Exploring the Impact of Housing on Health in England and Wales: A Visual Analytics Approach

2411243

May 8, 2025

Abstract

This project explores how housing conditions affect public health in England and Wales, using census data from 2011 and 2021, with predictions for 2031. Key housing variables such as overcrowding, heating, and housing type were analysed against self-reported health levels, using a constructed and normalised health index. We applied clustering and dimensionality reduction (LDA, UMAP) to reveal structural patterns, and used Bayesian Ridge Regression to forecast future health distributions with uncertainty estimates. The visualisation system, built in Tableau and guided by Munzner's task framework, supports comparison, discovery, and trend communication. Findings highlight strong housing–health associations, regional disparities, and potential improvements by 2031, offering insights for data-driven policy.

1 Introduction

Housing is a major socio-economic factor influences public health in England and Wales. Poor housing conditions—such as overcrowding, inadequate heating, and substandard structures—can negatively impact both physical and mental well-being.

This study investigates the relationship between housing and health using data from the 2011 Census, focusing on indicators such as housing type, occupancy density, heating availability, and overcrowding. These are analysed against self-reported health levels across regions. To assess changes over time, we compare the 2011 data with health data from the 2021 Census, and use health levels in 2031 by Bayesian modelling prediction to measure the trend.

The findings are intended to inform urban planners(identify spatial correlations between housing infrastructure and health outcomes), public health officials(allocate health interventions based on predicted risks), and policy researchers(detect health inequalities and support data-driven resource redistribution) by identifying areas where poor housing may be linked to declining health, providing a foundation for targeted interventions.

2 Data Preparation and Abstraction

2.1 Data Sources

This analysis uses official census datasets from England and Wales:

- **2011 Health Data:** `KS301EW.xlsx`: Includes the distribution of self-reported health levels across regions (very good, good, fair, bad, very bad).
- **2011 Housing Data:**
  - `KS401EW.xlsx`: Distribution of housing types (detached, terraced, flats, etc.).

  - `KS403EW.xlsx`: Metrics on housing density (e.g., average rooms per household, overcrowding) and heating coverage.

- **2021 Health Data:** `TS037_Health.zip`: Extracted at the Lower Tier Local Authority (LTLA) level and above, with a structure comparable to 2011 health data.

2.2 Data Processing

For 2011 health and housing data:

- Filtered valid rows using `Area code` and propagated region names to avoid non-data information like data introduction.
- Standardized variable names and removed irrelevant fields (e.g., caregiving, disability).
- Selected key variables from the 2011 health and housing data, grouped as follows:

Variable(s)	Category	Type	Description
very_good, very_good_num	Health indicator	Proportion / Absolute	Share and count of residents reporting very good health
good, good_num	Health indicator	Proportion / Absolute	Share and count of residents reporting good health
fair, fair_num	Health indicator	Proportion / Absolute	Share and count of residents reporting fair health
bad, bad_num	Health indicator	Proportion / Absolute	Share and count of residents reporting bad health
very_bad, very_bad_num	Health indicator	Proportion / Absolute	Share and count of residents reporting very bad health
health_index, health_index_norm	Health indicator	Derived	Composite score based on weighted health categories, and its normalised form (0–100)
pct_detached, num_detached	Housing type	Proportion / Absolute	Percentage and number of detached houses
pct_semi, num_semi	Housing type	Proportion / Absolute	Percentage and number of semi-detached houses
pct_terraced, num_terraced	Housing type	Proportion / Absolute	Percentage and number of terraced houses
pct_flat_built, num_flat_built	Housing type	Proportion / Absolute	Percentage and number of purpose-built flats
pct_flat_converted, num_flat_converted	Housing type	Proportion / Absolute	Percentage and number of converted flats
pct_flat_commercial, num_flat_commercial	Housing type	Proportion / Absolute	Percentage and number of flats in commercial buildings
pct_mobile_home, num_mobile_home	Housing type	Proportion / Absolute	Percentage and number of mobile or temporary homes
pct_heated, num_heated	Housing quality	Proportion / Absolute	Percentage and number of households with heating
pct_overcrowded_rooms, num_overcrowded_rooms	Housing quality	Proportion / Absolute	Percentage and number of households with overcrowded rooms
pct_overcrowded_beds, num_overcrowded_beds	Housing quality	Proportion / Absolute	Percentage and number of households with overcrowded bedrooms
avg_household_size	Housing quality	Derived	Average number of people per household
avg_rooms	Housing quality	Derived	Average number of rooms per household
avg_bedrooms	Housing quality	Derived	Average number of bedrooms per household
Area code	Geographical	Categorical	Unique administrative code for region
Region	Geographical	Categorical	Region name
County	Geographical	Categorical	County name
District	Geographical	Categorical	District name
All_num	Population	Absolute	Total number of usual residents in the area

Table 1: Grouped variable table with descriptions

These variables were selected based on the project goals: to compare regions, detect housing–health patterns, and forecast health trends.

