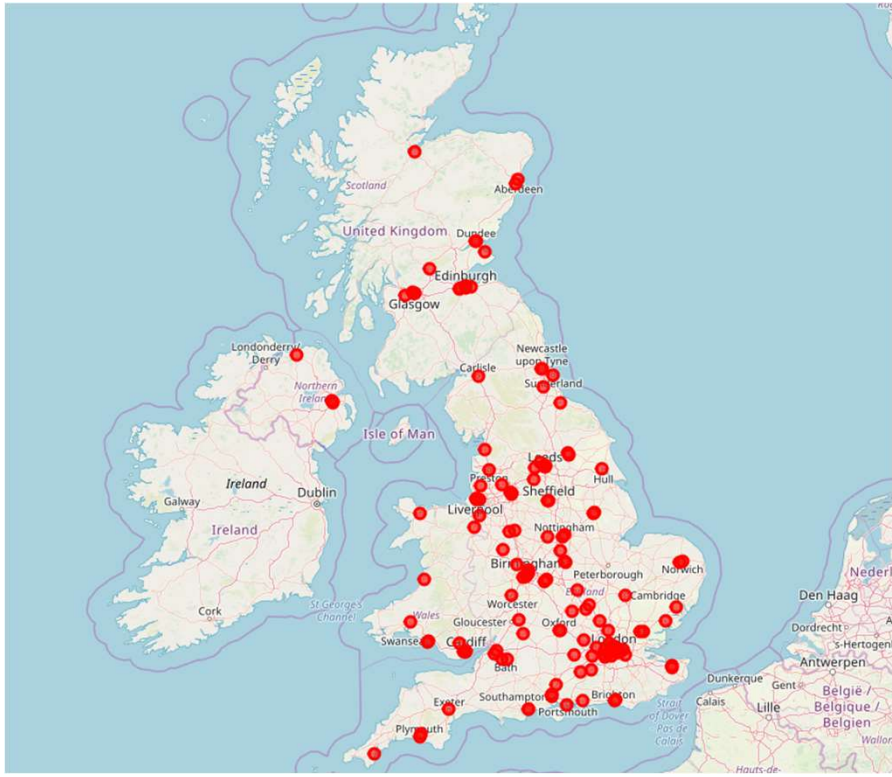
Outdoor spaces, culture, student satisfaction and research excellence at UK universities

Alistair Knock

Coursera / IBM Data Science Professional Certificate – Applied Data Science Capstone

Problem

- Can we help provide ‘soft’ information on the 160 UK universities through the combination of geodata, student satisfaction and research excellence data, so that these shortlists can be informed by a broader set of criteria?
- Do satisfied students tend to be concentrated in certain types of environment?
- Is world-class research related to the facilities and locale of the university?



Data sources

- NSS: Student satisfaction data is published by the UK's Office for Students through the National Student Survey (2018)
- REF: Research excellence data is captured through the Research Excellence Framework exercise administered by sector agencies, with the most recent data available from REF 2014
- Basic geographic data on UK universities including latitude and longitude is available from the UK Learning Providers website
- Detailed geographic data on 'positive facilities' and services in the vicinity of each universities is available through the Foursquare API

Assumptions and decisions

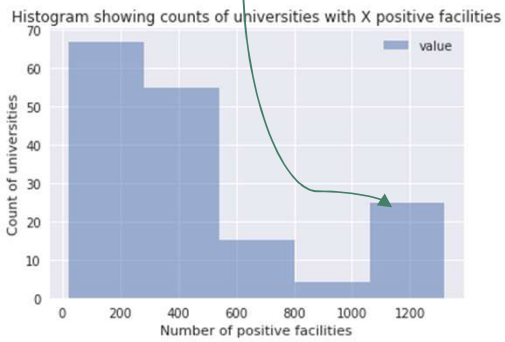
- Geographic data is limited to a five kilometre radius around the university as a proxy for ‘one hour’s walking’
- Foursquare venue categories were manually selected as those felt to be ‘positive’ – fits the criteria of including natural/culture provision which creates a healthy and positive environment
- Research excellence metric chosen as proportion of outputs graded as 3* and 4*
- Student satisfaction metric chosen as percent who agree with the statement ‘overall, I am satisfied with the quality of the course’

