Exploring Regional Disparities in Graduate Employment (2011–2021): A Visual Analytics Approach

Abstract

This project explores regional disparities in graduate employment across England and Wales from 2011 to 2021 using UK census data. A visual analytics approach was adopted to support policymakers, prospective graduates, and higher education institutions. Datasets from both census years were harmonised at the Local Authority District level to derive indicators on qualification levels, employment outcomes, and industry structure.

Two interactive Tableau storylines were developed to address tasks defined by Munzner's taxonomy, including comparison, ranking, clustering, and forecasting. Visualisations included choropleth maps, ranked and diverging bar charts, and PCA-based clustering. Bayesian Ridge Regression was used to forecast graduate employment in 2030.

Peer evaluation confirmed the visualisations were intuitive and effective for identifying spatial trends and regional outliers. The project highlights how theory-informed visual analytics can uncover socio-economic patterns and support evidence-based planning.

1 Introduction

Regional disparities in graduate employment outcomes persist across the UK, despite rising levels of higher education participation. These differences are influenced by factors beyond qualification rates, including local industry composition, economic conditions, and demographic characteristics. Such variation has important implications for social mobility, regional development, and the planning of education and labour policies.

This project examines graduate employment patterns across Local Authority Districts (LADs) in England and Wales between 2011 and 2021, drawing on UK census data. The aim is to identify where employment disparities are most pronounced, how they have shifted over time, and which structural variables may contribute to these patterns. A forecasting component is included to estimate how graduate employment may evolve by 2030.

The analysis is intended to support three key user groups: prospective graduates evaluating regional job prospects, policymakers designing targeted labour strategies, and higher education institutions monitoring graduate outcomes. A visual analytics approach is used to enable users to explore employment patterns, detect regional outliers, and inform strategic planning.

Key research questions addressed include: Which regions have the highest and lowest graduate

employment rates? How have these changed over the past decade? Are there areas with high graduate attainment but low employment? Which sectors dominate in the top- and bottom-performing regions? And what employment levels might be expected by 2030?

2 Data Preparation and Abstraction

2.1 Datasets Selected

Seven datasets from the 2011 and 2021 UK Censuses were selected to support analysis of graduate qualifications, employment outcomes, industry structure, and demographics across Local Authority Districts (LADs) in England and Wales. These tables were chosen to ensure alignment across time and geography, enabling direct regional comparison. A summary is shown in Table 1.

Table 1: Summary of census datasets used in the project.

Year	Table Code	Description
2011	QS501EW	Highest level of qualification
2011	QS601EW	General economic activity (16–74 population)
2011	QS104EW	Age by sex
2011	QS605EW	Industry of employment (simplified to 9 sectors)
2011	DC5601EWla	Employment and unemployment by qualification level
2021	RM048	Highest qualification by economic activity
2021	RM063	Industry by sector (9-group version)

All datasets were obtained from the ONS census portal and were available at the Local Authority District level to support geographic analysis and cross-regional comparison [1].

2.2 Data Cleaning and Attribute Semantics

Each dataset was filtered to retain only valid Local Authority records. Irrelevant rows were removed, and column headers were standardised for consistency. Raw counts were converted into percentage-based indicators, including the proportion of residents with Level 4+ qualifications, graduate employment, unemployment, and inactivity rates. The share of individuals employed across nine simplified industry sectors was also calculated. Where applicable, gender distribution was extracted. All percentage values were rounded to three decimal places to ensure uniform formatting.

Following Munzner's data abstraction framework [2], the final dataset was tabular, with each row representing a region and each column an attribute. Spatial analysis was supported via updated geographic identifiers. Most attributes were quantitative (e.g., percentage employed), while industry sectors were categorical, and qualification levels were treated as ordinal. These classifications informed visual encoding and modelling decisions.

2.3 Spatial Alignment and Dataset Integration

Cleaned 2011 datasets were merged using original region codes, then harmonised with 2021 geography using an official lookup table from the Office for National Statistics [3]. This process

mapped legacy LAD codes to updated 2021 identifiers, enabling consistent spatial referencing across both census years. The 2021 datasets, already aligned to the updated schema, were merged separately and joined to the harmonised 2011 records. A total of 292 Local Authority Districts were matched across both years, with regions lacking one-to-one correspondence excluded.