- Created a health index using the formula `health_index = 2 × very_good + good - bad - 2 × very_bad` to reflect overall health. It was scaled to a 0–100 range using MinMaxScaler for easier comparison across regions.
- Merged health, housing type, and housing quality tables by `Area code` to ensure spatial alignment.
- Retained proportional variables as float type for modelling and dimensionality reduction.
- Removed national summary rows (e.g., `K04000001`), resulting in 395 subregional records.

For 2021 health data:

- Extracted multi-level data from TS037 (LTLA, UTLA, RGN, CTRY), and mapped all to the 2011 LAD11 regions using ONS-provided LAD21 → LAD11 mappings.
- Computed health proportions and applied the same formula and normalisation as in 2011 to generate a comparable `health_index` scaled to 0–100, to support direct comparison across regions and time.
- Prioritised the finest available spatial level for each area (LTLA > UTLA > RGN > CTRY), aggregating values where multiple regions mapped to one LAD11.
- Identified 11 regions (out of 359) with missing health data in all levels. These were imputed using Bayesian Ridge Regression, with 2011 health proportions as predictors.
- Only missing values were estimated—observed values remained unchanged to prevent distortion.

To construct the 2011–2021–2031 trend dataset:

- Combined cleaned 2011 and 2021 datasets to support temporal trend analysis. All regions were aligned to the 2021 `Area code` system.
- For each region, we constructed input features from 2011 and 2021 health proportions (`very_good`, `good`, ..., `very_bad`) to capture past trends. These 10 variables were standardised using `StandardScaler`.
- We applied Bayesian Ridge Regression to predict the distribution of health levels in 2031, estimating the proportion for each category independently. This model provides both point estimates and 95% credible intervals (`_lower` and `_upper` suffixes) for uncertainty-aware analysis.
- All predicted variables were suffixed with `_2031_pred`, and the outputs were stored in a structured dataset suitable for visualisation (e.g., stacked bar charts with error bars, prediction heatmaps).
- The final table includes `Area code`, region names, and all health indicators from 2011, 2021, and 2031, exported as `health_data_with_2021_2031_index.csv`.
- While the model predicts each health category separately, the sum may not always equal 100%. Future improvements could adopt multi-output regression or constrained models to preserve distributional consistency.

2.3 Variable Abstraction

We abstracted three main types of variables based on their semantic roles and analytical use:

- **Proportional Variables:** Represent category percentages within each area (e.g., `very_good`, `pct_detached`, `pct_overcrowded_rooms`, `pct_heated`). These are suitable for comparing regional structures and used in clustering and dimensionality reduction.
- **Absolute Numeric Variables:** Represent raw counts (e.g., `very_good_num`, `num_detached`). Though secondary to proportions, they support tooltips and allow toggling between absolute and relative values in visualisations.
- **Categorical / Geographical Variables:** Identify spatial units and hierarchies (`Area code`, `Region`, `County`, `District`). These are essential for geographic alignment and map-based analysis across years.

We also derived several composite indicators such as `health_index`, its normalized version `health_index_norm`, and the forecasted `_2031_pred` fields. These indicators aggregate multiple variables and capture temporal trends, making them well-suited for visual storytelling and predictive analysis.

3 Task Definition

This project follows the task taxonomy proposed by Tamara Munzner [1], which defines visualisation tasks from three perspectives: **why** the task is performed (motivation), **what** data is involved (target), and **how** users interact with it (action). Our visualisation supports three main high-level tasks:

- **Discover:** Identify spatial patterns, clusters, and outliers in the relationship between housing conditions and health.
- **Compare:** Examine differences in health status across regions and over time (2011 vs. 2021).
- **Present:** Communicate predicted health trends for 2031 to support policy planning and public communication.

The data used in this project includes item-level indicators such as the health index for each region, aggregated distributions like housing type percentages, and temporal trends across 2011, 2021, and predicted 2031 data. Interactions in the visualisation are designed to support tasks such as lookup, filtering, comparison, and aggregation, meeting the needs of different user groups including policy researchers, data analysts, and the general public.

