TEACH FOR AMERICA – CASE COMPETITION

Wharton People Analytics Conference 2017

TEAM DATA MAFIA

Sriraman R Krishnamurthy
Muralidhar Mutnuru
Raghavendiran Nagarajan

UCONN

1.0 Executive Summary

Problem Statement:

Teach for America seeks to recruit remarkable leaders from a broad spectrum of universities. In order to maximize our limited resources, we differentiate resource allocation for recruitment at individual campuses, based on the competitive dynamics at each university. For the purpose of this case study, we are solely focusing on campus-based recruitment for undergraduates, though we have separate strategies for graduate students and professionals.

This year, we have a three-tier strategy for undergraduates, represented by 1 – 3, where 1 represents the most resource investment. Last year, we used a two-tier strategy for undergraduates, represented as 2 and 3, where 2 represents the most resource investment.

Our question is simple: Our question is simple: What are the optimal tiers for our recruiting strategies? More? Fewer? Which schools should be in which tiers?

Approach:

We performed initial data cleaning with an objective to identify preliminary errors and discrepancies. In the next step, we identified the variables required for analysis. Then we performed univariate and multivariate analyses to determine the significant variables required for determining the tier of a school.

Results:

- Identified that there should be x tiers or categories for schools. They have the capacity to distinguish schools well enough for Teach for America (TFA) to make a proper decision.
- We created new variables from the given dataset by regrouping the levels into new buckets. This was performed in order to reduce uneven distribution of the data and derive insights.
- Transformation was performed on some of the variables. This was performed to reduce the skewness of the variables and treat the outliers.

1.1 Overview of the Data

The data consisted of basic information about every applicant over the past two years and about every university TFA recruits. Sheets 1 and 2 contained applicant data from 2016 and 2017. Sheet 3 contained information about each university.

The dataset has three different sheets, we planned to aggregate all the sheets based on the University Id. On our analysis, we found many variables in TIER2016 and TIER2017 were missing. All these missing values are imputed from TIER Variables present in 2016 and 2017 sheets. The default values in the Selectivity variable is just categorical in this case. Since, we could make sense of the data by understanding the order, we imputed all values like 1 for least selective to 5 for most selective. Further, for all the unknown values in the column are assigned 1(Least selective) by considering the fact that any reputed Universities would be ranked by U.S News.

The Tier variables are replaced by reverting from 1 to 3 and 3 to 1 for 2017 and 2 to 3 and 3 to 2 for 2016 for modelling purpose. The highest value signifies that university has highest rank. This replacement is effective for both the sheets. We were particularly interested in understanding the STEM and Non – STEM courses. From facts and information, STEM courses are highly paid than Non – STEM courses. This will be beneficial in understanding the students outlook for teaching. So, a column ISSTEM is added. TFAAwareness and General Awareness are made into separate columns. Considering UniversityData as the base table, all the tables are aggregated in one table for each year.

Kindly check the Python codes and all the sheets we have used to aggregate the data from other tables.

