Visual Analytics of Socio-Economic Trends in England and Wales (2011–2021)

Contents

1 Abstract

This report details a visual analytics project exploring socio-economic issues in England and Wales, specifically focusing on the interplay between housing tenure, occupation, accommodation type, and economic activity by comparing 2011 and 2021 census data. The project involved extensive data preparation, including cleaning, transformation, and merging of census datasets. Bayesian regression was employed to handle data inconsistencies arising from changes in geographical definitions between the two census periods. Tasks for visualization were defined using Munzner's taxonomy to guide the design of interactive Tableau dashboards. Visualization techniques include choropleth maps, bar charts, dot plots, and heatmap tables, each justified by principles of information visualization. Dimensionality reduction techniques—t-SNE and PCA—were implemented to reveal underlying structures in occupation and accommodation data respectively, with these projections integrated into the dashboards. Peer evaluation validates the effectiveness of the visualizations while highlighting areas for improvement. The analysis reveals shifting patterns in housing tenure across different occupational groups and economic statuses, with notable increases in private renting and changes in accommodation preferences between 2011 and 2021.

2 Introduction

The socio-economic landscape of England and Wales has undergone significant transformations over the past decade. Understanding these changes is crucial for policymakers, researchers, and the public. This project investigates key socio-economic issues, primarily focusing on the dynamics of housing tenure and its relationship with occupation, accommodation type, and economic activity, by analyzing and comparing data from the 2011 and 2021 censuses.

The primary goal of this project is to develop interactive visualizations that allow users to explore these complex relationships, identify trends, and discover patterns or disparities across different local authorities and over time. These visualizations are intended for users such as:

- Social researchers and academics studying socio-economic trends
- Urban planners and housing policy specialists
- Local government analysts and policy advisors requiring evidence for decision-making
- Students and informed members of the public interested in community changes

These users typically have an analytical background or specific interest in socio-economic data but may not be experts in raw data manipulation, hence benefiting from curated visual explorations.

3 Data Preparation and Abstraction

The analysis is based on census data for England and Wales from 2011 and 2021. The 2011 data was sourced from Excel files, while 2021 data was obtained from ONS census datasets.

3.1 Data Manipulation Steps

1. **Data Loading and Initial Cleaning:** Data for different socio-economic aspects (accommodation, economic activity, occupation) and tenure types were loaded from multiple Excel sheets or CSV files. Initial cleaning involved skipping metadata rows, renaming columns for consistency, and adding a 'Year' column.
2. **Standardization of Categories:** Categorical values across datasets and years were standardized. For example, accommodation types and tenure descriptions were mapped to consistent labels. Economic activity categories were also consolidated using a mapping dictionary.
3. **Handling "Does Not Apply" Values:** Irrelevant entries, such as 'Does not apply', were filtered out from datasets to ensure data quality for aggregation and analysis.
4. **Aggregation:** Data was grouped and aggregated where necessary. For instance, 2021 data on accommodation and tenure was grouped by 'Area Code', 'Area Names', 'Tenure', and 'Accommodation Type' to sum 'Counts'.

5. **Managing Geographic Changes and Bayesian Imputation:** A significant challenge was the change in Local Authority District (LAD) codes and boundaries between 2011 and 2021. A lookup table was used to identify these changes. For LADs that merged in 2021, their aggregated 2021 data needed to be disaggregated to match 2011 boundaries for consistent comparison. A Bayesian Ridge Regression model (from scikit-learn [4]) was trained on the 2011 data to predict the distribution of counts within these newly merged 2021 districts.
   The Bayesian Ridge Regression model estimates a probabilistic model of the regression problem. The prior for the regression coefficient β is given by a spherical Gaussian:
   $$p(\beta|\lambda) = \mathcal{N}(\beta|0, \lambda^{-1}I_p)$$
   The prior for α (the precision of the noise $\epsilon \sim \mathcal{N}(0, \alpha^{-1})$) is given by a gamma distribution. The linear model is:
   $$y(X, \beta) = X\beta + \epsilon$$
   The scikit-learn implementation of Bayesian Ridge Regression tunes the hyperparameters λ (precision of the weights) and α (precision of the noise) by maximizing the marginal log likelihood [7]. The features X for this model were one-hot encoded categorical variables representing 'Tenure' and 'Accommodation Type' from the 2011 data, and the target y was the corresponding count distribution within the constituent 2011 LADs. These predictions were then normalized and applied proportionally to the actual 2021 merged totals to estimate the counts for the constituent 2011 LADs.
6. **Limitations of the Disaggregation Approach:** While Bayesian Ridge Regression provides a robust framework, its application for disaggregating census data based on potentially sparse historical patterns has limitations. The accuracy of the predicted distributions for the constituent 2011 LADs from a merged 2021 LAD is dependent on the assumption that the relationships observed in the 2011 training data (i.e., how counts were distributed among various tenure and accommodation types within those original LADs) remained stable or changed in a way that the model could generalize. However, if the data for some of the original 2011 LADs that formed a new merged 2021 LAD was very sparse (i.e., very low counts for certain combinations of tenure/accommodation), the predictive power of the model for these specific sparse segments could be reduced. Literature suggests that applying ridge regression models in contexts with small or sparse datasets can sometimes lead to unstable coefficients or predictions if not carefully managed [6]. In this specific application, the "sparseness" would refer to the number of examples of disaggregation patterns available in the 2011 data for broadly similar LADs. If an LAD merged in 2021 was formed from 2011 LADs with very few historical data points for certain socio-economic categories, the model's ability to accurately disaggregate the 2021 total for those specific categories might be less precise. This implies that while the disaggregation aims to maintain consistency, the fine-grained accuracy for very small sub-groups within the disaggregated areas might be lower than for larger, more well-represented groups.
   It's worth noting that the datasets did not contain null values in the traditional sense. The main data quality challenge involved missing geographical locations due to boundary changes between census years, which were handled through lookup tables and the Bayesian disaggregation approach described above. The lookup table mapped 2011 LAD codes to their 2021 counterparts, allowing us to identify areas that had been merged, split, or renamed.
7. **Data Merging and Pivoting:** Cleaned and standardized data for 2011 and 2021 were concatenated to create unified datasets for comparison. For dimensionality reduction, data was pivoted to create matrices where rows represented entities and columns represented features.
8. **Dimensionality Reduction Data Preparation:** For t-SNE and PCA, data was pivoted and then scaled using StandardScaler before applying the dimensionality reduction algorithms. The results were appended back to the pivoted dataframes.

