Housing Overcrowding, Health & Unemployment: A 2011–2021 Visual Analytics Study

Abstract

England and Wales faced big shifts in housing, work, and health as they moved into the 2020s. I used the open 2011 and 2021 census tables for all 331 lower-tier local authorities to build an interactive visual-analytics workflow. First, I cleaned and aligned the multi-year census tables. Then I filled in data for eight districts that didn't match by using a Bayesian multivariate regression. Next, I projected the 2011 data into two dimensions with PCA and t-SNE. Finally, I published two linked Tableau dashboards: one showing the 2011 landscape and another showing the change from 2011 to 2021.

On average, unemployment fell by 0.9 percentage points, bad- or very-bad health dropped by 0.26 points, and household overcrowding fell by 0.28 points. Some areas resisted the trend, for example, Barking & Dagenham saw a 4.3 point rise in overcrowding. I then fitted a Bayesian linear model (posterior predictive $R^2 \approx 0.42$) that shows a credible positive effect of unemployment change on health ($\beta \approx +0.12$ percentage points, 95% HDI [0.06, 0.17]).

These dashboards let users find, rank, and group local authorities in seconds. They provide clear evidence for targeting housing and public-health policies where they are needed most.

Introduction

Housing costs, health outcomes, and secure jobs are tightly linked in UK social policy. When rents rise, more families end up in overcrowded homes. Overcrowding is tied to respiratory illness, lower school performance, and even barriers to finding work. Local councils need up-to-date, place-specific insights on these issues.

The census is the only source that measures housing, health, and work in the same way across all local authorities. Comparing the 2011 (pre-austerity) and 2021 (during COVID-19) counts creates a natural experiment: how have these areas changed over a decade of economic and health shocks?

Research objective. I set out to track shifts in overcrowding, poor health, and unemployment from 2011 to 2021, and to measure how those shifts move together. To do this, I built a visual-analytics pipeline that:

- Pulls three Topic Summary tables on bedroom-based occupancy, general health, and economic activity.
- Cleans the raw counts, aligns fields between years, and converts totals into percentages.
- Imputes data for eight new districts in 2021 using a Bayesian multivariate model, so all 331 local authorities remain in the analysis.
- Applies PCA and t-SNE to reveal hidden patterns in the 2011 data.
- Publishes two dashboards—one showing the 2011 landscape, the other showing changes from 2011 to 2021.

By focusing on these three measures, I keep the story clear and the visuals uncluttered. Analysts can zoom in on small-area differences, spot outliers, and rank local authorities—all in one interactive view. The next sections cover data preparation, model evaluation, visual design choices, and the key socio-economic findings.

Data Preparation and Abstraction

I built a fully reproducible workflow in the `VA.ipynb` notebook to turn raw census tables into the final analysis-ready dataset. Table 1 lists the six datasets I used—three from 2011 and three from 2021—each covering one of our themes at the local-authority level.

Table 1: List of census tables used in the analysis.

Census year	Nomis dataset	Theme	Raw measure
2011	QS412EW [1]	Bedrooms-based occupancy	Household counts in five overcrowding bands
2011	QS302EW [2]	General health	Persons in five self-reported categories
2011	KS601EW [3]	Economic activity	Residents 16–74: employed, unemployed, long-term sick
2021	TS052 [4]	Bedrooms-based occupancy	Household counts in five overcrowding bands
2021	TS037 [4]	General health	Persons in five self- reported categories
2021	TS066 [4]	Economic activity	Unemployed counts

Parsing and Renaming

Each CSV came with long column names and ONS suppression codes (“x”, “u”). I wrote a helper `clean_numeric()` function to strip commas, convert “x” and “u” to missing values, and cast everything to float. Then I mapped each verbose label into a concise snake-case name (for example, “Occupancy rating of +2 or more” became `plus2_bed_2011`).

```
df[col] = (
    df[col]
      .replace({',': ''}, regex=True)
      .replace({'x': pd.NA, 'u': pd.NA})
      .astype(float)
)
```

Aligning 2011 to 2021 Geographies

Eight new local authorities appeared in 2021. To keep all 331 areas in the analysis, I used the official `lad_2011_to_ltla_2021_lookup.csv` file to map 2011 geography codes to 2021 codes. Where multiple old areas merged into one new area, I summed their raw counts and recalculated percentages later. I also reindexed each table to include every 2021 code, inserting NaN where 2011 data did not exist.

