ADA Report

Contents **Page**

1. Abstract

This report aims to address the socio-economic problem which is to understand whether ethnicity influences an individual's occupational opportunities, educational attainment, and housing conditions, thereby creating disparities across different ethnic groups. Datasets have been gathered from the UK Census data 2021 and prepared for analysis through Tableau visualisations. Two dashboards have been created within Tableau that aim to address this socio-economic issue and are split into two different parts of analysis. Munzners' task taxonomy has been outlined to define the visualisations tasks. Justifications have been made for the visualisations that refer to principles of info vis, relevant aspects of human perception and cognition, with scientific literature to further validate my choices of visualisations. Data projection techniques including t-SNE and PCA have been used to aid the analysis. The main conclusion of the analysis was that ethnicity does not influence an individual's occupational opportunities, educational attainment, and housing conditions and disparities are not created across different ethnic groups. A lot has been learnt from this visual analytics task, primarily relating to the size and scale of these types of tasks.

2. Introduction

The socio-economic problem that I would like to address in the visualisations is to understand whether ethnicity influences an individual's occupational opportunities, educational attainment, and housing conditions, thereby creating disparities across different ethnic groups.

The first part of the analysis will focus on the interplay between ethnicity, occupation, and education. This will delve deeper into potential patterns and correlations between ethnic groups and their respective occupational and educational attainments. The objective of this part of the analysis is to explore whether disparities exist in the realms of occupation and education as a function of ethnic background. The second part of the analysis will concentrate on the relationship between ethnic groups and housing conditions. The objective of this part of the analysis is to determine if housing disparities exist across ethnic groups. The reason for separating the analysis into two parts is to provide a comprehensive understanding of the ways in which ethnicity influence these socio-economic factors. The findings from both parts of analysis will then be combined to provide a holistic view of ethnic disparities within these areas.

I am interested to understand and learn more about ethnic disparities and possible factors that could be contributing to barriers for ethnic groups. This analysis will help to uncover potential disparities and systematic biases that exist across different ethnic groups in these crucial life domains.

The audience for these visualisations will be mainly academics and researchers in fields such as sociology, demography and possibly even economics. This analysis will provide valuable insights for their research, particularly for those interested in ethnicity, social stratification, and inequality. The visualisations will include some data projections using dimensionality reduction techniques such as t-SNE and PCA, therefore knowledge of how to interpret these plots will be beneficial, however, this has been taken into consideration when designing the visualisations. Explanations will be provided in the visualisations to support users who do not have experience with such techniques.

3. Data Preparation and Abstraction

To conduct the analysis and create visualisations, multiple datasets were taken from the UK Census data 2021 (Office for National Statistics, 2021a). Two datasets were used for the analysis which were obtained using the 'add/remove variables' option on the ONS census data page (Office for National Statistics, 2021a). The first dataset contained the variables: 'Ethnic Group', 'Highest level of qualification', 'Occupation(current)' and 'Sex' (Office for National Statistics, 2021b). The second dataset contained the variables: 'Ethnic Group', 'Accommodation type' and 'Tenure of Household' (Office for National Statistics, 2021c). Both datasets contained two columns: 'Lower tier local authorities' and the corresponding area code which represents the geographical area where each 'Observation' was made. Not all UK Census data 2021 has been released to the public yet, and therefore when using multiple variables, data on some of the geographical areas was unavailable. Specifically, 265/331 areas were available for the first dataset and 320/331 areas were available for the second dataset. This was a limiting factor when collecting the data and it became apparent that the full dataset was not available for all geographical areas when selecting multiple variables. The intention was to include 'Sex' as a variable on the second dataset, however this was not possible without losing over half of the geographical areas.

After collecting both datasets for the visualisations, the next stage was to clean and format the datasets ready to be used as data sources in Tableau and for the data projections. Most of the data cleaning and formatting was completed in Excel. Columns that were not meaningful for the analysis and visualisations were removed. These columns included the label columns of each of the variables, for example 'Occupation' contains 10 categories and each category was given a corresponding label and this was not useful for the visualisations. Some variables contained a 'Does not apply' category which was removed by filtering out all observations that contained a 0.