Known issues

- Geographic data is from 2019, but NSS data is for students who started circa 2015, and REF data is from researchers who may have started working on outputs over 10 years ago
- A boycott of the NSS 2018 survey led to missing values for around 16 universities
- REF and NSS are samples – if we knew the unknown values for non-respondents, the results may be different
- Foursquare data may not have been captured with this type of analysis in mind (generally focused on commercial enterprises)

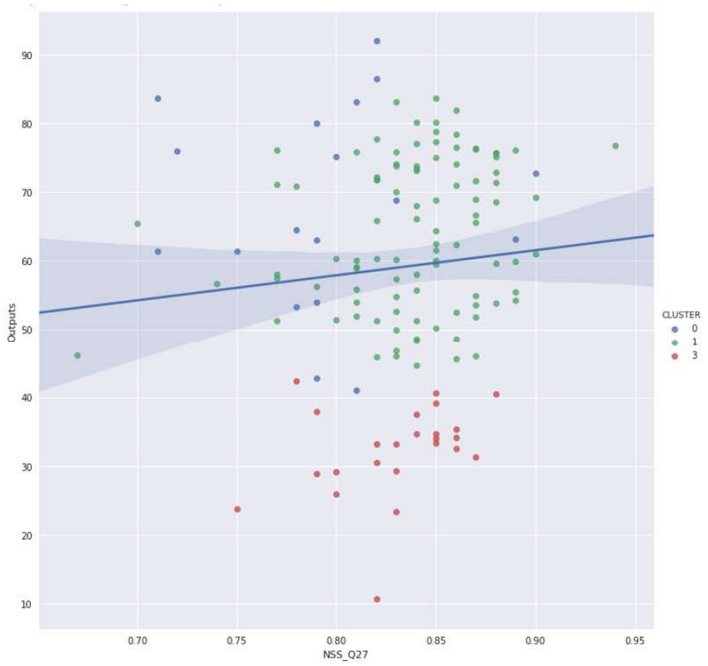
Methods and approach

- *Tools: Python, pandas, scikit-learn, Jupyter notebooks, numpy, matplotlib*
- The universities, NSS, REF were imported from flat files, the geodata from API calls, and all were merged into a single dataframe containing one record per university
- Descriptive statistics and visualisations (maps, histograms, box plots) were initially used to explore and understand the data
- Iterative – a discovery part-way through (that London universities have notably different geographic properties) led to a further iteration excluding a subset
- Start with a simple three feature set, then move to a detailed 300 feature set
- K-means clustering and DBSCAN were used to identify clusters of universities within both the simplified feature set and the detailed set