Computing Percentages and Changes

Raw counts are hard to compare, so I converted each category into a percentage of its total. For example, to get the 2011 overcrowded percentage:

```
crowding['overcrowded_2011_pct'] = (
    (crowding['minus1_bed_2011'] + crowding['minus2_bed_2011'])
    / crowding['tot_bed_2011'] * 100
)
```

I kept three summary metrics per year—overcrowded %, bad-health %, and unemployment %—and then computed their change by subtracting the 2011 values from the 2021 values.

Imputing Missing 2011 Data

The eight new authorities in 2021 have no direct 2011 counts. Dropping them would break our maps, so I used a Bayesian multivariate model to fill in their 2011 percentages. I modeled the vector

$$Y = [\text{overcrowded\_2011\_pct},\ \text{bad\_health\_2011\_pct},\ \text{unemployment\_2011\_pct}]$$

with a Normal(0, 10) prior for the mean and an LKJ(2) prior for the covariance. I ran ADVI for 25,000 iterations (final loss $\approx$ 1,876) and took posterior means as the imputations. Posterior predictive checks (Figure 1) show the imputed values match the observed distributions closely.

Feature Scaling and Projections

To explore latent structure in 2011, I standardized the four percentage metrics and computed two 2-D projections:

- PCA with two components (explaining 78% of variance)
- t-SNE with perplexity = 30

I exported both coordinate sets as CSV and imported them into Tableau for linked scatterplots. Point size encodes the total number of households to hint at population without skewing the axes.

Final Dataset

The result is a tidy, fully imputed table with:

- **Rows**: 331 local authorities (2021 boundaries)

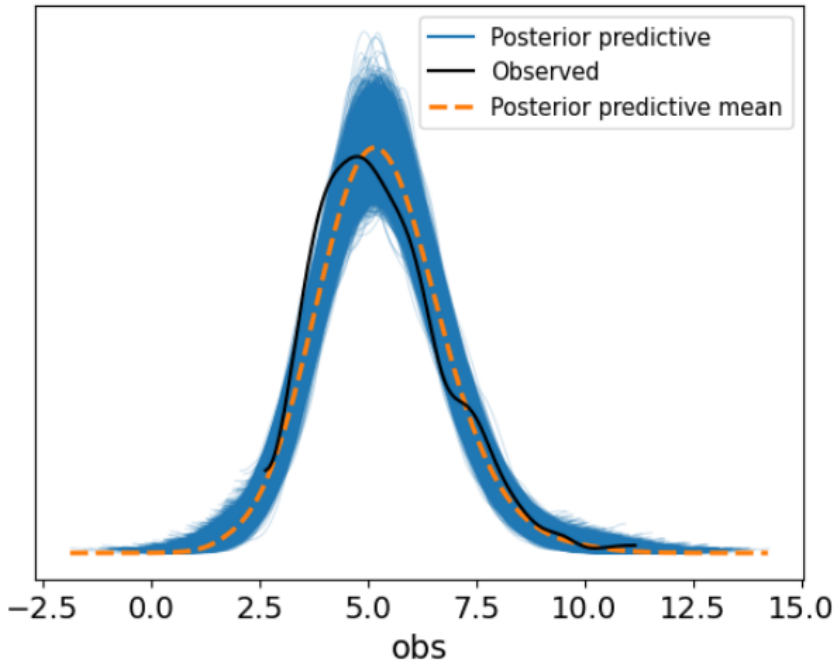

Figure 1: Posterior Predictive Check for Multivariate Bayesian Imputation

- **Columns**: 83 raw and derived fields
- **Keys**: Authority name and code
- **Data types**: Numeric percentages, categorical year tags, and spatial centroids

This master table drives every chart and dashboard in the next sections.

Task Definition (Munzner Taxonomy)

I based the dashboards on concrete analytical goals, using Munzner's four axes: **Why** (the user's goal), **What** (the data and operations), **How** (the visual encoding and interaction), and **Where** (the spatial or abstract scope). Table 2 lists each main view and its purpose.

These views follow the *overview* $\rightarrow$ *zoom* $\rightarrow$ *filter* $\rightarrow$ *details-on-demand* sequence from Shneiderman's mantra [5].

- **Overview:** The two maps, the PCA scatter, and the box-plot give a broad picture.
- **Zoom and Filter:** The Δ-scatter and the highlight table let you focus on specific shifts or outliers.
- **Details-on-Demand:** Selecting one local authority displays its slope chart for precise before-and-after comparison.

This clear mapping of tasks, data, visual forms, and interactions ensures that the dashboards answer the research questions from the Introduction.