3.2 Principal Data Types and Semantics

- **Area Code** (Categorical, Nominal): Unique identifier for Local Authority Districts.
- **Area Names** (Categorical, Nominal): Names of Local Authority Districts.
- **Year** (Categorical, Ordinal): Census year (2011, 2021).

- **Tenure** (Categorical, Nominal): Household tenure (e.g., "Owned: Owns outright", "Rented: Social rented").
- **Accommodation Type** (Categorical, Nominal): Type of housing (e.g., "Whole house or bungalow: Detached", "Flat, maisonette or apartment").
- **Economic Activity** (Categorical, Nominal): Economic status of individuals.
- **Occupation** (Categorical, Nominal): Occupation major groups.
- **Counts** (Quantitative, Ratio): Number of households or people fitting the combined categories.
- **tSNE1, tSNE2, PCA1, PCA2** (Quantitative, Ratio): Derived coordinates from dimensionality reduction.

4 Task Definition

The visualizations were designed to support a range of analytical tasks, framed by Munzner's task taxonomy:

4.1 Overarching Goal

To enable users to discover patterns and trends in housing and socio-economic data, and to present these findings.

4.2 Specific Tasks Supported by Dashboard Components

1. **Exploring Trends in Tenure and Occupation between 2011–2021:**
    - Analyze how tenure preferences vary across different occupations and how these have changed between 2011 and 2021.
    - Identify regions on the choropleth map showing significant changes in population density.
    - Compare tenure shifts across occupations and years (see Table 1).
2. **Observing Trends in Tenure and Accommodation Type:**
    - Understand how the distribution of accommodation types varies by tenure and how this has evolved.
    - Compare tenure counts/proportions between 2011 and 2021 (see Figure 3).
3. **Relationship between Economic Activity and Tenure:**
    - Analyze how economic activity status correlates with housing tenure and observe changes over the decade.
    - Compare economic activity breakdowns across tenure types .
4. **Data Projection Visualizations:**
    - Explore the high-dimensional occupation space to identify clusters of similar LADs or shifts in these clusters between 2011 and 2021.

5 Visualization Justification

The visualization techniques were chosen based on their effectiveness in conveying the specific data types and supporting the defined analytical tasks, adhering to principles of information visualization and human perception.

5.1 General Principles

- **Effectiveness & Expressiveness:** Visual encodings were matched to data types (e.g., color intensity for quantitative data on maps, position and length for quantitative data in bar charts).
- **Human Perception:** Preattentive attributes like color, size, and position are used to highlight key information.
- **Interactivity:** Tooltips provide details-on-demand. The dashboards are designed for linked interactions, where selections in one view can filter or highlight data in other views.

5.2 Specific Visualization Techniques

- **Choropleth Maps:** Used for showing spatial distributions and identifying geographical patterns (see Figure 3). For example, the dashboards visualize population density change or accommodation trends across Local Authority Districts, with color scales highlighting both positive and negative changes.