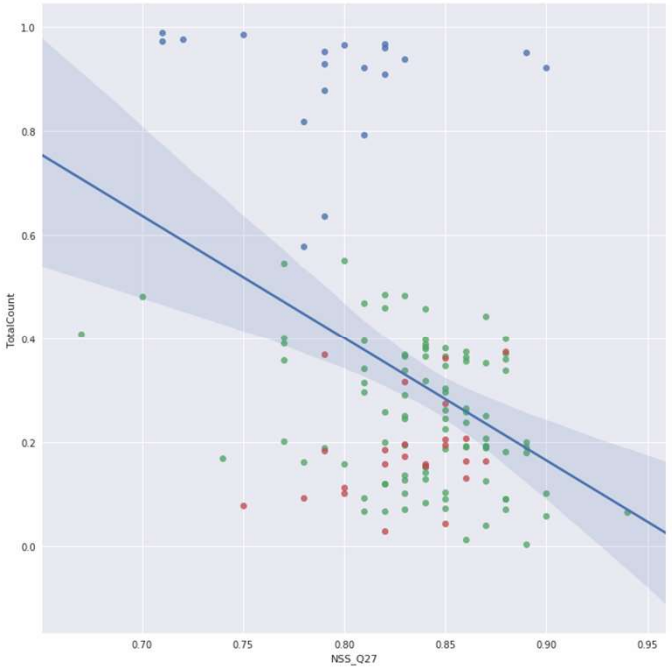


Observations – pairwise comparison

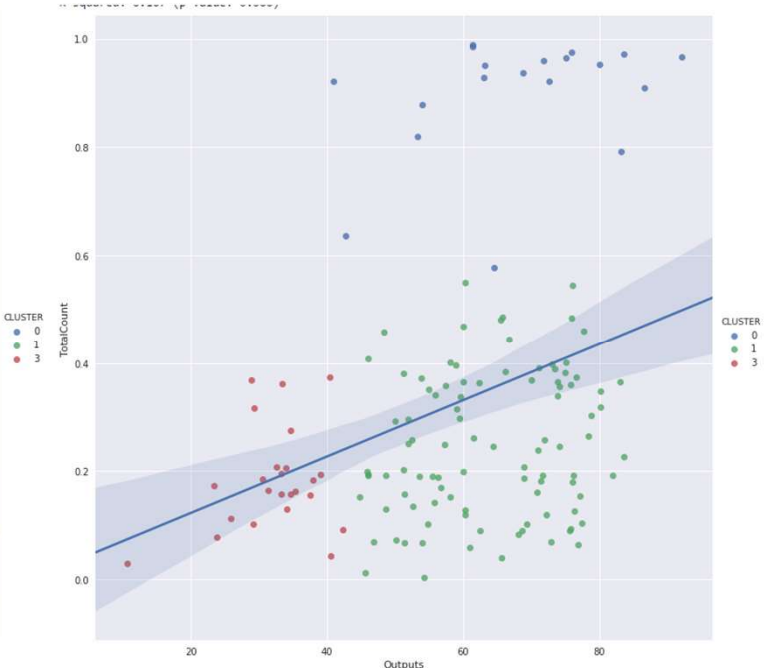
NSS vs REF



NSS vs positive facilities



REF vs positive facilities



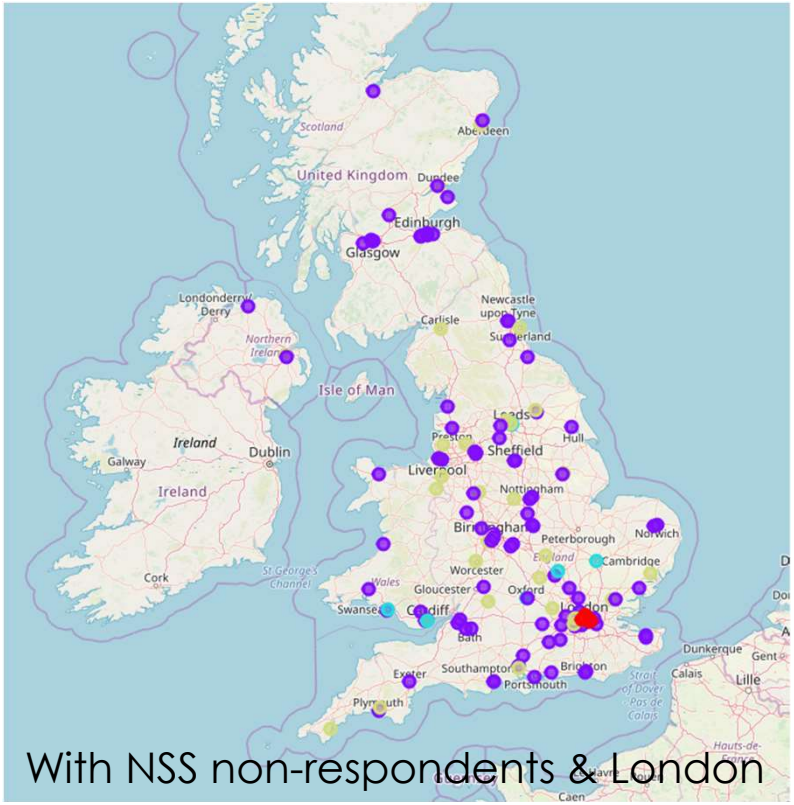
- Weak relationship between positive facilities and REF; weak inverse with NSS. No relationship between NSS and REF – and note low R-squared values, poor fit