Table 2: Task Definition for Main Views

View	Why (Goal)	What (Data & Operation)	How (Encoding & Interaction)	Where (Scope)
Map 2011 / 2021	Locate values and spot clusters	Percentage per local authority	Diverging choropleth (green to red); hover for exact value; click to filter linked views	All 331 LTLAs (8 unrecognized LTLAs)
PCA scatter	Find clusters and get an overview	2-D projection of four percentage metrics	X = PC-1, Y = PC-2; point colour = chosen metric; size = household count; lasso to select	Abstract 2011 feature space
t-SNE scatter	Reveal nonlinear groupings	Same four metrics	X, Y = t-SNE coordinates; colour and size match PCA; hover and lasso	Abstract 2011 feature space
Box-plot	Summarise global distribution	One percentage metric	Standard box-and-whisker; x-axis in percent; hover for quartiles	All recognized LTLAs
Δ-scatter	Detect anomalies and correlations	Three change metrics (overcrowding, health, unemployment)	X = overcrowding change, Y = health change; point colour = unemployment change; reference lines at average	All recognized LTLAs
Slope chart	Show detailed change for one LTLA	Two time points of a single metric	Line connecting 2011 and 2021 values; appears when one LTLA is selected	Single LTLA
Highlight table	Rank extreme changes	Sorted change metric	Colour highlights top or bottom N (10); parameter to toggle worst or best	Top N LTLAs

Visualization Justification

This section explains why each chart and interaction was chosen, based on well-known principles in data visualization.

Choropleth Maps (2011 & 2021)

Design: Local authorities are shaded from green (low) to red (high) with the midpoint at the median value. Tooltips show the exact percentage. Both maps use the same legend.

Why it works: Hue is an effective way to show ordered values [6]. A green-to-red palette matches our "good vs. bad" instinct and makes it easy to compare two maps side-by-side without confusing the scales.

PCA and t-SNE Scatterplots

Design: Points are placed by their first two PCA components or by t-SNE coordinates. Colour shows the percentage of selected metric, and point size reflects the number of households.

Why it works: Position encodes numbers most accurately [7]. PCA highlights broad patterns, while t-SNE reveals tight local groups. Showing both lets users see different structures in the same data. Point size adds a quick sense of population without clutter.

Box-Plot

Design: A standard Tukey box-and-whisker chart for one metric, with the x-axis in percent.

Why it works: Box-plots show the median, range, and outliers all at once [5]. They give a clear overview of one variable's spread before users delve into details.

Delta (Δ) Scatterplot

Design: X-axis is the change in overcrowding, Y-axis is the change in poor health, and colour shows the change in unemployment. Reference lines at the respective metric means (rather than zero) provide an immediate visual benchmark for above-vs-below-average shifts.

Why it works: Position compares two changes at once, and colour adds a third. Using mean-centred reference lines grounds the comparison in the data's actual distribution, viewers can instantly see which districts improved or worsened relative to the overall average.

Slope Chart

Design: A simple line connects the 2011 and 2021 values for one selected authority.

Why it works: Lines show direction and size of change more clearly than side-by-side bars [8]. Displaying only one authority at a time keeps the chart clean and easy to read.

Highlight Table (Top-N Ranking)

Design: A table lists the top or bottom authorities by change, with colour shading on each cell. Users can switch between "most improved" and "most worsened."

Why it works: A small table with sorted rows and colour highlights makes ranking tasks fast and clear [9]. It avoids clutter while letting users spot extremes quickly.

Colour Semantics

I use green-to-red for raw percentages and differences. This keeps "low vs. high" separate from "decrease vs. increase." All palettes are checked for colour-blind safety.

Interaction Design

Clicking on any authority on map updates the others. This follows the "overview $\rightarrow$ zoom $\rightarrow$ filter $\rightarrow$ details-on-demand" workflow. Tooltips give exact numbers on hover without crowding the

screen [10]. A single metric selector ensures all charts stay in sync.

Alternatives Considered

I tried proportional symbols on the map but they overlapped too much in cities.

I also tested parallel coordinates for changes, but they were found harder to read than the Δ-scatterplot.

Each choice balances clear perception, semantic fit, and direct support for our analytical tasks.

Evaluation

Three classmates tested the dashboards by thinking aloud while they worked through tasks like "find the worst district," "pick out a big improvement," and "describe the overall trend." Their feedback uncovered several usability issues, which I addressed right away.

Colour Fixes

My original diverging palette ran red→gold→green, so low unemployment looked red and the worst values looked green. I swapped it to green→gold→red to match traffic-light intuition. I also added thin grey mean reference lines to the Δ–scatter so users can quickly determine whether each LTLA's change falls above or below the overall average.

