Are Top Universities More Attractive To Tech Startups?

Introduction

Do we see more tech startups around the top Universities in the UK? There is evidence that tech startups cluster together and become incubators for innovation but does that also correlate with where certain Universities are located?

The purpose of this project is to find out if certain Universities have a higher proportion of tech startups in close proximity and the similarities between those Universities.

Data

To look at this problem we are going to need two data sets:

- A list of UK universities with ranking data
- A count of tech startups in close proximity to each University

For the university ranking data we will use The Times Higher Education World Rankings data

To get tech startup proximity data we will use the FourSquare API

UK university ranking data

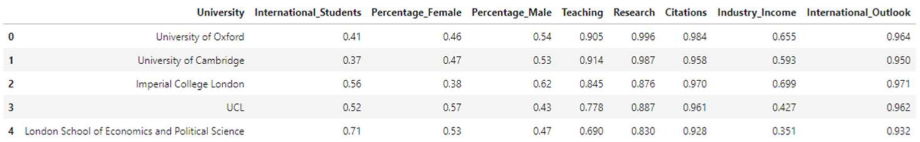
The data is not easy to scrape from the THE website, luckily this data has already been collected by a contributor on kaggle

The data can be found in https://www.kaggle.com/joeshamen/world-university-rankings-2020/download

Credit must be given to THE for producing this data and Joe Shamen for making it available in this form.

To make the data useful for our purposes we drop columns we are not interested in. Remove the rank field because we will use the index instead.

The data is formatted and then scaled by dividing each number by 100 so we can compare them later on. This leave us with a dataset that looks like this:

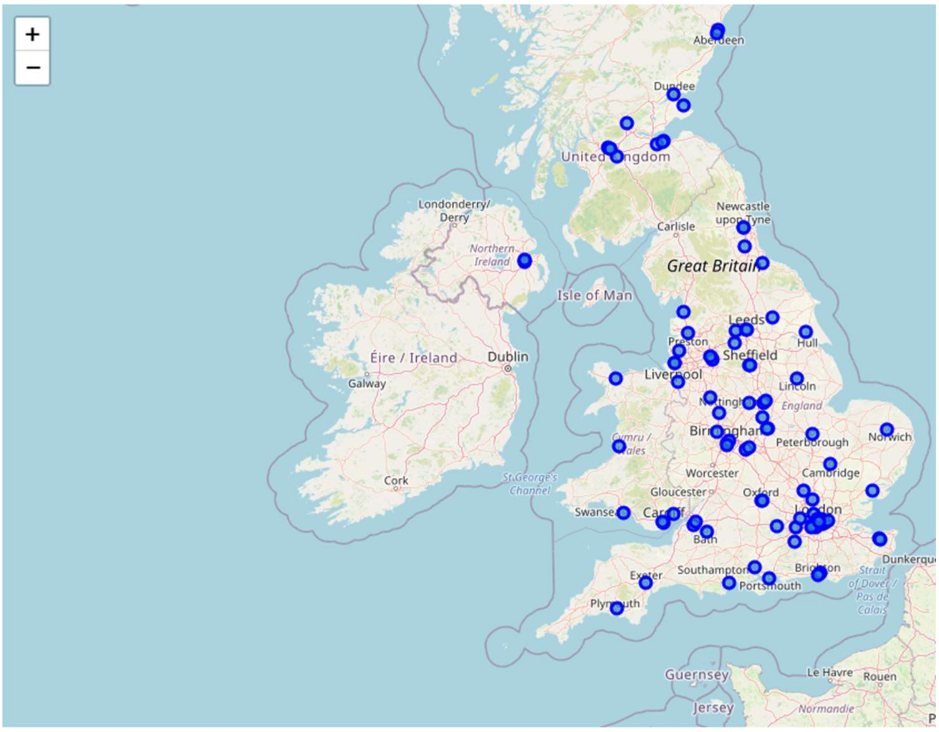


	University	International_Students	Percentage_Female	Percentage_Male	Teaching	Research	Citations	Industry_Income	International_Outlook
0	University of Oxford	0.41	0.46	0.54	0.905	0.996	0.984	0.655	0.964
1	University of Cambridge	0.37	0.47	0.53	0.914	0.987	0.958	0.593	0.950
2	Imperial College London	0.56	0.38	0.62	0.845	0.876	0.970	0.699	0.971
3	UCL	0.52	0.57	0.43	0.778	0.887	0.961	0.427	0.962
4	London School of Economics and Political Science	0.71	0.53	0.47	0.690	0.830	0.928	0.351	0.932