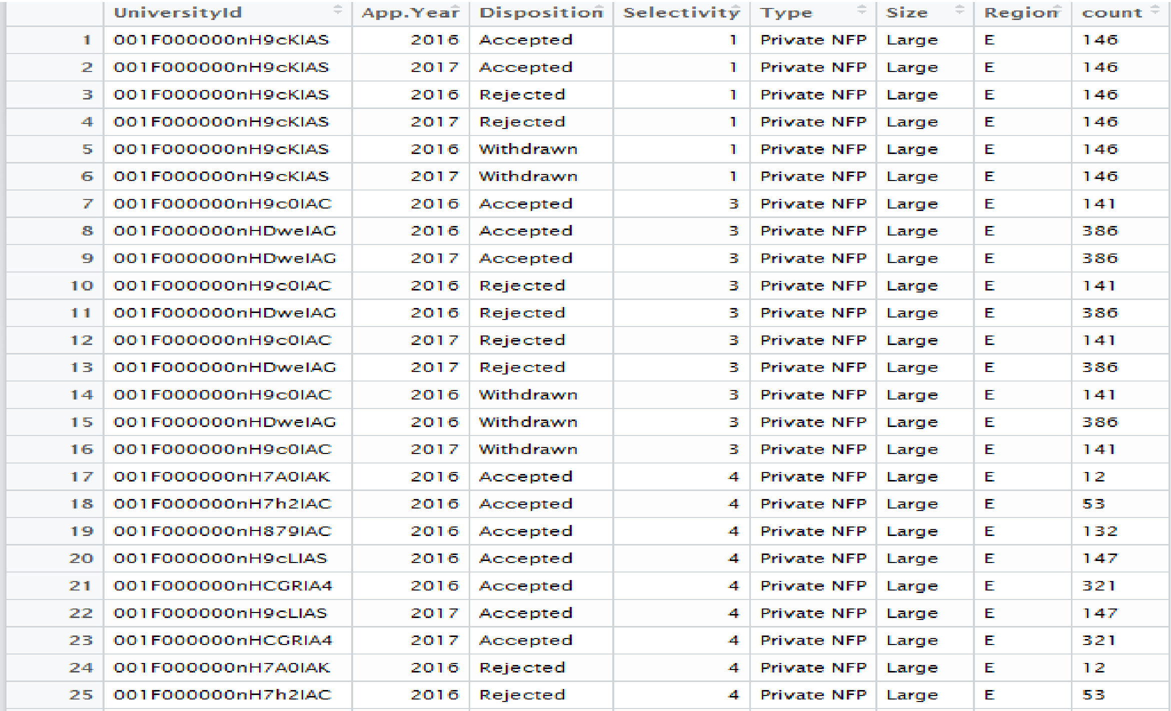
1.2 Research & Findings

Objective:

The objective of the analysis is to segment the data into various segments so that the segments best represents the data. We did not use the tier variable in the given data rather we tried to cluster the universities into various tiers with various clustering techniques

In order to cluster the data, we aggregated the data at the university level, and applied clustering techniques on the data. Below is the technique used to aggregate the data and the screenshot of the aggregated variables

```
aggdata <-aggregate(wharton_agg_orig$UniversityId,
by=list(wharton_agg_orig$UniversityId,wharton_agg_orig$App.Year,wharton_agg_orig$Disposition,wharton_agg_orig$Selectivity,wharton_agg_orig$Type,wharton_agg_orig$Size,wharton_agg_orig$Region), FUN=count)
```



	UniversityId	App.Year	Disposition	Selectivity	Type	Size	Region	count
1	001F000000nH9cKIAS	2016	Accepted	1	Private NFP	Large	E	146
2	001F000000nH9cKIAS	2017	Accepted	1	Private NFP	Large	E	146
3	001F000000nH9cKIAS	2016	Rejected	1	Private NFP	Large	E	146
4	001F000000nH9cKIAS	2017	Rejected	1	Private NFP	Large	E	146
5	001F000000nH9cKIAS	2016	Withdrawn	1	Private NFP	Large	E	146
6	001F000000nH9cKIAS	2017	Withdrawn	1	Private NFP	Large	E	146
7	001F000000nH9c0IAC	2016	Accepted	3	Private NFP	Large	E	141
8	001F000000nHDweIAG	2016	Accepted	3	Private NFP	Large	E	386
9	001F000000nHDweIAG	2017	Accepted	3	Private NFP	Large	E	386
10	001F000000nH9c0IAC	2016	Rejected	3	Private NFP	Large	E	141
11	001F000000nHDweIAG	2016	Rejected	3	Private NFP	Large	E	386
12	001F000000nH9c0IAC	2017	Rejected	3	Private NFP	Large	E	141
13	001F000000nHDweIAG	2017	Rejected	3	Private NFP	Large	E	386
14	001F000000nH9c0IAC	2016	Withdrawn	3	Private NFP	Large	E	141
15	001F000000nHDweIAG	2016	Withdrawn	3	Private NFP	Large	E	386
16	001F000000nH9c0IAC	2017	Withdrawn	3	Private NFP	Large	E	141
17	001F000000nH7A0IAK	2016	Accepted	4	Private NFP	Large	E	12
18	001F000000nH7h2IAC	2016	Accepted	4	Private NFP	Large	E	53
19	001F000000nH879IAC	2016	Accepted	4	Private NFP	Large	E	132
20	001F000000nH9cLIAS	2016	Accepted	4	Private NFP	Large	E	147
21	001F000000nHCGRIA4	2016	Accepted	4	Private NFP	Large	E	321
22	001F000000nH9cLIAS	2017	Accepted	4	Private NFP	Large	E	147
23	001F000000nHCGRIA4	2017	Accepted	4	Private NFP	Large	E	321
24	001F000000nH7A0IAK	2016	Rejected	4	Private NFP	Large	E	12
25	001F000000nH7h2IAC	2016	Rejected	4	Private NFP	Large	E	53

