Exploring Socio-Economic Trends in England and Wales: A Visual Analytics Approach

EMATM0065 - Visual Analytics Coursework (Spring 2025)

Student Name: Thrisha Rajkumar

Student Number: 2623567

1. Abstract

This coursework project explores how socio-economic trends across England and Wales between 2011 and 2021 were used in a visual analytical approach. The main focus and objective were to compare the 2011 and 2021 census data on topics such as education, employment, immigrants, economic activities, and occupational opportunities, correlating these vast features for target users such as international students, policy advisors, and education planners. The visual analytics dashboards provide an intuitive tool to assess regional opportunities and qualification-to-employment pathways, especially for non-UK-born residents in each Local Authority District (LAD), and to compare between the years 2011 and 2021.

The data preparation involved several preprocessing steps, LAD name and code mapping for 348 recognised LAD names extracted from the Tableau geocoding shapefile, and the use of Bayesian Ridge Regression to impute missing LAD education, occupation, and economic values in the 2021 census, ensuring complete comparability across both years with clean and complete data. The dimensionality reduction techniques—UMAP and t-SNE—helped to find the 2-dimensional reduction from 28 features, and the patterns were observed clearly on the choropleth maps for year filtering 2011 and 2021, with comparison to binned values as high, low, and mid for both UMAP and t-SNE dimensionally reduced values. The visualisation followed methods and approaches by Munzner's Chapter 4 taxonomy and visual evaluation.

Visualised through Tableau dashboards, the project was effective in identifying the trends, such as the national improvement in Level 4+ education and regional employment shifts, and high immigration in districts like Westminster and the City of London, with a high percentage of Level 4 and high-skilled occupational opportunities. By blending spatial analysis with calculated fields and interactive design using the tooltips and filtering, the dashboards help the users to explore meaningful insights and answer core questions about educational return, employment and occupational opportunities post-education, and regional development patterns over the decade.

2. Introduction

The main objective is to explore the Socio-Economic trends in England and Wales in 2011 vs 2021, targeting users like International students and graduates, University career services, Migration Policy Advisors & Analysts, Graduate Programme Designers at universities and Higher Education Strategy Planners. In recent years, the UK has been among the top popular places to pursue higher education degrees like postgraduate study and research. International students wanting to study abroad have many questions about the uncertainties to decide whether the investment is justified, particularly regarding the employment outcomes, future benefits and job opportunities. A recurring question and concern is:

"Will studying in the UK lead to employment opportunities?"

To help understand how different types of education qualifications, immigration from different continents and ethnicity result in occupation and employment across the Local area districts in England and Wales. We used 2011 census data and comparative insights from 2021 census data, the project identifies regional socio-economic patterns and highlights the intersection between **qualification levels (Bachelor's, Master's, PhD)** and **employment sectors** (e.g., professional occupation, Associates and technical, Managers and Senior Officers etc), particularly for **non-UK-born populations**.

Problem Statement and Motivation

For international students and graduates, many of whom invest substantial personal and financial capital, the post-study employment is often uncertain. Despite the interest of international students to pursue their dreams, there is limited data visibility which truly reflects the situation, which aligns with the actual job opportunities and placements.

Target Users and End-User Needs

- **International students and recent graduates** evaluating the return on educational investment and identifying regions of opportunity.
- **University career services** aim to support students' transition into the UK employment sector.
- **Migration policy analysts and advisors**
- **Graduate programme and higher education planners** – to help the students have a clear idea of the opportunities to explore the education in the higher qualification degrees that the universities are offering.

This project examines socio-economic trends from multiple angles, contrasting traditional census data with dimensionality-reduced visualisations to uncover complex relationships between features. The dashboard addresses five basic questions specifically, each tackled through distinct visual modules:

Key Questions Addressed by the Dashboards are:

1. How have education qualification levels (e.g., Level 4+) changed across regions between 2011 and 2021 in comparison to Employment opportunities and Occupation breakdown?
2. What are the trends in employment types (full-time, part-time, unemployed) from 2011 to 2021 in each Local Area District (LAD)?
3. The percentage of high-skilled occupations with Level 4 and above education in the top 15 highest immigrant population from a particular ethnicity (Continent)?
4. Where are immigrant populations most concentrated, and how does that relate to high-skill employment and education levels?
5. Can UMAP and t-SNE dimensionality reduction techniques clustering reveal regional groupings with similar socio-economic characteristics?
6. Which Local Authority Districts retain the highest percentage of skilled, foreign-born graduates?
7. Changes in the percentages of professional occupation and the Level 4 and above education increase in 2011 vs 2021 for each LAD across England and Wales

Visualisation Motivation

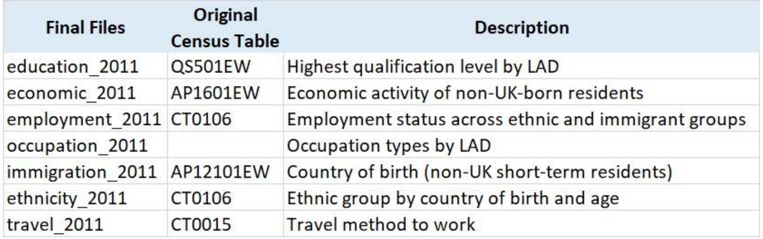
The aim is not only to focus on the above questions but to give the target users, particularly the international students, a visual and intuitive tool to analyse their chances and potential when moving abroad for their career aspirations, especially for the non-UK resident population. Comparison between the changes in 2011 vs 2021 to observe the trends over the decade.

The uses of the Dimensionality reduction techniques like UMAP and t-SNE on the education, occupation and employment target features only in both 2011 and 2021 provide additional insights into latent socio-economic clusters, offering a unique perspective not typically visible through raw statistical charts. This project's Tableau visualisation empowers users to make informed decisions and plan educational strategies.

3. Data Preparation and Abstraction ***(1.5–2 pages)***

Data Extraction:

This data preparation and abstraction is the most important part for loading the clean data into Tableau for visualisation. The original datasets downloaded from the “All Tables Excel file” from the Nomis census data are:



Final Files	Original Census Table	Description
education_2011	QS501EW	Highest qualification level by LAD
economic_2011	AP1601EW	Economic activity of non-UK-born residents
employment_2011	CT0106	Employment status across ethnic and immigrant groups
occupation_2011		Occupation types by LAD
immigration_2011	AP12101EW	Country of birth (non-UK short-term residents)
ethnicity_2011	CT0106	Ethnic group by country of birth and age
travel_2011	CT0015	Travel method to work

With over **1600 tables** available for 2011, we selected **7 specific tables** which was mostly relevant for the questions we are focusing on. Downloaded all the tables as CSV files for checking the calculations in the code easily. Also chose the raw numerical format and not as precomputed percentages or mixed textual formats to allow for **greater flexibility in calculations, standardisation, and ratio-based interpretation** during analysis.

The 7 tables of education, economic, employment, occupation, immigration, ethnicity and travel CSV files for the UK population at the Local Authority District (LAD) level give a wide range of data to work with, readily available and cleaned data. Working with raw numeric data allowed us to **normalise, merge, and project** the datasets easily and consistently, which was essential for further dimensionality reduction and comparison across 2011 and 2021.

Similarly, we extracted the data for education, economic, and occupation from the 2021 census in England and Wales, by Local Authority Districts (LAD), as CSV files in raw numeric format. With a total of 10 tables, we proceeded with the preprocessing of the data.

Data Cleaning and Preprocessing:

Firstly, the header and footer were removed manually, as they contained metadata about the table, and only the column headers and data were retained. After this step, the files - education, economic, employment, immigration, occupation, and travel census 2011 data had three separate columns for the regions' names; therefore, the columns were merged to have a single column, and the header was named **LAD Name**. All the files were saved in a **Clean_Files_2011** folder with the prefix **cleaned_(filename). csv**.

