Advanced Data Analytics Coursework Report

Student number: 2217555

Abstract

This report presents a visual analysis of population data focusing on the relationships between tenure types, NS-Sec classifications, gender, and general health conditions. There are four tasks defined in this study and explored by visualization techniques, yielding to several key findings, For example, it reveals that higher socio-economic status individuals concentrate in specific regions in the southern part of the UK, but the best general health outcomes are not necessarily associated with the highest socio-economic group. Moreover, it highlights the relationships between Tenure, General Health, NS-Sec, and Sex, with NS-Sec showing associations with gender and health status, and Tenure types exhibiting a stronger link with General Health. The coursework effectively uses visualization techniques, such as heatmap and scatterplots, to present data and gain valuable insights into socio-economic issues.

1 Introduction

This report explores the relationships between NS-Sec (National Statistics Socio-economic Classification), general health, household tenure, and sex in England and Wales. By analyzing data from 2011 and 2021, it is aimed to uncover the connections between these variables and examine potential health disparities.

The primary objective of this data visualization project is to examine the associations between NS-Sec, general health, household tenure, and sex. This study utilizes UMAP(Uniform Manifold Approximation and Projection) and t-SNE(t-Distributed Stochastic Neighbor Embedding) techniques for dimensionality reduction and employ Tableau for creating interactive visualizations.

The target audience for this visualization is policymakers, public health officials, researchers, and stakeholders. By presenting user-friendly visualizations, it is aimed to facilitate understanding and addressing socioeconomic and health disparities.

The upcoming sections will outline the steps taken to prepare the data, go over the datasets used, and showcase the visualizations produced. This analysis aims to inform policy decisions based on evidence and improve health outcomes for the population.

2 Data Preparation and Abstraction

2.1 Data Description

The original data used in this analysis comes from four different datasets, two from 2011 and two from 2021. The datasets are related to general health and socio-economic factors, specifically focusing on the NS-Sec (National Statistics Socio-economic Classification) and tenure (household ownership/rental) categories.

The first dataset from 2011 is titled ”General Health by NS-Sec by Sex by Age.” It provides information on the general health status of the usual resident population aged 16 and over in England and Wales, which is classified based on general health, NS-Sec, sex, and age categories. The purpose of this dataset is to allocate health resources and services, develop and monitor policies related to healthcare delivery, reduce health inequalities, and assess progress in improving the general health of the population.

NS-Sec is a socio-economic classification system used to indicate a person's socio-economic position based on their occupation and other job characteristics. It is an Office for National Statistics standard classification. NS-Sec categories are assigned based on a person's occupation, whether they are employed, self-employed, or supervising other employees. Full-time students are categorized as ”full-time students” regardless of their economic activity.

General health refers to a person's self-assessment of their overall health, ranging from very good to very bad. It does not necessarily reflect their health over a specific time period.

The second dataset from 2011 is called ”General Health by Tenure by Sex by Age.” It offers estimates from the 2011 Census, classifying usual residents in households in England and Wales based on general health, tenure, sex, and age. Tenure refers to whether a household owns or rents the accommodation it occupies. The dataset aims to provide insights into the relationship between general health, tenure, and demographic factors.

The datasets are organized by geography, specifically Local Authorities Districts.

2.2 Data Preprocessing

In the initial data processing step, each row of the raw data contains a unique key value representing the geographic location. However, different attributes are included within each row. In the 2011 datasets, the Household Tenure data includes General Health Type, Household Tenure Type, Age, and Sex. The NS-Sec dataset includes General Health, NS-Sec categories, Age, and Sex. This analysis is mainly interested in exploring the relationship between General Health, NS-Sec, Tenure, and Sex. Therefore, the data related to age is removed during the data processing stage.

Additionally, the General Health Type values differ across the datasets. In the 2011 Household Tenure and NS-Sec datasets, General Health Type has five categories ranging from ”Very good” to ”Very bad.” However, in the 2021 NS-Sec dataset, General Health has only three categories. Specifically, ”Very good” and ”Good” are grouped as ”Good Health,” while ”Very bad” and ”Bad” are categorized as ”Bad Health.” In the 2021 Household Tenure dataset, General Health has only two categories, where ”Bad Health” and ”Fair Health” are combined as ”Not good health.”

