Exploring Socio-Economic Trends in England and Wales

Abstract

This report looks at how housing conditions changed in England and Wales between the 2011 and 2021 Censuses. It uses visual tools and methods like PCA and t-SNE to understand complex data. Bayesian Gaussian Mixture clustering was used to create cluster with similar housing patterns. Four main groups were found, along with big changes in overcrowding and housing types.

The results showed that some areas are facing housing problems, while the others have more under-used homes. More people are renting privately and very few are owning the homes because of this city are under high pressure.

the dashboard shows how things change over time. the data methods used in this report help show patterns clearly, even when the data is complicated. The four connected dashboards make it easier to explore and explain the data, which can help people make better decisions about housing.

1. Introduction

This report investigates the changes in housing conditions across England and Wales between the 2011 and 2021 Censuses. Focusing on key metrics such as occupancy ratings and housing tenure, the analysis seeks to answer the following research questions:

1. How have patterns of overcrowding and under-occupation evolved across England and Wales between 2011 and 2021?
2. What relationships exist between housing tenure types and occupancy conditions?
3. Can local authorities be meaningfully grouped into distinct clusters based on their housing characteristics?
4. Are changes in housing conditions geographically clustered or dispersed?
5. How do room-based and bedroom-based measures of overcrowding differ in their portrayal of housing stress?
6. How are key housing characteristics, such as tenure distribution and overcrowding, forecast to change by 2030?

2. Data Preparation and Abstraction

- **Data Sources and Integration**

The analysis draws on Census datasets from 2011 and 2021, focusing on:

- Occupancy ratings for rooms and bedrooms
- Housing tenure types (owned, rented privately, rented socially, etc.)
- Geographic identifiers for local authorities

Three primary datasets were extracted, cleaned, and integrated:

1. Occupancy ratings for bedrooms (TS052-2021 and 2011 equivalent)
2. Occupancy ratings for rooms (TS053-2021 and 2011 equivalent)
3. Tenure composition (from both census periods)

The Office for National Statistics (ONS) uses occupancy ratings to show if a home has few or many rooms for the people living there. The room-based rating looks at how many rooms are there in the home compared to how many the household needs, based on the number of people and their relationships. If the rating is negative, the home is overcrowded and there are not enough rooms. If it's positive, the home has more rooms than needed. The bedroom-based rating works the same way but only counts bedrooms instead of all rooms.

Housing tenure shows the legal way someone lives in their home. The Census puts people into different groups based on how they live, like whether they own their home or rent it:

Owned outright (no mortgage)
Owned with mortgage or loan
Shared ownership

Social rented (from council or housing association)
Private rented
Living rent-free

- **Data Cleaning and Transformation**

  Data Restructuring: The Census data was changed into a better format by reorganizing occupancy ratings and tenure categories. This made it easier to compare different categories and years.

  Standardization of Geographic Units: Local authority codes were made consistent between 2011 and 2021, even with boundary changes. Some areas merged or changed boundaries, so the analysis focused on local authorities that stayed the same across both years. Areas with major boundary changes were left out to make sure the comparisons were valid.

- **Missing Value Analysis:** There were no missing values in the dataset.

- **Derivation of Key Metrics:**

  I. **Overcrowding Percentages:** The percentage of households with occupancy ratings of -1 or lower.
  II. **Under-occupation Percentages:** The percentage of households with occupancy ratings of +1 or higher.
  III. **Tenure Composition Percentages:** The percentage of households in each tenure category compared to the total number of households.
  These metrics make it easier to compare local authorities of different sizes by turning the data into percentages.
  IV. Okay, here is a shorter version, summarizing the dimensionality reduction based on the code:

- **Dimensionality Reduction:**

  To enable visualization and analysis of the complex dataset, the features were first scaled using StandardScaler, and then dimensionality reduction was performed. Both Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were applied to reduce the data to 2 dimensions. The trustworthiness scores of 0.934 (PCA) and 0.983 (t-SNE) indicate that these projections reasonably preserved the original data structure for subsequent analysis, such as clustering.

- **Future Analysis:**

  The predictive modeling was used to estimate housing stress and tenure indicators for the year 2030. Bayesian Ridge regression was chosen for this task.

  The model was trained on the available historical housing data for the local authorities (from 2011 and 2021). Key features influencing housing characteristics and overcrowding were used

as inputs to the model. The trained model was then used to predict the values of relevant housing stress and tenure metrics for each local authority for the year 2030.