Whereas the ethnicity file did not have **LAD Names**, it had the **ethnic group** as a column for each country. A new column for ethnic group was created, only country-wise rows were retained, and only the necessary columns were used. The file was saved as **ethnicity_2011_structured.csv**, as it does not have the LAD Name or code as a common header, which we are focusing on.

The number of rows containing regions and LAD Name in all the 2011 census data was 395, with no null values or rows. For 2021, we used education, occupation and economic files for comparison with 2011. All the 2021 census data had no null values present for the 331 rows of data, and no null rows were encountered when extracted.

LAD Code Extraction and Matching for Tableau Mapping

Matching the LAD Names and Codes to load the data and get full visualisation was a challenge. Firstly, opened the **FullMapData.hyper** file provided in Tableau, selected the LAD Code, and then extracted it through the Data tab as a CSV file. This CSV file had 182,549 rows of data with the England and Wales regional, district, and area code and name data. As we are working with the LAD Code, we only extracted the unique LAD Codes and Names, which resulted in **348 LAD rows** that contained the LAD Code along with the LAD Name, in the format recognised by Tableau.

This was a crucial step for the complete England and Wales map visualisation process. Now we used this FullMapData.hyper extracted file and saved the 348 LAD Names into a separate CSV file named – Full_Data_Unique_LAD.csv, this file will be used as a master reference for matching the LAD Names and codes before loading the data to Tableau.

Added a function named align_to_reference_clean () for LAD Names matching in 2011 and 2021 data. which renames the column headers "LAD Name" and "LAD Code" from "Geography" and "Geography code", which also drops the unnecessary columns if encountered, and adds the numeric columns after cleaning carefully and merging the data with the LAD Name and LAD Code as the master column primary keys.

Firstly, we applied this to the **2011 data** as it had a larger number of rows, so easier to clean and remove the unrecognised and unwanted rows by merging with the aggregated new LAD Names if necessary. We are not cleaning the ethnicity file here as it does not have the headers as LAD Name or code rather has the ethnic groups in each country. Applied the function on the clean_filename.csv files by referencing the master columns. And saving it to the Final_data column with the prefixes "final_(filename) _ 2011.csv. Final check to ensure there are no missing rows or values in the 2011 files, and ensured there are none. Check manually the sum of each of the files to ensure all the files have 348 LAD Names and LAD code.

Now we proceeded to the cleaning of **2021 data files**, which was quite challenging as there were only 331 rows and we had to map to 349 rows. For the 3 tables of education, economic and occupation, we will have to match and update the names and area codes.

First, we tried cleaning manually by mapping the unrecognised LAD Codes to the LAD Names in the education_2021 file by first renaming the main columns' headers and the metadata and retaining the year column and filtering to LAD Names present in the master reference, which led to **31 unrecognised LAD Names**. Filtered by finding the missing LAD references and manually mapped from the ONS changes and LAD merges from the master file. After filling in the rows and combining the cleaned file. The final 2021 file had 348 LAD Names. But without Bayesian, and using the concatenated filled value, gave a sum of 102,859,518 when the actual data sum was 97,132,682.

Bayesian Ridge Regression

Therefore, instead of calculating the mean or aggregate sum for the 31 missing LAD Names, we use a Bayesian model to predict the values. Now cleaning the 2021 education, economic and occupation files using the Bayesian Ridge regression modelling to predict the values. Filtering the valid LAD Names again and checking the missing LAD Names using the reference LAD Names.

Bayesian Ridge Regression was used to predict and fill in missing values for 31 Local Authority Districts (LADs) within the 2021 education dataset. The LADs were entirely missing in the original dataset, and thus, a statistical model was needed to predict their likely values based on trends in the data that were available.

Firstly, compare the list of expected LADs (from the master file Full_Data_Unique_LAD.csv, which contains 348 LADs) to the LADs present in education_2021.csv. 31 LADs were found to be completely missing in the 2021 file.

From the rows that were available, a training dataset was created. the X_train is the training dataset, which contains all columns (except the 31 LADs) and y_train = the values of the column being predicted. Since Bayesian Ridge Regression is sensitive to the scale of the input, StandardScaler was used to normalise the training data.

For each column, a Bayesian Ridge Regression was trained on the X_train and y_train. This model learns a probabilistic relationship between all the known LADs and the column to be predicted. The model predicted the 31 missing LADs by repeating the average values of the scaled features for all 31 LADs (as their exact features are unknown) and making mean predictions with standard deviations. The code subsequently computed 95% credible intervals for all predictions with lower-bound and upper-bound formulas.

Here we purposely scaled the predictions to match the total original data sum of 97,132,682 by using a simple calculation of the Scaling Factor, which is the difference between the sum of the total target value and the existing LAD Total divided by the total Imputed LAD value. This acts as a boundary to set the values bounded or scaled to the Actual population of the original data to avoid errors and miscalculations and present true values. This step was essential to predict by aligning within the boundary or target of the final dataset with the original 2021 sum (97,132,682).

The final_df dataframe had all the LAD known values with no null values or rows and was saved to the final_education_2021.csv file. The same steps were followed for fitting the Bayesian Ridge regression model to the 2021 data on economic and occupation. Similarly, the occupation data's original target value of 55,547,384 and the economic data’s target value of 201,246,425 were also predicted using the model and saved as final_occupation_2021.csv and final_economic_2021.csv. And manually checked the total sum of all the data files, and the Bayesian Ridge Regression well predicted all the data within the target boundary. Also, a few manual checks for the LAD Names to match the reference LAD Name.

Now we have all the 348 LAD Names and Code data for 2011 and 2021 with no null values or empty rows or columns.

Merging the Files for the Tableau Visualisation:

Loading all the final 2011 and 2021 cleaned files to the df_2011 and df_2021 dataframes and merging on the LAD Name and LAD Code as keys. First, individually merging the 2011 files on education, economic, employment, occupation, immigration and travel to the “merged_all_2011.csv” file and the 2021 final files – education, economic and occupation as “merged_all_2021.csv” file.

Now merging the files 2011 and 2021 together, adding a column “Years” in the 2011 and 2021 files and creating a union for easy data import in Tableau for visualisation and using filtering for each year.

Renaming the 2021 column names for education, economic and occupation manually for easy merging, as the 2011 merged file contains 6 tables of data, and we are merging only 3 from 2021. All columns present in the 2011 data but not present in the 2021 files were filled with 0 in the df_2021_aligned dataframe to ensure a consistent and clear merge. In df_2021, after renaming and creating a new dataframe df_2021_aligned to have only the renamed columns with the LAD Name, LAD Code and Year as the key columns to merge the values clearly to the respective column headers in the df_2011 dataframe. Now the df_final_correct dataframe contains the concatenated df_2011 and df_2021_aligned data, which is saved to the final_merge-2011_2021.csv file.

Dimensionality Reduction Techniques to visualise patterns or clusters in the data

UMAP Clustering (2011 vs 2021)

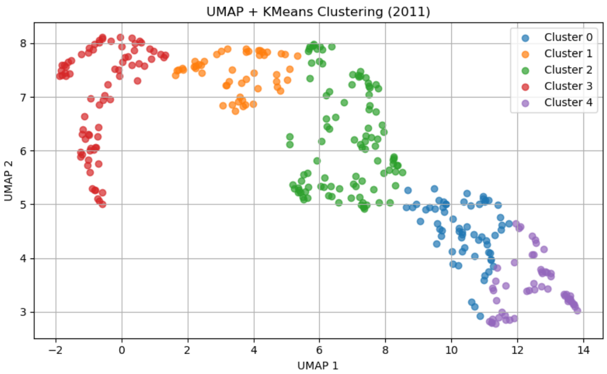
To understand the relationships between education, economic activity, and occupational outcomes across regions in England and Wales, we applied the dimensionality reduction technique - **UMAP (Uniform Manifold Approximation and Projection),** followed by **K-Means clustering** to differentiate the regions with similar relationships. For the dimensionality reduction, we used only the columns which were common in the 2011 and 2021 datasets for education, occupation and economic data. The selected 28 column headers treated as features have a wide range of variables covering the qualification levels, full/part-time work, unemployment, and detailed occupational breakdowns.