For geospatial visualisation, official 2021 shapefiles from the ONS were used instead of third-party geocoding packs such as the Information Lab's postcode files. This ensured compatibility with LAD21CD boundaries and allowed direct joins using official codes. Although a small number of unmatched regions appear blank in 2021 map views, this trade-off preserved accuracy and consistency across visual outputs.

All merged data were exported in wide-format CSV files for Tableau visualisation and forecasting, and in long format for use in dimensionality reduction models such as PCA and UMAP.

3 Task Abstraction

Analytical tasks in this project were defined using Munzner's taxonomy [2], which categorises actions based on user intent. Each task was considered in terms of what users aim to do, why they need that information, and how the visual analytics design supports their goals. Tasks ranged from simple lookups to complex pattern discovery and forecasting.

3.1 User Groups and Use Cases

The visual outputs were designed to support three key user groups. Policymakers require insights into underperforming regions to inform workforce planning and targeted interventions. Prospective graduates may use the findings to evaluate regional job prospects following higher education. Higher education institutions can assess where their graduates are succeeding or facing barriers, and which sectors are absorbing graduate labour.

Each group benefits from clear, regional, and time-aware visualisations. Accordingly, task design was guided by practical use cases aligned with broader socio-economic and planning objectives.

3.2 Task Types

Tasks were selected to match user goals across different levels of analysis. These ranged from low-level queries (e.g., retrieving values) to higher-level insights (e.g., clustering and forecasting). Each task was linked to a specific purpose, as summarised in Table 2.

These task types informed the choice of analysis techniques and visualisation strategies applied throughout the project. The specific visual encodings used to support them are discussed in the following section.

Table 2: Summary of task types based on Munzner's taxonomy.

Task Type	**User Goal (Why)**	**Example in Context**
Lookup	Retrieve a specific value	Find the percentage of graduates employed in Barnsley in 2021
Compare	Identify differences between regions or time points	Compare graduate employment in Adur between 2011 and 2021
Sort / Rank	Highlight best or worst performers	Display top and bottom 10 LADs by graduate employment in 2021
Cluster	Reveal structural similarity in multivariate data	Group LADs by education and industry structure using PCA and K-Means
Summarise	Show aggregate patterns or trends	Compare overall change in graduate employment across LADs using a diverging bar chart
Predict	Forecast future outcomes for decision-making	Estimate 2030 graduate employment in Ashfield using Bayesian Ridge Regression

4 Visualisation Design and Justification

4.1 Overview of Visualisation Approach

The visual narrative was structured around two complementary storylines, both developed using Tableau. The first focused on descriptive analysis of graduate employment rates and industry structure in 2011 and 2021. Visualisations such as filled maps, ranked bar charts, and diverging bar charts were used to support spatial comparison, ranking, and change detection across Local Authority Districts (LADs).

The second Tableau story extended the analysis by exploring regional change over time, identifying structural similarities through PCA, and presenting a forecast of graduate employment in 2030 based on Bayesian Ridge Regression. These visualisations addressed higher-level analytical tasks such as summarisation, clustering, and prediction. Together, the two stories provided a cohesive progression from observed trends to future projections.

Tableau was selected for its ability to support narrative-driven visual analytics while minimising cognitive load. Story points were preferred over dashboards to maintain a clear analytical flow and reduce visual clutter. According to Ware [4], limiting simultaneous visual stimuli improves interpretability, particularly for non-technical users such as policymakers and prospective graduates.

4.2 Encoding Choices: Marks and Channels

Visual encoding decisions were informed by the data types, the task abstractions defined earlier, and perceptual design principles from Munzner and Ware [2, 4]. Marks and channels were selected to enable accurate comparison, spatial reasoning, and pattern discovery.

Table 3 summarises the main visual encodings used in the project, along with their justification.