Table 1: Tenancy distribution over each Occupation (2011–2021)

Occupation	Owns Outright	Mortgage	Private Rent	Social Rent
Managers, directors	20.97%	56.64%	18.01%	4.38%
Professional	19.58%	54.93%	21.41%	4.08%
Associate prof.	17.78%	52.94%	22.99%	6.28%
Admin.	24.91%	44.40%	20.09%	10.60%
Skilled trades	22.00%	46.39%	20.77%	10.84%
Caring, leisure	18.59%	30.74%	26.77%	23.90%
Sales, customer	20.00%	31.47%	28.59%	19.94%
Process, plant	20.01%	41.65%	21.46%	16.88%
Elementary	17.58%	26.85%	29.87%	25.70%

- **Bar Charts and Dot Plots:** Used for comparing magnitudes across discrete categories (see Figure 2). These plots make it easy to compare, for instance, the shift in tenure for different occupation groups between 2011 and 2021, or to observe economic activity trends by tenure type. Excellent for comparing magnitudes across discrete categories. Dot plots are used for clarity when comparing multiple categories or specific data points between years. For instance, "Changes from 2011-2021 within different Occupations based on Tenure" clearly shows the shift for each occupation, with Professional occupations showing the highest increase at 18.17
- **Heatmap-like Tables:** The color encoding adds a visual layer to quickly identify high/low values along with the numerical labels to pinpoint the accurate values (see Table 2). For example, tenancy distribution tables show at a glance which occupation groups are more likely to rent or own, and which economic activity groups dominate each tenure type.
- **Scatter Plots for Data Projections:**Dimensionality reduction techniques are essential for exploring complex, high-dimensional datasets by revealing clusters, outliers, and relationships not apparent in the raw data [5] (see Figure 1 and the t-SNE plot described in Dashboard 1 navigation). Two main types of projections were used:
  - **t-SNE for Occupation Data:** The t-SNE visualization was implemented by first pivoting the data so that each row represented a unique combination of Area Code, Area Names, Tenure, and Year, and columns corresponded to the nine occupation categories. The occupation counts were standardized, and t-SNE was applied (with $n\_components = 2$ and a fixed random seed for reproducibility). The resulting coordinates ($tSNE1$, $tSNE2$) were visualized, as seen in the dashboard, to reveal clusters and separations between 2011 (blue) and 2021 (red) data points. This helps users spot shifts in occupation structure and similarities between different regions or tenure types.
  - **PCA for Accommodation Data:** Similarly, accommodation data was pivoted with rows for each Area Code, Area Name, Tenure, and Year, and columns for accommodation types. After standardization, PCA was applied to reduce the data to two principal components ($PCA1$, $PCA2$). The scatter plot shows how accommodation preferences cluster or shift over time, with 2011 and 2021 points colored distinctly. The first principal component often separates areas by predominant housing type, while the second may relate to the balance between houses and flats.

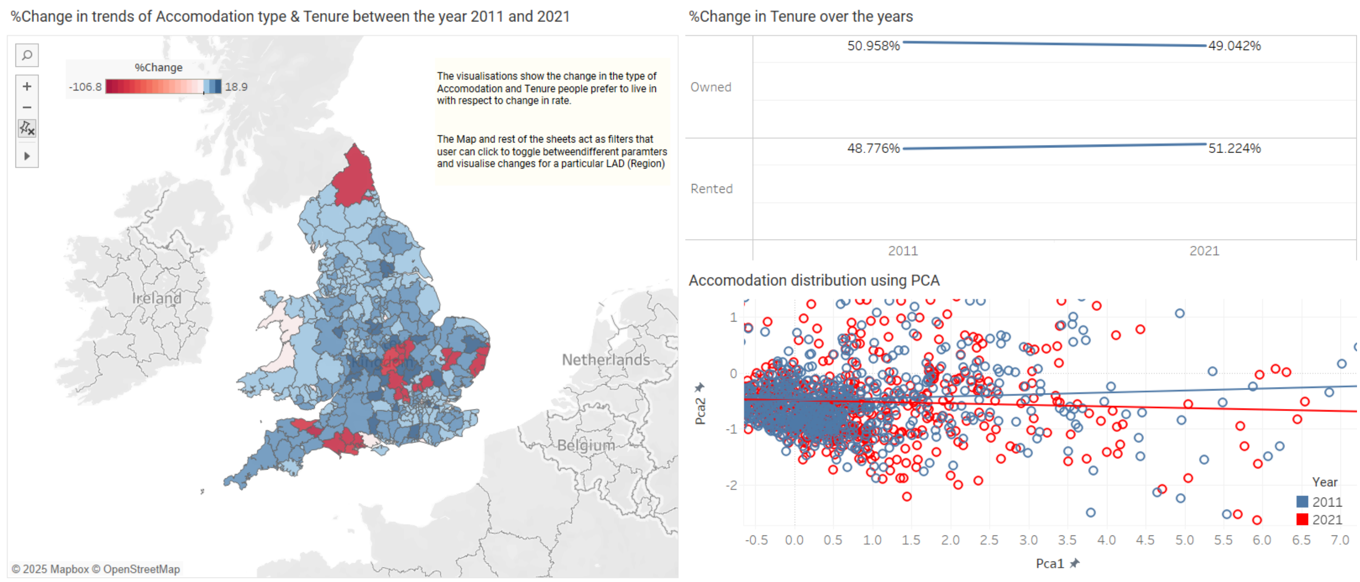

Figure 1: PCA scatter plot for accommodation distribution (2011–2021).

