Geospatial and cluster analysis of the best places for Data Science graduates to live in England and Wales

1. Abstract

MSc Data Science students will be leaving the University of Bristol in every September, so they must decide where to go and prepare for that in teaching block 3. This data visualization coursework aims to guide the users, which are MSc Data Science students, to select better areas to live in terms of residents' age and job opportunities. To accomplish the aim, three goals should be achieved from the data visualizations. Firstly, "Overview" dashboards were created, to provide an overview of geographic distribution of areas that perform better in the relevant qualities that users require. Secondly, dimensionality reduction and clustering were conducted, and the results were visualized to help users to identify areas that perform well in most of the relevant qualities. Lastly, with the help of target cluster that selects better areas for users to live in, users can identify the best area that suits their personal needs to live in. The result showed that London and its surrounding areas were the best area for users, while high proportion of population with similar age and low unemployment rate should not be overly pursued.

2. Introduction

The MSc Data Science programme will be completed in every September. Where to live and work is the decision of many students have to make in teaching block 3. Therefore, this visual analytics project aims to provide necessary and relevant information for the users (MSc Data Science students) for the decision process. Three goals would be achieved from the data visualizations:

- Provide an overview of the distribution of the qualities that users (MSc Data Science students) may considered by local authorities.
- Use dimensionality reductions and clustering methods to categorize the local authorities into clusters based on the related qualities, for users to understand the correlations between qualities and characteristics of each cluster. This could help them to select the target cluster.
- Select the best local authorities that perform better in most of the related qualities.

The target users of this data visualization work are MSc Data Science students. They would desire a community with more residents at similar age rage (20s-30s) and plenty of data science related job opportunities. As a result, relevant 2021 census in England and Wales data was used to provide up-to-date and useful information for the goals.

3. Data Preparation and Abstraction

3.1 Relevant tables

Since the information provided by 2011 census was obsolete and this data visualization work values actual usability, 2021 census in England and Wales data was used. The following tables from 2021 census were selected for the residents age and job opportunities requirements of MSc Data Science students.

- TS007 - Age by single year

  The variable 'Aged 25 to 34 years (%)' from this table is the relevant quality as it shows the proportion of usual residents in the community similar to users' age range.

- RM074 - Legal partnership status by sex by age

  The derived variable 'Never married: Aged 25 to 34 years (Grand Total%)' is the relevant quality as it shows the proportion of single residents in the community similar to users' age range. This variable was calculated from dividing the 'never married population aged 25-34 years' by total population aged 16 years and over in that community.

- TS063 - Occupation

  According to the coding index of Standard Occupational Classification (SOC) 2020 (1), which was used in census 2021, position titles ‘Data Engineer’ and ‘Data Scientists’ were mostly classified as ‘Professional occupations’, so the variable ‘2. Professional occupations (%)’ which shows the proportion of ‘professional occupations’ can reflect the relevant job opportunities in a community.

- TS060 - Industry

  According to the Standard Industrial Classification (2), ‘Computer programming, consultancy and related activities’ and ‘Information service activities’ were classified as ‘J: Information and communication’ industry, so ‘J: Information and communication (%)’ variable in this table was selected as it reflects how likely residents worked in relevant industry as the users’ in a community.

- RM024 - Economic activity status by sex by age
  The derived variable ‘Economically active (excluding full-time students): Unemployed (%)’ is the relevant quality as it shows the difficulty of getting job in a community. It was calculated from dividing the economically active but unemployed population by total economically active population. Full-time students were excluded.

Census 2021 provided tables at different geographical level, from Output Areas which were made up of between 40 and 250 households each, to Regions which consisted of 9 areas (10 if Wales was included). Choosing tables with lower geographical level could overwhelm users with thousands of choices and add difficulty in visualization, while higher geographical level could result in the loss of information for users to select specific areas that meet their requirements. Eventually, tables at 331 local authority districts level were used for a manageable number of records to be visualized while the goals of selecting the right area to live could still be achieved.

Apart from local authority districts that were originally available in the tables, county and region/country (for Wales) variables were also added (3) to provide higher level of geographic data, so that the local authorities can be located more easily.

