DATA SCIENCE FOR UNIVERSITY RECOMMENDATION

LOOKING BEYOND THE UNIVERSITY RANKINGS

SELECTING THE BEST UNIVERSITY: THE PROBLEM

- A Foreign Student’s Conundrum
  - Private or Public University?
  - High Tuition Fees with Part Time Job or Low Tuition with fewer prospects?
  - Higher ranked Expensive University or Lower ranked Cheaper University with similar salaries?
  - University Town vs Metropolis?
  - Pursue Research or Get a Job after Graduation?
  - Study vs Fun?

  …
- Many such questions confront a student when deciding on the colleges to apply to.
- Counselling tends to be subjective & expensive.
- University Rankings don’t help completely either due to narrow evaluation parameters set.
- Lack of a holistic view leads to suboptimal choice of University.

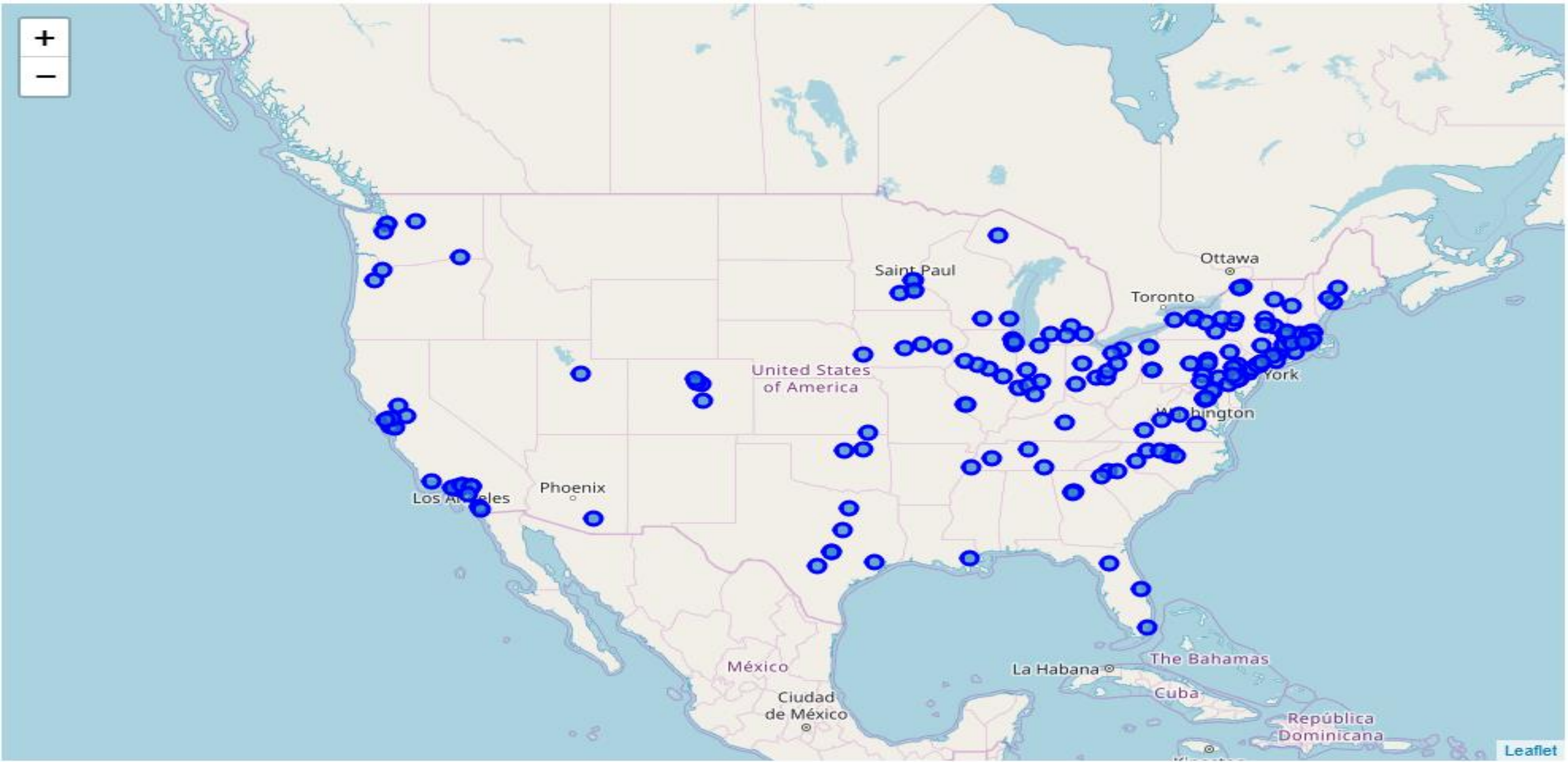
SELECTING THE BEST UNIVERSITY: CAN LOCATION DATA HELP?

- Take into account the venues of interest in the vicinity.
  - Universities in Urban areas may be expensive but offer plenty of part-time work opportunity
  - University Town's might offer peace and quite and overall cheaper education
  - Proximity to Offices may be important in landing internships & jobs
  - Social venues, Art Galleries, Museums etc. may help you experience the life in a foreign country.

- So, Location Matters!
  - Combine the Rankings with Location Data to make the best judgement
  - Location data such as venues, details can be obtained from providers such as Foursquare.
  - Cluster Analysis can then be run on the combined dataset to group similar Universities together.
  - Choose the cluster best suited to your preferences.
  - Choose the Best University from the chosen cluster.