To facilitate data dimensionality reduction and comparison across datasets, the data with five General Health Type categories is transformed into three categories. Furthermore, to enable a comparison between the 2021 Tenure data and the 2011 Tenure data, an additional Household Tenure dataset from 2011 with only two General Health Type categories is included.

To explore the relationships between different sexes and their general health status concerning NS-Sec and Tenure, the labels are created for data dimensionality reduction. which include combinations of different sexes and general health statuses.

The processed data can be described as multidimensional tables(Munzner, 2015) with multiple key values and attributes:

- **Local Authorities District:** Categorical geographical data used for mapping purposes.
- **General Health Type:** An ordinal attribute comprising three categories (Very good or good health, Fair health, Bad or very bad health).
- **Sex:** A categorical attribute with two categories (Females, Males).
- **NS-Sec:** An ordinal attribute with 15 different socio-economic position types (L1-L15).
- **Household Tenure:** A categorical attribute with four different types.
- **Population Count:** A continuous numerical attribute representing the population count.

2.3 Data Projections

After preparing the data, the next step is to apply data projection techniques for dimensionality reduction and create derived data(Munzner, 2015) that extends beyond the original attribute set for further visualization. This analysis mainly focuses on the NS-Sec data and Household Tenure data projections with UMAP and t-SNE techniques, respectively.

As Table1 shows, both datasets share the same labels.

Labels	Sex	General Health Type
0	Females	Very good or good health
1	Males	Very good or good health
2	Females	Fair health
3	Males	Fair health
4	Females	Bad or very bad health
5	Males	Bad or very bad health

Table 1: Labels For Projections

2.4 Data Blending

In Tableau, the processed tables are integrated by performing inner joins. Since there are variations in geographic data between 2011 and 2021, an inner join is performed using the geographic location and general health as key values. The resulting physical table serves as the basis for final visualization.

Additionally, location data obtained from data projection is joined with NS-Sec and tenure using Local Authorities and labels as key values, creating a physical table for precise data positioning. Logical connections are established between NS-Sec and Household Tenure data. The merged data is then ready for visualization design.

3 Task Definition

The following section outlines the task definitions for four dashboards designed to investigate various aspects of demographic data, focusing on NS-Sec (National Statistics Socio-economic Classification), tenure, general health, gender, and their interrelationships. Each dashboard is designed for one task and each task falls under Munzner's task taxonomy and covers querying, comparing, summarizing, browsing, analyzing, and exploring.(Munzner, 2015)

3.1 Task 1 - Focus on NS-Sec by General Health by Sex in 2011

This task mainly focuses on four sub-tasks as follows:

- The objective of this task is to explore how NS-Sec categories are distributed across various districts in the year 2011. The focus is on identifying regions where people with the highest social status are most prevalent.
- In addition, the task aims to analyze the relationship between gender and NS-Sec.
- It also involves determining the most prevalent NS-Sec category with each health type.
- Furthermore, the task involves summarizing the top 10 districts with the highest proportions of specific NS-Sec categories under different health types.

3.2 Task 2 - Focus on Household Tenure by General Health by Sex in 2011

This task mainly focuses on the exploration of the relationship between tenure types, general health, and gender, which includes four sub-tasks:

- The task aims to explore the population distribution of different tenure types across regions, with a particular emphasis on identifying districts with a high proportion of rented accommodation.
- Additionally, the task involves to determine the tenure type with the highest population proportion for each health type.
- It is also required to compare the population sizes of different tenure types. The male-to-female ratio within each tenure type will be assessed to gain further insight into the distribution of tenancy types.
- Moreover, the task includes browsing the top 10 countries where ”Owned outright” households have the highest population proportions within each district.

3.3 Task 3 - Tenure, NS-Sec, and General Health Relationships

This task mainly focuses on exploring the relationships between tenure, NS-Sec, and general health, it involves two sub-tasks as follows:

- The task involves analyzing the correlation between tenure and NS-Sec, identifying any relationships between specific tenure types and NS-Sec categories.
- In addition, the task aims to explore the relationships between NS-Sec, general health, and tenure through the plots or data projections techniques after encoding the original data. These techniques enable the investigation of clustering patterns based on different genders and health conditions.

3.4 Task 4 - The Comparison and Changes between 2011 and 2021

This task emphasis on comparing the demographic data between the years 2011 and 2021.