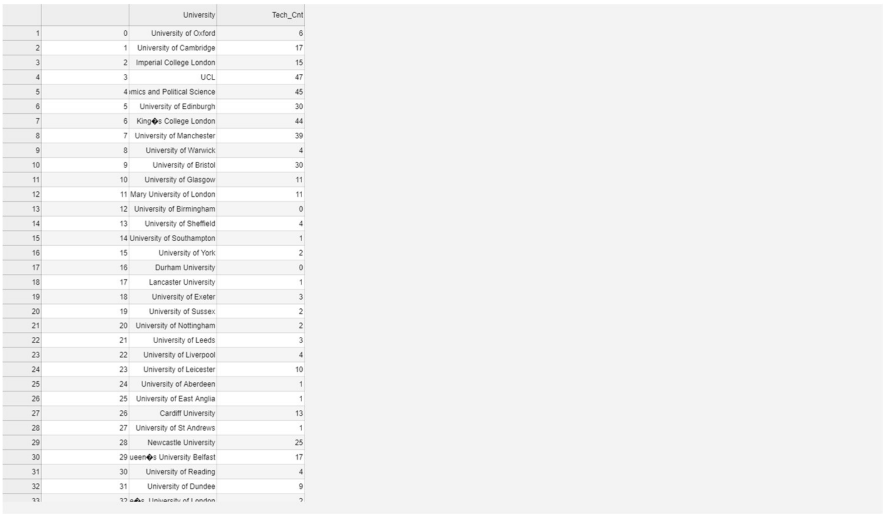
Tech Startup Data

Before we can get the count of tech startups near each University, we need to find the location data for each University. For that we use the geopy package and the Nominatim geocoder.

To ensure the location data is valid, two levels of validation were performed. First we check if any rows have null data for location, and that was found. The missing data was entered manually. Secondly we validate the locations are correct by presenting the Universities on a folium map as shown below.



Using the location data, we use the FourSquare search API to find tech startups within a kilometre of each University. Any Universities that returned an empty venues list were updated with a count of 0. This count is then stored in a new dataframe.



		University	Tech_Cnt
1	0	University of Oxford	6
2	1	University of Cambridge	17
3	2	Imperial College London	15
4	3	UCL	47
5	4	›mics and Political Science	45
6	5	University of Edinburgh	30
7	6	King�s College London	44
8	7	University of Manchester	39
9	8	University of Warwick	4
10	9	University of Bristol	30
11	10	University of Glasgow	11
12	11	Mary University of London	11
13	12	University of Birmingham	0
14	13	University of Sheffield	4
15	14	University of Southampton	1
16	15	University of York	2
17	16	Durham University	0
18	17	Lancaster University	1
19	18	University of Exeter	3
20	19	University of Sussex	2
21	20	University of Nottingham	2
22	21	University of Leeds	3
23	22	University of Liverpool	4
24	23	University of Leicester	10
25	24	University of Aberdeen	1
26	25	University of East Anglia	1
27	26	Cardiff University	13
28	27	University of St Andrews	1
29	28	Newcastle University	25
30	29	ueen�s University Belfast	17
31	30	University of Reading	4
32	31	University of Dundee	9
33	32	e�s University of London	2

We now have the datasets required in formats that we can use in the analysis phase. To avoid having to repeat the exercise the datasets were saved to csv files.

Methodology

The data will be analysed from two perspectives.

One is by looking at the correlation between the number of tech startups and attributes of each University. For this we will look for linear relationships between the values and their significance through statistical measures such as pearson correlation coefficient.

Secondly the Universities will be clustered using kmeans clustering from there we will look if there is a relationship.