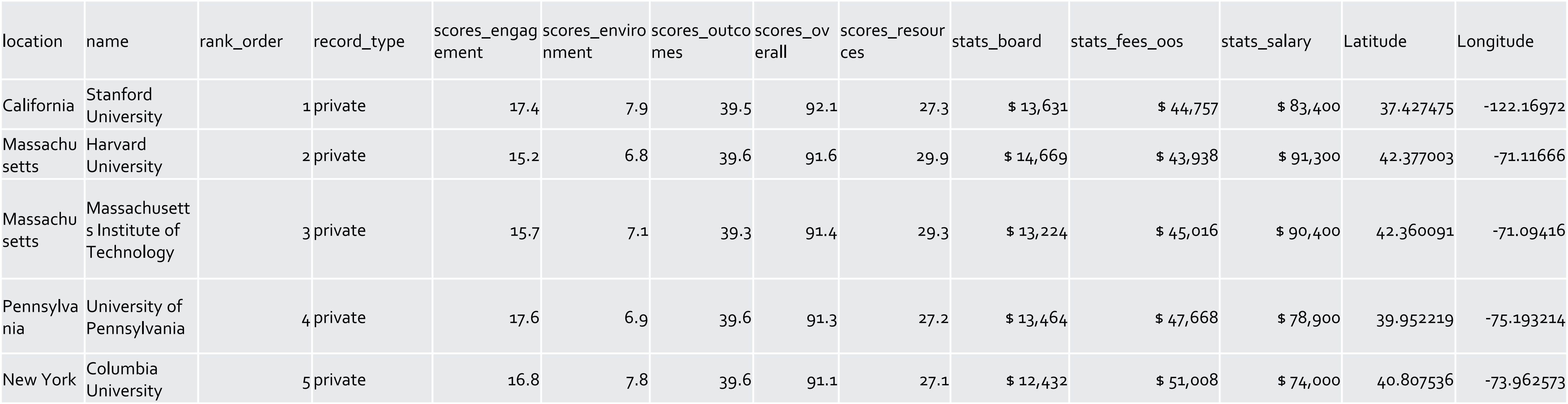
SELECTING THE BEST UNIVERSITY: TEST CASE USA

- Analyze the groupings among the Top 200 Universities



SELECTING THE BEST UNIVERSITY: RANKINGS DATA

- Scrape THE/WSJ 2017 Rankings Table here: https://www.timeshighereducation.com/rankings/united-states/2017#!/page/0/length/50/sort_by/rank/sort_order/asc/cols/scores
- Retain the numeric columns (scores on various parameters, dollar values), drop the rest.
- Add locations data: Use Google Geocode API to fetch coordinates for each University.
- Sample of the final dataframe.



location	name	rank_order	record_type	scores_engagement	scores_environment	scores_outcomes	scores_overall	scores_resources	stats_board	stats_fees_oos	stats_salary	Latitude	Longitude
California	Stanford University	1	private	17.4	7.9	39.5	92.1	27.3	$ 13,631	$ 44,757	$ 83,400	37.427475	-122.16972
Massachusetts	Harvard University	2	private	15.2	6.8	39.6	91.6	29.9	$ 14,669	$ 43,938	$ 91,300	42.377003	-71.11666
Massachusetts	Massachusetts Institute of Technology	3	private	15.7	7.1	39.3	91.4	29.3	$ 13,224	$ 45,016	$ 90,400	42.360091	-71.09416
Pennsylvania	University of Pennsylvania	4	private	17.6	6.9	39.6	91.3	27.2	$ 13,464	$ 47,668	$ 78,900	39.952219	-75.193214
New York	Columbia University	5	private	16.8	7.8	39.6	91.1	27.1	$ 12,432	$ 51,008	$ 74,000	40.807536	-73.962573

SELECTING THE BEST UNIVERSITY:
ADD LOCATIONS DATA FROM FOURSQUARE

- Identify Venue Categories and Sub-Categories of most interest and the corresponding Foursquare IDs