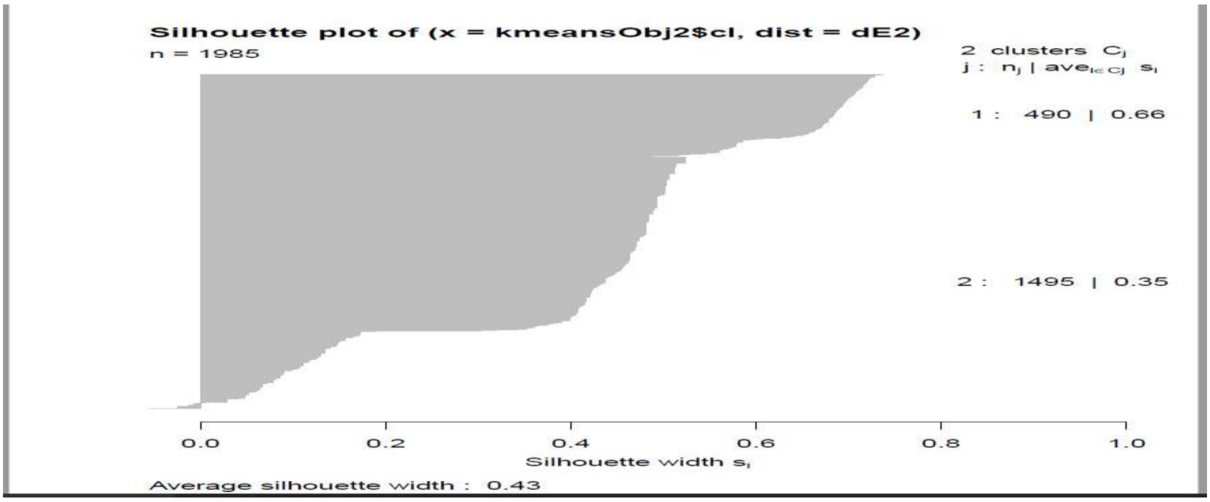
From the above data we try to segment The universities into various clusters.

We used Kmeans Clustering technique to cluster data into various groups and find the best cluster.

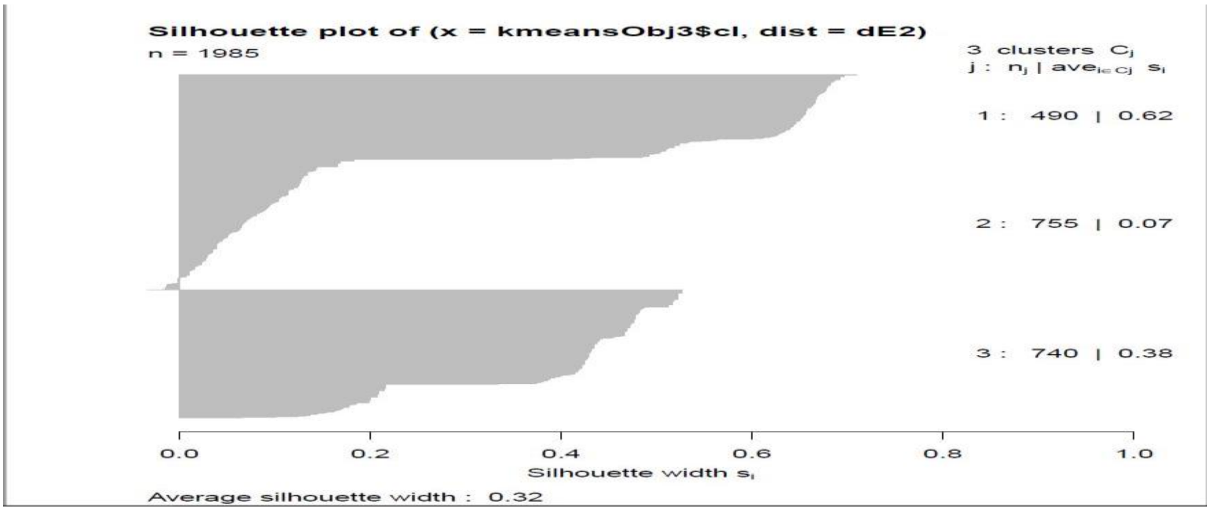
We tried clustering the data with various K values and choosing the best optimal K values by plotting the data

We iterated K values between 2 and 7 and plotted Silhouette plots. The Average Silhouette width shows how good the clusters are integrated.