Then the tables were merged to be one dataset with the code representing each local authority (Lad22Cd) as the key. The dataset type of this dataset is a table, with local authority districts as key/item and other variables as attributes.

3.2 Geocoding package

To plot the census 2021 data on a map, a corresponding up-to-date geocoding package was required. A GeoJSON file containing local authority districts’ location and shape data was downloaded from the ONS Geography Linked Data site (4), and the local authority code (Lad22Cd) was also used as the key to merge this spatial dataset with the previously prepared table dataset within Tableau.

3.3 Missing value:

No missing value was found from the tables and geocoding package which were all downloaded from ONS.

4. Visualisation Justification

4.1 Dimensionality Reduction and Clustering

The following five variables that represents users’ requirements for considering where to live in were the input for the dimensionality reduction, to form two low dimension representation variables for visualization:

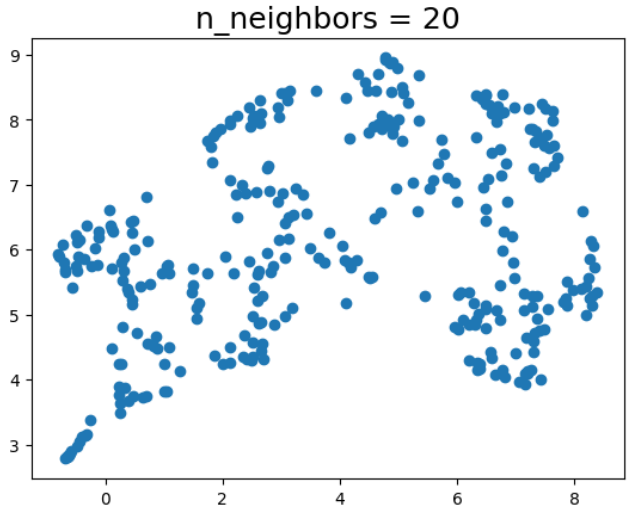
- The variable ‘Aged 25 to 34 years (%)’
- ‘Never married: Aged 25 to 34 years (Grand Total%)’
- ‘2. Professional occupations (%)’
- ‘J: Information and communication (%)’
- ‘Economically active (excluding full-time students): Unemployed (%)’

Both t-SNE and UMAP methods for dimensionality reduction were used, and then clustering by Gaussian Mixture Model method was conducted on their low dimension representation variables to decide which dimensionality reduction would be used for analysis.

UMAP

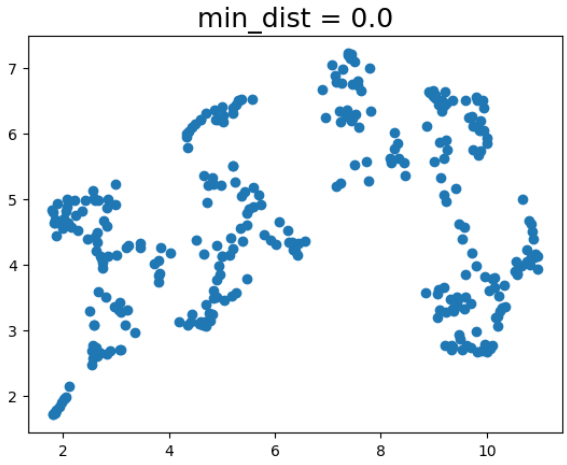
There were 2 hyperparameters to be optimized: number of neighbours and min-dist.

Firstly, a list of number of neighbours (5, 10, 15, 20, 30, 50, 100, 200) was used to conduct UMAP, and the two resulting variables were visualized by scatterplot, to see which number of neighbours could create the most identifiable structure for clustering. It was observed that number of neighbours=20 created the most well-defined distribution for clustering.