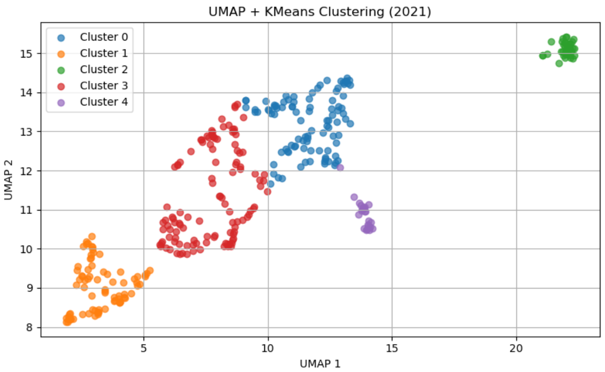
The purpose of using UMAP was to **reduce high-dimensional socio-economic data** into a two-dimensional space while preserving the local structure and proximity between LADs (Local Authority Districts). UMAP helps capture **nonlinear manifolds** that traditional methods like PCA may not handle well, which is essential for identifying hidden clusters based on human and socio-economic behaviour.

The selected **28 features** are related to qualification levels, employment status, like in employment, unemployed, self-employed, etc, and occupational types, which were cleaned, normalised using StandardScaler(), and then reduced using umap.UMAP().

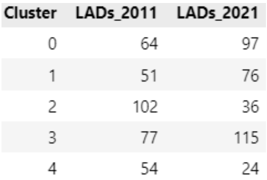
Once reduced to two latent dimensions (UMAP_1, UMAP_2), we applied **K-Means clustering** to group LADs into **five regional clusters** to observe the similar socio-economic patterns. The number of clusters (n=5) was chosen based on visual separability in the UMAP scatter plot and iteratively adjusted until meaningful interpretation was possible.

UMAP is used instead of PCA and other dimensionality techniques because it preserves both **local and global structure** and can extract the non-linear structures and **K-Means** was used for its simplicity and interpretability when grouping similar LADs in the latent space, which helps to visualise **socio-economic**





groupings across England and Wales. Below are a few observations from the clustering using the scatter plot with the K-means Clustering of the 2-D UMAP_1 and UMAP_2:



Cluster	LADs_2011	LADs_2021
0	64	97
1	51	76
2	102	36
3	77	115
4	54	24

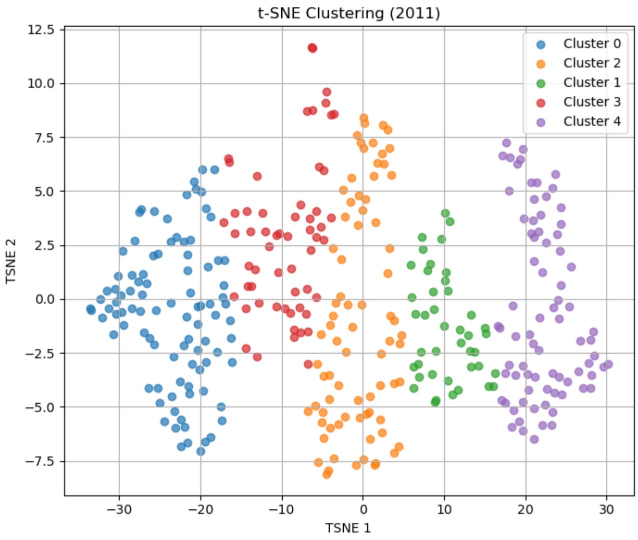
In 2011, Cluster 2 had the highest number of LADs of 102, which shows that the mid-range socio-economic balance, while in Cluster 4, there were the fewest LADs of 54, likely representing outlier LADs like coastal or rural districts. In 2021, Cluster 3 had the highest LAD count, 115, which may be due to the regional shifts in employment and education since 2011.

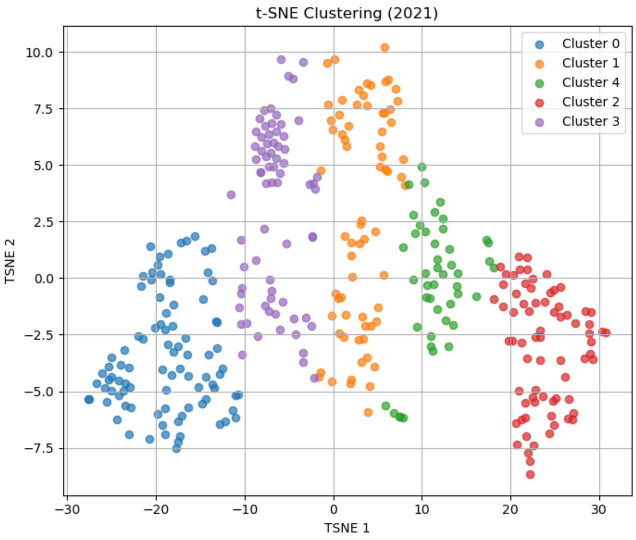
Also computed the **cluster-wise average values** for all 28 indicators and saved them to cluster_feature_averages_2011.csv and cluster_feature_averages_2021.csv files. These tables were crucial to understand and observe what each cluster represents (Cluster 0 had the highest % of full-time professionals and level 4+ qualifications) and to compare the socio-economic changes in the years 2011 and 2021.

Clustering with t-SNE and K-Means (2011 & 2021)

To help compare the UMAP analysis, we also implemented t-distributed Stochastic Neighbour Embedding (t-SNE), and similarly used the K-Means clustering with 5 groups on the education, economic activity, and occupation data across all Local Authority Districts (LADs) in England and Wales for both 2011 and 2021. This dimensionality reduction technique is especially well-suited to preserving local structure, which helps in identifying clusters of LADs with similar socio-economic characteristics.

Similar to the UMAP we used the 28 columns headers as the input features and standardised using StandardScaler() to ensure equal weighting before dimensionality reduction and plotted as a scatter plot to see the groupings.





In 2011, the clustering was more spread out and smooth, with moderately distinct groups. This suggests relatively less regional polarisation across socio-economic variables at that time. Whereas in 2021, closer structures and cluster 2 and cluster 4 have replaced their places in the plotting, implying sharper regional divergence over the decade, likely driven by economic shifts post-Brexit and during the pandemic.

Unlike PCA or UMAP, t-SNE preserves the local distances, making it ideal for detecting fine-grained socio-economic similarities in the education, occupation and economic patterns among LADs. It allows us to identify clusters of LADs with similar patterns and groupings, and detect outlier LADs possibly those undergoing rapid socio-economic change and compare clustering behaviour over the years.

We also saved these t-SNE cluster coordinates and assignments to tsne_kmeans_2011.csv and tsne_kmeans_2021.csv, then merged them with UMAP results using LAD_Code, LAD_Name, and Year to form final_cluster_data_with_year_column.csv.

The full final data with the socio-economic variables for both years with UMAP and t-SNE were merged with the key columns LAD Code, LAD Name and Year. The final_cluster_data_with_year_column.csv and the final_merge-2011_2021.csv were merged to have the final file – FINAL_2011_2021_MERGED.csv file.

The FINAL_2011_2021_MERGED.csv file contains a total of 137 columns and 696 rows (348 LADs for each year 2011 and 2021)

4. Task Definition (Munzner's task taxonomy)

Using Munzner's task taxonomy framework, the visualisations are built to support both high-level exploratory tasks and targeted analytical tasks for different types of users: international students, graduate planners, career services, and policy analysts. Muzner's point of view explains that the **What–Why–How** model's output is used as an input to the next question, showing chained sequence dependencies, which is highly important.

We applied the **What–Why–How** breakdown as suggested in Munzner's model:
Why:

This question is important as it is the basis of why this visualisation is necessary and why it is a need, and why the user needs the visualisation tool.