These decisions aimed to avoid common perceptual issues, such as misleading colour gradients or overuse of area-based encoding.

Table 3: Summary of visual encoding strategies used in the project.

Visualisation	**Mark**	**Channel(s)**	**Justification**
Choropleth maps	Area	Colour (Hue)	Enables spatial comparison across LADs
Ranked bar charts	Bar	Aligned length	High perceptual accuracy for value comparison
Diverging bar chart	Bar	Length + Colour	Shows direction and magnitude of change
PCA/UMAP scatterplots	Point	2D Position + Colour	Reveals multivariate structure and clusters
Forecast bar chart	Bar	Length + Error bars	Communicates prediction and uncertainty

4.3 Justification of Tableau Visualisations

Each Tableau visualisation was selected to support a specific analytical task while adhering to established principles of visual perception and encoding. The use of story points, rather than dashboards, helped guide user attention through a focused, sequential narrative.

Choropleth maps were used to display graduate attainment and employment in 2011 and 2021. Colour was used as the primary channel to highlight regional variation. Although less precise than length, colour effectively supports spatial comparison and pattern detection. A consistent blue-teal scale was applied across years to enable direct visual comparison, in line with Munzner's guidance on channel expressiveness [2].

Ranked bar charts visualised the top and bottom LADs by graduate employment. Aligned horizontal bars, sorted by value, were used to maximise accuracy and support ranking and lookup tasks. This approach follows Cleveland and McGill's findings on perceptual effectiveness.

Diverging bar charts illustrated the percentage-point change in graduate employment between 2011 and 2021. These used symmetric axes and a red-green colour scale to highlight both positive and negative shifts, following Ware's recommendations for bipolar encoding [4].

Stacked bar charts were used to present the sectoral composition of employment in the top and bottom 10 LADs. While not ideal for comparing individual categories across groups, they effectively showed part-to-whole relationships and highlighted structural differences between regions. Consistent colour assignments were used across years to support continuity.

Choropleth and diverging bar charts were also used in Story 2 to highlight changes over time. The map provided a spatial overview, while the diverging bars made individual LAD-level differences more legible. Together, these views supported comparison, summarisation, and outlier detection.

The PCA scatterplot provided a structural view of graduate employment patterns across regions. PCA was used to reduce multivariate data into two dimensions, and clusters were colour-coded using K-Means results. Although static, the plot aligned conceptually with the Tableau story and helped reveal latent groupings among LADs.

The forecast bar chart visualised predicted graduate employment in 2030, based on Bayesian Ridge Regression. Horizontal bars were used to display the posterior mean forecasts, ranked from highest to lowest. Although Tableau does not natively support visual error bars with uncertainty intervals, the 95% credible intervals were incorporated into tooltips rather than shown directly on the chart. This design choice helped minimise visual clutter while still providing access to uncertainty information for users seeking more detailed insights.

All visual forms were grounded in principles of expressiveness and effectiveness [2], and each was matched carefully to its underlying data type and analytical purpose. Sequential story points, consistent encodings, and appropriate mark-channel pairings helped ensure the outputs were interpretable, even for non-technical audiences.

4.4 PCA Clustering Design and Output

Principal Component Analysis (PCA) was used to reduce 23 graduate-related and industry-based indicators to two principal components, enabling structural comparison between Local Authority Districts (LADs). As shown in Figure 1, LADs were projected onto a 2D scatterplot using position to encode multivariate similarity. Colour was used to represent cluster assignments generated by K-Means clustering. Both PCA and K-Means were implemented using the `scikit-learn` library in Python [5].