3.1 Dimensionality Reduction

This module helps users explore how housing conditions relate to regional health levels. We first applied KMeans clustering ($k = 4$) to group regions based on their health index. Then, two dimensionality reduction techniques—Linear Discriminant Analysis (LDA) and Uniform Manifold Approximation and Projection (UMAP)—were used to compress multi-dimensional housing features into two dimensions, allowing users to visually examine structural differences.

The visualisation supports several tasks. Users can discover patterns in housing structure linked to different health clusters, compare how LDA and UMAP present these relationships,

and locate specific regions to examine their housing profiles and health groupings. Each point in the projection represents a region, coloured by its health cluster. In the LDA view, axis directions also reflect which variables contribute to cluster separation. Interaction in Tableau includes hover tooltips to identify regions and side-by-side views for comparing the two projection methods.

3.2 Overview of 2011 Health Data

This dashboard combines a heat map and stacked bar chart to provide a general view of 2011 health levels. Following Munzner's task model [1], it supports *Present* (show known data) and *Compare* (regional differences) tasks. The heat map shows spatial variation in health index values, highlighting higher levels in southern and urban areas, and lower levels in the north and rural regions. The stacked bar chart displays the composition of five health categories per region, with a toggle between counts and percentages. These views support both item-level and aggregate-level understanding of regional health patterns.

3.3 Relationship Between Housing and Health

This module allows users to explore statistical relationships between various 2011 housing features and health levels. A bar chart in the top-left shows the linear correlation between each housing variable and the normalised health index. According to Munzner's taxonomy [1], this task supports *Discover* and *Compare*, involving both aggregate-level and feature-level analysis.

On the right, two linked heat maps show the spatial distribution of the health index (top) and a selected housing variable (bottom). Users can switch between variables using a dropdown and explore values via tooltips. These views support lookup and comparison tasks, helping users detect spatial correlations and structural contrasts between housing patterns and health levels.

3.4 Analysis of Housing Types and Health with Specific Variable

To better understand how housing types may influence health outcomes, we developed a set of variable-focused visualisations. Users can explore a stacked bar chart that shows the proportion and count of different housing types (e.g., detached, semi-detached, terraced, flats) across regions. A linked heat map on the right displays the spatial distribution of the health index, enabling a side-by-side comparison. These views support *Discover* and *Compare* tasks at both item and region levels.

We then focus on two housing variables with the strongest correlations: `pct_semi` (semi-detached housing, negatively correlated) and `pct_flat_built` (purpose-built flats, positively correlated). Users can examine their relationships with the health index through scatter plots and grouped box plots, which reveal trends in median health across value ranges. The spatial pattern of each variable is also shown on a map for geographic context. Overall, these views support lookup, comparison, and trend identification, helping users detect structural associations and spot regional outliers.

3.5 Health Trend Comparison (2011–2021–2031)

This module enables users to compare health levels across three years: 2011, 2021, and 2031. A bar chart on the left shows the standardised health index (Health Index Norm) for each region across the three years, helping users identify rising, declining, or fluctuating trends. On the right, a heat map visualises the spatial distribution of health for each year, which can

be toggled to observe structural shifts. According to Munzner's framework [1], this design supports *Compare* and *Present* tasks at both item and region levels, making it suitable for temporal analysis and policy communication.

The second page shows a side-by-side map comparison of actual health changes (2011–2021) and predicted changes (2021–2031). Map colours indicate direction of change—blue for improvement and red for decline—allowing users to identify trend reversals, consistent deterioration, or unexpected future shifts. This setup supports tasks such as identifying change, zooming into specific regions, and exploring spatial trend differences.

4 Visualisation Justification

Our visualisation design is task-oriented, combining spatial mapping, dimensionality reduction, and temporal trends to support different analytical goals. We followed key perceptual principles from information visualisation, such as prioritising position over colour for quantitative data [1]. Techniques like heat maps, gradient colours, and interactive scatter plots were used to match the data's structure and meaning, making patterns easier to interpret.

4.1 Dimensionality Reduction

To reveal structural links between housing features and health levels, we applied two complementary dimensionality reduction methods: Linear Discriminant Analysis (LDA) and Uniform Manifold Approximation and Projection (UMAP). LDA focuses on maximising class separation and explaining feature contributions, while UMAP preserves non-linear local structures, making it more suitable for visualising organic clustering. However, other methods like PCA was not used due to its limited ability to separate classes, and t-SNE was avoided as it does not preserve global structure well [3].