- Discover: User wants to explore the job opportunities and the changes in the Local Area Districts, and get an idea of whether I can settle after a high-level qualification attained in the UK. To explore latent clusters and socio-economic groupings (using the dimensionality reduction techniques UMAP and t-SNE).
- Present: Support decisions and analyse which occupation is most benefited after a high-level degree in each LAD and the changes in the opportunities in correlation to qualifications, employment, and occupations in a decade (2011 vs 2021)
- Identify: Highlight high-performing LADs in terms of skilled immigrant employment.
- Compare: 2011 vs 2021 education, employment and occupation across LADs and clusters.

What:

Based on Munzner's task taxonomy, the following primitive tasks are implemented across the Tableau visualisations and Python-generated cluster outputs. The data user visually receives is the socio-economic trends in education levels, occupations and employment (Also provides the non-UK residents socio-economic changes).

- Derived Data Cluster groups (0–4), % Level 4+ education, % full-time employed, % in professional jobs, change between years.
- Attributes LAD Name, LAD Code, Year, Continent of Immigrant Origin.
- Quantities Count and percentage measures from census datasets (e.g., 34.1% full-time employment).

How (Visual Encoding & Interaction):

This question is about visual encoding and interaction, how the data is presented, answering "Why" and "What".

- Tableau Dashboards: Feature interactive filters, hover tooltips, and toggle by year (2011/2021) and cluster.
- UMAP & t-SNE Visualisations: Colour-encoded scatter plots by cluster, dimensionally reduced from 28 socio-economic indicators in the code, and the clustering visually using choropleth maps in Tableau.

- Final Dataset: All LADs and features combined into a unified file (Final_2011_2021_merged.csv) with 696 rows and 137 columns, allowing for scalable and comparative analysis.
- Dynamic Filtering: Enables year-wise insights and cross-feature observations between education, immigration, and employment type.

5. Visualisation Justification

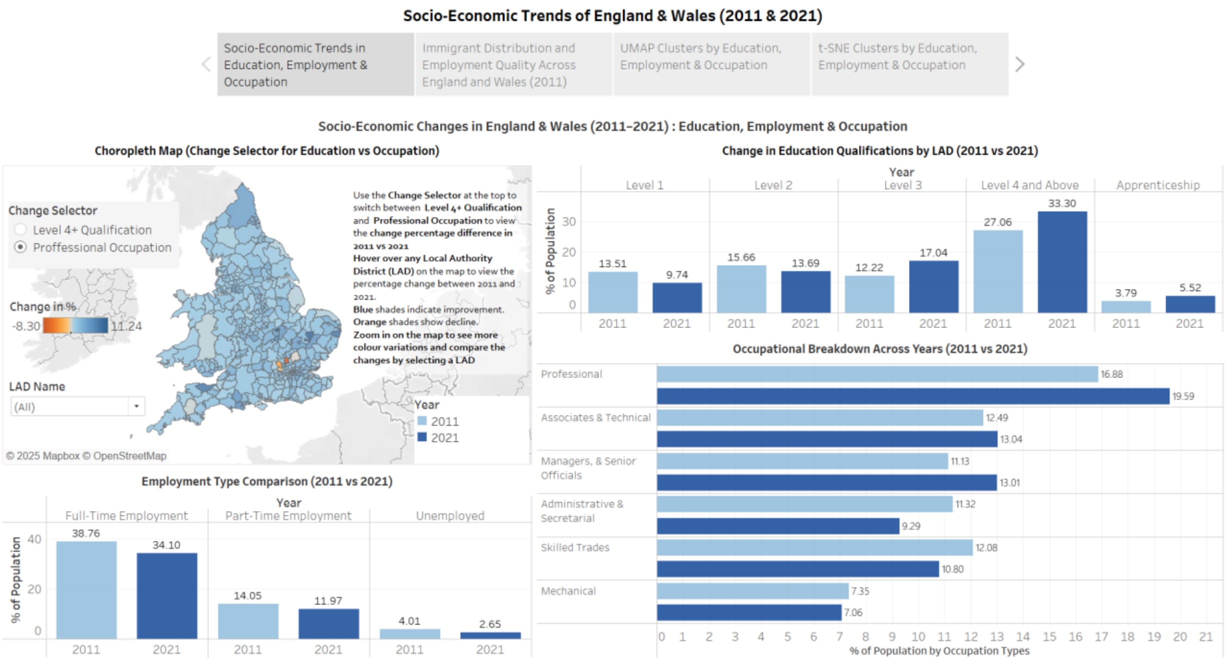
We have visualised 4 dashboards, each answering part of the story we set out to explore. The final story narrative brings together spatial, demographic, and clustering perspectives:

- Dashboard 1: Socioeconomic changes across 2011 and 2021, covering education, employment and occupation.
- Dashboard 2: Immigrant population distribution and employment quality.
- Dashboard 3: UMAP clustering of LADs.
- Dashboard 4: t-SNE clustering showing embedded similarities between LADs.

5.1 Visualisation and Modelling Approaches

5.1.1 Dashboard 1: Socio-economic Change Across 2011 and 2021 on occupations, education and employment.

This dashboard visualises using a choropleth map and grouped bar charts to show how education, employment, and occupation changed across LADs between 2011 vs 2021. We chose these visual forms based on the type of data and the spatial tasks we needed to support.



Use the **Change Selector** at the top to switch between **Level 4+ Qualification** and **Professional Occupation** to view the **change percentage difference in 2011 vs 2021**
Hover over any Local Authority District (LAD) on the map to view the percentage change between 2011 and 2021.
Blue shades indicate improvement.
Orange shades show decline.
Zoom in on the map to see more colour variations and compare the changes by selecting a LAD

Socio-Economic Trends of England & Wales (2011 & 2021)
Socio-Economic Trends in Education, Employment & Occupation
Immigrant Distribution and Employment Quality Across England and Wales (2011)
UMAP Clusters by Education, Employment & Occupation
t-SNE Clusters by Education, Employment & Occupation
Socio-Economic Changes in England & Wales (2011–2021) : Education, Employment & Occupation
Choropleth Map (Change Selector for Education vs Occupation)
Change Selector
○ Level 4+ Qualification
● Proffessional Occupation
Change in %
-8.30
11.24
LAD Name
(All)
© 2025 Mapbox © OpenStreetMap
Year
2011
2021
Change in Education Qualifications by LAD (2011 vs 2021)
Year
Level 1
Level 2
Level 3
Level 4 and Above
Apprenticeship
% of Population
0
10
20
30
13.51
9.74
15.66
13.69
12.22
17.04
27.06
33.30
3.79
5.52
2011
2021
Occupational Breakdown Across Years (2011 vs 2021)
Professional
16.88
19.59
Associates & Technical
12.49
13.04
Managers, & Senior Officials
11.13
13.01
Administrative & Secretarial
11.32
9.29
Skilled Trades
12.08
10.80
Mechanical
7.35
7.06
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
% of Population by Occupation Types
Employment Type Comparison (2011 vs 2021)
Year
Full-Time Employment
Part-Time Employment
Unemployed
% of Population
0
20
40
38.76
34.10
14.05
11.97
4.01
2.65
2011
2021

- Data Abstraction:

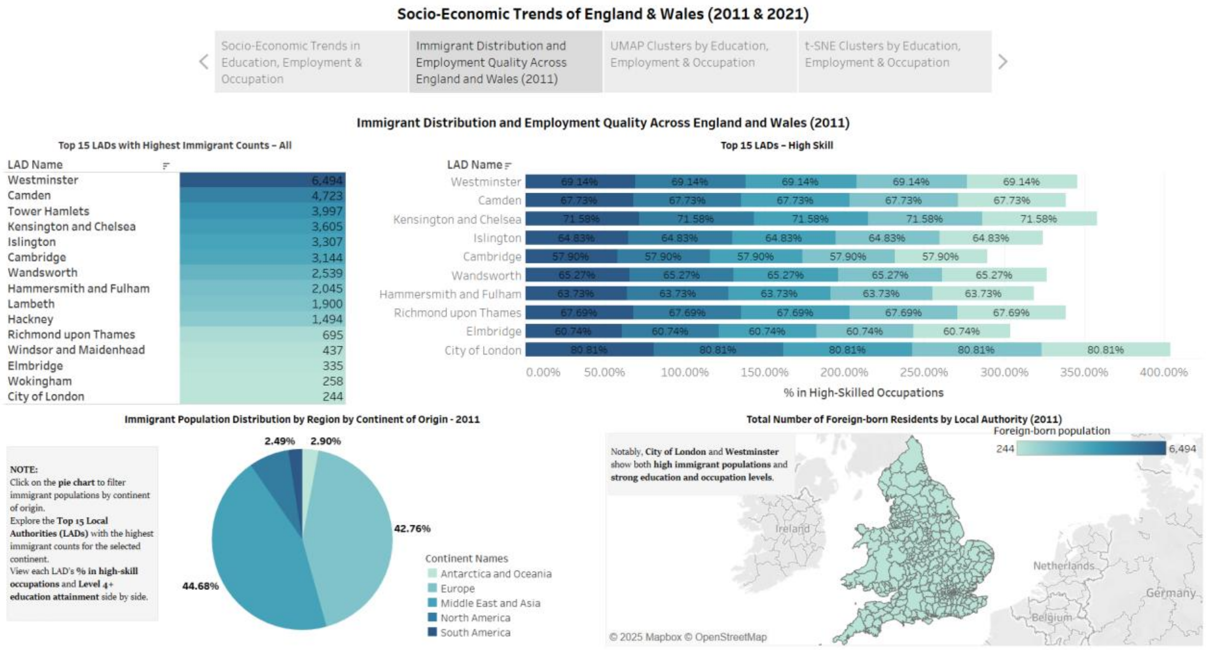
  To enable comparisons, we created many calculated fields for representing the change in higher education between 2011 and 2021 in percentages instead of raw numbers showing the population. The calculated fields are for each Local area district and by year filtering, which helps to reduce noise and prioritise useful comparisons instead of raw numeric values. Several calculated fields for percentages for Professional occupations, Associates and Technical, Managers and senior officers, administrative and Secretarial, Skilled trades and mechanical occupations for occupations, and In employment, part time and unemployment for the employability factors analysis and for qualifications – Level 1, Level 2, Level 3, Level 4 and above and Apprenticeship percentages calculations.

- Task Abstraction (Munzner):
  The main user tasks are:
  - Locate: Finding the LADs of interest on the map
  - Identify: look for the LADs which has positive changes and an increase in the percentages for the socio-economic factors which they are looking for.
  - Compare: 2011 vs 2021
  - Summarise: Overall changes at the regional or national scale
    These are well supported through spatial encodings and grouped bars with year-wise splits.
- Visual Encoding:
  The choropleth map uses hue and saturation to encode magnitude and direction of change. This supports pop-out effects and quick comparisons. The choropleth map was used not only to act as a filter but also to show the LADs changes in professional occupations and Level 4 and above qualifications with the use of the **Change selector Parameter**.

  We extensively used the bar charts to visualise the occupations, education and employment changes in each lad and in each year. Ensured colour scales are consistent, readable, and colour-blind friendly.
- Justification for Layout:
  The dashboard follows the "overview first, zoom and filter, details on demand" principle. Users first see overall patterns on the map, then can interactively filter to specific LADs and explore more with the use of the tooltips or bar breakdowns.
- Interaction & Modelling:
  Filters help break down by LAD and allow deeper insights into how specific categories changed over time. This directly improves user engagement and supports both lookup and explore actions. The tooltip has very well-structured information about the visuals to ensure the data reaches the audience or target user promptly. Also, the notes with instructions and details of the dashboard data help to better understand.

5.1.2. Dashboard 2:

In Dashboard 2, the focus was to see the economic, occupation, and education trends for the immigrant population and non-UK residents by continent. We used a bar chart, a highlight table chart, a choropleth map, and a pie chart to represent different facets of immigrant data in 2011. The highlight table chart was used for its accuracy in comparing LADs by immigrant counts for the TOP 15 LADs with the highest immigrant counts and the percentage in high-skilled jobs. The horizontal bar chart shows the TOP 15 LADs immigration population with a corresponding comparison of Level 4 and above qualifications in the tooltip, with the percentage calculated field for high-skilled occupations—professional, managers and senior officers, and associate and technical as aggregated percentage value.



Socio-Economic Trends of England & Wales (2011 & 2021)

Socio-Economic Trends in Education, Employment & Occupation
Immigrant Distribution and Employment Quality Across England and Wales (2011)
UMAP Clusters by Education, Employment & Occupation
t-SNE Clusters by Education, Employment & Occupation

Immigrant Distribution and Employment Quality Across England and Wales (2011)

Top 15 LADs with Highest Immigrant Counts – All

LAD Name	
Westminster	6,494
Camden	4,723
Tower Hamlets	3,997
Kensington and Chelsea	3,605
Islington	3,307
Cambridge	3,144
Wandsworth	2,539
Hammersmith and Fulham	2,045
Lambeth	1,900
Hackney	1,494
Richmond upon Thames	695
Windsor and Maidenhead	437
Elmbridge	335
Wokingham	258
City of London	244

Top 15 LADs – High Skill

Immigrant Population Distribution by Region by Continent of Origin - 2011

NOTE:
Click on the **pie chart** to filter immigrant populations by continent of origin.
Explore the **Top 15 Local Authorities (LADs)** with the highest immigrant counts for the selected continent.
View each LAD's **% in high-skill occupations** and **Level 4+ education attainment** side by side.

Total Number of Foreign-born Residents by Local Authority (2011)

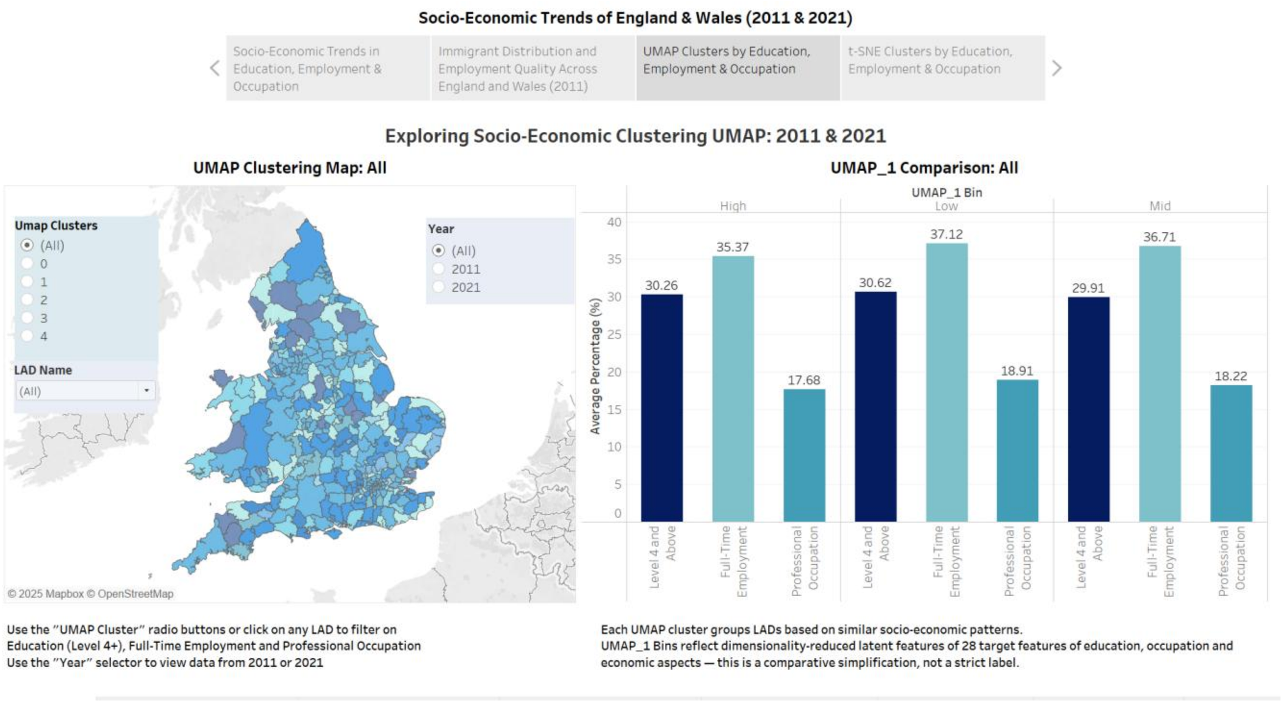
Notably, **City of London** and **Westminster** show both **high immigrant populations** and **strong education and occupation levels.**