Table 2: Distribution of Accommodation Types among different Tenure types

Accommodation Type	Owns Outright	Mortgage	Private Rent	Social Rent
Flat, maisonette	7.77%	9.61%	43.31%	44.99%
Detached house	36.91%	26.90%	10.33%	3.46%
Semi-detached	34.80%	36.41%	21.16%	28.15%
Terraced	20.52%	27.08%	25.20%	23.99%

6 Dashboard Navigation Guide

This section provides a brief guide on how to navigate and interact with the Tableau dashboards presented in this project. The dashboards are designed to be interactive, allowing users to filter data and explore various socio-economic dimensions.

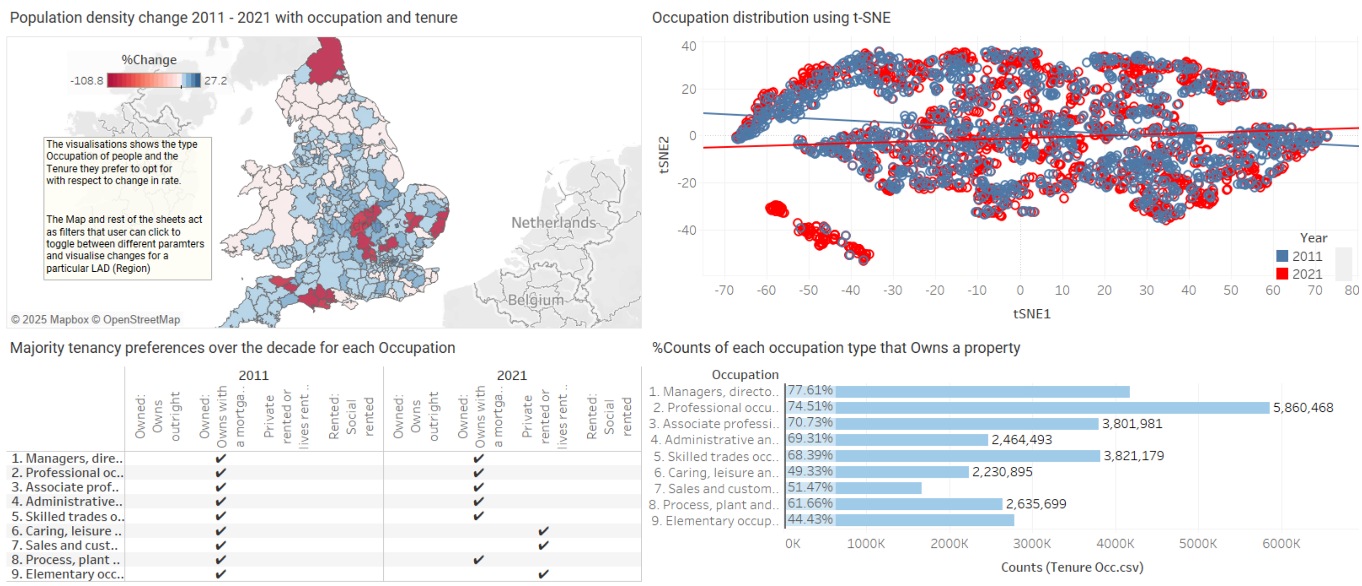

Figure 2: Choropleth map and dot plot showing population density change and tenure-occupation trends (2011–2021).

6.1 Dashboard 1: Occupation, Tenure, and Population Density

This dashboard (visual components are primarily shown in Figure 2) focuses on the interplay between occupation types, housing tenure, and changes in population density between 2011 and 2021.

- **Interactive Choropleth Map (Top-Left):**
  - Displays "Population density change 2011 - 2021 with Occupation and Tenure."
  - The color intensity on the map represents the percentage change in population density, with a legend ranging from negative (e.g., -108.8%, typically red) to positive (e.g., +27.2%, typically blue).
  - *Interaction:* Clicking on a specific Local Authority District (LAD) on this map will filter the other visualizations on the dashboard to show data relevant only to that selected region. This allows for a localized analysis of occupation and tenure trends.
  - The text box indicates that the map and other sheets act as filters, enabling users to toggle between different parameters and visualize changes for a particular LAD.
- **t-SNE Scatter Plot (Top-Right):**
  - Shows "Occupation distribution using t-SNE."
  - Points are colored by year (Blue for 2011, Red for 2021), visualizing the distribution and potential clustering of occupation profiles.
  - This plot helps identify shifts or stability in the multi-dimensional occupation space over the decade.
- **Tenancy Preferences Table (Bottom-Left):**
  - Titled "Majority tenancy preferences over the decade for each Occupation."
  - This table uses checkmarks to indicate the predominant tenure type (Owned outright, Owned with a mortgage, Private rented, Social rented) for each of the nine major occupation groups in both 2011 and 2021.
- **Ownership by Occupation Bar Chart (Bottom-Right):**
  - Displays "%Counts of each occupation type that Owns a property."
  - Horizontal bars show both the percentage and absolute count of individuals within each occupation category who own their property (either outright or with a mortgage).