Analysis

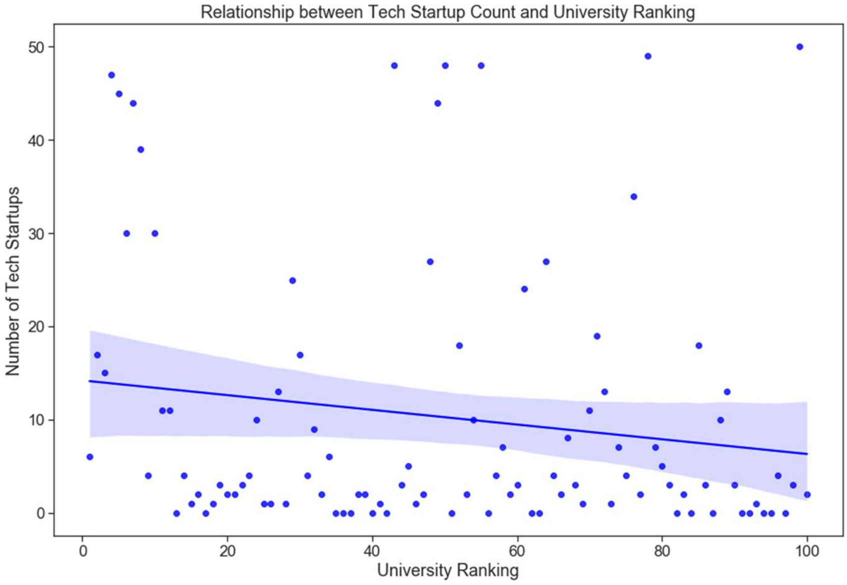
Correlation Analyis

Each data point in this section is correlated against number of tech startups to under understand the relationship between the two variables. We visually check for correlation and then check mathematically using pearson correlation coefficient.

Ranking vs Tech Startup Count

The ranking feature in the dataset is the rank of the University as given by THE in the UK. Does University rank correlate with the number of tech startups?

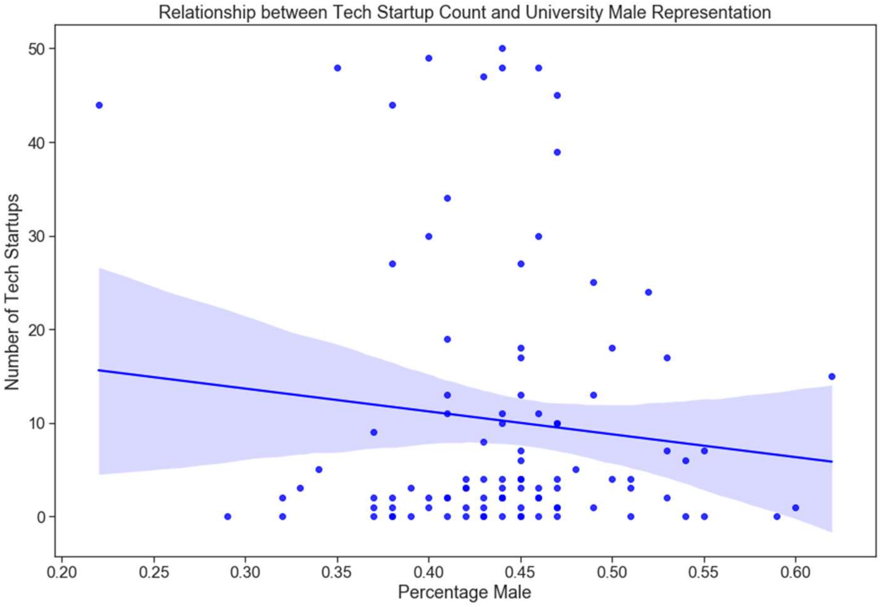
First lets visualise it using a regression plot.



There doesn’t appear to be a relationship here. This is confirmed by pearson correlation coefficient which gives us Correlation:: -0.15989384537789045, Significance:: 0.11204284425635871

Male Representation vs Tech Startup Count

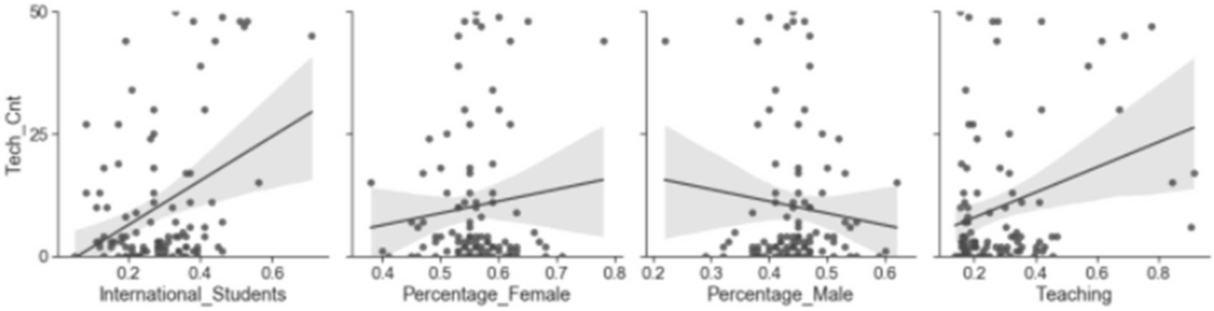
The dataset includes the male and female population ratios in each University. Do we see male bias when it comes to tech startups?