A side-by-side horizontal bar chart was placed alongside the highlighted table chart, which is used instead of stacked bars to produce a cleaner and more straightforward comparison of education and occupational levels. This facilitates and supports absolute judgment more efficiently on the LAD wise.

The choropleth map illustrates the spatial distribution of the immigrants using the sequential scale colouring (bluish hues), indicating the population intensity per continent to promote spatial thinking and exploration. The interactive LAD or origin continent filters enable the user to zoom in, filter, and get on-demand information. The main pie chart shows the calculated percentages of the continents, and then becomes the master filter when clicking on a continent, all the charts in the dashboard are updated to the respective continent.

5.1.2. Dashboard 3 & 4:

Comparing both the dashboards, which visualise the UMAP and t-SNE dimensionality reduction techniques to visually interpret the clustered groups of the LADs. Both the dashboards used choropleth maps to show the clustering using k-means by dimensionality reduction of the UMAP and t-SNE clustering outcomes across England and Wales. The use of a choropleth map over the scatterplot is to show the clustering and help the target user interpret the areas more easily and see the trends and changes in each cluster and each LAD in that cluster with the help of a tooltip. Each cluster groups Local Authority Districts (LADs) with similar socio-economic profiles based on features like education level, full-time employment, and professional occupation share. The clusters also have colour coding with the shades of blue to support the visual pattern discovery and geographic comparisons. UMAP and t-SNE were utilised for their ability to reveal structure in high-dimensional socio-economic information. These methods reduce 28 indicators into two latent dimensions, which are then discretised into clusters and space visualised.

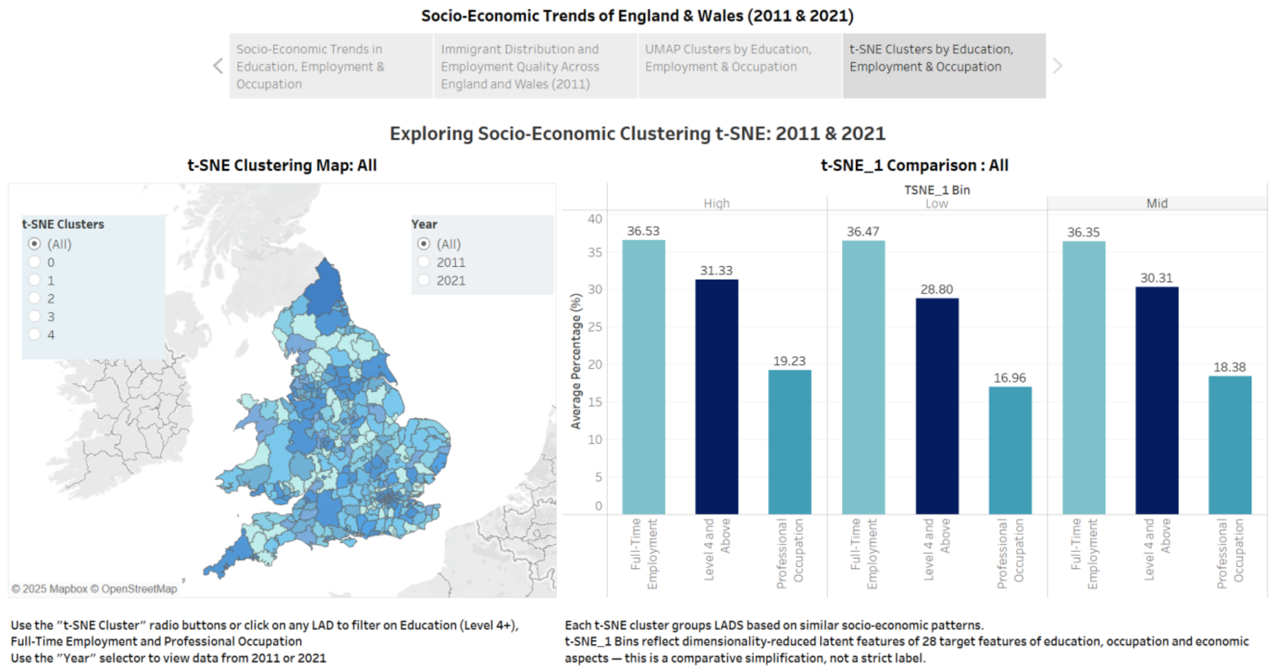

Socio-Economic Trends of England & Wales (2011 & 2021)
Socio-Economic Trends in Education, Employment & Occupation
Immigrant Distribution and Employment Quality Across England and Wales (2011)
UMAP Clusters by Education, Employment & Occupation
t-SNE Clusters by Education, Employment & Occupation
Exploring Socio-Economic Clustering UMAP: 2011 & 2021
UMAP Clustering Map: All
Umap Clusters
● (All)
○ 0
○ 1
○ 2
○ 3
○ 4
LAD Name
(All)
Year
● (All)
○ 2011
○ 2021
© 2025 Mapbox © OpenStreetMap
UMAP_1 Comparison: All
UMAP_1 Bin
High
Low
Mid
Average Percentage (%)
40
35
30
25
20
15
10
5
0
30.26
35.37
17.68
30.62
37.12
18.91
29.91
36.71
18.22
Level 4 and Above
Full-Time Employment
Professional Occupation

Use the "UMAP Cluster" radio buttons or click on any LAD to filter on Education (Level 4+), Full-Time Employment and Professional Occupation
Use the "Year" selector to view data from 2011 or 2021

Each UMAP cluster groups LADs based on similar socio-economic patterns.
UMAP_1 Bins reflect dimensionality-reduced latent features of 28 target features of education, occupation and economic aspects — this is a comparative simplification, not a strict label.

Bar charts are used to compare the UMAP dimension label calculated field with High, Low, and Mid binning for the Level 4 and above qualifications, full-time employment and professional occupation. Similarly, for the t-SNE dimensionally reduced latent feature from the 28 column headers as features. Also, the filters for clusters help to visualise each cluster's UMAP or t-SNE comparison and binning. also tooltip to each tad in the cluster will highlight each LADs percentage data.

The visual idioms follow Munzner's design principles: maps for spatial structures, bar charts for comparing quantitative measures, and radio selectors for switching views. Colour hue is used to denote clusters (categorical), and length/position to compare, achieving maximum perceptual accuracy.