6.2 Dashboard 2: Accommodation Type and Tenure Trends

This dashboard (visual components are primarily shown in Figure 3, which displays 'dashboard2.png') centers on trends in accommodation types and housing tenure. The PCA scatter plot for accommodation data (Figure 1 is also best understood in conjunction with this dashboard's theme.

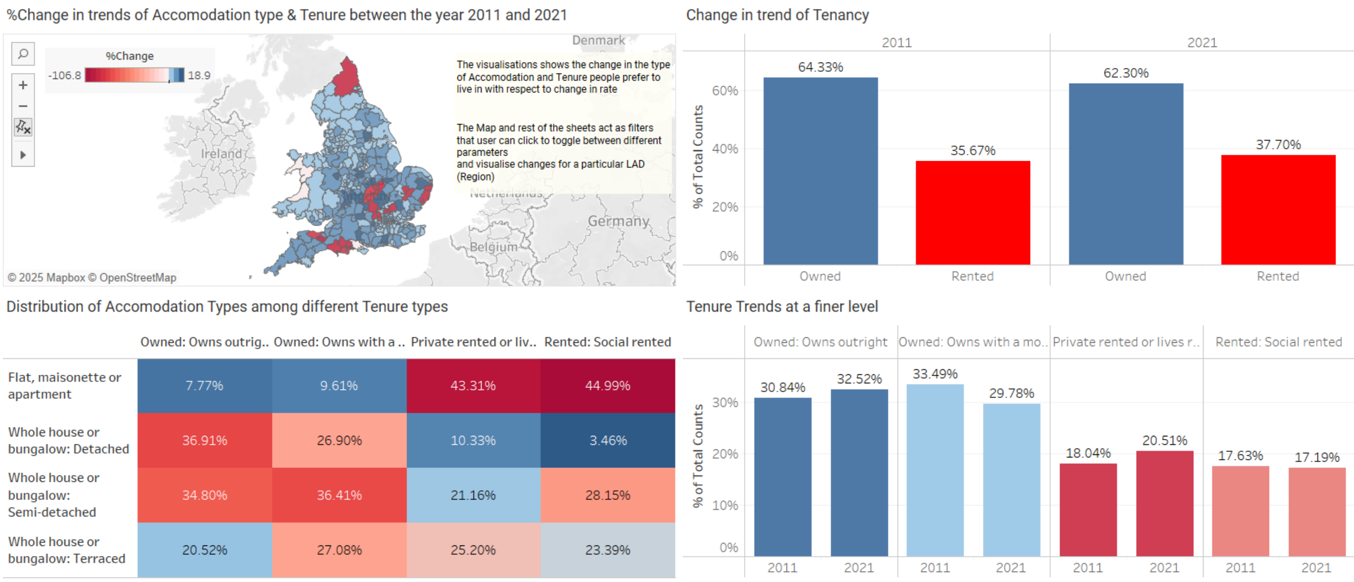

Figure 3: Change in trend of tenancy and accommodation type distribution (2011–2021).

- **Interactive Choropleth Map (Top-Left):**
  - Presents “%Change in trends of Accomodation type & Tenure between the year 2011 and 2021.”
  - Color intensity reflects percentage change, with a legend (e.g., +106.8% to -18.9%). Zoom controls are available.
  - *Interaction:* Similar to the first dashboard, clicking on an LAD on this map filters all other charts on this dashboard to focus on that specific region’s accommodation and tenure trends.
  - The text box again highlights that the map and other sheets act as interactive filters.
- **Tenancy Trend Bar Chart (Top-Right):**
  - Titled “Change in trend of Tenancy.”
  - Compares the overall percentage of “Owned” versus “Rented” properties for 2011 and 2021, showing shifts such as the decrease in owned from 64.33% to 62.30% and increase in rented from 35.67% to 37.70%.
- **Accommodation Distribution Heatmap Table (Bottom-Left):**
  - Shows “Distribution of Accommodation Types among different Tenure types.”
  - Rows represent accommodation types (Flat, Detached house, etc.), and columns represent tenure categories.
  - Cells contain percentages, with background color intensity providing a heatmap effect to quickly identify concentrations (e.g., flats being predominantly privately or socially rented).
- **Finer Tenure Trends Bar Chart (Bottom-Right):**
  - Titled “Tenure Trends at a finer level.”
  - Provides a more detailed breakdown of tenure types (Owned outright, Owns with a mortgage, Private rented, Social rented) as a percentage of total counts for both 2011 and 2021.
- **PCA Scatter Plot for Accommodation Data (Figure 1, displaying ‘dashboard3.png‘):**
  - This plot visualizes the principal components derived from accommodation type data.
  - It helps in understanding the primary dimensions along which accommodation profiles differ across LADs and tenure types, and how these profiles may have shifted between 2011 (blue points) and 2021 (red points).