- Specifically, it involves analyzing the changes in the population proportions of different NS-Sec and tenure types between the two years, identifying the increasing or decreasing categories.
- In addition, the task also aims to exploring the changes in population quantities for different NS-Sec categories and health conditions, as well as different tenure types and health conditions between 2011 and 2021.

4 Visualization Justification

4.1 Dashboard 1 - Focus on NS-Sec by General Health by Sex in 2011

For the first question of Task 1, a map is leveraged as a visualization tool to examine how different NS-Sec categories are distributed across various districts in 2011. It also aims to identify the areas where individuals with the highest social status are present. Here is the visualization justification:

1. **Map**: Using a map as the mark type is a great way to display geographical data. It allows for the representation of spatial relationships and provides an intuitive understanding of regional variations. This makes it an appropriate choice for the visualization task at hand.
2. **Data Description and Derived data**: The data description involves presenting the population values and calculating the percentage of NS-Sec population for each district. Using the percentage calculation helps remove the influence of varying total population sizes across different districts, enabling fair comparisons of NS-Sec populations within each district. This derived data(Munzner, 2015) is valuable for highlighting the proportion of specific NS-Sec categories relative to the total population in each district.
3. **Visualization Interactivity**: Three filters - Sex, NS-Sec, and Display Type (for data presentation) - are incorporated into the visualization to empower users with the ability to select and explore different combinations of variables. This interactivity enhances user control and enables deeper insights into the relationships between NS-Sec, gender, and geographic distribution.
4. **Tooltip**: The tooltip feature provides additional contextual information when hovering over a specific district. It displays the Local Authorities district name, the selected NS-Sec category, the chosen gender, the selected data presentation mode (absolute values or percentage), as well as the population count and percentage under the current filter conditions. Moreover, it includes a bar chart showing the distribution of health conditions in the selected district, with different colors representing each health condition.
5. **Color Legend**: The color legend employs an orange-blue color scheme, where higher values lean towards orange and lower values lean towards blue. This color encoding allows for visual differentiation and facilitates the perception of relative magnitudes.(Ware, 2008)
6. **Title and Annotations**: In the visualization, the title and annotations are clearly displayed and draw the eye to the overall percentage and population count of the selected NS-Sec category. These elements are large and bold, making it easy to quickly understand the main findings of the data.

7. **Zoom Functionality**: The map includes zoom-in and zoom-out functionality, enabling users to focus on specific regions or explore the data at different levels of granularity. This feature enhances the flexibility and scalability of the visualization.(Munzner, 2015)

For the second and third questions of Task 1, a Box-and-Whisker plot is utilized as the visualization to determine the most prevalent social status within each health category and gender category, respectively. Here is the justification:

1. **Box-and-Whisker**: Using a Box-and-Whisker plot is suitable for comparing distributions and identifying key statistics within different categories. It provides a compact representation of the data, including quartiles, outliers, and the median.
2. **Derived Data**: The visualization computes the current value as a percentage of the total, enabling fair comparisons across different categories. This derived data highlights the relative prevalence of each social status within the specified health or gender types.
3. **Color**: The color encoding in the Box-and-Whisker plots is utilized to distinguish between different NS-Sec types. This color choice aligns with Munzner's recommendation of using visual variables(Munzner, 2015), such as color, to represent categorical data. By assigning distinct colors to each NS-Sec type, it enhances visual separation and allows users to easily identify and compare the social status distributions within each health or gender category.
4. **Marks**: The mark type used is circle, representing the percentage of total population values for each general health type or sex. The color encoding is based on the NS-Sec dimension, providing additional details.

For the last question of the Task 1, the Heatmap is used as the visualization, here is the justification:

1. **Heatmap**: The Heatmap is an effective visualization choice for comparing proportions across different NS-Sec categories and health types in different Local Authorities Districts. It aligns with Munzner's recommendation of using heatmaps to represent multivariate data with two categorical key attributes and one quantitative value attribute(Munzner, 2015), where color encodes the values and enables easy visual comparison. Based on the justification above, the orange-blue color scheme is used to convey magnitude differences.