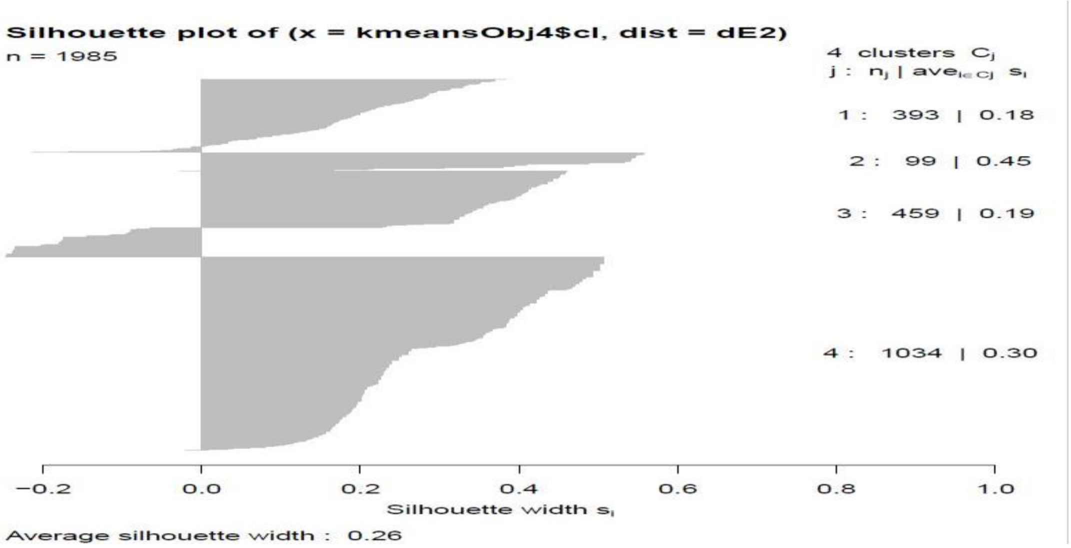
K=2



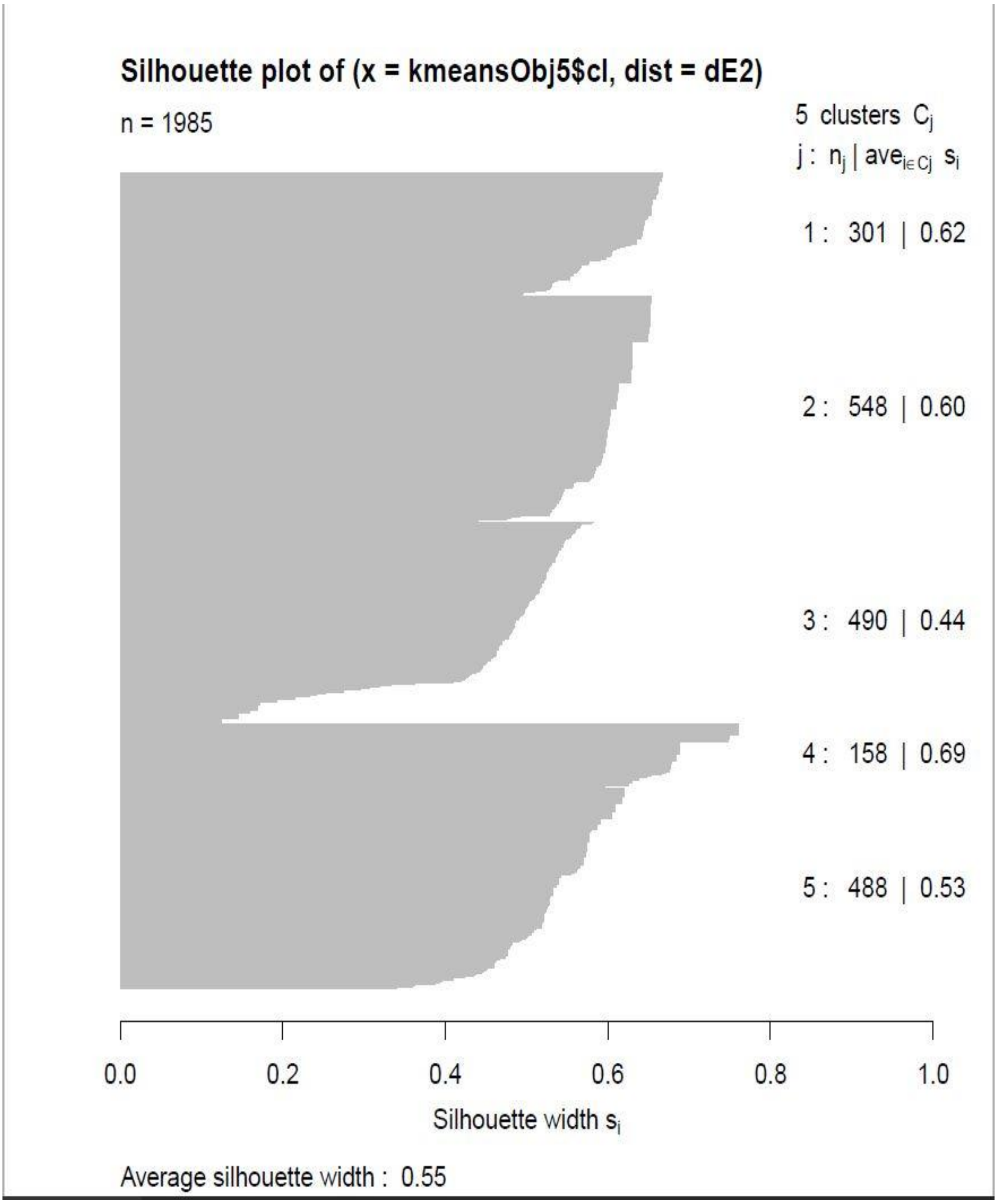
K=3


Silhouette plot of (x = kmeansObj3$cl, dist = dE2)
n = 1985
3 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 490 | 0.62
2 : 755 | 0.07
3 : 740 | 0.38
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.32