After completing the data cleaning and formatting steps for both datasets, they could be imported into Tableau for analysis. The steps taken for data preparation are crucial to following a sound visual analytics process since it ensures the reliability of the analysis, which involves several data types. Most are nominal categorical variables such as Ethnicity, Occupation, Accommodation type and Sex, that are self-explanatory variables that provide information about an individual's background. There was one ordinal categorial variable, 'Highest level of education', that could be used to provide insight into an individual's educational attainment, being an indicator of their socio-economic status. The only numerical variable is 'Observation' which represents the number of observations for the combination of different categorical variables. A brief task analysis was then completed to gain insights and identify specific tasks that the analysis required in order to create relevant visualisations. This provided an insight into the type of questions that the user could have when looking at the visualisations, therefore ensuring that the appropriate visualisation methods and data projections were used. Careful data preparation and abstraction has provided the foundation for a robust and insightful analysis.

4. Task Definition

Munzner’s task taxonomy defines the visualisation tasks across three main dimensions: the ‘why’, ‘how’ and ‘what’ of a task. The first task analyses the relationship between ethnicity, occupation, education and sex and the second task analyses the relationship between ethnicity and housing conditions. Whilst both tasks are similar, they explore different relationships. The reason **why** the specific visualisations were selected for the tasks was to enable the user to **discover** something new such as trends, patterns, and relationships within the data. They allow the user to analyse the data in ways that give depth and detail to the visualisations that help answer specific questions. **How** the user will be able to interact with the visualisations will be through encoding data attributes such as colour, size and position. The user will be able to interact with the data in many ways that includes search, select, filters and navigation techniques that can help highlight specific trends or correlations in the data. It will further give the user the ability to hone in on details of certain visualisations to gain a deeper insight on a specific datapoint. The visualisations will allow users to derive new data such as percentages or averages which gives a different view of the data to help show relationships more intuitively. The **what** component of Munzner’s task taxonomy focuses on the structure of the dataset. Since the datasets are contained in an Excel file, the data structure falls under the **tables** category where each datapoint is represented by a combination of variables. Understanding the structure of the data is a key component for effective data visualisation and will provide an understanding of the type of visualisation methods to use. The dataset further contains special data which is represented by geographical data. Understanding this type of data will enable the creation of visualisations such as maps that have highlighters for specific areas in the UK.

5. Visualisation Justification

Before discussing the individual visualisations used in the analysis, this section will explain the reasons for the choices of background colour and text. Two dashboards were used to cover the analysis, with each dashboard representing a specific dataset. Dashboard 1 uses the first dataset and Dashboard 2 uses the second dataset. Black was selected as the background colour in both dashboards. This colour choice gives a focus with a natural aesthetic appeal to the audience. A black background with white text helps to create a striking contrast, making the visualisations stand out. The high contrast draws focus to the visualisations and brings them to the centre of attention. It allows the user to easily read labels, text, titles and other elements especially if they are small or intricately placed. It also allows different colours to stand out against the black background, making it easier to distinguish different hues and intensities.

In Ware’s book Information Visualization: Perception for Design (Ware, 2004), he highlights the importance of a visual hierarchy. The choice of white on a black background helps immediately establish this hierarchy and focuses the user’s attention to the most important elements of the visualisation. A consistent font style throughout the visualisations, Tableau Bold, offers high legibility and readability. This is a clean, straightforward and attractive font that allows users to remain focused on the content of the visualisations. According to Ware, using the bold font style enhances readability even at smaller sizes which is crucial where text elements such as labels and titles need to be easily readable. This establishes a visual hierarchy allowing the text to be more prominent when larger sizes are used in titles and headings. It also provides an indirect benefit for users that are familiar with Tableau’s aesthetic, contributing to brand consistency.

Two geographical maps were used in the visualisations, 1 map for each dashboard. Geographical maps are necessary for the analysis since the datasets include geographical areas in the UK, providing granularity at the lower tier local authorities level. Maps are one of the most intuitive and commonly used methods for visualising data that has a special or location-based aspect. They are excellent at providing special context which is important in the analysis since this can illustrate the geographical distribution of different ethnic groups and the correlation with factors such as education, occupation and housing conditions. Ware explains that humans have excellent spatial perception abilities which is used to better understand and interpret data. Ware discusses that positioning data points in a space can help create a mental map in the users’ mind, making it easier to understand relationships and patterns in the data. The maps used are based on lower local tier authorities and the observations are colour coded using a light-dark colour palette. This use of the colour palette allows the user to immediately recognise areas that have higher observations or lower observations based on the scale. This allows for easy comparison of different areas on the map.

Filters for the maps were added based on ethnic group and other factors such as Sex and Tenure of household. This allows users to look specifically at details required, giving them the freedom to explore many different relationships. According to Edward Tufte in his book: The Visual Display of Quantitative Information (Tufte, 2001), he argues that having a high data-ink ratio is desirable for good visualisations. He explains that it is not about removing ink entirely from the visualisation but more about avoiding unnecessary embellishments that do not contribute to understanding the data. This is why using geographical visualisations help to minimise the data-ink ratio, leaving more room for the actual data representation.