We first used KMeans to cluster regions based on the 2011 health index. The silhouette score suggested $k = 4$ as the optimal number of clusters (Figure 1), with a balanced size distribution (Cluster 1: 56, Cluster 2: 125, Cluster 3: 124, Cluster 4: 79). These clusters were then used as labels for supervised LDA, and later as a reference for validating UMAP structure.

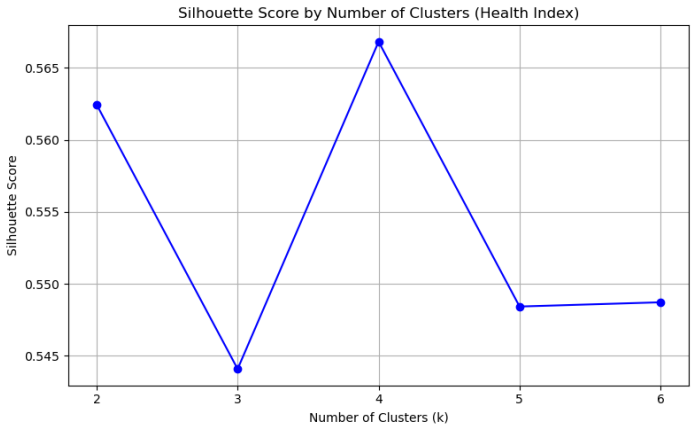


Figure 1: Silhouette Score by Number of Clusters (k); $k = 4$ is optimal

All housing features were standardised before analysis to ensure equal contribution across variables. LDA produced a clear 2D projection with visible separation between health-based clusters (Figure 2). We also extracted feature contributions to each discriminant component,

identifying key housing indicators that distinguish health levels (Figure 3).Since UMAP does not show variable contributions directly, we pair it with LDA to retain interpretability of the original features.