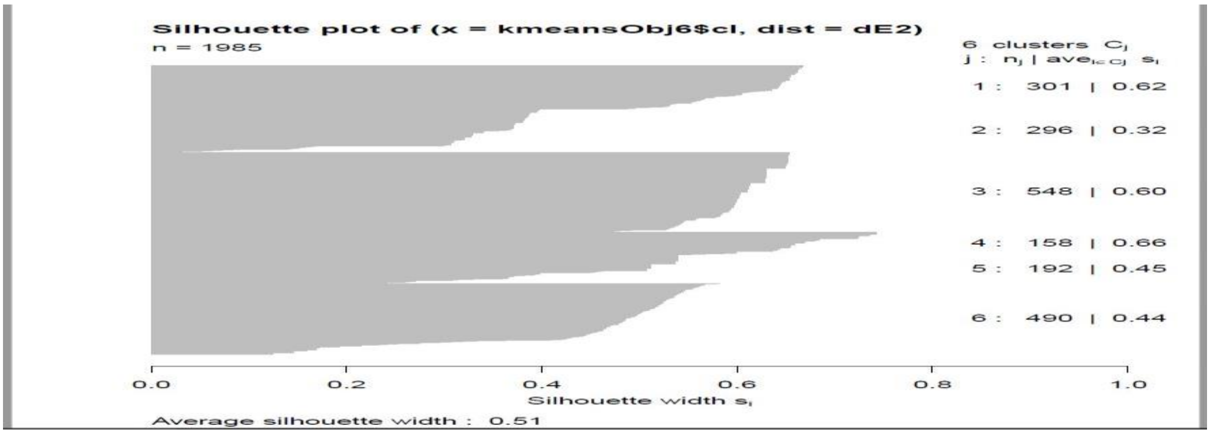
K=4


Silhouette plot of (x = kmeansObj4$cl, dist = dE2)
n = 1985
4 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 393 | 0.18
2 : 99 | 0.45
3 : 459 | 0.19
4 : 1034 | 0.30
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.26