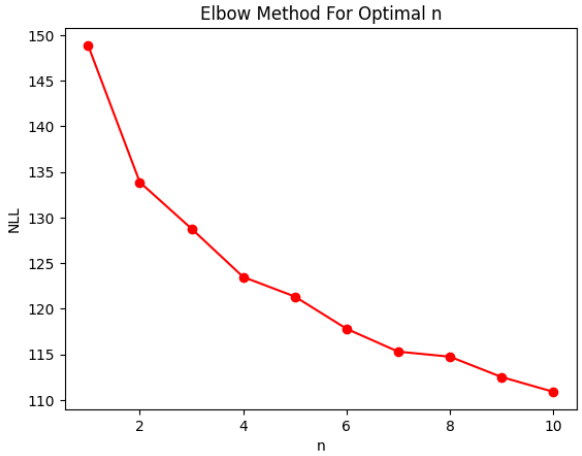
UMAP, neighbours=20, min_dist=0.1

Then a list of min-dist (0.0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.8, 0.99), which determines how closely points can be packed together, was used. From the visualization of the resulting variables, min-dist=0 can provide the most identifiable structure for clustering.



UMAP, neighbours=20, min_dist=0.0

Finally, Gaussian Mixture Model was used for clustering, and the hyperparameter to be optimized was the number of clusters. Number of clusters from 1 to 10 were tested, and the average total negative log likelihood from the validation sets (k-fold cross validation was used) at each number of clusters was plotted to find the elbow point.

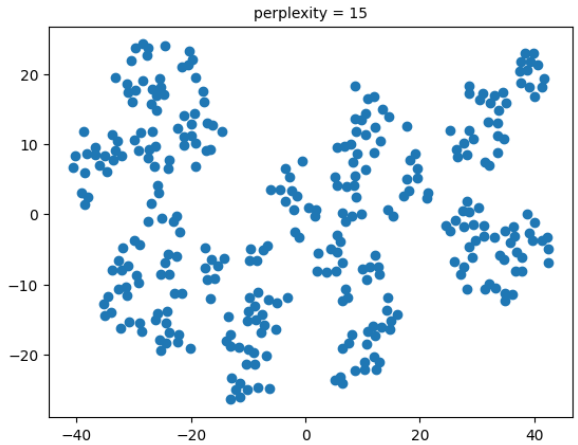


It was seen that the decrease of average total negative log likelihood slighted slowed down after n=4, so 4 was chosen as the optimal number of clusters for UMAP.

t-SNE

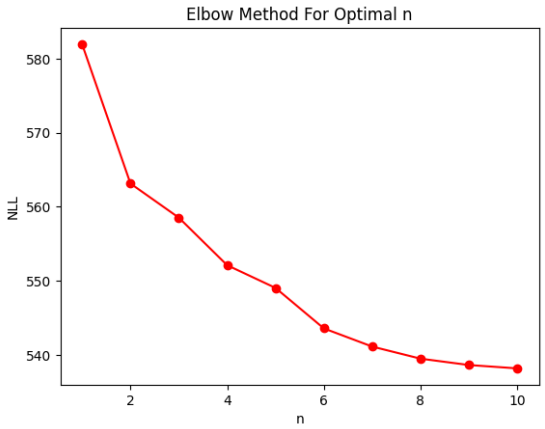
There was one hyperparameter to be optimized: perplexity, which can be interpreted as a smooth measure of the effective number of neighbours.

A list of perplexity (2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50) was used to conduct t-SNE. From the visualization of the resulting variables, perplexity=15 can provide the most identifiable structure for clustering.

