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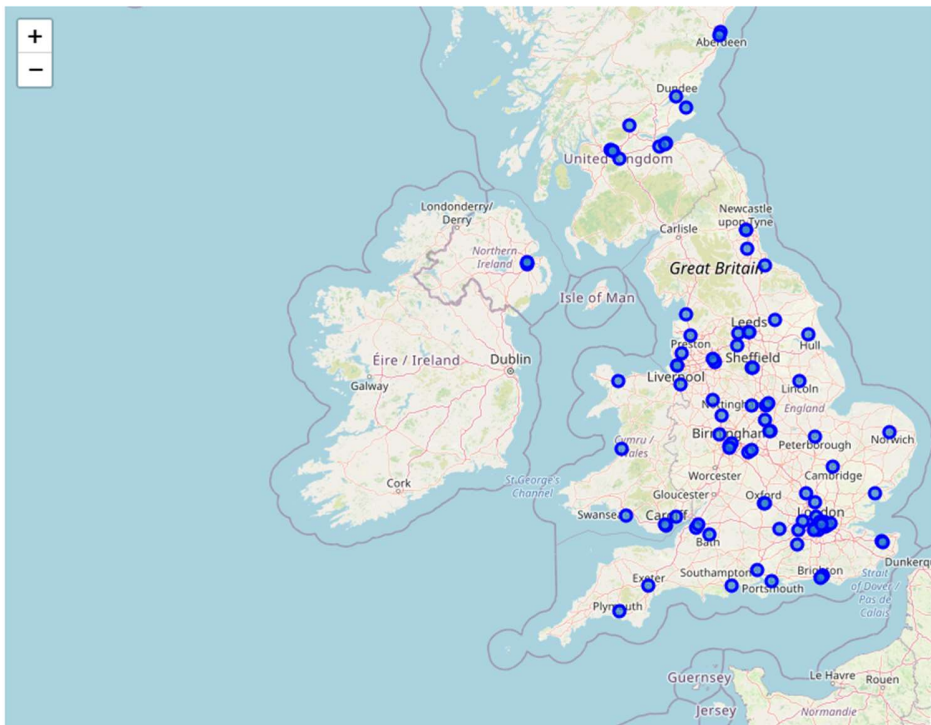
The data is formatted and then scaled by dividing each number by 100 so we can compare them later on. This leave us with a dataset that looks like this:

	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	s�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.