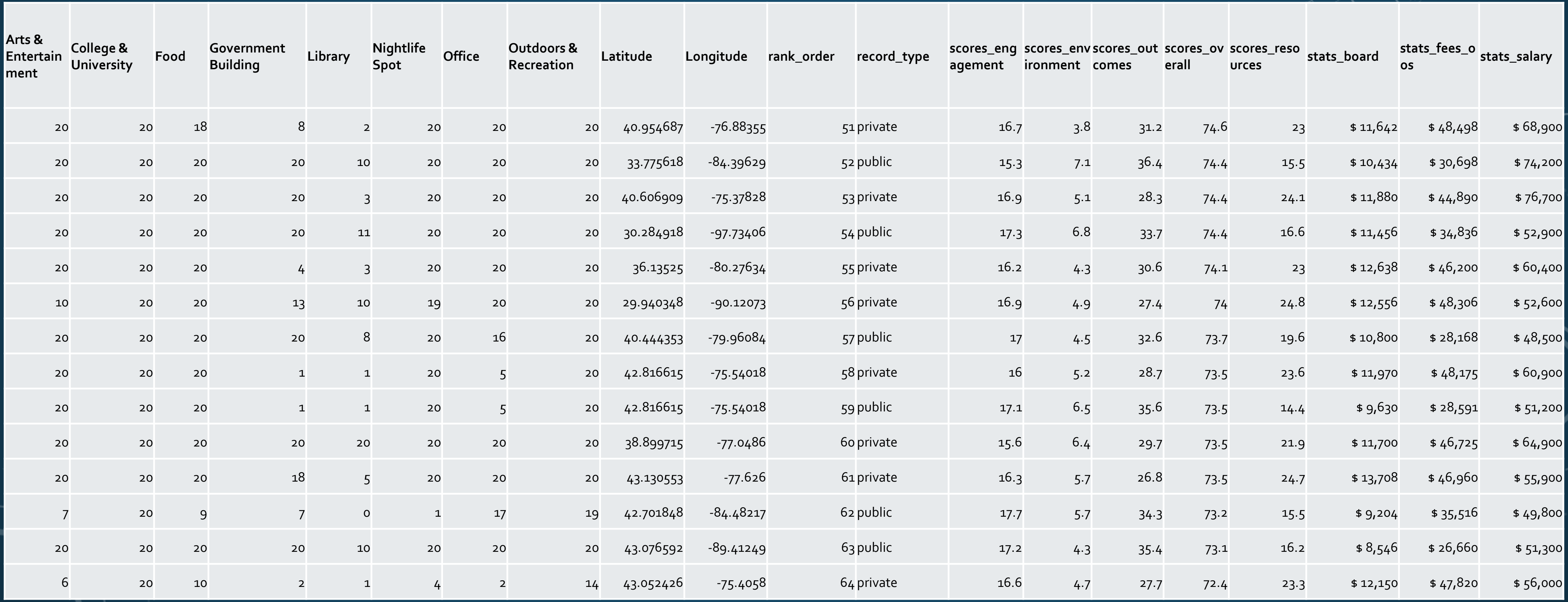
cat_name	**cat_id**
Office	4bf58dd8d48988d124941735
Shop & Service	4d4b7105d754a06378d81259
Food	4d4b7105d754a06374d81259
Arts & Entertainment	4d4b7104d754a06370d81259
Outdoors & Recreation	4d4b7105d754a06377d81259
Library	4bf58dd8d48988d12f941735
College & University	4d4b7105d754a06372d81259
Nightlife Spot	4d4b7105d754a06376d81259
Residence	4e67e38e036454776db1fb3a
Government Building	4bf58dd8d48988d126941735
Travel & Transport	4d4b7105d754a06379d81259

- Add these Venue Category Columns to the dataset created previously. Initialize the venue counts with Nulls.

Arts & Entertainment	College & University	Food	Government Building	Latitude	Library	Longitude	Nightlife Spot	Office	Outdoors & Recreation	Residence	Shop & Service	Travel & Transport	location	name
NaN	NaN	NaN	NaN	37.4275	NaN	-122.16972	NaN	NaN	NaN	NaN	NaN	NaN	California	Stanford University
NaN	NaN	NaN	NaN	42.377	NaN	-71.11666	NaN	NaN	NaN	NaN	NaN	NaN	Massachusetts	Harvard University
NaN	NaN	NaN	NaN	42.3601	NaN	-71.09416	NaN	NaN	NaN	NaN	NaN	NaN	Massachusetts	Massachusetts Institute of...
NaN	NaN	NaN	NaN	39.9522	NaN	-75.193214	NaN	NaN	NaN	NaN	NaN	NaN	Pennsylvania	University of Pennsylvania
NaN	NaN	NaN	NaN	[illegible]	NaN	[illegible]	NaN	NaN	NaN	NaN	NaN	NaN	New York	Columbia

SELECTING THE BEST UNIVERSITY:
ADD LOCATIONS DATA FROM FOURSQUARE

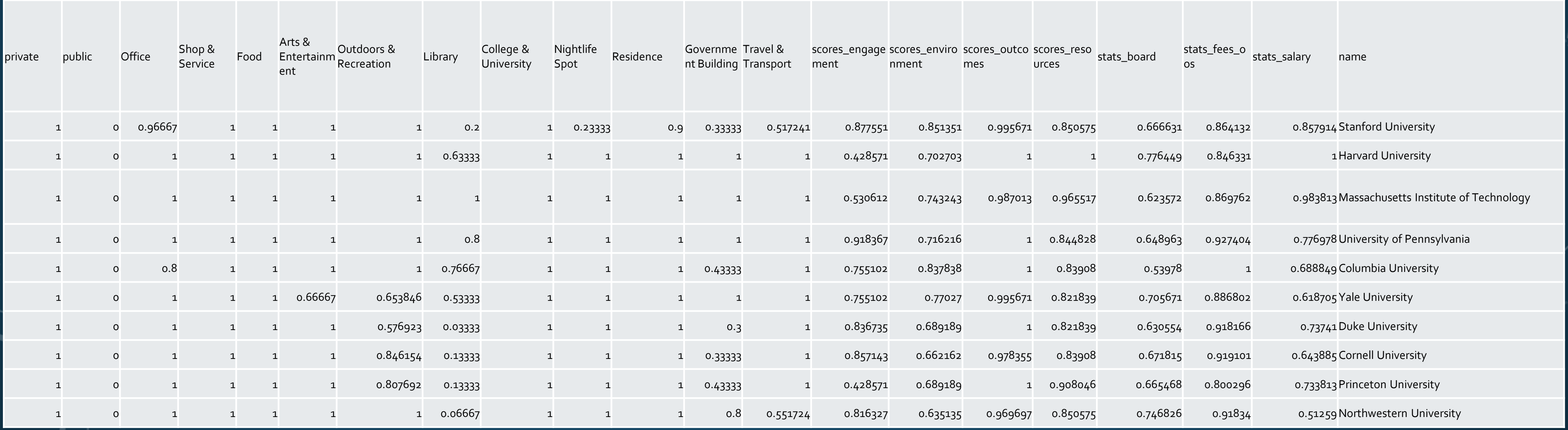
- Get the Venue Counts from Foursquare & populate the relevant columns in the data frame
- Sample of the final dataframe:



