis

While not specifically tied to the Research Questions a couple of initial data visualizations have been created

The graph to the right represents the distributions of HEIs by Stateand US Territory.



Initial Exploratory Analysis

While not specifically tied to the Research Questions a couple of initial data visualizations have been created

The graph to the right represents the distributions of HEIs by type of local environment they are set into (urban vs rural vs sub-urban)

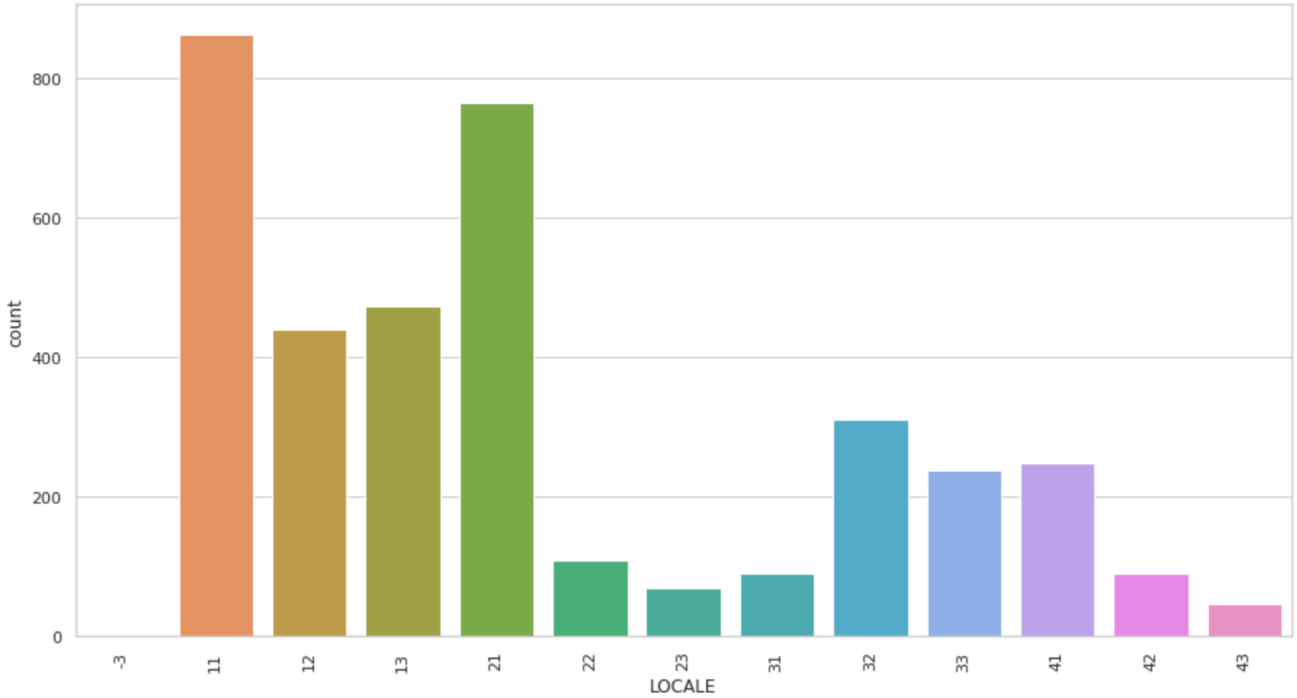
Code -3 = Territories Outside the Continental US

Codes 11-13 = City (Large, Midsize, Small)

Codes 21-23 = Suburb (Large, Midsize, Small)

Codes 31-33 = Town (Fringe, Distant, Remote)

Codes 41-43 = Rural (Fringe, Distant, Remote)

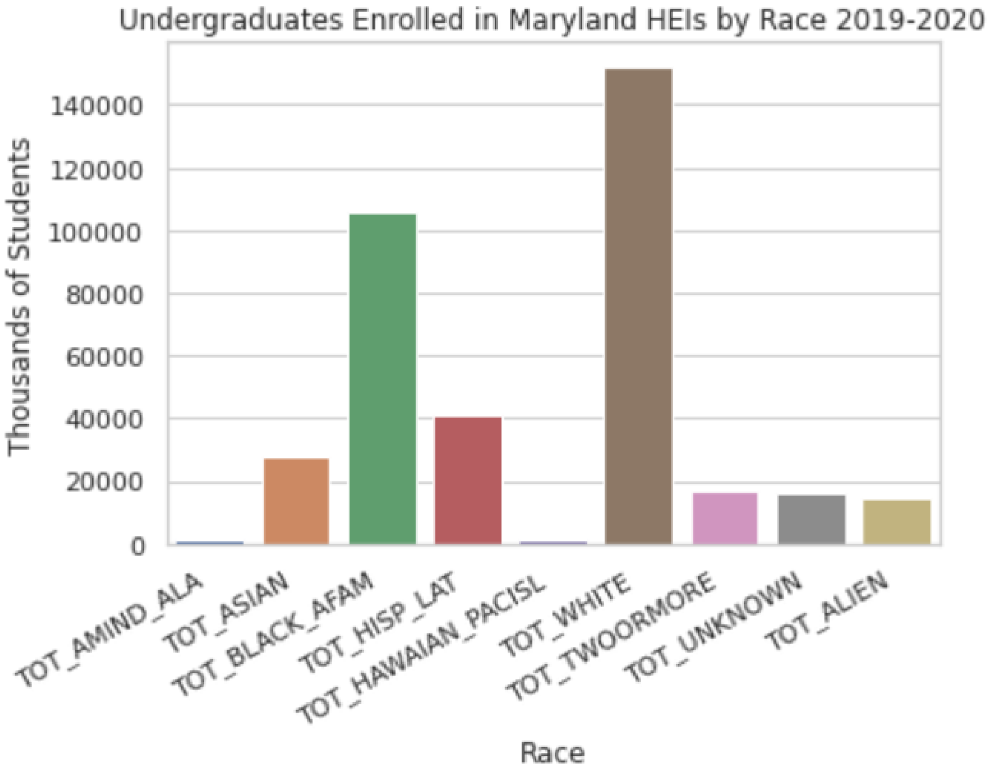


Initial Exploratory Analysis

While not specifically tied to the Research Questions a couple of initial data visualizations have been created

The graph to the right represents the distributions of Student Enrollment in HEIs who offer Undergraduate Programs by ethnicity

- American Indian and Alaska Native
- Asian
- Black and African American
- Hispanic or Latino
- White
- Two-or more races
- Undisclosed Ethnicity
- Non-Resident-Alien)



Intended next steps

- Verifying data scales

- Separating out a test set to avoid increasing the models’ bias

- Continue with some additional data visualizations

- Correlation Analysis of the features

- Completing final data preparation to ensure that the ML Algorithms run smoothly (for example using scaling techniques, labeling categorical data, dealing with outliers etc.)

- Select the best ML models (Regression Model; possibly DBSCAN for clustering; Feature Selection optimization via Principal Component Analysis)

-Extracting the shortlisted Institutions’ data across different years