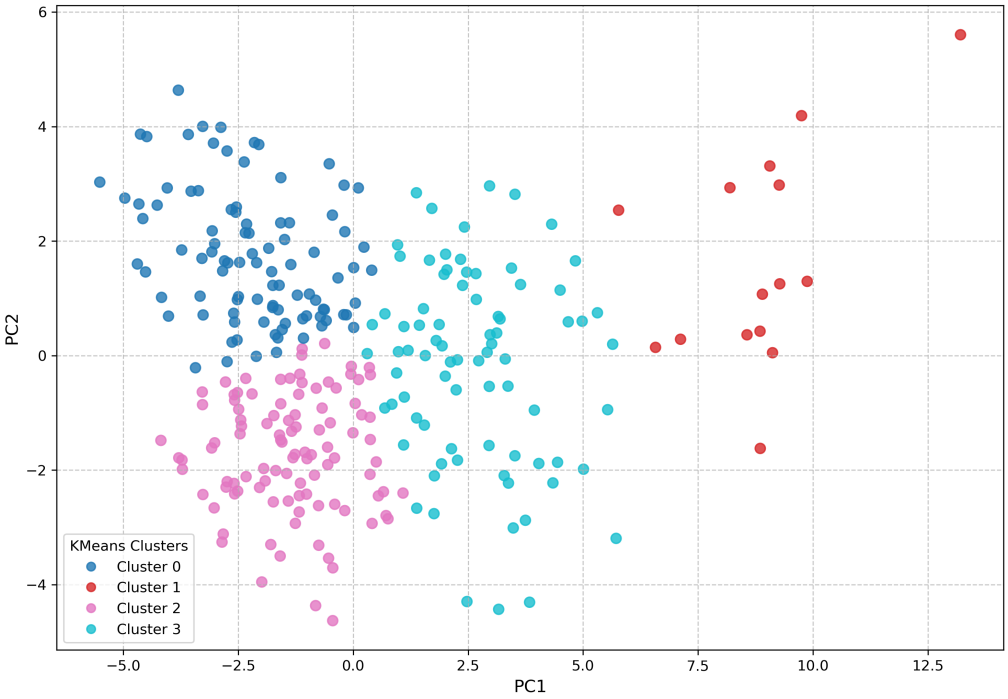

Figure 1: 2D PCA scatterplot showing structural similarity across LADs. Each point represents a LAD, coloured by K-Means cluster label.

The first two principal components captured approximately 59.5% of the total variance in the dataset, with PC1 explaining 42.6% and PC2 accounting for 16.9%. This level of explained variance provided sufficient dimensionality reduction to support meaningful structural interpretation in two dimensions.

The number of clusters ($k = 4$) was selected using the elbow method, based on the inflection point in the total within-cluster inertia curve (Figure 2). This configuration balanced interpretability and separation, allowing for coherent regional groupings to emerge without overfitting.

A sample of the PCA output is shown in Table 4, which includes LAD codes, names, and their corresponding PC1 and PC2 values. These coordinates served as input to both the clustering model and the 2D projection.