Arts & Entertainment	College & University	Food	Government Building	Library	Nightlife Spot	Office	Outdoors & Recreation	Latitude	Longitude	rank_order	record_type	scores_engagement	scores_environment	scores_outcomes	scores_overall	scores_resources	stats_board	stats_fees_oos	stats_salary
20	20	18	8	2	20	20	20	40.954687	-76.88355	51	private	16.7	3.8	31.2	74.6	23	$ 11,642	$ 48,498	$ 68,900
20	20	20	20	10	20	20	20	33.775618	-84.39629	52	public	15.3	7.1	36.4	74.4	15.5	$ 10,434	$ 30,698	$ 74,200
20	20	20	20	3	20	20	20	40.606909	-75.37828	53	private	16.9	5.1	28.3	74.4	24.1	$ 11,880	$ 44,890	$ 76,700
20	20	20	20	11	20	20	20	30.284918	-97.73406	54	public	17.3	6.8	33.7	74.4	16.6	$ 11,456	$ 34,836	$ 52,900
20	20	20	4	3	20	20	20	36.13525	-80.27634	55	private	16.2	4.3	30.6	74.1	23	$ 12,638	$ 46,200	$ 60,400
10	20	20	13	10	19	20	20	29.940348	-90.12073	56	private	16.9	4.9	27.4	74	24.8	$ 12,556	$ 48,306	$ 52,600
20	20	20	20	8	20	16	20	40.444353	-79.96084	57	public	17	4.5	32.6	73.7	19.6	$ 10,800	$ 28,168	$ 48,500
20	20	20	1	1	20	5	20	42.816615	-75.54018	58	private	16	5.2	28.7	73.5	23.6	$ 11,970	$ 48,175	$ 60,900
20	20	20	1	1	20	5	20	42.816615	-75.54018	59	public	17.1	6.5	35.6	73.5	14.4	$ 9,630	$ 28,591	$ 51,200
20	20	20	20	20	20	20	20	38.899715	-77.0486	60	private	15.6	6.4	29.7	73.5	21.9	$ 11,700	$ 46,725	$ 64,900
20	20	20	18	5	20	20	20	43.130553	-77.626	61	private	16.3	5.7	26.8	73.5	24.7	$ 13,708	$ 46,960	$ 55,900
7	20	9	7	0	1	17	19	42.701848	-84.48217	62	public	17.7	5.7	34.3	73.2	15.5	$ 9,204	$ 35,516	$ 49,800
20	20	20	20	10	20	20	20	43.076592	-89.41249	63	public	17.2	4.3	35.4	73.1	16.2	$ 8,546	$ 26,660	$ 51,300
6	20	10	2	1	4	2	14	43.052426	-75.4058	64	private	16.6	4.7	27.7	72.4	23.3	$ 12,150	$ 47,820	$ 56,000

SELECTING THE BEST UNIVERSITY: PRE PROCESSING THE DATASET

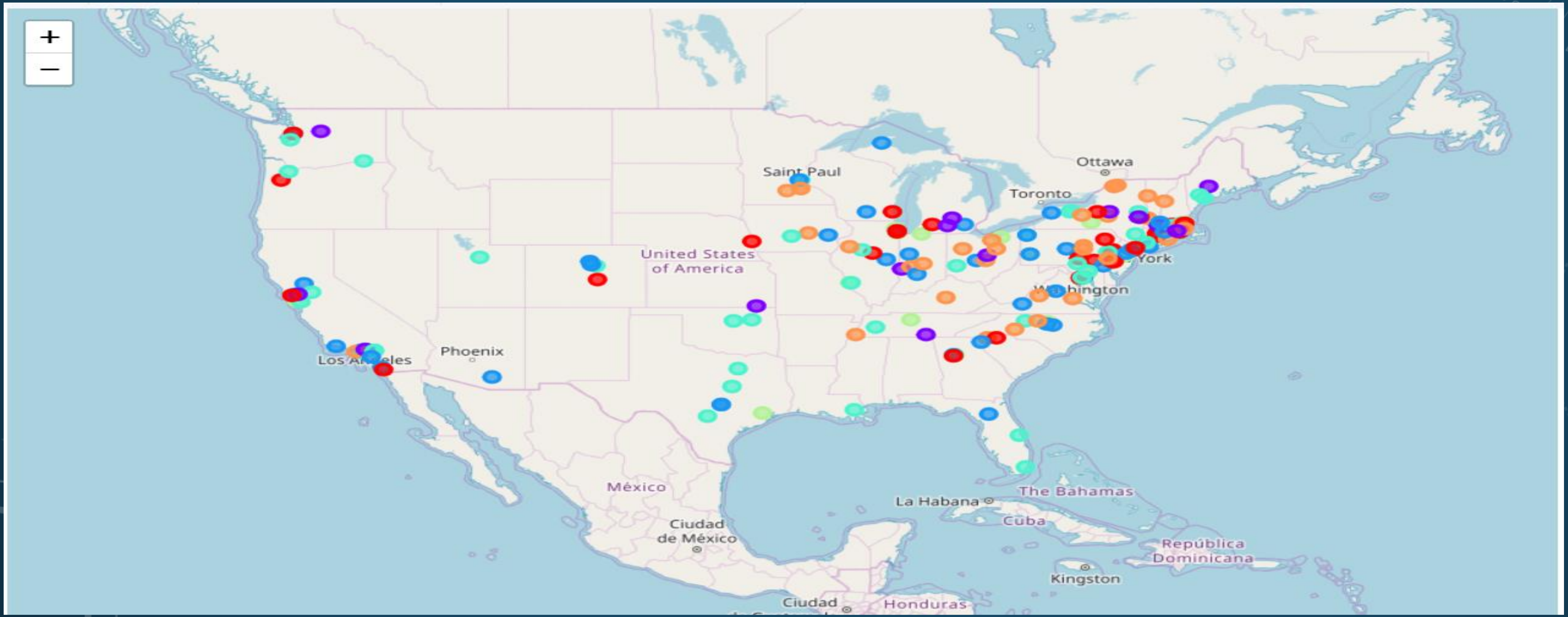
- Convert Values to Numeric (Int64/Float64)
- Check for & substitute any Null Values with Mean/Median of their respective columns
- One-hot encode the categorical column
- Min-Max scale the values between 0 & 1.
- Sample of the Pre-processed dataset:



private	public	Office	Shop & Service	Food	Arts & Entertainment	Outdoors & Recreation	Library	College & University	Nightlife Spot	Residence	Government Building	Travel & Transport	scores_engagement	scores_environment	scores_outcomes	scores_resources	stats_board	stats_fees_oos	stats_salary	name
1	0	0.96667	1	1	1	1	0.2	1	0.23333	0.9	0.33333	0.517241	0.877551	0.851351	0.995671	0.850575	0.666631	0.864132	0.857914	Stanford University
1	0	1	1	1	1	1	0.63333	1	1	1	1	1	0.428571	0.702703	1	1	0.776449	0.846331	1	Harvard University
1	0	1	1	1	1	1	1	1	1	1	1	1	0.530612	0.743243	0.987013	0.965517	0.623572	0.869762	0.983813	Massachusetts Institute of Technology
1	0	1	1	1	1	1	0.8	1	1	1	1	1	0.918367	0.716216	1	0.844828	0.648963	0.927404	0.776978	University of Pennsylvania
1	0	0.8	1	1	1	1	0.76667	1	1	1	0.43333	1	0.755102	0.837838	1	0.83908	0.53978	1	0.688849	Columbia University
1	0	1	1	1	0.66667	0.653846	0.53333	1	1	1	1	1	0.755102	0.77027	0.995671	0.821839	0.705671	0.886802	0.618705	Yale University
1	0	1	1	1	1	0.576923	0.03333	1	1	1	0.3	1	0.836735	0.689189	1	0.821839	0.630554	0.918166	0.73741	Duke University
1	0	1	1	1	1	0.846154	0.13333	1	1	1	0.33333	1	0.857143	0.662162	0.978355	0.83908	0.671815	0.919101	0.643885	Cornell University
1	0	1	1	1	1	0.807692	0.13333	1	1	1	0.43333	1	0.428571	0.689189	1	0.908046	0.665468	0.800296	0.733813	Princeton University
1	0	1	1	1	1	1	0.06667	1	1	1	0.8	0.551724	0.816327	0.635135	0.969697	0.850575	0.746826	0.91834	0.51259	Northwestern University