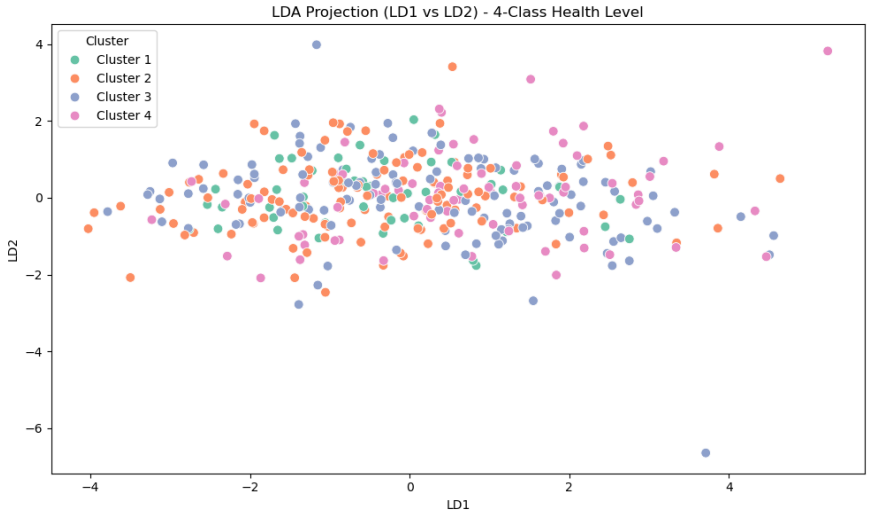


Figure 2: LDA projection (LD1 vs LD2) showing class separation by health cluster

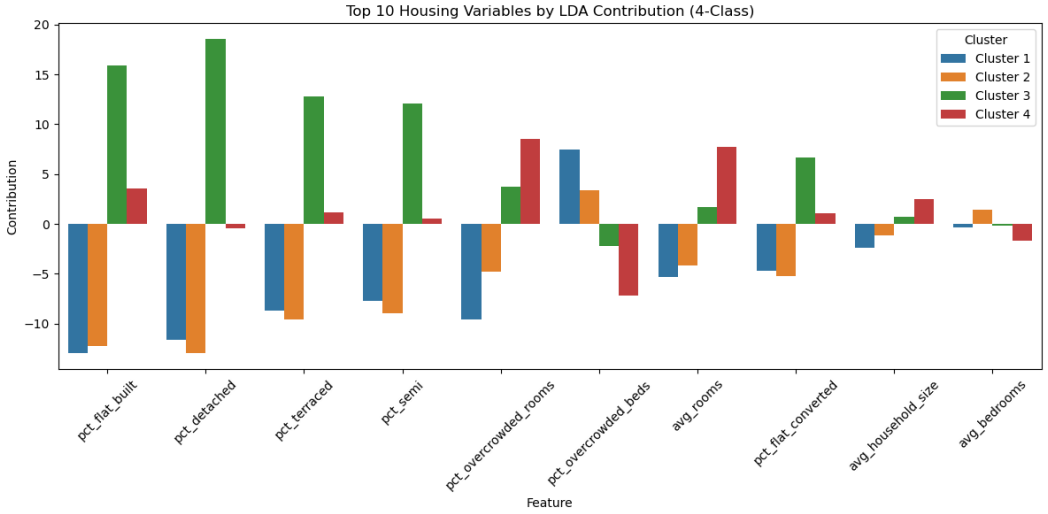
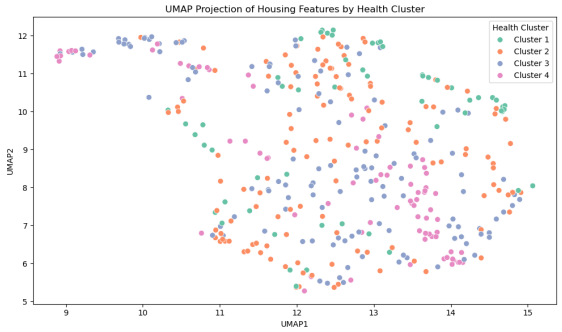
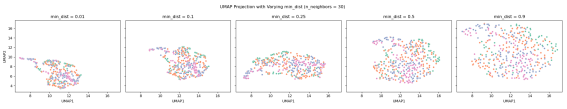


Figure 3: Top 10 housing variables contributing to LDA class separation

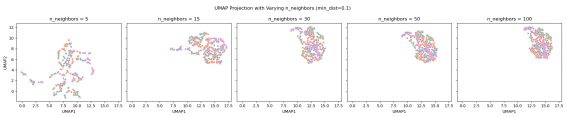
To explore non-linear structures, UMAP was applied in an unsupervised setting. After tuning hyperparameters, we selected `n_neighbors = 30` and `min_dist = 0.1` as the optimal configuration. The final UMAP plot (Figure 4a) revealed smooth transitions and tight local groupings, with health clusters forming distinguishable shapes suitable for interactive analysis in Tableau. Parameter sensitivity tests (Figures 4b and 4c) showed how layout varies with different values.



(a) Final UMAP projection (health clusters)



(b) Varying `min_dist` (fixed `n_neighbors=30`)



(c) Varying `n_neighbors` (fixed `min_dist=0.1`)

Figure 4: UMAP projection comparison. (a) Final layout with tuned parameters. (b–c) Parameter sensitivity tests showing the effect of `min_dist` and `n_neighbors`.

4.2 Spatial and Structural Health Visualisation

This module combines a geographic heat map and a stacked bar chart to show both the spatial distribution and internal composition of health levels. A blue-to-red gradient encodes the standardised health index (`health_index_norm`), following perceptual guidelines that favour colour intensity for quantitative values [2]. The heat map enables intuitive spatial analysis, while the bar chart breaks down five health categories per region, with a toggle for viewing counts or percentages. The design supports clear comparison and is suited for quick insight and communication by analysts and policymakers.

4.3 Correlation and Linked Spatial Views

This module links variable-level correlation analysis with spatial comparison to explore relationships between housing variables and health. On the left, a bar chart presents Pearson correlations between housing variables and the health index, using position and length to encode values for perceptual clarity [1]. Value labels and tooltips allow users to quickly see variable order and correlation direction.

On the right, two coordinated maps show spatial patterns: a red–blue scale for health index and a single-hue gradient for the selected housing variable, avoiding channel conflict. Both maps are linked by shared area codes, supporting simultaneous comparison. Tooltips display region names and values to aid spatial reasoning.

This design integrates statistical interpretation with spatial context, helping users detect patterns, compare variable impacts, and explore regional differences relevant to housing and health policy.

4.4 Variable Analysis and Spatial Coordination

This module combines multiple views—bar charts, scatter plots, box plots, and heat maps—to link variable-level patterns with spatial distribution. In the housing type overview, a stacked bar chart uses colour to represent different housing categories, with a toggle between counts and percentages to support comparison. A linked health heat map (red–blue gradient for `Health Index Norm`) enables joint analysis of structural and spatial patterns.