6.3 Dashboard 3: Economic Activity and Tenure

This dashboard (visual components are primarily shown in Figure 4) is dedicated to exploring the relationship between economic activity status (e.g., employed, unemployed, retired, student) and housing tenure, and how these dynamics have changed between 2011 and 2021.

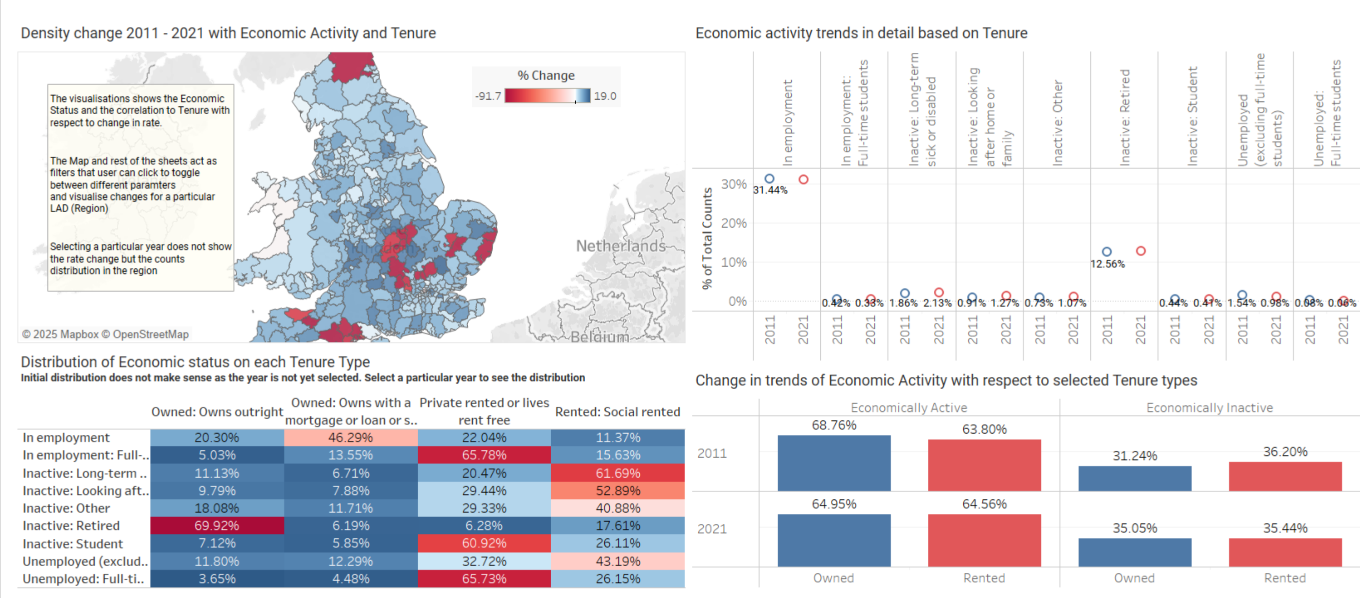

Figure 4: Economic Activity - Tenure trends (2011-2021)

- **Interactive Choropleth Map (Top-Left):**
  - Titled "Density change 2011 - 2021 with Economic Activity and Tenure."
  - The color intensity on the map represents the percentage change in density, with a legend ranging from negative (e.g., -91.7%, typically red) to positive (e.g., +19.0%, typically blue).
  - *Interaction:* The text box explains that the map and the rest of the sheets act as filters. Users can click to toggle between different parameters and visualize changes for a particular Local Authority District (LAD) or region. Selecting a particular year does not show the rate change but the counts distribution in the region.
- **Detailed Economic Activity Dot Plot (Top-Right):**
  - Titled "Economic activity trends in detail based on Tenure."
  - This dot plot shows the percentage of total counts for various detailed economic activity statuses (In employment, In employment: Full-time students, Inactive: Long-term sick or disabled, Inactive: Looking after home or family, Inactive: Other, Inactive: Retired, Inactive: Student, Unemployed: Excluding full-time students, Unemployed: Full-time students) for both 2011 and 2021.
  - Users can observe shifts in these specific categories over the decade. For example, the plot shows "Inactive: Retired" at approximately 12.56% in 2021.
- **Distribution of Economic Status Heatmap Table (Bottom-Left):**
  - Titled "Distribution of Economic status on each Tenure Type."
  - The instruction "Initial distribution does not make sense as the year is not yet selected. Select a particular year to see the distribution" guides the user.
  - Once a year is selected (implicitly through other filters or a dedicated year filter not visible in the static image), this table shows the percentage breakdown of economic activity statuses for each of the four main tenure types (Owned: Owns outright, Owned: Owns with a mortgage or loan or s..., Private rented or lives rent free, Rented: Social rented).