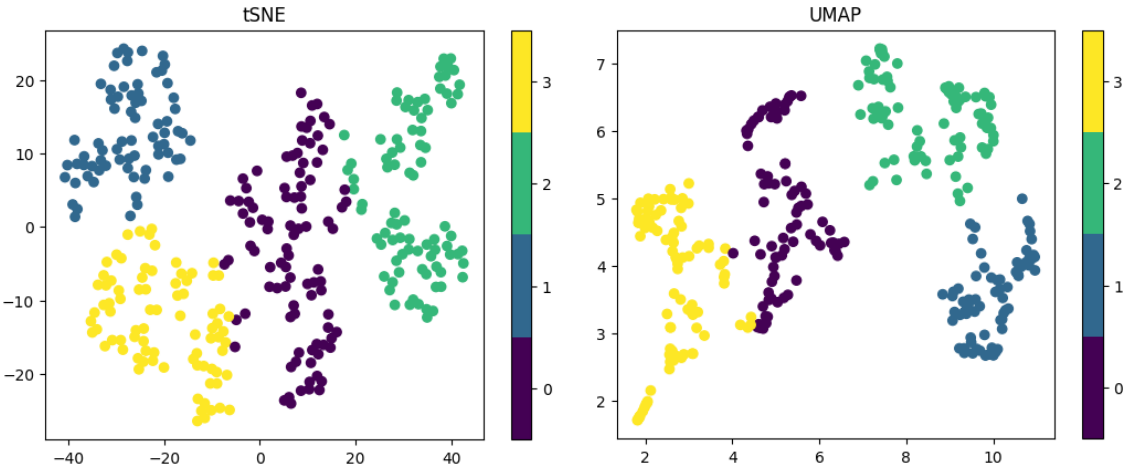


t-SNE, perplexity=15

Then number of clusters from 1 to 10 were tested to find the optimal number of clusters for Gaussian Mixture Model. From the plot of average total negative log likelihood from the validation sets (k-fold cross validation was used) at each number of clusters, it was observed that n=4 and 6 were the potential elbow points. Since 6 clusters might be too many for analysis, number of clusters=4 was selected for the Gaussian Mixture Model.



It was observed the resulting clustering from t-SNE method (left figure) looked similar to the clustering from UMAP method (right figure). A 180-degree rotation could make the positions of clusters from one method coincide with the position of clusters from another method.



t-SNE, perplexity=15, number of clusters=4

UMAP, neighbours=20, min_dist=0.0, number of clusters=4

UMAP vs. t-SNE

		t-SNE clusters			
		0	**1**	**2**	**3**
UMAP clusters	**0**	82	0	0	1
	1	1	71	0	0
	2	5	0	0	81
	3	5	0	85	0

From the table above, which shows the classification of 331 local authorities in clusters from UMAP and clusters from t-SNE, it was observed that the results from the two dimensionality reduction methods were nearly identical, as the highlighted cells represent 96.4% of all local authorities. Therefore, choosing either one of the dimensionality reduction methods would not affect the analysis result. From the perspective of visualization, t-SNE was preferred as the points are more spread out while the clustering structure is still preserved.

Characteristics of clusters

From the table below which shows the simple averages of the five relevant variables by clusters, it was observed that cluster 2 had the best performance except higher unemployment rate. Cluster 1 and 3 had very little relevant job opportunities for users. Cluster 1 had higher proportion of 25-34 population but the highest unemployment rate. Cluster 3 had low unemployment rate but low proportion of 25-34 population. Cluster 0 was the cluster with medium performance on all of these five variables.

t-SNE Cluster	Age 25-34 %	Never married & Age 25-34 %	Professional occupations %	Information and communication %	Unemployed %
0	12.3	10.5	20.1	4.1	4.1
1	13.5	11.8	15	2.8	5.3
2	15.1	13.4	25.9	7.5	4.9
3	10.7	9.1	16.1	2.7	3.9

Therefore, cluster 2 should be the target for the users to live in, and the characteristics of the clusters were summarized in the table below:

t-SNE Clusters	Cluster Label	Age 25-34 %	Never married & Age 25-34 %	Professional occupations %	Information and communication %	Unemployed %
0	Non-target 1	Medium	Medium	Medium	Medium	Low
1	Non-target 2	Higher	Higher	Low	Low	High
2	Target	High	High	High	High	Higher
3	Non-target 3	Low	Low	Low	Low	Low

Low Dimension Representation Variables

It was observed that the x-axis of the t-SNE scatterplot is related to relevant job opportunities, as higher % of resident working in professional occupations and information and communication industries would result in higher value in x-axis.

On the other hand, y-axis of the t-SNE scatter plot is related to age and unemployment, as higher proportion of population aged 25-34, higher proportion of unmarried population aged 25-34, and higher unemployment rate could result in higher value in y-axis.

4.2 Other visualization techniques

Tooltip: On the "Overview" and "Clustering" dashboards, a bar chart/table pop ups immediately when the cursor hovers above the local authority on the map. The bar chart/table contains distribution/values about the local authority to be compared with the bar chart on the same dashboard but at overall/cluster level. The aim of using tooltip is to provide a quick and intuitive way of comparison to lower the cost of memory, instead of requiring users to memorise figures across figures/dashboards for comparison.