For selected variables (`pct_semi` and `pct_flat_built`), scatter plots with trend lines show linear relationships with health, box plots reveal grouped variations, and maps highlight geographic patterns. This "chart + map" design helps users explore trends, clusters, and outliers across both statistical and spatial dimensions.

Overall, the design supports causal reasoning and geographic comparison, balancing visual clarity with analytical depth.

4.5 Time Comparison and Forecast Visualisation

This module visualises health index changes across three key years—2011, 2021, and 2031—using a grouped bar chart and interactive heat maps. The bar chart uses colour-coded bars by year to show trends along a common axis, helping users identify turning points and long-term patterns. The heat maps encode health values using a blue gradient (darker = healthier), and interactively link to bar chart selections when switching between years.

The second page presents side-by-side maps of health index change (differences between years), using a diverging red–blue scale to represent direction and magnitude. This colour scheme follows best practices for visualising data centered around a neutral baseline, aiding in the detection of both improvement and decline.

The overall design supports trend tracking and spatial comparison, making it useful for forecasting, identifying at-risk regions, and informing public health policy.

5 Evaluation: Visualisation Assessment

We evaluated our visualisation system using Tamara Munzner's nested validation model [1], combining qualitative and quantitative methods targeting both interaction and perception levels.

Qualitative validation involved observing user interactions and collecting feedback on colour use, layout, and interface flow. Quantitatively, we designed typical analysis tasks and recorded task accuracy, time, and comprehension.

Five students participated in the testing: three from data science, one from engineering, and one from humanities. Each explored the Tableau dashboard to complete typical visual tasks:

Task Type	**Example Question**	**Observation**
Trend detection	Which regions are predicted to improve between 2021 and 2031?	Users easily identified changes using the heatmap. Year toggling and zoom were responsive.
Causal exploration	Which housing variable relates most to poor health?	`pct_semi` stood out via correlation bars and scatter plots. Tooltips supported understanding.
Geographic comparison	Which two cities differ most in housing structure and health?	Users compared bar charts across regions effectively; toggling count/percentage helped interpretation.

Table 2: Summary of typical tasks and user observations

Overall, users found the system intuitive, with smooth interaction, consistent colour encoding, and helpful tooltips. However, we identified several improvement points:

- Some heatmaps lacked clear legends—adding simple labels (e.g., "blue = healthier") would help.
- Variable names like `pct_flat_built` were unclear to non-technical users; readable aliases are recommended.

- Users struggled to infer how housing types ratios impact health. Adding trend charts or correlation matrices may help.
- Multivariable interactions were not supported. Future designs could add faceted or multi-axis plots.

6 Conclusion

This project has deepened our understanding of how housing conditions impact public health. Based on detailed 2011 census data, we identified significant correlations between various housing factors—such as overcrowding, heating, and housing types—and regional health levels. For instance, a higher proportion of semi-detached houses was associated with lower health, while purpose-built flats showed a positive relationship. Spatially, health outcomes were generally better in southern and urban areas, with lower levels observed in parts of the north and rural regions. Over time, health declined in several regions between 2011 and 2021, but our Bayesian regression forecast suggests potential improvements in some areas by 2031. These findings highlight housing improvements as a key lever for enhancing regional public health, with strong policy relevance.

From a visualisation perspective, we gained hands-on experience in data modelling, interaction design, and visual encoding. We explored both LDA and UMAP for dimensionality reduction, understanding their strengths in explaining variables versus revealing structure. Applying Munzner's Why–What–How framework and multi-level validation helped us see that visualisation is not just about displaying information, but about bridging user goals, cognitive load, and task efficiency. Design choices such as colour channels, chart linking, tooltip content, and variable naming proved crucial for user comprehension and exploration. We also recognised the value of interactivity and guidance in supporting analytical thinking.

In addition, this project reinforced the importance of data preparation. From cleaning multi-source datasets to standardising features and filling missing values, we built a solid foundation for effective visual analysis. Using Tableau, we learned to coordinate multiple views, enabling users to move fluidly between overview and detail. This experience improved both our design thinking and our ability to structure complex data for user-driven insight.

References

[1] Tamara Munzner. *Visualization Analysis and Design.* CRC Press, 2014. Chapters 3–5: Task, Data, and Goal Abstractions.

[2] Tamara Munzner. *Visualization Analysis and Design.* CRC Press, 2014. Chapter 10: Marks and Channels.

[3] Wattenberg, M., Viégas, F., & Johnson, I. (2016). *How to Use t-SNE Effectively.* Distill. https://distill.pub/2016/misread-tsne