Table 2: Pearson's correlation co-efficient and R-squared values for each combination:

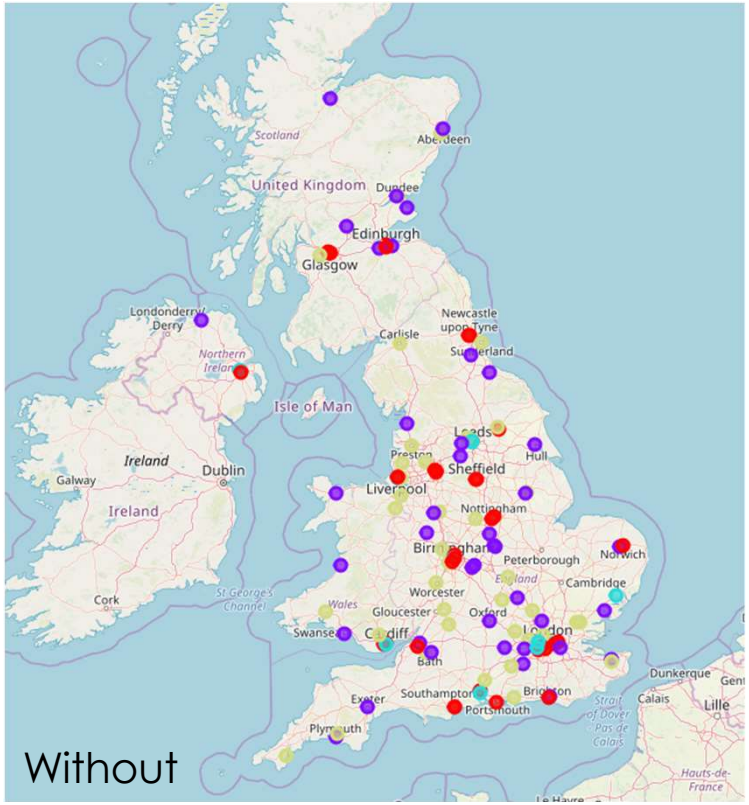
Combination	Pearson's	R-squared	p-value
NSS Q27 and outputs	0.096	0.009	0.259
NSS Q27 and positive facility count	-0.390	0.152	<0.001
Outputs and positive facility count	0.327	0.107	<0.001

Observations – k-means clustering

- Clustering the full dataset into 4 clusters identified a group of universities which did not have NSS results (light blue), and a cluster with very high positive facilities all located in London. Removing these led to wide geographic distribution of clusters. Initial DBSCAN analysis did not easily identify density clusters