3. Data Types and Organization

The analytical foundation of this report is built upon Census data from 2011 and 2021, specifically sourced from datasets related to occupancy ratings (TS052, TS053 equivalents) and housing tenure. This raw data, initially structured across various files for each census year, underwent a rigorous cleaning, restructuring (including pivoting occupancy data), and merging process as implemented in the accompanying code. The result is an integrated dataset where the Local Authority serves as the primary unit of observation, facilitating comparative analysis across regions and over time.

The consolidated dataset incorporates several distinct types of data:

- **Time-based Data:**

  Census year (2011 or 2021): Acts as a crucial ordinal variable, differentiating observations temporally.

- **Geographic and Categorical Data:**

  Local authority names and codes (geography, geography code): Serve as unique identifiers, anchoring the data spatially.

  Cluster assignments (Cluster): Represent categorical labels (0 through 3) derived from the Bayesian Gaussian Mixture clustering analysis applied to the dimensionality-reduced data.

- **Quantitative Data:**

  This encompasses both raw counts and derived metrics calculated per Local Authority:

  Absolute Counts: Direct counts from the Census datasets, including household numbers based on granular occupancy ratings (e.g., '-2 or less', '-1', '0', '+1', '+2 or more' for both rooms and bedrooms) and detailed tenure categories (e.g., 'Owned outright', 'Private rent').

  Derived Percentages: Key metrics calculated by the script to enable standardized comparison across Local Authorities of different sizes. These include:

  - % Overcrowded (Rooms) and % Overcrowded (Bedrooms): Percentage of households with occupancy ratings of -1 or lower.
  - % Under-occupied (Rooms) and % Under-occupied (Bedrooms): Percentage of households with occupancy ratings of +1 or higher.
  - Aggregated Tenure Percentages (Owned, Social Rent, Private Rent, Shared Ownership, Other): The proportion of households within each broad tenure category relative to the total households.

Projected Coordinates (Interval Scale): Reduced-dimension representations of the combined feature space:

- PC1 and PC2: The first two principal components resulting from PCA applied to the scaled quantitative features.
- tSNE1 and tSNE2: The two dimensions resulting from the t-SNE algorithm applied to the scaled quantitative features.

The dataset is logically organized with Local Authorities as the central focus, situated within the broader geographical hierarchy of England and Wales, although the analysis and visualizations primarily operate at the Local Authority level using the integrated data table structure.

3. Task Definition

- Discover Tasks: Identify areas with extreme overcrowding conditions
  Example: Locate local authorities with >5% overcrowding
  Implementation: Color intensity mapping on Geographic Map of Overcrowding.

- Locate: local authorities showing significant changes between 2011-2021
  Example: Find areas where overcrowding increased by >5 percentage points
  Implementation: Change bars with diverging color scale

- Browse: patterns of association between tenure types and occupancy
  Example: Explore relationship between private rental percentage and overcrowding
  Implementation: Interactive scatterplots with parameter selection

- Explore: clusters of local authorities with similar housing characteristics
  Example: Identify which authorities share housing profiles with Birmingham
  Implementation: Color-coded projection plots;

4. Visualization Justification

1. Dashboard 1: Housing Stress and Overcrowding Patterns (2011 vs 2021)

This dashboard is made to look at housing problems, especially focusing on how overcrowding has changed from 2011 to 2021. The different charts were chosen to show many views, including how the problem looks on a map, how different things are related, how things have changed over time, and how different ways of measuring compare.

Geographic Map of Overcrowding (Choropleth Map)

Justification: I used a choropleth map (a map with colors) to show how overcrowding is different in each area. This kind of map is good at showing where overcrowding is high or low. The different colors make it easy to see which places have more crowded homes. It gives a clear picture of the housing problem in different regions.

How it's set up:

- The map uses an equal-interval color scale to show the real differences in overcrowding percentages, instead of a quantile-based scale.
- A semi-transparent layer is added to the map to make the boundaries of each region easier to see.
- Users can hover over different regions to see the authority's name and the exact overcrowding percentage for that area.

Overcrowding vs. Under-Occupation Scatterplot

This scatter plot shows the connection between overcrowding and under-occupation. Each dot stands for a place, either in 2011 or 2021. It helps compare the situation in these two years. The line in the middle (trend line) shows the overall pattern. For example, if a place has more overcrowded homes, it might have fewer under-occupied ones. This helps us see if the pattern changed in 10 years.

The scatterplot also includes:

- **Authority Markers:** The size of the markers shows how important each area is, based on the total number of rooms.
- **Hover Functionality:** You can hover over the points to see more details about each area.
- **Reference Lines:** Lines that show the national averages for overcrowding and under-occupation, giving extra context to the data.