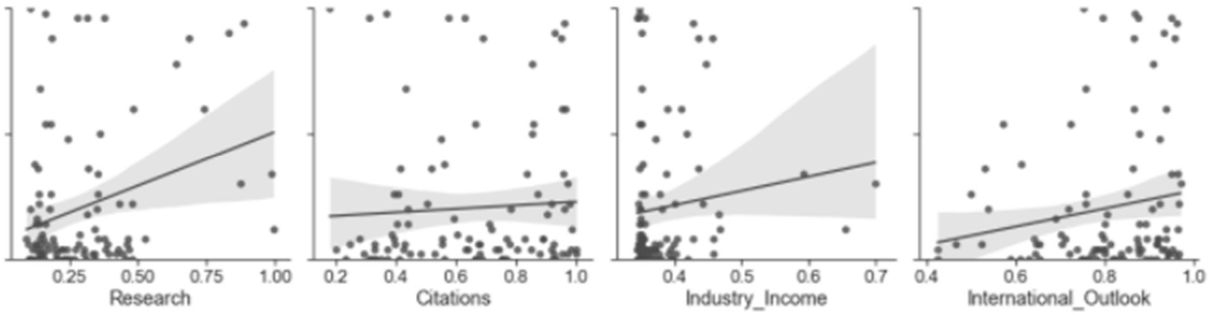


Again there doesn’t appear to be a relationship here. This is confirmed by pearson correlation coefficient which gives us Correlation:: -0.10689824071206178, Significance:: 0.2897913801355428

Multiple value correlation

Instead of going through each variable in turn lets graph the variables simultaneously so we can spot any relationships





Looking at the graphs the strongest corelations appear to be:

- Ratio of International Students
- Teaching Score
- Research Score
- International Outlook Score

Lets run these through pearson to see if these are statistically significant

International Students

Correlation:: 0.3725851074495698, Significance:: 0.00013485931494028897

Teaching

Correlation:: 0.29666362220215314, Significance:: 0.00272546569137422

Research

Correlation:: 0.2987968924629771, Significance:: 0.002530209708622019

International Outlook

Correlation:: 0.16969820209989445, Significance:: 0.09142801878333925

All of the correlations are quite weak. P values under 0.05 indicates statistical significance so all have significance except International Outlook.

Cluster Analyis

Lets cluster the universities based on there attributes and lets see if particular clusters have a higher number of tech startups

We are going to use kmeans as our clustering algorithm. We will start with a k of 4 and then go from there.

For us to be able to get a good set of clusters we need drop the following variables when fitting the data:

Ranking – Unique, continuous value so it would distort the similarity tests

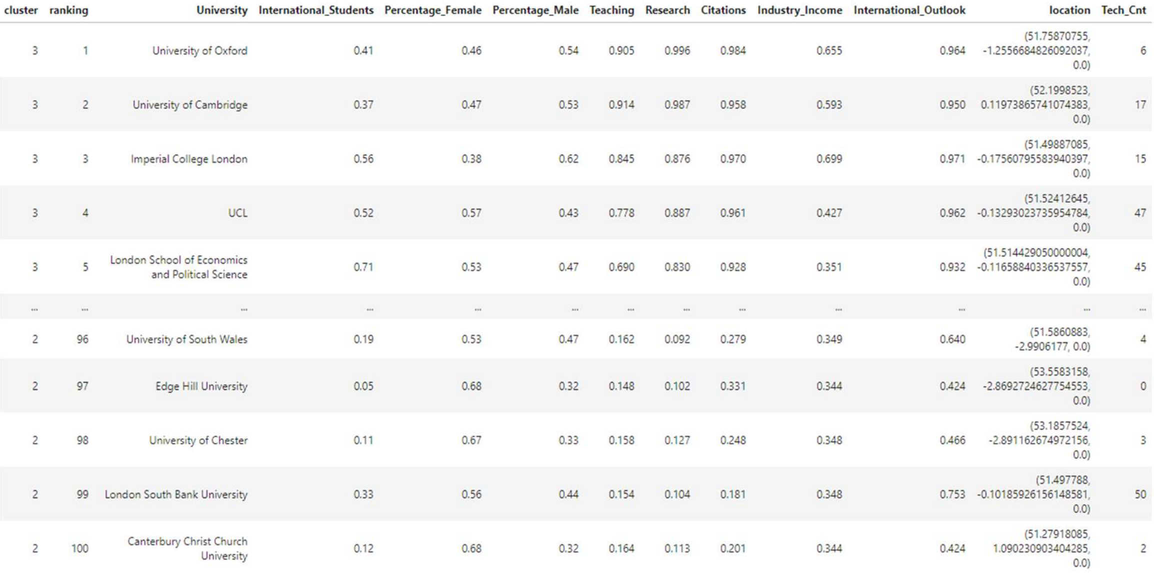
University – Dropped as this text field cannot be used in the distance calculation