2. **Labeled Marks and Stacked Marks**: The Heatmap utilizes square marks to represent each combination of NS-Sec and Local Authorities District. The marks are labeled with the display measures, allowing users to directly observe and interpret the values associated with each square. Stacked marks are used to facilitate comparisons within each NS-Sec category, providing a clear visual separation between different health types within the same district.

3. **Filtered Data and Interactive Exploration**: The Heatmap incorporates various filters, such as NS-Sec, Local Authorities District, General Health Type, and Sex, to enable interactive exploration of the data. Users can select specific combinations of filters to focus on particular subsets of the population and gain insights into the top 10 districts with the highest proportions for each NS-Sec category and health type.

Interactive filters are incorporated into the dashboard to enable user-driven exploration and analysis. Users can select different dimensions, such as NS-Sec, General Health Type, Sex, and Display Type, to dynamically update the visualizations and focus on specific subsets of the data.

By following the reduction theory and eliminating unnecessary elements, the overall layout of Dashboard 1 achieves a streamlined and focused presentation of the data. The clean design allows users to easily comprehend the visualizations, make comparisons, and gain insights without visual distractions.(Munzner, 2015)

4.2 Dashboard 2 - Focus on Household Tenure by General Health by Sex in 2011

For the second task, Dashboard 2 is created to explore the relationship between Tenure, General Health, and Sex. Although the data is different, the overall structure and exploratory goals of the data are similar to those of the first task. As a result, Dashboard 2 follows a nearly identical layout and design approach as Dashboard 1. Here is a brief justification for Dashboard 2:

1. **Visualization Choices**: Dashboard 2 adopts almost the same visualization types as Dashboard 1, including Map, Box-and-Whisker plots, and Heatmap. Apart from that, the Bar Chart is also used in this dashboard to indicate distribution of different tenure types and the percentage of each gender. Using stacked bar charts is a good choice for displaying and comparing quantitative values. The length of the bars makes it easy to compare the different categories of Tenure.
2. **Layout and Structure**: Dashboard 2 has a similar layout and structure to Dashboard 1. The visualizations are arranged in a compact manner, and unnecessary titles and axes are removed to keep the interface clean and focused. This design decision is based on the reduction theory principles that were explained in the visualization justification of Dashboard 1.
3. **Color**: Dashboard 2 uses the same color palette as Dashboard 1, following Ware's color perception principles(Ware, 2008) . Consistent coloring helps to understand data patterns and maintains visual coherence between the two dashboards.
4. **Interactive Filters**: Dashboard 2 is similar to Dashboard 1 in that it also has interactive filters that allow users to explore and analyze data on their own. By selecting dimensions like Tenure, General Health, and Sex, users can update the visualizations in real-time and concentrate on specific subsets of the data. This interactive feature helps users with their exploration and engagement with the data.

4.3 Dashboard 3 - Tenure, NS-Sec, and General Health Relationships

For the first question of the Task 3, the heatmap is used to explore the correlation between tenure types and NS-Sec social status with the correlation formula. Here is the justification:

1. **Heatmap**: The heatmap is chosen to represent the correlation between Tenure Type and NS-Sec categories. Heatmaps are a useful tool for analyzing correlations between categorical variables, as they effectively display the strength and direction of the relationships.
2. **Color Encoding**: Color encoding is employed to represent the correlation between Tenure Type and NS-Sec categories. An orange-blue color scale is selected that visually displays the correlation's strength and direction. The chosen gradient moves from low to high correlation values, which follows Ware's principle of using color gradients for quantitative values.(Ware, 2008)
3. **Axes and Labels**: In this chart, the rows indicate different Tenure Types, while the columns represent NS-Sec categories. The labels on the axes are correctly displayed to provide precise details about the variables being compared, so users can accurately interpret the correlation values in the heatmap.
4. **Data Filtering**: To better understand the correlations between Tenure Type and NS-Sec, the data has been filtered based on Sex and General Health Type dimensions. This filtering capability allows for a more focused analysis of specific subgroups, and provides valuable insights within the context of sex and general health.
5. **Marks and Labels**: In the heatmap, the square marks indicate the correlation values. Specific correlation values are labeled automatically, which makes it easy for users to interpret and compare the strength of correlations between different pairs of Tenure Types and NS-Sec categories.

For the second question of Task 4, a scatterplot is used to explore the relationship between NS-Sec, General Health, and Tenure.