Socio-Economic Trends of England & Wales (2011 & 2021)
Socio-Economic Trends in Education, Employment & Occupation
Immigrant Distribution and Employment Quality Across England and Wales (2011)
UMAP Clusters by Education, Employment & Occupation
t-SNE Clusters by Education, Employment & Occupation
Exploring Socio-Economic Clustering t-SNE: 2011 & 2021
t-SNE Clustering Map: All
t-SNE Clusters
● (All)
○ 0
○ 1
○ 2
○ 3
○ 4
Year
● (All)
○ 2011
○ 2021
© 2025 Mapbox © OpenStreetMap
t-SNE_1 Comparison : All
TSNE_1 Bin
High
Low
Mid
Average Percentage (%)
40
35
30
25
20
15
10
5
0
36.53
31.33
19.23
36.47
28.80
16.96
36.35
30.31
18.38
Full-Time Employment
Level 4 and Above
Professional Occupation

Use the "t-SNE Cluster" radio buttons or click on any LAD to filter on Education (Level 4+), Full-Time Employment and Professional Occupation
Use the "Year" selector to view data from 2011 or 2021

Each t-SNE cluster groups LADS based on similar socio-economic patterns.
t-SNE_1 Bins reflect dimensionality-reduced latent features of 28 target features of education, occupation and economic aspects — this is a comparative simplification, not a strict label.

5.2 Systematic Use of Methods

For all four dashboards, the systematic use of methods was grounded in Munzner's task abstraction model and a consistent analytical pipeline by calculating fields for percentages when needed and showing the count in numeric format and dashboard objective breakdown with visual encoding and interactivity. Multiple filters were used to help the users, especially international students or policy analysts, explore LAD trends, compare decade shifts, and uncover hidden socio-economic patterns through a clean and interactive interface.

Muzner's why, what and how:

The dashboards were focused to be designed to answer the question **"why"**, which helps to uncover spatial patterns, regional differences, and socio-economic groupings and present the communication decade-wise trends and identifying and comparing between LADs the various features of education, employment and occupation and observing the changes and trends in the years.

"What" we are presenting in the data abstraction: Quantitative data: counts and derived percentages (e.g., % Level 4+, % professional jobs, aggregate sum of associated and technical, managers and senior officers and the professional occupations' aggregated percentage) with Categorical data: LAD Name, Continent of origin. Derived attributes like percentage change, binning labels (High/Mid/Low), cluster labels (0–4 from KMeans) for easier interpretation instead of raw numeric counts.

Custom calculated fields were created in Tableau for % change (2011 vs 2021), professional occupation shares, and high-skilled immigration values—ensuring consistent and meaningful comparisons across all visuals.

"How":

- In Dashboard 1 uses choropleth maps and grouped bar charts were creatively used to show regional change from 2011 to 2021 comparison parallel and to enhance the value of the choropleth map further, the changes in the percentages between the 2011 and 2021 level 4 and above qualifications and professional occupations were highlighted. The tooltips, when hovering, provide extensive information details that help the viewer to engage and analyse well. With the thorough use of text boxes, notes and captions, the user can easily navigate and effectively utilise the visualisation.
- Dashboard 2 combines highlight tables, horizontal bars, pie charts, and maps. To present data in an easier format and visually allure the user to interact was also a key note in this dashboard. With the use of the notes and instructions, the users can select a continent and immediately see spatial and skill-level breakdowns of immigrants. Tooltips and coordinated filters allow "zoom and filter, then details on demand."
- Dashboards 3 & 4 visualise the dimensionality reduction techniques UMAP and t-SNE clusters using choropleth maps and cluster-wise (on selection, even LAD-wise) bar charts. UMAP and t-SNE may not be familiar knowledge to an international student or analyst; therefore, several captions and tooltips were used, and also, instead of a scatterplot, which will not give a very concrete analysis or interpretation, a choropleth map is used by filtering on each cluster grouping. The binned comparisons (e.g., High/Mid/Low bins for % Level 4+) help to understand the bifurcations of each cluster group or LAD.

5.3 Justification

All four dashboards were designed using a consistent visual strategy that supports spatial, categorical, and temporal comparisons without introducing clutter. For Dashboard 1, the choropleth maps and grouped bar charts are used to visualise socio-economic changes (education, employment, occupation) across LADs from 2011 to 2021. Colour encodes % change, allowing fast scanning of improvement/decline, while grouped bars make year-on-year comparisons clearer. We avoided stacked bars as they make it hard to judge the change; only the base segment aligns. The grouped bars show exact shifts, especially in Level 4+ education (e.g. +6.02%) and a drop in full-time work (−4.6%).

Dashboard 2 focused on immigrants by LAD and continent. A highlight table with horizontal bars helped rank the Top 15 LADs by immigrant count and their share in high-skilled jobs, with a side-by-side bar graph. The pie chart was used mainly for filter interaction and to visualise the percentage of the population's immigrant or non-UK born residents, and the choropleth map showed spatial intensity of immigrant concentration, like the City of London, Westminster, and Camden stand out with all the immigrant ethinicities being high in educational qualifications and integration into skilled roles.

Dashboards 3 & 4 used UMAP and t-SNE to uncover hidden patterns using dimensionality reduction. Both dashboards mapped clusters spatially to retain real-world interpretation. K-means groups LADs with similar education, economic, and occupational profiles—these were summarised using bar charts (binned High/Mid/Low by % Level 4+, professional jobs, etc.). Interactivity allows switching between years and clusters. Usage of grouped bars here as they are justified as they align better for multi-year and multi-cluster comparisons.

Each design supports Munzner's core tasks:

- **Locate** (using maps),
- **Compare** (2011 vs 2021 or across clusters),
- **Summarise** (via bar breakdowns),
- **Filter** (with interactivity),
- **Identify** (like Ashfield's +52% rise in education or City of London's high immigrant skill %).

6. Evaluation

Peer Evaluation:

Group discussions helped to get the storyline-related tables well aligned to address the questions set out for the narrative. Further discussions on extracting the data by regions, district-wise, etc, were another important discussion. During the peer evaluations, discussions were set out to identify the right design of the visuals across all the dashboards. It was crucial to validate through peer discussions and observation of user interactions, and self-reflection. The evaluation focused on whether the visualisation supported intended questions answered, how well viewers could interpret and interact with each view, and whether the overall design facilitated meaningful insights. As recommended in Chapter 4 of Munzers (Validation methods), in identifying the visual encodings issues early through the user's iteration and iterations.

Below is a structured evaluation on each dashboard:

In Dashboard 1, the initial use of stacked bar charts to compare between the categories using both year comparisons did not derive meaningful insights falling short in data/task abstraction. Then switched to the grouped bar chart as suggested, which aligned with Chapter 4.2, Muzner's principle or guidance on the channel effectiveness that the position on a common scale is more interpretable than stacked lengths. This change helps with a clear understanding of the visualisation and usability. The colour was first set out with different shades, which were harmonised and neutralised to use a consistent palette. Tooltip enhancements with simplified and clear addressing and font sizes, and labelling right necessary feature. The tooltip enhancements in interpretability are also included in the validation recommendation in Chapter 4.

In Dashboard 2, the immigrant population was first visualised using a stacked area chart, which was not necessary and overloaded the attention on the chart more than the data presented. This corresponds with Munzner's definition of "idiom appropriateness" -the stacked area chart was visually intriguing but not suitable for categorical comparisons. Peer's evaluation found it difficult to isolate the population percentage from the continents as well. The use of the highlight bar graph was well appreciated, but the number of LAD Names shown for the immigrant count was reduced. The dual-axis bar chart was confusing and was replaced by clean side-by-side bars for professional occupations and Level 4+ education.

In dashboard 3 & 4, with interesting discussions with the peers, the scatter plotting was not necessary, deriving good insights and trends could be analysed thoroughly with the scatter plot. ased on feedback and Munzner's recommendation for idiom-task fit (Chapter 4.1–4.2), cluster data was re-encoded as choropleth maps, enabling spatial tasks like Locate and Identify. Peers were finding the choropleth maps enhancement on filter by clusters very interesting and a useful analysis. The binning idea with the UMAP and t-SNE's comparison was also very much appreciated.

Self-Evaluation:

Throughout the project, the data extraction and preprocessing were given the utmost importance to ensure well-structured data to work ensuring data accuracy and completeness. Validated through multiple data cleansing processes and checking for null values and verifying the LAD Names and codes, and correct mapping with the master file data to ensure consistency. The visualisations evolved significantly, as initially the focus was on trying to utilise as many different types of charts and aesthetics, which did not lead to easy interpretation of data. Several tweaks and re-evaluations to focus on the interpretability.