  - Color intensity in the cells acts as a heatmap, highlighting concentrations. For example, for "Owned: Owns outright," "In employment" might be 20.30

- **Economic Activity Trends by Tenure Bar Chart (Bottom-Right):**
  - Titled "Change in trends of Economic Activity with respect to selected Tenure types."
  - This grouped bar chart compares the percentage of "Economically Active" versus "Economically Inactive" individuals within "Owned" and "Rented" tenure categories for both 2011 and 2021.
  - For example, in 2011, 68.76

General Interaction Notes (Applicable to all Dashboards):

- **Filtering:** The primary mode of interaction across all dashboards is typically through clicking on geographical areas on the maps, which then act as filters for the rest of the dashboard. Selections in one chart may also filter others if designed with linked actions.

- **Tooltips:** Hovering over data points in charts or regions on maps will usually reveal tooltips with more detailed information (e.g., specific counts, percentages, area names).

- **Legends:** Pay close attention to legends provided for color encodings (on maps and charts) and year differentiations to correctly interpret the visualizations.

7 Evaluation

This section outlines the planned approach for evaluating the visualizations with peers from the discussion group, as per Munzner's validation framework [2].

7.1 Methodology

- A brief introduction to the dashboards' purpose and the socio-economic context was provided.
- Participants were given a set of representative analytical tasks to perform using the dashboards.
- Observations of their interaction, time taken (informally), points of confusion, and 'think-aloud' comments were noted.
- A short questionnaire or discussion, similar in nature to focus group methods used for evaluating information visualization applications [1], was done to gather qualitative feedback on clarity, usability, and insights gained.

7.2 Peer Evaluation Summary

- **Visual Appeal and Design:** 60% of peers rated the dashboards 5/5, and 40% rated them 4/5. The color scheme and interactivity were praised, with many peers enjoying the layout, plots, and innovative animation feature.
- **Ease of Use:** 75% of peers experienced the fully functional animated heatmap on the Working Hours Dashboard. Two peers noted the absence of a legend explaining the colors, which was subsequently added. One peer disliked the darker heatmap colors, but these were retained for their strong contrast and accessibility for color-blind users. The splitting of geographic data into separate regions was appreciated, and all peers found the select boxes and other interactive elements easy to use.
- **Application of Visualization Theories:** To assess the application of visualization theories, the "Golden Ratio" principle was tested by asking if participants looked at the small text boxes around the graph at the top of the Industry Dashboard first. 60% answered "Yes" and 40% "No." This raised questions about the clarity of the Industries per Region graph for those who did not follow this viewing pattern, potentially compromising visual encoding. However, for those who did, their viewing behavior aligned with the Rule of Thirds, consistent with Djamasi et al., given the graph's visual prominence.
- **Insightfulness:** Most peers rated the visualizations 4/5 or 5/5 for insightfulness; one peer selected 3/5 but commented that the dashboards were intuitive and information-rich. This feedback highlights the challenge of fully understanding the audience's needs within the domain situation.

- **Technical Challenges:** Some performance issues were noted, including slow dashboard processing during interactivity, non-functional hovering, and one instance of Tableau crashing. Additionally, a data loading error caused the top visualization on the Industry Dashboard to disappear once. As a contingency, the Line Bar Combo worksheet was duplicated using an alternative variable, allowing for quick substitution if the issue recurs.

7.3 Validation through Munzner's Framework

The validation process described below follows the tiers proposed in Munzner's nested model for visualization design and validation [2].

7.3.1 Domain Problem & Data/Task Abstraction Validation (Tiers 1 & 2)

The high ratings for "insightfulness" (all but one peer rating 4/5 or 5/5) and comments such as the dashboards containing "interesting information" suggest that peers found the chosen socio-economic problem of housing tenure and its links to occupation and the economy both relevant and well-defined. The ability of the dashboards to present this "wealth of information" in an "intuitive" manner indicates that the abstracted tasks were meaningful and effectively supported by the visualizations.

7.3.2 Visual Encoding and Interaction Design Validation (Tier 3)

Peers generally found the choropleth maps intuitive for identifying regional disparities. The color schemes were considered visually appealing, though the need for a legend to clarify color meanings was noted. Bar charts, dot plots, and tables were effective for comparing magnitudes and distributions across different socio-economic categories.

However, some interactive elements presented challenges, particularly with processing speed and filter loading times. The animated heatmap, while appreciated for its innovation, posed performance challenges. Tooltips were considered valuable for providing details-on-demand, but their clarity could be enhanced with more explicit labels.

7.3.3 Algorithm Design Validation (Tier 4)

The effectiveness of the t-SNE plot for occupation data was mixed. While some peers found it useful for high-level pattern spotting and identifying shifts between 2011 and 2021 data points, its interpretability was limited for others. The challenge stemmed from the difficulty in relating the abstract t-SNE dimensions directly back to the original occupation variables.

The PCA plots, while generated, were not explicitly assessed in the peer feedback, so their impact remains unclear in this evaluation. Similar interpretability considerations would apply: PCA components should be related back to the accommodation types to make the visualization meaningful.