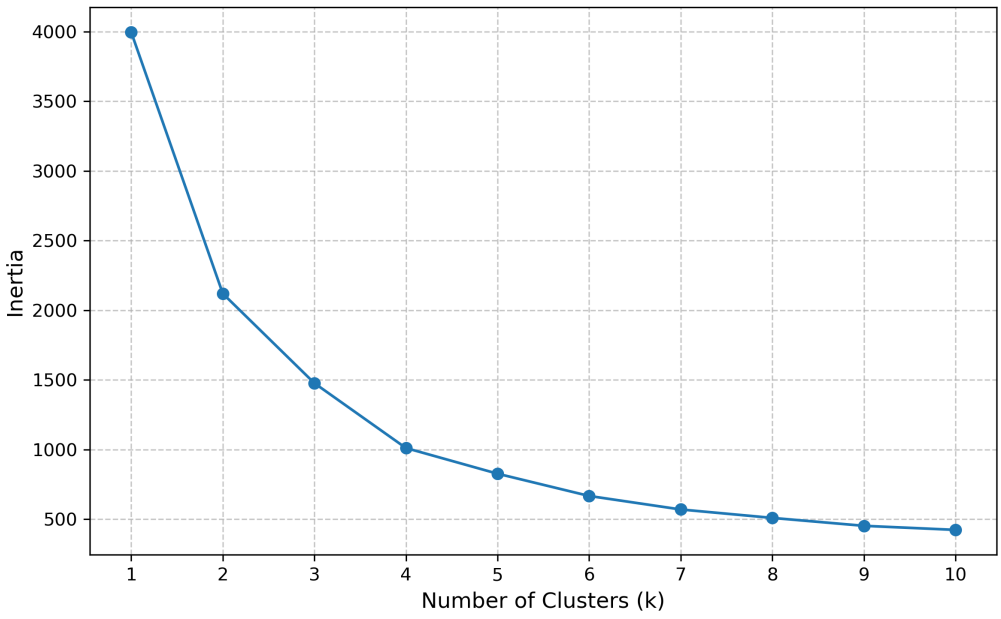

Figure 2: Elbow plot used to determine the number of K-Means clusters. The inflection point at $k = 4$ guided final selection.

Table 4: Sample of PCA output used for clustering and 2D projection.

LAD21CD	**LAD21NM**	**PC1**	**PC2**
E08000037	Gateshead	-0.406	-1.788
E08000021	Newcastle upon Tyne	1.112	-0.725
E08000022	North Tyneside	0.118	-0.419
E08000023	South Tyneside	-1.244	-1.247
E08000024	Sunderland	-1.603	-1.388

4.5 UMAP Projection Design and Output

Uniform Manifold Approximation and Projection (UMAP) was applied to the same 23 standardised features used in the PCA workflow to explore potential nonlinear structures in the data.

Unlike PCA, which focuses on global variance, UMAP preserves local neighbourhoods, making it useful for revealing subtle regional similarities.

While clustering was based on PCA, UMAP was included in the report as a complementary, non-interactive visualisation. It was excluded from the Tableau dashboard to avoid redundancy but provided additional context for structural variation in the written analysis.

The 2D UMAP projection was generated using the `umap-learn` library with `n_neighbors = 15`, `min_dist = 0.1`, and `random_state = 42`. Position encodes multivariate similarity, while colour reflects PCA-derived cluster labels (Figure 3). A sample of the coordinates is shown in Table 5.

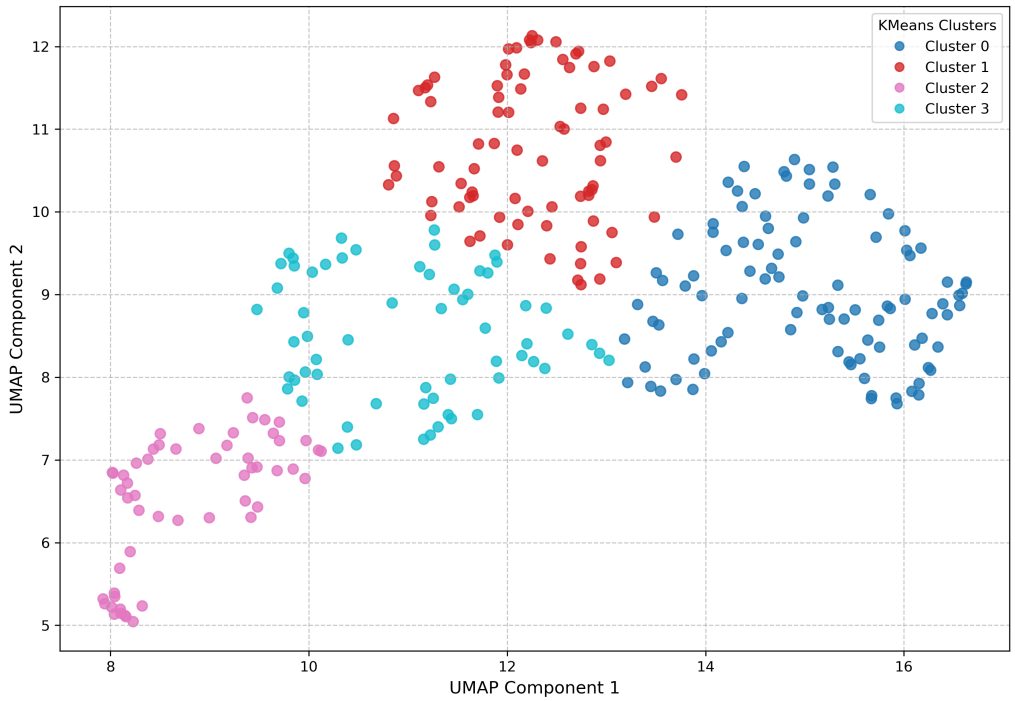

Figure 3: 2D UMAP projection of LADs. Position encodes structural similarity; colours represent cluster membership (from PCA).