Location – This will lead to noise in the data. We are not interested In this data for this report

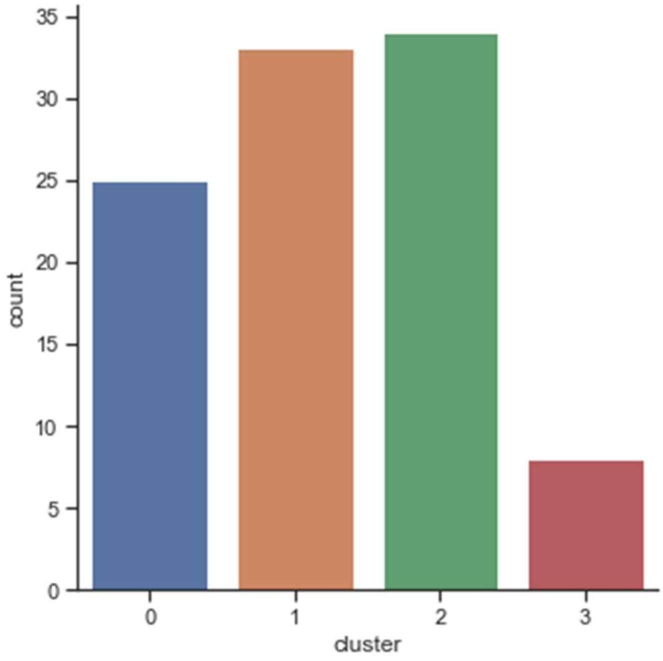
Tech Count – This is the variable that we want to check the clusters against later on

KMeans Cluster with K of 4

After fitting the data to the kmeans algorithm we get a grouping as shown below



cluster	ranking	University	International_Students	Percentage_Female	Percentage_Male	Teaching	Research	Citations	Industry_Income	International_Outlook	location	Tech_Cnt
3	1	University of Oxford	0.41	0.46	0.54	0.905	0.996	0.984	0.655	0.964	(51.75870755, -1.2556684826092037, 0.0)	6
3	2	University of Cambridge	0.37	0.47	0.53	0.914	0.987	0.958	0.593	0.950	(52.1998523, 0.11973865741074383, 0.0)	17
3	3	Imperial College London	0.56	0.38	0.62	0.845	0.876	0.970	0.699	0.971	(51.49887085, -0.17560795583940397, 0.0)	15
3	4	UCL	0.52	0.57	0.43	0.778	0.887	0.961	0.427	0.962	(51.52412645, -0.13293023735954784, 0.0)	47
3	5	London School of Economics and Political Science	0.71	0.53	0.47	0.690	0.830	0.928	0.351	0.932	(51.514429050000004, -0.11658840336537557, 0.0)	45
...	...	...	...	...	...	...	...	...	...	...	...	...
2	96	University of South Wales	0.19	0.53	0.47	0.162	0.092	0.279	0.349	0.640	(51.5860883, -2.9906177, 0.0)	4
2	97	Edge Hill University	0.05	0.68	0.32	0.148	0.102	0.331	0.344	0.424	(53.5583158, -2.8692724627754553, 0.0)	0
2	98	University of Chester	0.11	0.67	0.33	0.158	0.127	0.248	0.348	0.466	(53.1857524, -2.891162674972156, 0.0)	3
2	99	London South Bank University	0.33	0.56	0.44	0.154	0.104	0.181	0.348	0.753	(51.497788, -0.10185926156148581, 0.0)	50
2	100	Canterbury Christ Church University	0.12	0.68	0.32	0.164	0.113	0.201	0.344	0.424	(51.27918085, 1.090230903404285, 0.0)	2