Overall, the high ratings for "insightfulness" suggest that peers did find value in the overall analytical approach, even if some specific visualizations presented challenges in direct interpretation.

8 Conclusion

8.1 Insights on Socio-Economic Trends

The visual analysis of 2011 and 2021 census data for England and Wales provided several important insights into housing tenure and socio-economic patterns:

- **Shifts in Housing Tenure:** Overall, there has been a slight decrease in owned properties from 64.33% in 2011 to 62.30% in 2021, with a corresponding increase in rented properties from 35.67% to 37.70%.
- **Occupation and Tenure Relationships:** The visualizations revealed significant variation in tenure preferences across occupational groups. Professional occupations showed the highest increase (18.17%) in representation between 2011 and 2021, while other occupations showed more modest changes. Managers and professionals have the highest rates of outright ownership (20.97% and 19.58% respectively)

and mortgage ownership (56.64% and 54.93%). In contrast, elementary occupations have much higher rates of private renting (29.87%) and social renting (25.70%).

- **Accommodation Type Patterns:** There are clear relationships between tenure type and accommodation preferences. Flats are predominantly rented (43.31% privately rented, 44.99% socially rented), while detached houses are mostly owned (36.91% outright, 26.90% with mortgage). Semi-detached houses show a more balanced distribution but still lean toward ownership (34.80% outright, 36.41% with mortgage).
- **Economic Activity and Housing:** The visualizations demonstrate strong correlations between economic activity and housing tenure. Those in employment are most likely to own with a mortgage (46.29%), while those looking for work predominantly live in social rented accommodation (52.89%). Students show high rates of private renting (60.92%), and the retired population has the highest rate of outright ownership (69.92%).
- **Geographical Variations:** The choropleth maps revealed significant regional disparities in population density changes and housing tenure patterns. Some regions in northern England show population decreases (red on the map), while certain areas in the south show increases (blue). These patterns likely reflect broader economic and migration trends within England and Wales.

8.2 Information Visualization Learnings

This project provided valuable insights into the practical application of information visualization principles:

- **The Power of a Structured Framework:** Applying Munzner's "What-Why-How" taxonomy [3] was invaluable in systematically defining data abstractions, user tasks, and appropriate visual encodings. This ensured that design choices were purposeful and user-centric.
- **Data Preparation is Critical:** A significant portion of the effort went into data cleaning, transformation, and especially handling the complexities of changing geographical boundaries. The use of Bayesian regression to disaggregate 2021 data for consistency with 2011 geographies was a critical step, demonstrating how statistical methods can support robust visualization.
- **Dimensionality Reduction for Insight:** Implementing t-SNE and PCA highlighted their utility in uncovering latent structures in high-dimensional data. The challenge lies in making these abstract projections interpretable for the user, often through linking and brushing with other familiar data representations.
- **Interactivity Enhances Exploration:** Features like tooltips and linked views are crucial for moving from overview to detail and for exploring relationships across different facets of the data. However, the technical challenges encountered with processing speed highlight the importance of optimizing performance for interactive visualizations.
- **Evaluation is Essential:** The peer evaluation provided practical feedback that went beyond theoretical justification. It revealed aspects of the visualizations that worked well (visual appeal, overall insightfulness) and areas for improvement (legends, performance issues), demonstrating the importance of user testing in the visualization design process.
- **Balance Between Complexity and Accessibility:** Creating visualizations that are both sophisticated enough to reveal complex patterns and accessible enough for users to understand is a delicate balance. This project highlighted the need to provide appropriate context and guidance, especially for more abstract visualization techniques like t-SNE and PCA.

In conclusion, this project demonstrated how visual analytics can effectively reveal patterns and relationships in complex socio-economic data, providing insights that would be difficult to discern through tabular data alone. The interactive dashboards enable users to explore these relationships from multiple perspectives, supporting both broad overview analysis and detailed investigation of specific aspects of housing tenure and socio-economic patterns in England and Wales.

References

[1] Ricardo Mazza. Evaluating information visualization applications with focus groups: the coursevis experience. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods For information Visualization*, pages 1–6, 2006.

[2] Tamara Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[3] Tamara Munzner. *Visualization Analysis and Design*. CRC Press, Boca Raton, FL, 2014.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python, 2011.

[5] Nicolai Peremezhney, Colm Connaughton, Gianfranco Unali, Evor Hines, and Alexei A. Lapkin. Application of dimensionality reduction to visualisation of high-throughput data and building of a classification model in formulated consumer product design. *Chemical Engineering Research and Design*, 90(12):2313–2323, 2012.

[6] Charlotte Porzelius, Martin Schumacher, and Harald Binder. To tune or not to tune, a case study of ridge logistic regression in high-dimensional and low-sample size settings. *BMC Medical Research Methodology*, 21(1):202, 2021.

[7] Scikit-learn developers. Bayesian ridge regression. `https://scikit-learn.org/stable/modules/linear_model.html#bayesian-ridge-regression`. Accessed: May 08, 2025.