The modifications improved task abstraction and idioms of interaction, ensuring the charts were aligned with users' goals. According to Chapter 4 of Munzner, validation usage through observation, self-testing, and iterative redesign cleaned and preprocessed data presentation and allowed for a clearer and more analytical narrative.

7. Conclusion

7.1 Key Insights on Socio-Economic Trends (2011–2021) from the Tableau Dashboards

In conclusion, with this extensive 2011 and 2021 census data, we performed various statistical methods, such as Bayesian ridge regression, to predict the missing data and used dimensionality reduction techniques to observe the patterns in socio-economic changes across England and Wales. Explored trends and patterns of the socio-economic changes with a main focus on education, employment, occupation, and immigration.

From the visualisation, there were several insights drawn from the observation and analysis. The percentage of the population with Level 4 and above educational qualifications increased significantly in several districts such as Leeds, Bristol, Corby, Wellingborough, Westminster, etc., with a national average increasing from 27.06% to 33.30%. With occupational breakdowns, we can observe that there is an increase in professional occupations from 16.88% to 19.59% and a slight increase in Associates and Technical and Managers & Senior Officers; however, there is a decrease in the Administrative, Skilled Trades, and Mechanical job opportunities over the decade. These changes reflect a transition toward knowledge-based employment.
In employment, there is a steady decrease in full-time employment, part-time employment, and unemployment, with approximately 2%. The analysis on the decline of full-time employment from 38.76% to 34.10% reflects that even after educational growth, employment patterns altered, which is likely due to circumstances like Brexit and the pandemic. Education Level 3 and Apprenticeship also show a positive increase. In the choropleth map, each district's changes over the decade are clearly visible within the range. The range helps us understand that in Level 4 and above education, the lowest declined change is -12.50% and the highest is 8.93%, but on the contrary, there are mainly blue shades present, showing us that there was improvement across the UK.

In Dashboard 2, the immigrant non-UK-born residents from the continents give a clear view of further enhancing the understanding of the employment, education, and occupational changes in 2011. The highest immigrant population is from the Middle East & Asian immigrants, with 44.68%, then Europe with 42.76%. The continents North America, South America, and Antarctica & Oceania have the least populations of 7.17%, 2.49%, and 2.90%. Across all the immigrant ethnicities, Westminster, Camden, City of London, Cambridge, and Wokingham had the highest immigrant populations and also had strong education and employment integration. For example, City of London had 80.81% of its foreign-born residents in high-skilled jobs—showing strong alignment between qualification and occupation.

The immigrant dashboard showed that boroughs like Westminster, Camden, and Tower Hamlets had the highest immigrant populations and also had strong education and employment integration. For example, City of London had 80.81% in high-skilled occupations and 68.36% in Level 4 and above qualifications, Westminster with 69.14% in high-skilled occupations and 50.29% in Level 4 and above education of its foreign-born residents in high-skilled jobs, showing strong alignment between qualification and occupation.

The UMAP and t-SNE dashboards provided a completely different way to look at LADs—not just by location but by socio-economic similarity. Cluster 0 and Cluster 2 LADs have the highest and mid-level UMAP binning and had the highest average values in full-time employment, Level 4+ education, and professional jobs in both the years 2011 and 2021. The binning approach (High, Mid, Low) helped unpack the trends inside each latent feature, which would have been hard to identify otherwise. Cluster 2 had improvement from mid-level binning to high-level binning, and also a significant increase in percentages of changes. The LADs in Cluster 2 in 2011 are similar to those in Cluster 3 in 2021, showing there was a shift increase from binning from mid to high with improvements in these LADs in education, employment, and occupational changes.

Similarly, in t-SNE binning comparisons – the Cluster 0 in 2011 and 2021 have similar LADs prominently and also an increase from low-level LAD to mid-range showing positive improvements. But in Cluster 2, the high and mid-level binning from 2011 is reduced to mid-level, and both clusterings have many LADs in common. Cluster 3 from 2021 and Cluster 4 from 2011 have common clustering of LADs, showing an increase from low to high binning for all the features. Both these dimensionality reduction techniques help to have a quick overview of the districts with the highest changes as it is clustering based on similar patterns and reduced the 28 latent features to a 2-dimensional view, helping to gain insights on the clusters and each particular district the target user is researching about for the trends in 2011 and 2021 for education, employment, and occupation.

7.2 Reflections on Information Visualisation and Design

To give perspective on the information, visualise the challenges faced during the project were to engage the user to interact effectively. Learnings from Munzner's visual encoding and task abstraction principles, the major takeaways were the need and importance of the choice of the visual encoding. It is not necessary to use different visualisation charts; rather, it helps convey the meaning of the data to the user without any information transfer loss from the visualisation to the user. The main objective is to perceive oneself as the target user while building the visualisations and re-evaluating to make the necessary changes to derive meaningful insights and understanding. The visual changes from a stacked bar chart for year differences to a grouped bar chart changed the interaction and understanding drastically. Similarly, the stacked area chart was used to a simple pie chart to break down the immigrant population, which helped with the simplification of the data being presented. The choropleth maps and highlight tables helped capture the clustering and the changes using the colour hues and saturation to understand the range of the differences, which was very insightful.

Interaction through filters, tooltips, and radio buttons also made a big difference in usability. Instead of raw comparisons, each dashboard now follows a structured path: overview → filter → drilldown, supporting users like students and policymakers in drawing their own insights. The use of Bayesian ridge regression to impute the missing values within a set boundary target—with the use of this statistical method, we could effectively give the predicted values and not an aggregated sum or mean, which would not answer the necessary questions. The use of dimensionality reduction techniques and visualising in the choropleth maps was more effective than a scatter plot generated, as we can see the regions and clusters across England and Wales show similar patterns in socio-economic trends in 2011 and 2021.

The main analytical learning and information visualisation learning was that visualisation is to be presented to the target user in a simplified and effective way, and to help them to understand easily what is the expected outcome. Here, the dashboards convey all the questions in the storyline narrative to help international students and aren't just about making things look nice—it's about telling the right story with the right data and designing visuals that actually match the user's task. From checking LAD codes and predicting missing rows using Bayesian regression, to restructuring every chart based on peer evaluation, every change helped get closer to a meaningful, interpretable, and task-focused final product.

Overall, the coursework helped in understanding the importance of Munzner's taxonomy – What, Why, and How – to help present what is necessary with a narrative, and also visual encoding and re-evaluations. The use of Munzner's Chapter 4 – Validation Approaches – helped to present the visualisation with the aim of effectively conveying the large census data to the target user in a simplified and effective way. Understanding the vast census data on socio-economic trends in England and Wales in the decade 2011 and 2021, Tableau visualisation helped gather and analyse the meaningful insights creatively.

References

1. Munzner, T. (2014). *Visualisation Analysis and Design*. A K Peters/CRC Press.
2. Ware, C. (2020). *Information Visualization: Perception for Design* (4th ed.). Morgan Kaufmann.
3. Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press.
4. Fry, B. (2007). *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. O'Reilly Media.
5. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
6. Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (3rd ed.). Packt Publishing.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
8. Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
9. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
10. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.
11. van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579–2605.
12. Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1(Jun), 211–244.
13. Nomis Official Census Data 2011 & 2021
    https://www.nomisweb.co.uk/ (Used for education, occupation, economic activity, immigration, ethnicity, and travel data at LAD level)
14. ONS - Office for National Statistics
    https://www.ons.gov.uk/ (Used for verifying LAD codes, names, and boundary changes between 2011 and 2021)
15. ONS Geography Codes and Lookups
    https://geoportal.statistics.gov.uk/ (Used to align LAD codes with Tableau visualisation mapping)
16. Tableau Desktop (2025) – for building the four dashboards