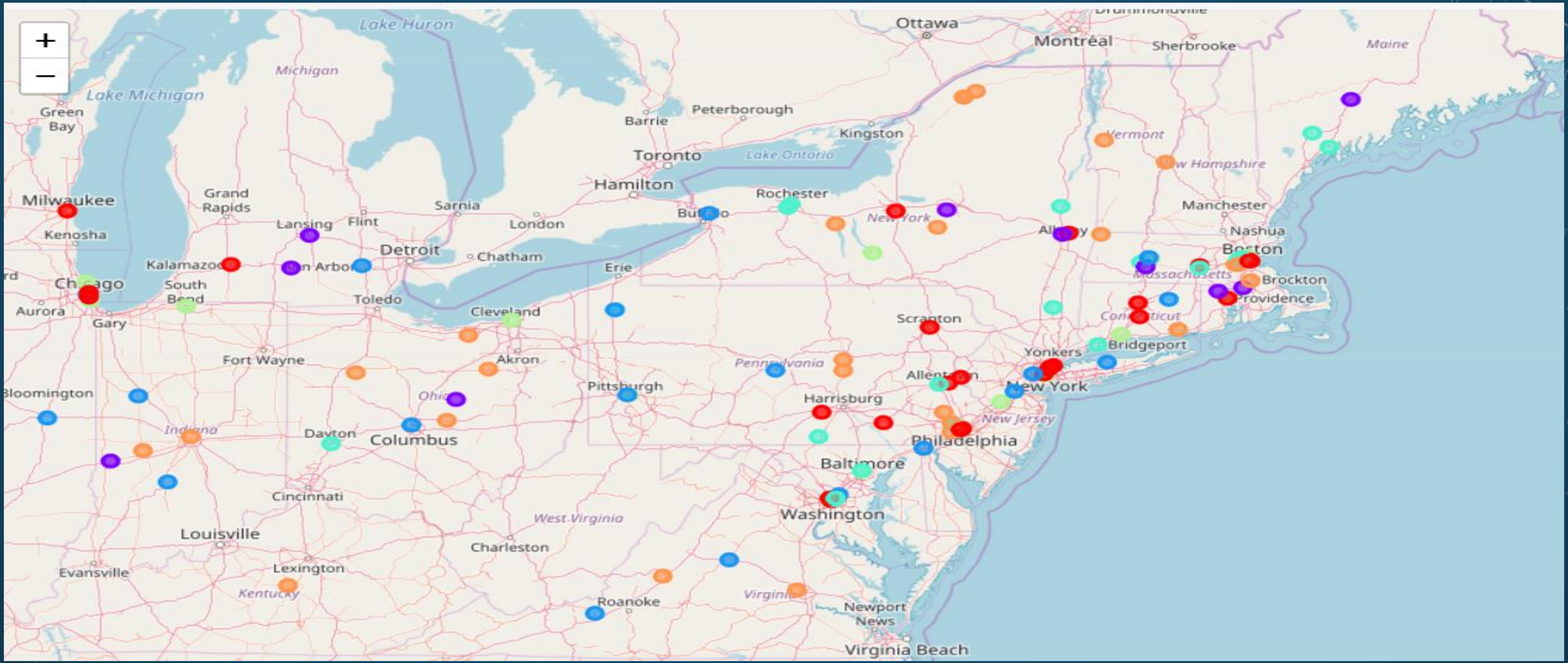
SELECTING THE BEST UNIVERSITY:
RUN K-MEANS CLUSTERING ON THE PROCESSED DATASET

- Set number of clusters = 6
- Run k-means clustering: kmeans = KMeans(n_clusters=6, random_state=0).fit(df[columns])
- Check the clusters identified:



SELECTING THE BEST UNIVERSITY:

- NOTICE AT A GLANCE:
  - No clear Geographic Separation between the clusters.
  - This means, a University meeting Student preferences can be found in almost any State/City.
  - Focus on the Mid & East part of US as Max Universities in the Top 200 are located here:



SELECTING THE BEST UNIVERSITY: INTERPRET THE CLUSTERS

Cluster 1: (Mid-to-Low ranking Private Universities , Urban, Average Fees & Costs)

This cluster is formed of **Mid-to-Low ranking Private Universities** located in **major urban centers**. The neighborhoods of these Universities are packed with all the amenities. While these Universities don't score that well on THE-parameters, by dint of their location they are able to command a starting salary of ~60K +/- 11K for their Graduates. Out-of-state Tuition Fees & cost of living are average at 42K & 12.5K respectively.

Cluster 2: (Low ranking Private Universities, Semi-Urban, Average Fees, Low Cost of Living)

This cluster is formed of, barring a few exceptions, **mostly Low ranking Private Universities located either in Tier 2 cities** or in main cities but farther away from population centers. The neighborhoods are sparsely populated indicated by low no. of Offices, food joints, arts & entertainment venues & nightlife spots. The Tuition & boarding costs are relatively low compared to cluster 1 Universities at 40K & 11.5K per year. Median salaries are significantly lower than cluster 1 at 51K.

SELECTING THE BEST UNIVERSITY: INTERPRET THE CLUSTERS

Cluster 3: (Public Universities , varying ranks, Urban centers , Tuition - Cheap, cost of living - low)

This cluster is perhaps of most interest to foreign students. These are predominantly **Public Universities** of wildly **varying ranks** in **Urban centers** with great access to part-time job opportunities (there is no lack of any public amenities within the close proximity of the campus). The highest ranked universities are within top-30 while most of them are ranked decently in the range 50-150. The **Tuition is Cheap** @ 30K, **cost of living is low** @11K and the salaries are better than the more pricey Cluster 2 universities @53K. Depending on eligibility and score, a student with limited means should consider applying to some of the high ranking colleges in this cluster

Cluster 4: (Similar to Cluster 3 but with Private Universities, Costlier than Cluster 3)

This cluster is very **similar to Cluster 3 but with the difference that most Universities in this group are Private**. The rankings vary from low 30s to below 150. The tuition & boarding are costlier than Cluster 3 while the median salary is the same.

Prefer Cluster 3 over Cluster 4

SELECTING THE BEST UNIVERSITY: INTERPRET THE CLUSTERS

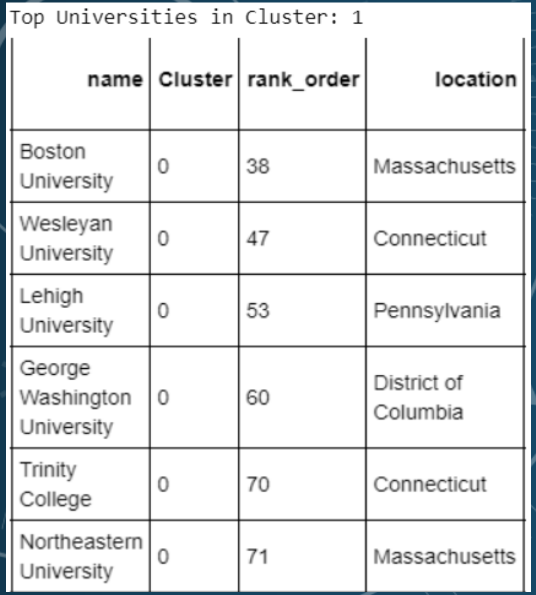
Cluster 5: (Top Ranked Universities, Urban, Expensive, High Salaries)

This is the top ranked cluster! All of them Private & ranked 1-30. Most expensive tuition (46K), high cost of living (~13.5K), but significantly higher median Salaries (~70K+). Prime locations. Plenty of internship & part-time job opportunities. Resource rich Universities.