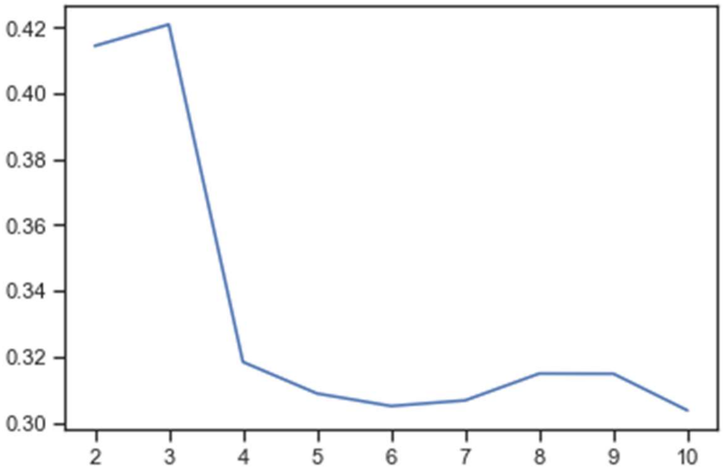


Finding the best k for KMeans

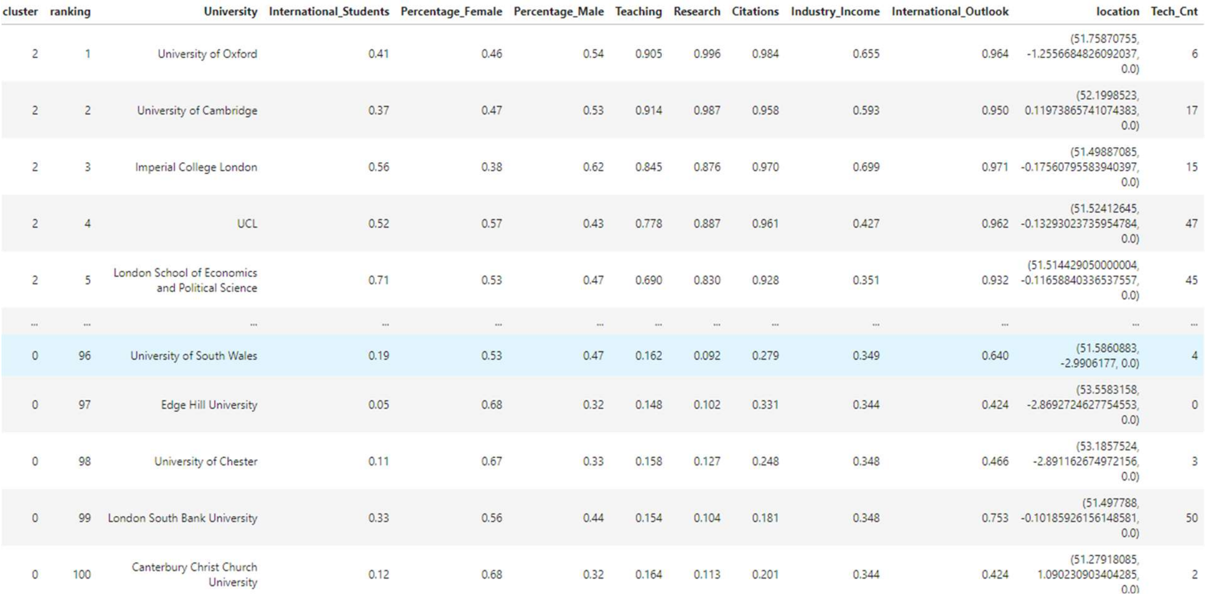
We arbitrarily took 4 as a starting point for kmeans but lets now take a more methodical approach before going further.

There are several methods available to iteratively run kmeans to find the best value for k. One of those is the silhouette method which uses a score based on how similar the points are in each cluster and dissimilar to other clusters. The score is between –1 and +1, the highest score for a given range of k will give us the best k to use.