Another useful feature that further enhances the usefulness of maps as a visualisation is ‘highlights’ or ’search’. This feature was included for the maps since it allows the user to search on an area of interest. This provides efficiency as the user now has the ability to search for an area without having to navigate the whole map, especially as the map contains a higher level of granularity. It provides user engagement, focus and efficiency as discussed by Tufte and Ware. As stated in the data preparation, some geographical areas for both maps are unavailable. The use of a black background for the map along with the colour choices of green and blue on each map, make it very easy to distinguish which areas are missing from the map so that the user can clearly see without having to filter with additional steps.

This section will focus on development of the first dashboard which uses the first dataset. Since most of the data is categorical, it was important to focus on visualisations that would be able to explore the relationships between the categorical variables. One of the visualisations used was a treemap. This visualisation is an effective method of displaying hierarchical data especially within a limited space. Each rectangle in the treemap displays the percentage total of each occupation. A classification was added for each occupation based on whether it was considered a white or blue-collar job. This adds a further useful dimension to the treemap as it allows the user to compare occupation type based on the classification. For example, the user might want to investigate if there are a higher proportion of white-collar jobs in a specific area for a certain ethnic group. In a study by Wang et al (Wang et al, 2006) called Evaluating the Effectiveness of Tree Visualization Systems for Knowledge Discovery they analysed how the use of treemaps can be used for effective visualisation. By using a treemap, this aligns with the tasks that they found to be effective such as comparing different sizes of categories (occupation). The treemap design also adheres to the aesthetic criteria they have suggested such as space and clarity representation.

Another visualisation technique used was a heatmap. The heatmap created looks at the relationship between the highest level of qualification and occupation. The heatmap uses a diverging colour palette based off the % of total observations. This allows the user to easily recognise highest and lowest values for a specific combination of categories without having to manually highlight specific cells within the heatmap. The heatmap colour palette used was Green-Gold since there is a greater distinction in colours making it easier to observe lower and higher values that will be immediate points of interest to the user. A key reason for using this type of visualisation is due to the relationship between occupation and the highest level of education, when comparing ethnic groups. When using heatmaps it is important to understand that they will not work unless the data is normalised especially for large datasets where data may be skewed. Since the dataset contains significantly more observations for white ethnic groups, this was normalised by using a percentage of total observation. The heatmap created is completely interactive with the filters in the dashboard such as Sex and Ethnic Group and allow the user to drill down to specific details as required.

A list is another visualisation that was used. A list is a simple, yet effective visualisation that allows the ranking to be ordered by the highest level of qualification. The list is easy for users to interpret, and the percentage for each qualification, based on the filters and per geographical area, is visible by hovering over the map. The list is great for analysis since it allows users to quickly compare education levels and to highlight any disparities. From a cognitive perspective, a ranked list is naturally understood and is straightforward to observe the highest and lowest values. This cognitive pattern is known as the ‘top-to-bottom’ processing that makes a list universally intuitive. It reduces the cognitive load by presenting the data in a structured format and eliminates the need for the user to sort the data. Using a list as a visualisation is excellent for memory recall due to the ‘serial position effect’ where people tend to remember the first and last items of a list more easily. These ideas around human cognition and perception are highlighted by Card et al (Card et al, 2017) in The Psychology of Human-Computer Interaction where, whilst they do not discuss list visualisations specifically, their ideas inform the design of lists in general. There is one further visualisation on the first dashboard which is a t-SNE plot, and this visualisation will be discussed in a later section around data projection methods.

The second dashboard uses the second dataset and looks to explore the relationship between ethnic group and housing conditions. This dashboard also utilises visualisations that allow categorical data to be explored. The first visualisation used on this dashboard was a Marimekko chart. Marimekko charts (also known as mekko charts) are a more advanced type of visualisation and can be more difficult to interpret. The Marimekko chart is great in visualising multiple dimensions at once and helps provide an overview of the analysis on this dataset. The Marimekko chart displays multiple dimensions, where the x-axis position highlights the ethnic group, with the width showing the total observations for the whole dataset in that ethnic group – note that White: English, Welsh etc will have a very large bar since there are many observations for that specific ethnic group. The y-axis or height of the bar indicates the total percentage of observations based on accommodation type, with the colour representing a specific accommodation type. The Marimekko chart is fully interactive with the geographical map and will change based on the area highlighted. This type of visualisation is helpful for analysis since it provides a great comparison for all types of variables in the dataset and gives a proportional display across ethnic groups. The Marimekko chart utilises pre-attentive processing which involves the brain's ability to identify visual properties such as colour and size rapidly and accurately. This chart follows two specific principles based on Stephen M. Kosslyn's book: Graph Design for the Eye and Mind (Kosslyn, 2006) which are the principles of capacity limitations and proximity compatibility. Related categories are grouped together in the visualisation to avoid making the chart too cluttered or complicated, so that it does not overwhelm the users cognitive processing abilities.