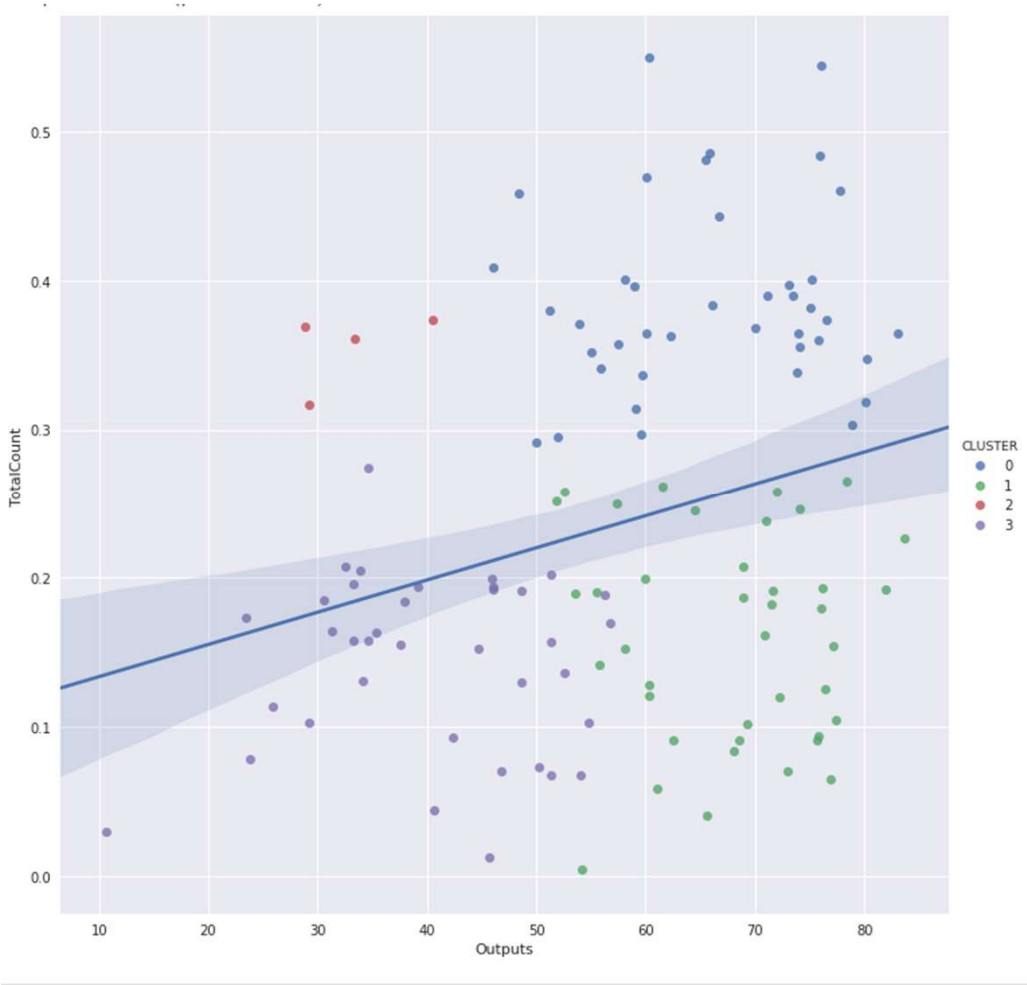
With NSS non-respondents & London



Without

Observations – REF vs positive facilities, subset

- It is interesting to note however the separation of universities into three main discrete clusters here
- However, this may be nothing to do with the features selected and may instead be due to other explanatory factors such as some universities being research-intensive and some consisting of mainly teaching provision.



Observations – detailed cluster

- When 340 individual features are used rather than the 3 initially selected features, some overlap between features is visible
- e.g. original cluster 0 equates to new cluster 1
- Notable that original clusters 1 and 3 have each split in half, with these halves combining to form new clusters 0 and 3 (yellow highlights)
- This distinction between new cluster 0 and cluster 3 is likely to be something in the detailed dataset other than the three original features

Table 6: Counts of universities in simple clustering (4 way) vs detailed clustering (5 way)

Simple \ Detailed	0	1	2	3	4
High positive facilities, low NSS, high REF	1	16	-	1	2
Low positive facilities, high NSS, high REF	55	-	-	42	-
High positive facilities, no NSS, high REF	-	4	7	-	5
Low positive facilities, high NSS, low REF	15	-	-	18	-

Conclusions and recommendations

- **Other geographic features** (e.g. London) complicated the analysis and should be taken into account in future either by exclusion or by separating/standardising the geographic areas used
- **No strong or significant relationship** between the three datasets
- Further exploration may be beneficial regarding REF and positive facilities (and converse for NSS), i.e. the expectations and needs of staff may not be the same as those of students.
- Further exploration may be beneficial of the division which emerged in the detailed clustering analysis
- If re-run, clearer decisions should be made on what to do **with missing results** and with a **narrower business problem**
- **Conclusion:** diversity of UK university sector represents a rich and mature education landscape which cannot be reduced down to a few performance metrics.