This scatterplot helps viewers understand how overcrowding and under-occupation are connected and shows the national trends, as well as differences across areas.

Overcrowding Change Bar Chart

Justification: This bar chart shows how overcrowding has gone up or down in different places. The bars go left or right from the middle, showing whether things got better or worse. The names of the places are shown clearly on the side. This makes it easy to see which areas had big changes, either positive or negative, between 2011 and 2021.

Implementation Details:

- The bars will be sorted by size to show the areas with the biggest changes.
- Color(light red for decrease, dark red for increase) will show if overcrowding is getting worse or better.
- The chart will show the all local authorities with the biggest changes at top for information.

Room vs. Bedroom Metric Comparison

We used box plots to compare overcrowding measured in two ways—by rooms and by bedrooms—in both 2011 and 2021. A box plot shows how the data is spread out: where the middle is, what the range is, and if there are any values far from the rest (outliers). This helps us see how the difference between these two measures changed or stayed the same over time.

Implementation details:

- Side-by-side arrangement for direct comparison
- Outlier points are easily visible.
- Interactive functionality to display full distribution on demand

The design follows the principle of "multiple coordinated views," which suggests that showing data in different visual forms at the same time improves understanding. The coordination between these views ensures that users can stay oriented while exploring different aspects of the data.

2. Dashboard 2: Tenure Landscape

This dashboard looks at how housing tenure types have changed over time and how they are related to other housing issues like overcrowding. The chosen graphs help to show trends, changes in different places, and how things are connected in the tenure data.

Tenure Composition (Bar Chart):

Justification: This graph shows the share and total number of different housing tenure types (Owned, Private Rent, Social Rent, etc.) from 2012 to 2022. By stacking the tenure types, it clearly shows how the housing situation is made up and how each type has changed over time. This helps us see trends, like which tenure types are growing or shrinking, and how the whole housing system has changed.

Implementation details:

- A consistent color scheme is used across all tenure visualizations for clarity.
- Data points are labeled directly on the bars for easy reading.
- The 2011 and 2021 charts are placed next to each other to enable straightforward comparison.

Tenure Change Map (Choropleth Map):

A diverging color scale is used for showing positive and negative changes. A diverging color palette works best for showing changes from a clear middle point, like zero change. It helps people easily see differences from this reference point. Studies on maps have shown that diverging color schemes are very effective for showing how things move away from a middle value.

Implementation details:

- Blue color show the direction and size of change
- A control to switch between different tenure types

Tenure and Overcrowding Correlation (Scatter Plot):

Justification: This scatter plot looks at the link between average tenure value (likely about property value or cost) and average overcrowding (in rooms). It shows data for two years (2011 and 2021) to see if there is a connection between these two things and if that connection has changed. This graph helps us find out if housing cost affects overcrowding, giving ideas about what causes housing problems.

Implementation Details:

- Choose specific tenure types to compare
- Used diverging color to show data from 2011 and 2021

Top Authorities by Tenure Type (Bar Chart):

Justification: A bar chart is used to show the top authorities (places) with the highest value for a certain tenure type. It might be filtered or split by tenure type. Bar charts are good for comparing values between different groups. This graph helps us see which places have the highest (or lowest) values for a certain tenure type, making it easy to focus on the most important areas.

Ranking Approach
By focusing on the top-ranking areas, the chart highlights the most important cases. This makes it easier for users to spot key patterns and explore more details if they want to.

Implementation Details

- Users can see different tenure types by hovering over bars.
- Bars for 2011 and 2021 are shown side by side for easy comparison.
- The chart shows only the top 15 areas to avoid too much information and keep it clear.

The dashboard follows the principle of "overview first, zoom and filter, then details-on-demand." It starts with a national overview using stacked bar charts, shows the geographic distribution on a map, allows users to explore patterns through a scatterplot, and ends with detailed information by ranking local authorities.

3. Dashboard 3: Cluster Profiles

PCA and t-SNE Projections

Position and Color Encoding of Cluster Assignment
Dimensionality reduction techniques like PCA and t-SNE help turn complex, high dimensional data into a format that's easier to understand visually. They preserve important patterns and relationships in the data, which makes it easier to detect meaningful clusters.