Another type of visualisation used on the second dashboard is a horizontal bar chart. Two bar charts parallel to each other are used on the dashboard where they display the ranking order of Tenure of Household and Accommodation based on the percentage of total observations, filtered by Ethnic Group and are fully interactive with the geographical map. This type of visualisation allows the user to intuitively compare the percentage of total observations for each type of category and shows a clear ranking from biggest to smallest based on the percentage, which clearly communicates the proportion of each category for easy comparison. As per Ware's research, humans are generally good at comparing the lengths of bars, and this is recognised by the use of a horizontal bar chart. Horizontal bar charts allow the user to efficiently utilise space on the dashboard where the width can be expanded without it becoming cluttered that might be the case if using a vertical bar chart. This type of visualisation is similar to the list visualisation on the first dashboard except there is a greater visual aspect to it, which reduces the cognitive load. The choice of displaying the visualisation parallel to each other makes it easier to compare the relationship between the two variables, by quickly glancing back and forth to check for certain trends or relationships between accommodation type and tenure of household.

For the analysis, two different types of data projection were used. The t-Distributed Stochastic Neighbour Embedding (t-SNE) was used for the for the first dashboard and Principal Component Analysis (PCA) was used for the second dashboard. For the t-SNE plot, multiple variables were used that include Lower tier local authorities, Ethnic group, Highest level of occupation, Sex and Observation. Each of the categorical variables were encoded with a label encoder so that they can be used for the dimensionality reduction along with the single numerical variable. In particular, t-SNE is a non-linear technique that is very good for the visualisation of high-dimensional data. Several variables were used in this method, which resulted in a large number of dimensions to reduce. Furthermore, t-SNE preserves the local structure of the data. This is useful since it keeps instances that are close in the high-dimensional space close in the low-dimensional space. Clusters of instances in the data can be visualised and can help to show relationships between the variables and possible underlying patterns that are not immediately apparent through other types of visualisations.

For the PCA plot, only two categorical variables for the data projection were used, being Lower tier local authorities and Accommodation type, along with the numerical variable observation which is necessary for the projection. PCA can help identify key patterns, for example, the first principal component will indicate the accommodation type that accounts for the most variance across local authorities. This can help to identify which accommodation types are more relevant in distinguishing between different local authorities. In addition, PCA has the ability to detect outliers or anomalies in the data. Outliers in PCA are useful in representing differences from the norm. For example, an anomaly in a projection can help to identify local authorities that have a distribution of accommodation types that are different from the norm.

6. Conclusion

Through the visualisations the user will be able to learn about the socio-economic issues influenced by ethnicity. The visualisations will enable the user to understand if ethnicity influences an individual's occupational opportunities, educational attainment, and housing conditions, thereby creating disparities across different ethnic groups.

From the visualisations it can be concluded that there is no significant indication that ethnicity influences an individual's occupational opportunities or educational attainment. Firstly, by filtering through different ethnic groups, this shows that all ethnic groups have more than 39% of their population obtaining a level 4 qualification or above, with the lowest being White: English, Welsh etc, however, they have a higher percentage with level 2 and 3 qualifications than other ethnic groups. All ethnic groups have lower than 17% of their population with no qualifications, with Other Ethnic Groups having the highest percentage. This shows that all types of ethnic groups have fair opportunities in the UK to be able to achieve a decent level of education.

It is noticed that for some ethnic groups who have a higher percentage of no qualifications, that they tend to have a higher percentage that hold a level 4 qualification. This shows that all ethnic groups clearly have the opportunity to gain that higher level of education, but some prefer not to as a choice. Similar findings can be seen for occupation type. Professional occupations hold the highest percentage for every type of ethnic group. This shows that there is a correlation between education and occupation where the higher level of education achieved, the higher likelihood of obtaining a professional occupation or at least a white-collar job. This relationship, along with others can be clearly identified in the heatmap.