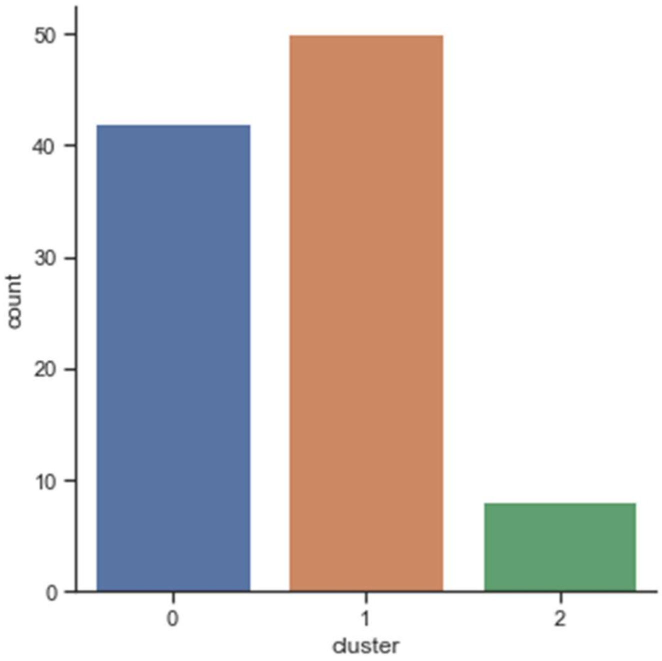
The results of the analysis are shown below. We can see that k = 3 is the optimal value



Re-running kmeans for the new optimal gives us



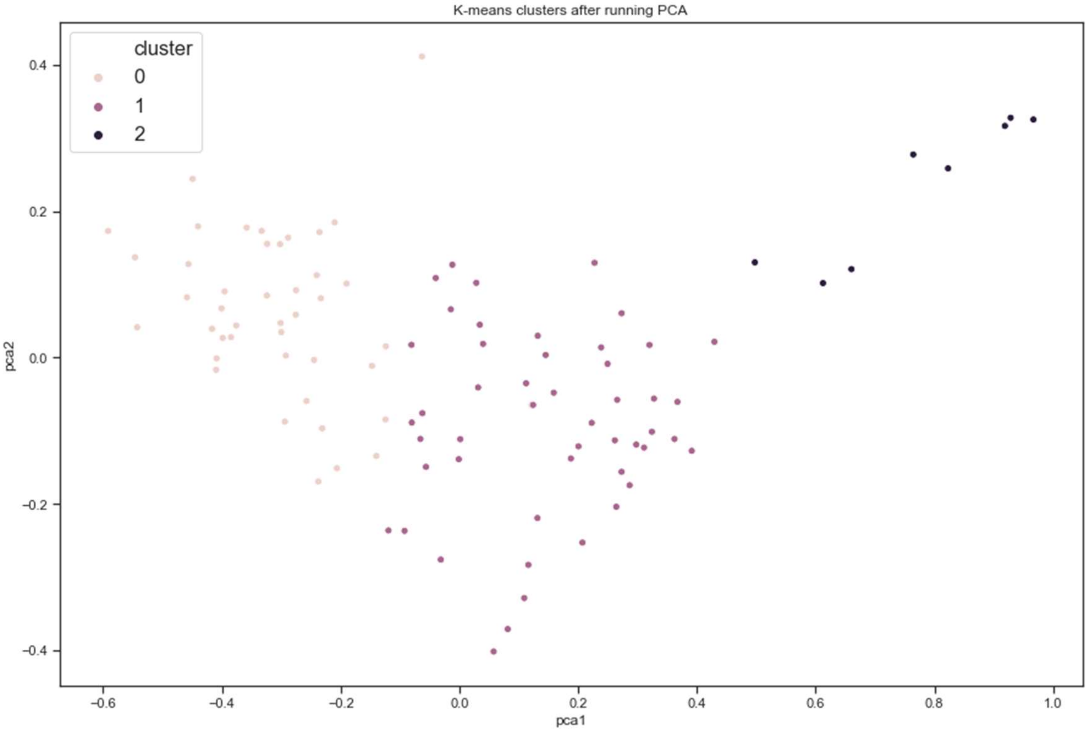
cluster	ranking	University	International_Students	Percentage_Female	Percentage_Male	Teaching	Research	Citations	Industry_Income	International_Outlook	location	Tech_Cnt
2	1	University of Oxford	0.41	0.46	0.54	0.905	0.996	0.984	0.655	0.964	(51.75870755, -1.2556684826092037, 0.0)	6
2	2	University of Cambridge	0.37	0.47	0.53	0.914	0.987	0.958	0.593	0.950	(52.1998523, 0.11973865741074383, 0.0)	17
2	3	Imperial College London	0.56	0.38	0.62	0.845	0.876	0.970	0.699	0.971	(51.49887085, -0.17560795583940397, 0.0)	15
2	4	UCL	0.52	0.57	0.43	0.778	0.887	0.961	0.427	0.962	(51.52412645, -0.13293023735954784, 0.0)	47
2	5	London School of Economics and Political Science	0.71	0.53	0.47	0.690	0.830	0.928	0.351	0.932	(51.514429050000004, -0.11658840336537557, 0.0)	45
...	...	...	...	...	...	...	...	...	...	...	...	...
0	96	University of South Wales	0.19	0.53	0.47	0.162	0.092	0.279	0.349	0.640	(51.5860883, -2.9906177, 0.0)	4
0	97	Edge Hill University	0.05	0.68	0.32	0.148	0.102	0.331	0.344	0.424	(53.5583158, -2.8692724627754553, 0.0)	0
0	98	University of Chester	0.11	0.67	0.33	0.158	0.127	0.248	0.348	0.466	(53.1857524, -2.891162674972156, 0.0)	3
0	99	London South Bank University	0.33	0.56	0.44	0.154	0.104	0.181	0.348	0.753	(51.497788, -0.10185926156148581, 0.0)	50
0	100	Canterbury Christ Church University	0.12	0.68	0.32	0.164	0.113	0.201	0.344	0.424	(51.27918085, 1.090230903404285, 0.0)	2