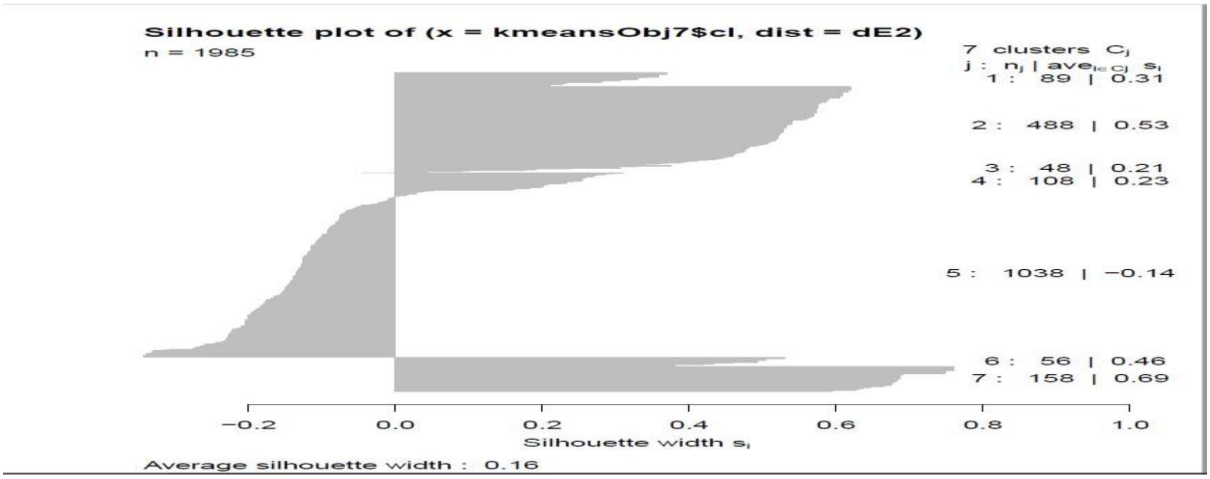
K=5


Silhouette plot of (x = kmeansObj5$cl, dist = dE2)
n = 1985
5 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 301 | 0.62
2 : 548 | 0.60
3 : 490 | 0.44
4 : 158 | 0.69
5 : 488 | 0.53
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.55

K=6



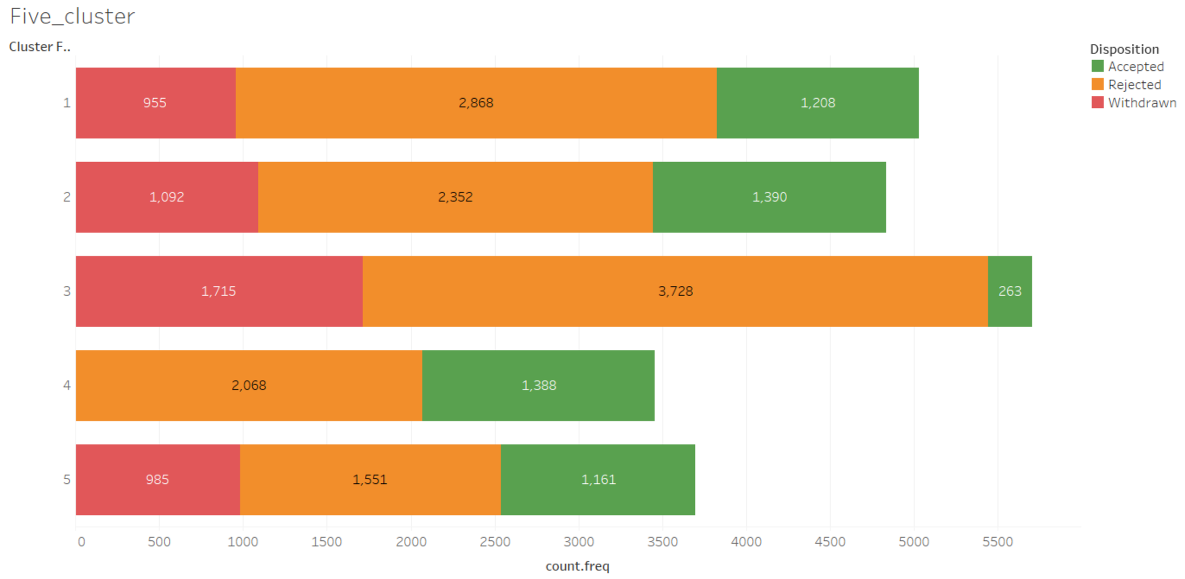
K=7



Looking at the values of the Width of the silhouette plots we find that clusterinf the data into 5 provides optimal ways to classify the data.