Cluster 6: (Average Ranked (30-120), Private, Semi-Urban, Average Fees & Cost, Average Salaries)

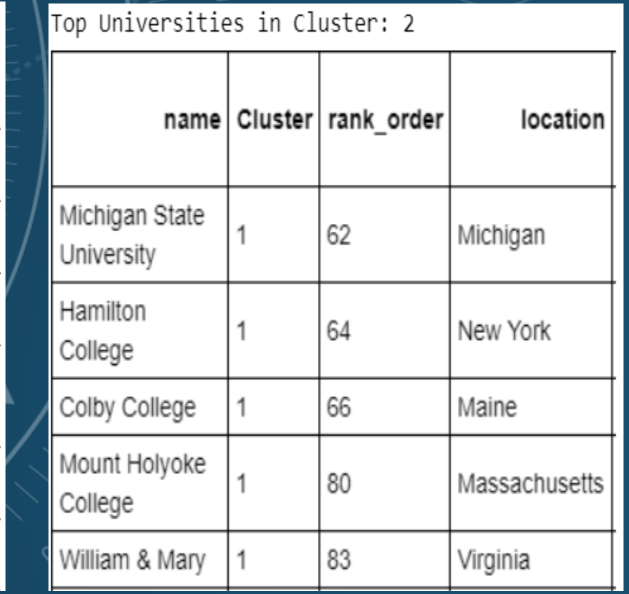
This is the "cluster of averages". Average location, Average Fees, Average Cost of living, Average Salaries, predominantly private. The rankings are good though with many of them ranked between 30 - 120. These may be good compromise choices.

SELECTING THE BEST UNIVERSITY:
TOP RECOMMENDATIONS FROM EACH CLUSTER



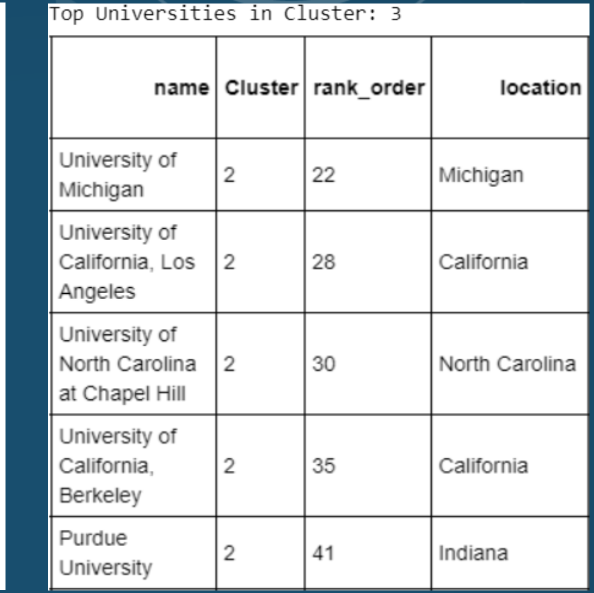
Top Universities in Cluster: 1

name	Cluster	rank_order	location
Boston University	0	38	Massachusetts
Wesleyan University	0	47	Connecticut
Lehigh University	0	53	Pennsylvania
George Washington University	0	60	District of Columbia
Trinity College	0	70	Connecticut
Northeastern University	0	71	Massachusetts



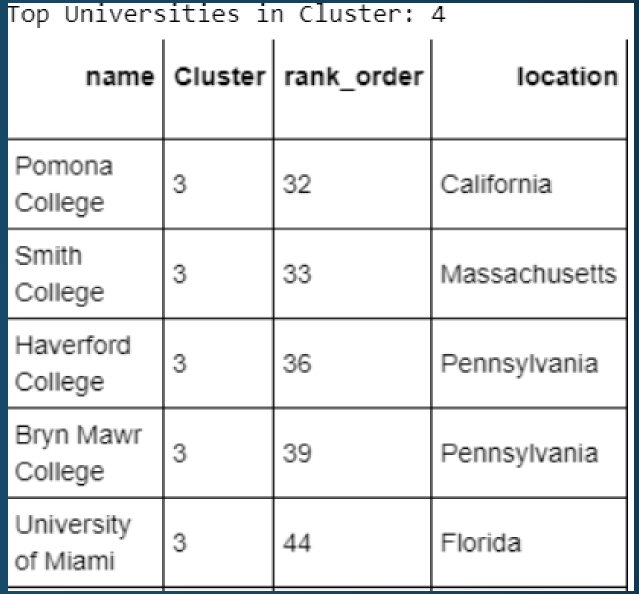
Top Universities in Cluster: 2

name	Cluster	rank_order	location
Michigan State University	1	62	Michigan
Hamilton College	1	64	New York
Colby College	1	66	Maine
Mount Holyoke College	1	80	Massachusetts
William & Mary	1	83	Virginia