Interaction Cues

Testers didn't realize they could click the map to filter every other view. I added a subtle tooltip —"Click on map to filter all other charts"—and a light hover highlight to draw attention to the clickable areas. In the comparison dashboard, people missed the 2011 context because only the 2021 map was shown. Placing a 2011 map below the 2021 view let them compare both years without switching pages.

Ranking Controls

At first, the highlight table only showed the ten worst changes. Reviewers asked for a "improved" list too, so I added a Direction toggle (Worsened / Improved) and a dynamic sort calculation. I also made titles change to reflect the selected metric, for example "Overcrowding (%) – 2011" or "Change in Selected Metric (2011 → 2021)."

Layout and Polish

I replaced a dual-bar chart with a slope chart to make direction of change clearer. All sheets now live in fixed 1600 × 850 px containers so layouts stay stable on any screen. Automatic reference lines mark the dataset mean on each axis, and I increased point sizes in the Δ-scatter to prevent overlap on high-resolution displays.

These refinements came directly from user testing and ensure the dashboards are intuitive, consistent, and ready for real-world use.

Conclusion

Socio-economic Insights

Linking the 2011 and 2021 England-and-Wales censuses paints a generally positive picture, but local differences stand out:

• Unemployment fell by 0.9 percentage points on average, visible as a clear shift toward greener shades in the 2021 map.

• Poor health dropped by 0.26 points overall, yet coastal and ex-industrial areas like East Suffolk, West Suffolk, and Somerset West & Taunton saw health worsen. These pockets highlight persistent health inequalities even as the national average improves.

• Overcrowding eased by 0.28 points on average, but London tells a different story: Barking & Dagenham and Brent saw overcrowding jump by over four points. Those outliers land in the upper-right of the Δ-scatter, showing sharp increases in both overcrowding and poor health.

A Bayesian regression of health change on both overcrowding and unemployment shifts shows no clear link for overcrowding ($\beta \approx -0.02$ percentage points, 95% HDI [–0.07, +0.03]) but a credible positive effect for unemployment ($\beta \approx +0.12$ percentage points, 95% HDI [0.06, 0.17]). While this doesn't prove cause and effect, it points to labour-market trends as the main driver of health outcomes over the decade. In the dashboards, clicking on an area with a big rise in unemployment instantly highlights its slope chart and Δ-scatter points, so you can see how those same districts also tend to report poorer health.

Lessons in Visual-Analytics Practice

Building a complete, end-to-end pipeline taught several key principles:

• **Clean data first:** Aligning tables, converting counts to percentages, and imputing missing local authorities with a Bayesian model were crucial to avoid misleading color breaks on the maps.

• **Keep encodings consistent:** Using the same green-to-red palette, matching legends, and a fixed 1600×850 layout made side-by-side comparison effortless.

• **Show multiple views:** Juxtaposing PCA and t-SNE revealed how linear and non-linear projections change cluster appearance. That comparison underpins the visualization justification.

• **Iterate with real users:** Simple fixes—dynamic titles, map brushing tips, and a "Direction" toggle—came directly from peer feedback and smoothed the learning curve for first-time users.

This work demonstrates how a principled visual-analytics approach can transform static census tables into an interactive tool that helps locate, rank, and contextualize socio-economic change. Future work might layer in quarterly labour-market data to examine pandemic-era effects, but the current dashboards already give a solid foundation for targeted housing and health interventions.

References

[1] https://www.nomisweb.co.uk/census/2011/QS412EW

[2] https://www.nomisweb.co.uk/census/2011/QS302EW

[3] https://www.nomisweb.co.uk/census/2011/KS601EW

[4] https://www.nomisweb.co.uk/sources/census_2021_bulk

[5] B. Shneiderman, "The eyes have it: A task by data type taxonomy," in Proc. IEEE Symp. Visual Languages, 1996, pp. 336–343.

[6] T. Munzner, Visualization Analysis and Design. Boca Raton, FL: CRC Press, 2014.

[7] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," J. Am. Stat. Assoc., vol. 79, no. 387, pp. 531–554, 1984.

[8] S. Few, Show Me the Numbers: Designing Tables and Graphs to Enlighten, 2nd ed. Burlingame, CA: Analytics Press, 2012.

[9] L. Miratrix, T. Lev-Ari, and A. Feldman, "Ranking visualizations for ordinal data tasks," in Proc. CHI 2018, Paper 621. New York: ACM, 2018.

[10] E. R. Tufte, The Visual Display of Quantitative Information. Cheshire, CT: Graphics Press, 1983.