Hence we Recommend classifying the universities into 5 tiers,each tier as recommended by the K means clustering.

We visualized all the 5 clusters to find rank the clusters , below is the screenshot for the same

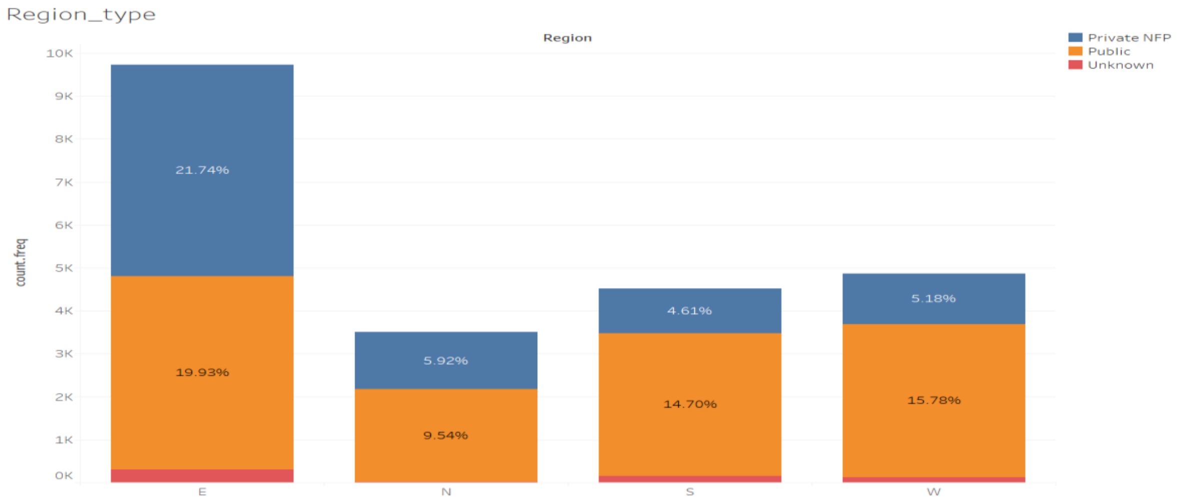


Sum of count.freq for each Cluster Five. Color shows details about Disposition. The marks are labeled by sum of count.freq.

From this visualization we see that **cluster 2** has the most number of people accepted followed by **cluster4,cluster1,cluster5 and cluster3**. Here we see that the universities in cluster 3 are the ones where maximum people apply or show interest rather only very few percentage get selected.

Thus we recommend allocating many universities for resource under cluster3 , because the likelihood of finding a match is very less.

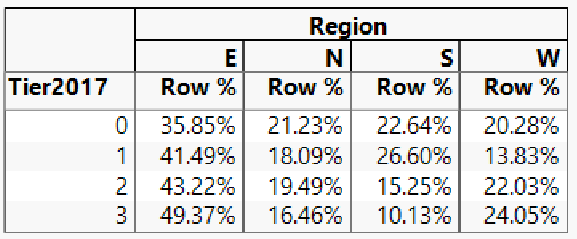
Our second recommendation is to decide the tier for each school based on the region in which it is located. This comes on the back of some thorough analyses and the very unique insights derived from it and is the following:



Sum of count.freq for each Region. Color shows details about Type. The marks are labeled by % of Total count.freq. The view is filtered on Region, which keeps E, N, S and W.

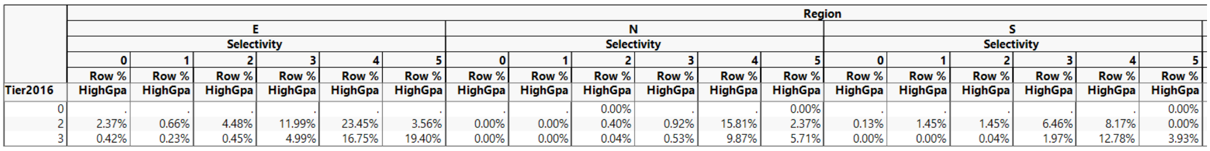
Specific Insights:

- If you are choosing a non-STEM course for a school, do not invest in the West region