1. **Scatterplot**: The scatterplot is chosen to depict the relationships between NS-Sec, General Health, and Tenure. Scatterplots are effective in showing the distribution and relationships between two

continuous variables (Display Measures) while using additional visual encodings to represent categorical variables (Sex and General Health Type). The Display Measure can be the population count or the percentage of population.

2. **Visual Encodings**:
   (a) **Shape**: The shape of the marks is used to encode the General Health Type dimension. Different shapes represent different health types, allowing users to discern the health category of each data point.
   (b) **Color**: The color of the marks is used to encode the Sex dimension. Distinct colors represent different sexes, enabling users to differentiate data points based on sex.
3. **Axes and Labels**: The scatterplot shows the relationship between Display Measures-Tenure (on the y-axis) and Display Measures (on the x-axis). The axis labels are clear and indicate the variables being compared, making it easier to accurately interpret the relationships between NS-Sec, General Health, and Tenure.
4. **Data Filtering**: The data is filtered based on NS-Sec and Tenure Type dimensions. This filtering allows users to focus on the specific combinations of NS-Sec and Tenure Type that are of interest.
5. **Trend Lines**: In order to better understand the relationship between Display Measures-Tenure and Display Measures, trend lines are added to the scatterplot. These lines help to visually illustrate the overall trend and direction of the relationship, since both factors are significant.

For the last question of the Task 3, two scatterplots with tooltip information is used to explore the clustering of population quantities based on UMAP and t-SNE data projection techniques.

Initially, Principal Components Analysis (PCA) was considered for dimensionality reduction. However, the correlation matrix revealed that the data did not meet the linearity assumption, and experimental results showed that PCA did not perform well in reducing the dimensions of this dataset.(Jaadi, 2023) Therefore, this study primarily utilizes two data projection algorithms: t-SNE and UMAP.

T-SNE (t-Distributed Stochastic Neighbor Embedding) is an iterative non-linear algorithm(van der Maaten, 2023) that projects high-dimensional data into a lower-dimensional space.(WATTENBERG, 2016) It is used for the Household Tenure by Sex by General Health dataset, reducing it to two dimensions based on the population counts for the four tenure types.

UMAP (Uniform Manifold Approximation and Projection) is another non-linear algorithm that provides fast and direct projection on raw data without PCA preprocessing.(UMA, 2018) It is applied to the NS-Sec by General Health by Health dataset, reducing it to two dimensions based on the population counts for the different NS-Sec types.

Here is a brief justification for the scatterplots:

1. **Scatterplot**: The scatterplot is chosen to represent the reduced dimensions obtained through UMAP and t-SNE. Scatterplots are a great way to show the connections and groupings between data points from multiple dimensions. They make it easy to analyze and cluster data visually.

2. **Color**: The ”Label” attribute is represented by different colors for the marks. These colors correspond to specific label categories, making it easier for users to identify grouping patterns based on gender and health conditions.

3. **Tooltip**: To establish the connection between the reduced variables and the original data, tooltips are used to display additional information for each data point. The tooltips include a bar chart representing the population quantity of each NS-Sec or Tenure category associated with the current data point. Furthermore, the tooltip provides the location information (Local Authorities District) and label information (gender and health condition) for the selected data point.

With the scatterplot featuring tooltip information, users can easily examine the clustering patterns of population quantities based on gender and health conditions. By analyzing the position, shape, and color of the marks, users can easily identify patterns and clusters within the UMAP and t-SNE reduced dimensions. Additionally, the tooltip information offers more details about each data point, including the population quantities for various NS-Sec or Tenure categories, along with location and label information.

Through the visualization, users can observe how different gender and health condition labels contributes to the clustering population quantities, allowing for a deeper insight of the relationships between these variables in the dataset.