Highlighting: On the "Overview" dashboards, hovering on the local authority on the map will highlight the corresponding local authority on the top/bottom bar chart (if that local authority is included in the bar chart), and vice versa. On the "clustering" dashboard hovering on the local authority on the map, or the point on the scatterplot will highlight the corresponding local authority on the map, on the scatterplot, and the corresponding cluster on the bar chart. The extensive use of highlighting is to guide user to know what figures/location to find corresponding information and figures to compare, which lower the cost of attention.

Filtering: The "Overview" dashboards provide an option for users to switch between %/absolute values. Also, after knowing which cluster is the target in the "Clustering" dashboards, users can go back to the "Overview" dashboards and select the target cluster in the cluster filter to narrow down the choices of local authorities to help their decision making. The use of filter in the dashboard is to control the number of dashboards and charts for easier navigation without losing information.

Top/bottom bar chart: The bar chart shows the top and bottom local authorities of the corresponding relevant variable, which save users' time, memory and attention to search for top local authority on the map.

Colour: The values on the map are encoded in diverging colourmaps with orange/red representing high. As high value is the most important for users to decide where to live in/avoid, using the most distinct colour to represent the value can draw user attention.

5. Conclusion

6.1 Socio-economic aspect

1. The local authorities in the target clusters are mostly located in London/South East regions, which consists of 55 out of 85 (65%) local authorities in the target clusters (London=29, South East=26). If East of England is also included, the proportion is increased to 66 out of 85 (78%). It could be concluded that top local authorities for the users were concentrated around London area.

2. It might not be worthwhile to live in local authorities in the target clusters that are far from London. These local authorities do not have many neighbouring local authorities in the same target cluster (the highest number is 4 in Greater Manchester), so users may not be benefited from relevant job opportunities and have friends of similar age range from neighbouring local authorities, while they can have these advantages in local authorities around London.

3. Low unemployment rate and higher proportion of population in similar age range should not be overly pursued. From one of the low dimension representation variables, it was discovered that unemployment rate was positively correlated with the proportion of population aged 25-34. Therefore, a balance between the two requirements should be achieved, as a local authority that did well in one aspect was likely to be awful in another aspect.

6.2 Information visualisation aspect

1. Data should be selected carefully. Originally health, education and living conditions of the residents were planned to be included in the analysis. However, it was then realized that including too many variables in the analysis would increase the elements to be visualized, thus diverting users' attention away from the core qualities to decide where to live in. As a result, only five tables were selected for analysis.

2. Data quality may be more important than analysis method. The data quality for this visualization task is good. The tables were all downloaded from ONS, so no missing value was found in the tables. Also, only the core variables that represent users' requirements were selected. Using this dataset as input, t-SNE and UMAP provided almost identical results. Therefore, as long as the data quality is ensured and the 'right' method is used, the effect of using which of the 'right 'methods may not be significant.

3. Tooltip and highlighting help guiding users where to pay attention. One of the main concerns for designing visualization is that users may omit some charts, so that that charts cannot fulfil their designed use. Tooltip and highlighting make the charts stand out when the occasions to use these charts come, thus ensuring the visualization serve the original designed purpose.

REFERENCES

(1) UK THE OFFICE FOR NATIONAL STATISTICS. SOC 2020 Volume 2: the coding index and coding rules and conventions [Online]. [viewed 18/05/23]. Available from: https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020/soc2020volume2codingrulesandconventions

(2) UK THE OFFICE FOR NATIONAL STATISTICS. UK SIC 2007 [Online]. [viewed 18/05/23]. Available from: https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007

(3) UK THE OFFICE FOR NATIONAL STATISTICS. Ward to Local Authority District to County to Region to Country (December 2021) Lookup in United Kingdom [Online]. [viewed 18/05/23]. Available from: https://geoportal.statistics.gov.uk/datasets/ons::ward-to-local-authority-district-to-county-to-region-to-country-december-2021-lookup-in-united-kingdom/about
(4) UK THE OFFICE FOR NATIONAL STATISTICS. Local Authority Districts (May 2022) UK BFE V3 [Online]. [viewed 18/05/23]. Available from: https://geoportal.statistics.gov.uk/datasets/196d1a072aaa4882a50be333679d4f63_0/explore?location=55.215451%2C-3.313875%2C6.67