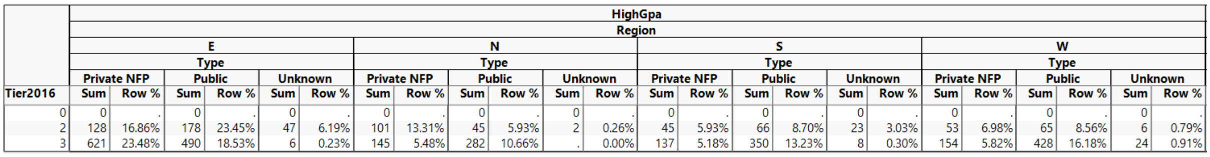
	Region			
	E	N	S	W
Tier2017	Row %	Row %	Row %	Row %
0	35.85%	21.23%	22.64%	20.28%
1	41.49%	18.09%	26.60%	13.83%
2	43.22%	19.49%	15.25%	22.03%
3	49.37%	16.46%	10.13%	24.05%

- Another observation was the most selective schools did not end up being in the best tiers. A school with more of a mid-level selectivity would make more sense.



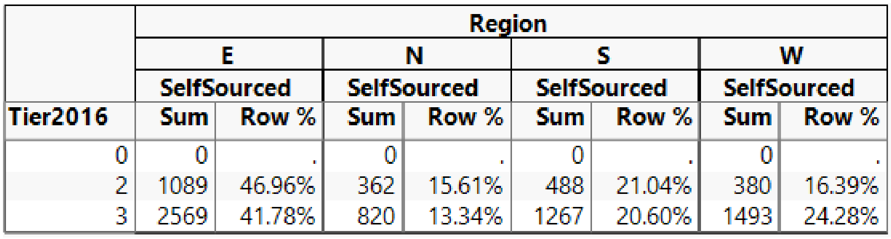
	Region																	
	E						N						S					
	Selectivity						Selectivity						Selectivity					
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %	Row %
Tier2016	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa	HighGpa
0	.	.	.	.	.	.	.	.	0.00%	.	.	0.00%	.	.	.	.	.	0.00%
2	2.37%	0.66%	4.48%	11.99%	23.45%	3.56%	0.00%	0.00%	0.40%	0.92%	15.81%	2.37%	0.13%	1.45%	1.45%	6.46%	8.17%	0.00%
3	0.42%	0.23%	0.45%	4.99%	16.75%	19.40%	0.00%	0.00%	0.04%	0.53%	9.87%	5.71%	0.00%	0.00%	0.04%	1.97%	12.78%	3.93%

- However, the schools are more selective in the Eastern region (20% of top tier + 17% from the remaining tiers) – roughly 18% of total staff can be from the Eastern region. This is followed by South, West and the North regions in decreasing order.
- If High GPA is your significant factor, then select only the schools in East or North regions.



	HighGpa																							
	Region																							
	E						N						S						W					
	Type						Type						Type						Type					
	Private NFP		Public		Unknown		Private NFP		Public		Unknown		Private NFP		Public		Unknown		Private NFP		Public		Unknown	
Tier2016	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %
0	0	.	0	.	0	.	0	.	0	.	0	.	0	.	0	.	0	.	0	.	0	.	0	.
2	128	16.86%	178	23.45%	47	6.19%	101	13.31%	45	5.93%	2	0.26%	45	5.93%	66	8.70%	23	3.03%	53	6.98%	65	8.56%	6	0.79%
3	621	23.48%	490	18.53%	6	0.23%	145	5.48%	282	10.66%	.	0.00%	137	5.18%	350	13.23%	8	0.30%	154	5.82%	428	16.18%	24	0.91%

- If self-sourced is your significant factor, then Eastern region is more self-sourced (at least 3 times more than others) followed by the schools from West especially if you are considering schools from the 3rd tier – west is preferred over east and south.



	Region							
	E		N		S		W	
	SelfSourced		SelfSourced		SelfSourced		SelfSourced	
Tier2016	Sum	Row %	Sum	Row %	Sum	Row %	Sum	Row %
0	0	.	0	.	0	.	0	.
2	1089	46.96%	362	15.61%	488	21.04%	380	16.39%
3	2569	41.78%	820	13.34%	1267	20.60%	1493	24.28%

- If public universities are your significant factors, then again East schools have a very large size and hence they should be preferred. South is in the second place