Visualising the clusters

As we are using multiple variables we can't easily visualise the clusters, there are too many dimensions. So we need to reduce the dimensions by using a method like Principal Component Analysis (PCA).

PCA was run against the data used for fitting against kmeans, with the PCA component number set to two so that we can graph by two dimensions. The data points were then overlayed with the cluster labels so that we can distinguish them.



The clusters can be easily identified. We can see that cluster 2 which is the high ranking Universities has mini clusters effectively and cluster 0 has an outlier which is far from all the clusters.

Our original analysis has shown that this k for kmeans produces the best similarity

Analyse the clusters

Lets run some simple statistics across the clusters so we can compare them (note that values have been rounded to two decimal places)

Field	Cluster	Min	Max	Mean
Ranking	0	55	100	79.24
	1	9	62	33.72
	2	1	8	4.50
International Students	0	0.05	0.53	0.22
	1	0.14	0.51	0.30

	2	0.37	0.71	0.48
Percentage Female	0	0.47	0.71	0.58
	1	0.40	0.78	0.54
	2	0.38	0.62	0.52
Percentage Male	0	0.29	0.53	0.42
	1	0.22	0.60	0.45
	2	0.38	0.62	0.48
Teaching	0	0.15	0.41	0.19
	1	0.14	0.47	0.30
	2	0.57	0.91	0.75
Research	0	0.09	0.38	0.15
	1	0.14	0.53	0.32
	2	0.64	1.00	0.83
Citations	0	0.18	0.68	0.42
	1	0.52	1.00	0.80
	2	0.85	0.98	0.95
Industry Income	0	0.34	0.37	0.35
	1	0.34	0.47	0.38
	2	0.35	0.70	0.50
International Outlook	0	0.42	0.89	0.69
	1	0.72	0.97	0.88
	2	0.91	0.97	0.95
Tech Count	0	0	50	8.71
	1	0	48	8.22
	2	6	47	30.38

We can see that on average the Universities in cluster 2 have more tech startups near them. These are also the same Universities that have higher values for the features that we saw earlier which had a weak correlation to tech startup count.

Results and Discussion

The analysis has shown that there is some correlation between clusters of Universities and the number of tech startups.

What we found was that individual features did not have a strong correlation but a combination did seem to have an effect. The top Universities clustered together, as an average had a higher tech startup count and interestingly at an average they had higher values for:

- Ratio of International Students
- Teaching Score
- Research Score

During the analysis we found that these features individually had a weak correlation to the number of tech startups but the correlation was statistically significant.

While there does appear to be a correlation we cannot be sure it is causation. Other areas to look at is location, from the data we could see that the tech startup count was higher around particular cities so this is an area for more investigation.

Conclusion

The purpose of this report is to help understand the relationship between Universities and tech startups. We have found a correlation with the top Universities but there are data points such as location that should be investigated further.

Hopefully this will be helpful for tech startups trying to find the best location to start their business.