Table 5: Sample of UMAP output used for exploratory clustering visualisation.

LAD21CD	LAD21NM	UMAP1	UMAP2
E08000037	Gateshead	12.003	9.601
E08000021	Newcastle upon Tyne	11.552	8.938
E08000022	North Tyneside	12.393	8.835
E08000023	South Tyneside	12.398	9.831
E08000024	Sunderland	12.080	10.162

4.6 Bayesian Forecast Design and Output

Bayesian Ridge Regression was used to forecast the percentage of employed graduates in 2030 across Local Authority Districts (LADs), based on patterns observed in the 2011 and 2021 census data. This model was selected for its ability to incorporate uncertainty through probabilistic inference and to produce both point estimates and credible intervals for each LAD.

The results were first visualised in Python using a horizontal dot plot (Figure 4), showing the top 10 and bottom 5 LADs by their predicted 2030 graduate employment rates. Each point represents the posterior mean forecast, with horizontal error bars indicating the 95% credible interval. A red dashed line marks the national average for reference.

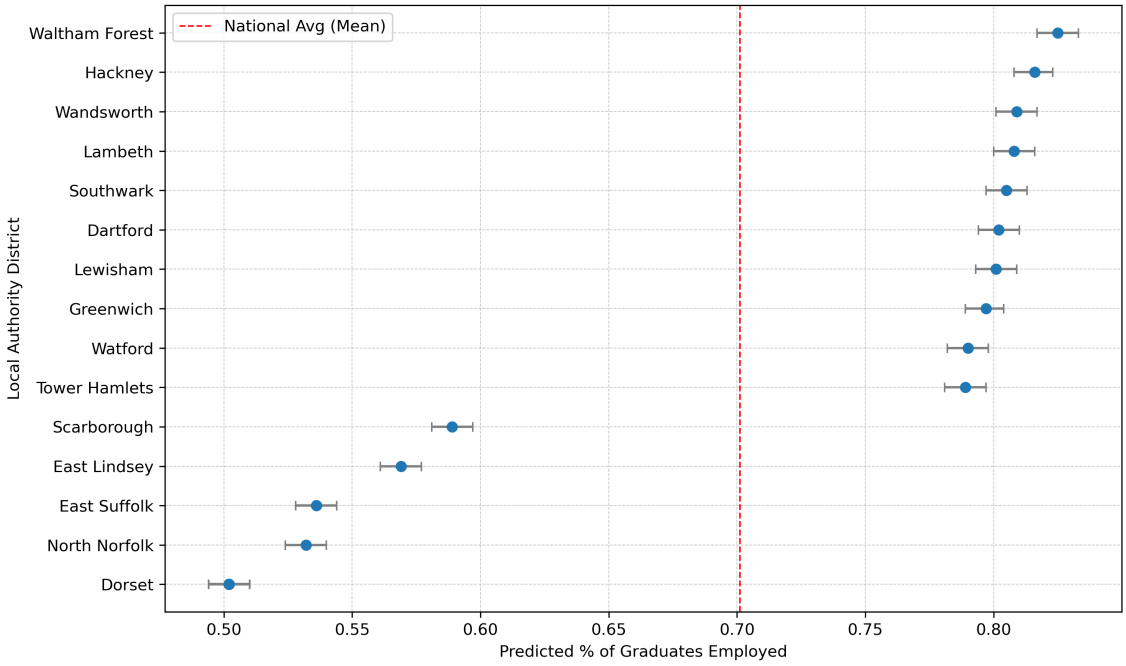

Figure 4: Forecasted graduate employment rates in 2030 for the top 10 and bottom 5 LADs, based on Bayesian Ridge Regression. Error bars represent 95% credible intervals (shown in tooltips).