Choice of Projection Methods
Both PCA and t-SNE were chosen carefully based on what each method is best at showing:

Principal Component Analysis (PCA):

- Highlights the overall structure of the data
- Keeps the distances between different groups of data accurate
- Uses straight-line (linear) transformations
- Helps explain the most variation in the dataset in a simple way

t-Distributed Stochastic Neighbor Embedding (t-SNE):

- Focuses on keeping nearby points close together
- Uses curved-line (non-linear) transformations to find patterns that PCA might miss
- Better at showing tight clusters in data
- Captures more detailed, local structures

Using both PCA and t-SNE gives two helpful perspectives:

- PCA shows the main directions in which the data varies.
- t-SNE brings out clear clusters that may be hidden in PCA's view.

Trustworthiness Scores:

- PCA: 0.934
- t-SNE: 0.983

These values show that both methods kept important relationships from the original data. The slightly higher score for t-SNE matches its goal of preserving small, local patterns.

Implementation Details:

- The same color is used for each cluster in all visualizations
- You can hover over points to see the local authority name

Cluster Characteristics Summary

Grouped Bar Chart of Cluster Characteristics

Grouped bar charts are effective for comparing multiple quantitative values across different categories (in this case, clusters). They allow for quick visual comparison of average metric values for each cluster.
Implementation Details:

- Each cluster's average key metric values are shown in a grouped bar chart format.
- Bars are colored to match the cluster groups.
- Important features are visually comparable to help differentiate clusters.

Cluster Distribution Map

Categorical Color Encoding of Clusters
Using different colors for each cluster on a map makes it easy to see where similar types of areas are located. This shows whether similar housing patterns are grouped in certain regions or spread out.

Implementation Details:

- Each cluster is shown with a different, easy-to-tell-apart color
- Hovering shows the name of each local authority

By combining PCA and t-SNE projections, cluster group bars, and maps, the visualizations offer a full picture of how housing and living conditions vary. These views work together to reveal patterns that might not be visible in just one type of chart.

4. **Dashboard 4: Forecasting the Future: Predicting the 2030 Census**

This dashboard visualizes the predicted housing landscape for the year 2030 using Bayesian Ridge regression forecasts. It provides insight into potential future tenure composition and the reliability of these predictions.

Geographic Map of Predicted Tenure Type (2030)

- Color Encoding: Categorical colors represent the dominant predicted tenure type for each local authority in 2030.

- Effectively shows the spatial distribution of predicted future tenure patterns, aiding in identifying regional trends.

Prediction Confidence Visualization

This chart shows the predicted value and confidence range for the selected tenure type per authority.

Implementation Details:

- Position on the x-axis shows the predicted value and confidence range endpoints where as colors differentiate value and range.
- Authorities are listed on the y-axis; coordinates with the tenure type selection.

This dashboard serves as a forward-looking tool for exploring predicted housing changes and understanding the confidence in those forecasts.

5. Evaluation:

In our group discussion, we did a small study to test how well visual tools helped student's complete tasks, following Munzner's framework. We asked the students to identify overcrowded areas using choropleth maps and find connections between types of homeownership and occupancy levels using scatterplots. The results showed that students were able to spot important patterns on the maps and finish the tasks faster than when using raw data tables. I noticed that everyone found the maps easy to understand, but they needed more explanation on the cluster projection views. After finishing the tasks, students said they liked how the maps and bar charts worked together to show geographic patterns but suggested making the legends clearer. They thought the overcrowding vs. under-occupation scatterplot was the most helpful, but they found the t-SNE projections harder to understand because they were more abstract. This feedback helped show that geographic and comparison visualizations worked well, but the more complex projections could be improved.

6. Conclusion:

What I have learned about the socio-economic problem that was the basis of the visualization:

This project has helped me understand the socio-economic problems related to housing in England and Wales. By analyzing data on housing types, ownership, and overcrowding, I learned how social and economic factors like income, access to affordable housing, and regional differences impact housing conditions. For example, I found that rented houses, especially in urban areas, are more likely to have overcrowding, which often affects people with lower incomes. This shows that housing policies need to focus on reducing overcrowding and making affordable housing more accessible. It also highlights the need for better support for people living in areas with higher levels of poverty and housing problems.

What I have learned about information visualization from doing the coursework:

This coursework taught me how powerful information visualization can be in making complex data easier to understand. I learned how to choose the right type of visualization for different kinds of data. For example, choropleth maps helped me show the differences in housing conditions across regions, and scatterplots helped me see the relationship between overcrowding and housing ownership. I also learned how dimensionality reduction techniques like t-SNE can simplify complex data and help find patterns. I realized that good design, like choosing the right colors and layout, is important to make the visualizations clear and easy to understand. This project helped me appreciate how visualizations can not only show data but also tell a story that helps people understand the key points quickly and clearly.