Overall, the easiest way to get an overview of the relationship between ethnic group, education, occupation and sex is from the t-SNE plot. The t-SNE plot shows instantly, that there is no correlation and that datapoints are scattered. When looking at the different types of ethnic groups through the colour legend, it can be seen that many datapoints for each ethnic group overlap with other ethnic groups. This indicates that there are no significant disparities, otherwise there would be clear clusters of some ethnic groups separated from others. By filtering through the different occupation types, it is clear that there are some outliers, but this is true for all ethnic groups and there are not enough outliers to significantly identify disparities between ethnic groups.

An analysis of the second dashboard shows that there is no significant indication that ethnicity influences housing conditions. There are some slight differences between ethnic groups in terms of the accommodation type and tenure of household. Ethnic groups such as Asian, Mixed, White: English etc and White: Irish tend to live in a house or bungalow rather than a flat or apartment, however, this can be explained due to number of observations of that ethnic group on the map. These ethnic groups are more populated in areas outside of cities such as Cornwall where price houses tend to be lower than some places like London that is more densely populated with Black and other ethnic groups that tend to live in flats or apartments. Similar can be determined for tenure of household where the same ethnic groups that live in houses tend to own their accommodation type compared to renting. This is highlighted through the map where areas in London will have more flats and apartments that are more likely to be rented rather than owned. For all ethnic groups, the lowest observations for tenure of household are socially rented and for accommodation type is caravans or mobile homes.

This coursework on information visualisation has been invaluable in my learning since it has enabled me to complete the full cycle of a visual analytics project. One key area to highlight is the importance of Data Projection techniques. When dealing with large datasets with numerous variables, it is important to be able to capture the relationships between them without being overwhelmed with information. Using data projections such as t-SNE and PCA has allowed me to understand the importance of dimensionality reduction since they are a useful tool for highlighting characteristics of the data without overwhelming the user. They can help to give an overview of the relationships between the variables by identifying clusters that can often be obvious to an average user.

I have learnt about the importance of using clear visualisations. This is critical for effective analysis and needs to be aligned with techniques that work with the relevant features of human cognition and perception that allow the user to be able to understand the visualisations in a way that does not require a lot of thought. This can include using visualisations that are interactive and using effective colour and sizing to highlight key points in the dataset. A key learning area has been through the data preparation and abstraction. I underestimated how long this process can take and how vital it is for ensuring accuracy and reliability when undertaking the analysis. I now have a better understanding that this is one of the key aspects of a visual analytics project and should not be taken lightly. Gathering the data can sometimes take as long as actually creating the visualisations and I have learnt that it is important to understand exactly the analysis required so that the correct datasets can be gathered and used accordingly. This project has given me a major insight into information visualisation as a whole and has improved my knowledge, understanding and use of visualisation methods that will be of significant value when applied in the real world.

7. References

Card, S., Moran, P.M. and Newell, A. (2017). *The Psychology of Human-Computer Interaction.* Available at: https://www.pdfdrive.com/the-psychology-of-human-computer-interaction-e176223153.html (Accessed 5 May 2023)

Kosslyn, S.M. (2006). *Graph Design for the Eye and Mind.* Available at: https://academic.oup.com/book/11432?login=false (Accessed 10 May 2023)

Office for National Statistics (2021a). *Census*. Available at: https://www.ons.gov.uk/census (Accessed 20 April 2023)

Office for National Statistics (2021b). *Census*. Available at: https://www.ons.gov.uk/filters/fccc9562-d767-4aa4-bd33-fc926bb75ab4/dimensions/change?f=browse#dimensions--added (Accessed 20 April 2023)

Office for National Statistics (2021c). *Census*. Available at: https://www.ons.gov.uk/filters/449a2327-63de-45f9-a4e5-37b011a27fa2/dimensions (Accessed 23 April 2023)

Tufte, E.R. (2001). *The Visual Display of Quantitative Information – Second Edition.* Available at: http://www.econ.upf.edu/~michael/visualdata/tufte-aesthetics_and_technique.pdf (Accessed 30 April 2023)

Wang, Y., Teoh, S.T. and Ma, K.L. (2006). *Evaluating the Effectiveness of Tree Visualization Systems for Knowledge Discovery.* Available at: https://diglib.eg.org/bitstream/handle/10.2312/VisSym.EuroVis06.067-074/067-074.pdf?sequence=1&isAllowed=y (Accessed 30 April 2023)

Ware, C. (2004). *Information Visualization: Perception for Design: Second Edition*. Available at: https://www.researchgate.net/publication/224285723_Information_Visualization_Perception_for_Design_Second_Edition/link/53ce92780cf24377a65dd024/download (Accessed 28 April 2023)