In Tableau, the same posterior means were presented using horizontal bars, while the credible intervals were included as tooltips rather than drawn directly. This design choice reduced visual clutter while still allowing users to access uncertainty information. Both visualisations supported comparison and prediction tasks in a statistically grounded manner.

A sample of the forecast output is provided in Table 6, showing LAD identifiers, predicted values, and corresponding credible interval bounds. These results were generated using scikit-learn's `BayesianRidge()` implementation and applied to the merged 2011–2021 dataset.

Table 6: Sample of Bayesian forecast output for 2030 graduate employment.

LAD21CD	LAD21NM	Forecast_2030	Lower_CI	Upper_CI
E08000037	Gateshead	0.712	0.704	0.720
E08000021	Newcastle upon Tyne	0.687	0.680	0.695
E08000022	North Tyneside	0.727	0.719	0.735
E08000023	South Tyneside	0.674	0.667	0.682
E08000024	Sunderland	0.669	0.661	0.677

5 Evaluation

The Tableau visualisations were evaluated during a class-based review session, where peers were asked to interpret selected story points without prior explanation. The goal was to assess whether key tasks—such as comparison, ranking, and change detection—could be achieved using only the visual encodings provided.

Choropleth maps in Story 1 were widely understood. Participants accurately identified LADs with high graduate attainment and employment, such as Camden and Hackney, and noted underperformance in rural and coastal areas. The use of a consistent colour scale across years supported comparison, though one participant suggested placing the legend closer to reduce eye movement.

Ranked bar charts were described as intuitive due to their horizontal layout, sorted order, and clear labels. Users easily identified extremes and retrieved values via tooltips. Stacked bar charts showing sectoral employment were also positively received, though a few noted that category labels could be more prominent. The contrast between top and bottom performing LADs in sector structure was considered especially insightful.

In Story 2, the diverging bar chart and choropleth map effectively highlighted changes in graduate employment between 2011 and 2021. Participants appreciated the red-green colour scheme for showing direction of change and found the bar chart easier to interpret than the map for precise LAD comparisons.

Overall, participants could answer questions on regional performance, change over time, and structural variation using the visualisations alone. Suggestions included enhancing label visibility and adding contextual annotations in complex views. Feedback affirmed the value of consistent encoding, clear layout, and perceptual design for supporting interpretability among non-expert users.

6 Conclusion

This project examined regional disparities in graduate employment across England and Wales between 2011 and 2021 using UK census data. Through Tableau-based visual storytelling and Python-based modelling, it showed how qualification levels, sectoral structure, and geographic context shaped graduate outcomes over time.

Visualisations were aligned with specific tasks—comparison, ranking, clustering, and forecasting—and informed by design theory. Evaluation showed that key outputs were intuitive and effective, especially for spatial and temporal comparison. Peer feedback highlighted the benefits of consistent encodings and a structured narrative approach.

Some limitations were identified, including the lack of direct uncertainty encoding in Tableau and the need for improved annotations in complex views like PCA. Nonetheless, the workflow successfully uncovered persistent inequalities and demonstrated the potential of visual analytics in supporting regional planning and policy.

Future work could incorporate additional dimensions, such as city-level granularity, demographic filters, or external indicators like income and housing. Overall, the project illustrates how theory-driven visual analytics can generate meaningful insights from complex socio-economic data.

References

[1] Office for National Statistics. Uk census data 2011 and 2021, 2023. `https://www.ons.gov.uk/census`.

[2] Tamara Munzner. *Visualization Analysis and Design.* CRC Press, 2014.

[3] Office for National Statistics. Local authority district to local authority district lookup (2011 to 2021), 2022. Available at `https://geoportal.statistics.gov.uk`.

[4] Colin Ware. *Information Visualization: Perception for Design.* Morgan Kaufmann, 2013.

[5] Pedregosa et al. scikit-learn: Machine learning in python, 2011. `https://scikit-learn.org/`.