4.4 Dashboard 4 - The Comparison and Changes between 2011 and 2021

For the first question of the Task 4, the bar chart visualization is employed in Dashboard 4 to investigate the changes in the proportion of different NS-Sec social status categories between 2011 and 2021. Here is the justification:

1. **Bar Chart**: A bar chart was chosen to show the proportion of each NS-Sec category in 2011 and 2021, based on Munzner's theory that bar charts are useful for displaying quantitative values. The bar length allows for easy comparison between different NS-Sec categories and their proportion changes.
2. **Color Encoding**: The color encoding is used to differentiate between the two measures, 2011 and 2021. The color legend orange-blue is also used to allow users to distinguish between the two time periods easily.
3. **Stacked Marks**: The bar chart shows the combined proportion of NS-Sec categories for each year using stacked marks. This arrangement helps users see how the overall proportion changes and the relative contribution of each category for that year.
4. **Axis and Labels**: The NS-Sec categories are represented on the x-axis, while the measure values (proportions) are on the y-axis. The axes are labeled clearly and concisely for user understanding. To avoid clutter, the bars are automatically labeled, following Munzner's reduction theory.(Munzner, 2015)

For the second question of the task 4, there are two scatterplots used in the Dashboard 4 to compare the population changes based on NS-Sec social status or Tenure Type with general health status.

1. **Scatterplot**: The scatterplot is chosen to represent the relationship between two quantitative variables: population values in 2011 and population values in 2021, as scatterplots effectively display the distribution and correlation between variables.
2. **Shape Encoding**: Shape encoding is employed to represent the different general health types. The use of various shapes for each category makes it easier to visually distinguish them and enables comparisons to be made between the different general health types within the scatterplot.
3. **Axes and Labels**: The x-axis shows the total population in 2011, and the y-axis shows the total population in 2021. The labels on the axes help to make the comparison of variables clear.
4. **Trend Lines**: To show the connection between variables, trend lines are included in the scatterplot. This helps to identify the general trend and provides a visual aid for understanding how population values change over time.
5. **Data Filtering**: By filtering the data based on relevant dimensions, like NS-Sec, Tenure Type, and General Health Type, the users can focus on specific categories or combinations of categories. This feature enhances the exploration and analysis of population changes, making it easier to identify trends and patterns.

5 Conclusion

The visualization analysis has revealed several key findings. Firstly, from Task 1, we observe that higher socio-economic status individuals tend to concentrate in specific regions in the southern part of the UK. However, the highest socio-economic group does not necessarily correspond to the best general health outcomes. Individuals in the intermediate socio-economic group exhibit a relatively high proportion of good health.

Secondly, Tasks 2 and 3 highlight the relationships between Tenure, General Health, NS-Sec, and Sex. NS-Sec show associations with gender and health status, while tenure types exhibit a stronger link with General Health rather than gender. The heatmap visualization in Task 2 demonstrated a significant correlation between tenure types and NS-Sec classifications.

Lastly, Task 4 reveals changes between 2011 and 2021. Both the highest and lowest socio-economic groups increases in proportion, and there is a rise in rented accommodation. These findings underscore the complex interplay between socio-economic factors and health outcomes, emphasizing the need for further exploration.

This coursework has showcased how different visualization techniques and theories can be used to present information effectively. For example, the heatmap visualization, which follows Munzner's principles, displays the correlation between NS-Sec and tenure types in a clear and understandable way. Meanwhile, the scatterplots, which incorporates Ware's principles of visual perception, allowed for easy exploration of socio-economic factors and their clustering patterns.

In conclusion, through this coursework, I have gained valuable insights into the socio-economic problem being investigated. I have learned how information visualization can help to gain a deeper understanding of complex datasets. It emphasizes the significance of selecting appropriate visualizations and utilizing visualization theories to effectively communicate insights and support data-driven decision-making.

References

Zakaria Jaadi. 2023. A step-by-step explanation of principal component analysis (pca). Available: `https://builtin.com/data-science/step-step-explanation-principal-component-analysis`, Mar.

Tamara Munzner. 2015. *Visualization analysis and design*. CRC Press, Taylor Francis Group, CRC Press is an imprint of the Taylor Francis Group, an informa business.

2018. Umap: Uniform manifold approximation and projection for dimension reduction. Available: `https://umap-learn.readthedocs.io/en/latest/`.

Laurens van der Maaten. 2023. t-sne. Available: `https://lvdmaaten.github.io/tsne/`.

Colin Ware. 2008. *Visual Thinking for Design*. Elsevier Science Technology.

MARTIN WATTENBERG. 2016. How to use t-sne effectively. Available: `https://distill.pub/2016/misread-tsne/`, Oct.

Appendix

1. `https://www.nomisweb.co.uk/sources/census_2021_bulk`
2. `https://www.nomisweb.co.uk/sources/census_2011`