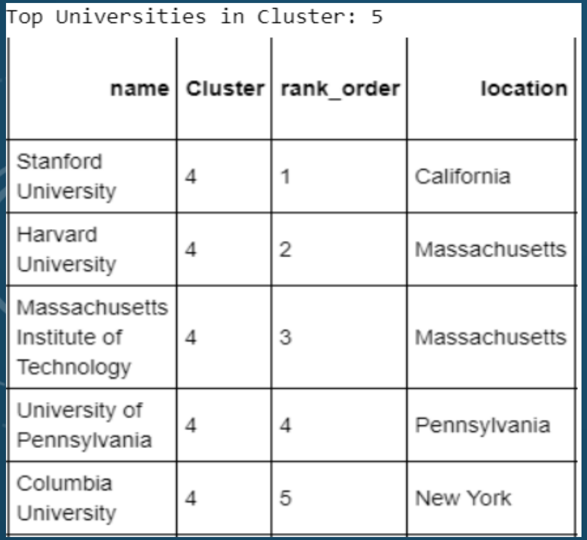
Top Universities in Cluster: 3

name	Cluster	rank_order	location
University of Michigan	2	22	Michigan
University of California, Los Angeles	2	28	California
University of North Carolina at Chapel Hill	2	30	North Carolina
University of California, Berkeley	2	35	California
Purdue University	2	41	Indiana



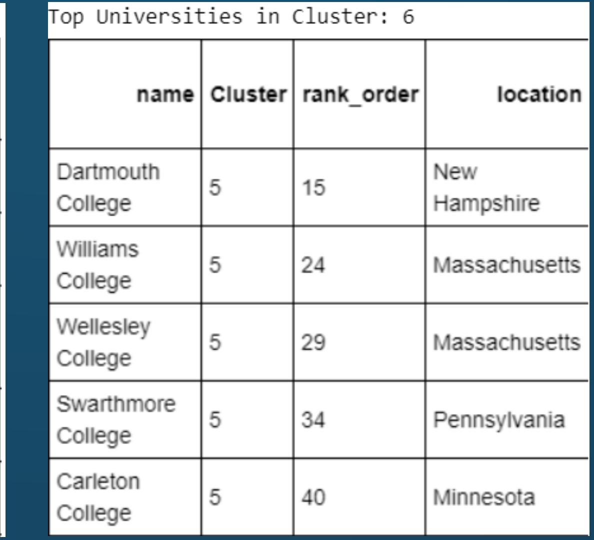
Top Universities in Cluster: 4

name	Cluster	rank_order	location
Pomona College	3	32	California
Smith College	3	33	Massachusetts
Haverford College	3	36	Pennsylvania
Bryn Mawr College	3	39	Pennsylvania
University of Miami	3	44	Florida



Top Universities in Cluster: 5

name	Cluster	rank_order	location
Stanford University	4	1	California
Harvard University	4	2	Massachusetts
Massachusetts Institute of Technology	4	3	Massachusetts
University of Pennsylvania	4	4	Pennsylvania
Columbia University	4	5	New York



Top Universities in Cluster: 6

name	Cluster	rank_order	location
Dartmouth College	5	15	New Hampshire
Williams College	5	24	Massachusetts
Wellesley College	5	29	Massachusetts
Swarthmore College	5	34	Pennsylvania
Carleton College	5	40	Minnesota

SELECTING THE BEST UNIVERSITY: IMPROVING THE ANALYSIS

- We are analyzing location goodness quantitatively not qualitatively: While the count of venues matter, there may be cases where size and quality of establishments may play a bigger part. For e.g. how many small diners are equal to 5 great restaurants? A huge museum vs several small theatres? Such an analysis is possible by obtaining venue details from Foursquare but since its not free, we don't pursue it.

- We are assuming that THE rankings and scores are objective and a true reflection of the ground truth! Rankings of Universities vary across agencies. A better way would be to aggregate rankings from multiple sources, scale and average the scores and perform the rest of the analysis as we have done.

- We are assuming that students meet minimum eligibility criteria and that the only consideration is to identify the best university for them from a group of 200. For e.g. consider a student with 750 on GMAT - if the financial burden is too much, he/she may choose to ditch top ranked Private Universities for a high ranked Public one.

- We are not taking into account the placement statistics for Universities. For e.g. a University obtains good offers for its students but only a small fraction of student population get selected. This is an important factor to consider and should be explicitly built into the dataset instead of relying on obscure "outcomes" score as provided by THE.

- Finally, crime statistics for the city & locality could be included in the analysis.

SELECTING THE BEST UNIVERSITY: CONCLUSION

- This exercise demonstrates how the choice of University question can be formulated and solved in a purely objective manner by taking into account enough parameters and data from varied sources. This is an approach that University Rating Agencies could take themselves while deciding on rankings.

- The cluster analysis step could also be a precursor to a "University Recommender" ML model where students key in their preferences and the model spits out a valid list of Universities.

- Finally, this analysis can be made use of by Universities themselves! Private universities seek out foreign students as they usually pay full fee. To make itself standout in comparison with competing Universities, a University could consider shoring up areas in which it doesn't score that well - slash the tuition fees, ensure sufficient part-time work availability, boost transportation facilities (more buses) etc.