EMATM0067: Visual Analytics Coursework

Exploratory analysis of the relationship between economic activity and education in England & Wales

Student ID: 2353759

Contents

Abstract

In the process of demonstrating that the UK's ageing population has had an impact on the country's economic activity between 2011 and 2021, few outliers indicated that age might not be the only factor in the decrease in the proportion of the employed population. One factor that was hypothesised to explain this decrease in the outliers was education due to its intrinsic link with work. This inspired an exploratory investigation to understand the relationship between economic activity and education. Using the 2021 England and Wales census data, and Munzer's task taxonomy, a Tableau storyboard was constructed to not only explore the age/gender relationship between economic activity and qualification level but also explore the relationship between working conditions and qualification level of the population. It was found that whilst the younger generation (below 50) is more likely to have a higher level of qualification than the older generation, overall more women than men are degree-level qualified. Moreover, it was found that those with degree-level qualifications are more like to work from home than travel to work which could suggest the high-skilled are less likely to be location dependent. Regardless of the qualification level, most people in England and Wales work between 31 to 48 hours per week.

Introduction

The UK has a growing ageing population (Bayliss & Sly, 2010), which is a threat to the economic output of the country. With this fact, it was hypothesised that this demographic shift would suggest a reduction in the working population. Preliminary analysis of the 2011 and 2021 census data (Office for National Statistics, 2011, 2021) for England and Wales validated the initial hypothesis to an extent; however, there were several regions which suggested that age was not the only reason for a decrease in the employment rate. As such it was thought that education might be a contributing factor due to its intrinsic link with work (Levin, 1987). This gave rise to the overarching question for exploratory analysis: what is the relationship between economic activity and education in England and Wales in 2021? To answer the main question, two sub-questions were curated based on the available census data:

1. What is the age/gender demographic relationship between economic activity and qualification level?
2. What is the relationship between working conditions and qualification level?

Using Munzer's task taxonomy (Munzer, 2014) a Tableau storyboard, consisting of three dashboards, was produced to demonstrate the preliminary analysis and the two sub-questions. The visualisation was made with the consideration that the primary user would be a non-domain specialist in the given socio-economic areas who may also use the storyboard to answer personal research questions or for exploration. As such the three key requirements were:

- Visualisations must be easy to interpret in terms of what each view shows
- Filters/interactivity must be kept intuitive
- Dashboards can be used as standalone visualisation tools for personal exploration

This report follows Munzer's 'What-Why-How' structure (Munzer, 2014) to explain the steps taken for building the visualisation. 'Data Preparation and Abstraction' addresses 'What' is shown to the user. 'Task Definition' addresses 'Why' the user is looking at what they see. 'Visual Justification' addresses the 'How' by justifying the visual encoding/interaction of idioms using information visualisation principles. Additionally, the 'Evaluation' section summarizes the results of a small validation study to assess the suitability of the visualisation for the primary user.

2 Data Preparation and Abstraction

The dataset type used in the exploratory analysis was only in the form of static tables. Moreover, geographical data obtained for the datasets was selected to have a Middle Layer Super Output Areas (MSOA) level of detail. The reason for choosing this level of granularity was to ensure that the data was representative of a maximum area in England and Wales, which in turn would make conclusions more reliable.

2.1 Preliminary Analysis

To conduct the preliminary analysis, the 2011 and 2021 data underwent significant data preparation. To acquire the economic activity data for both years, the information was queried from the Office for National Statistics (ONS) website and the Nomis website (Office for National Statistics, 2011, 2021).

The data acquired for 2011 data was extracted as a single CSV file with three tables, each referring to a different level of economic activity (total, employed, unemployed) and in their pivot form with age range categories as columns. The tables for 'total' and 'employed' economic activities were extracted manually in two separate CSV files. The MSOA code was separated from the MSOA name and two additional columns were added; one stating the year and the other stating the economic activity. Both tables were unpivoted to group observation counts by age range so that the tables are in their long format. Next, the

tables were merged by stacking one over the other. Figure 2.1 summarises the above cleaning tasks at a high level.

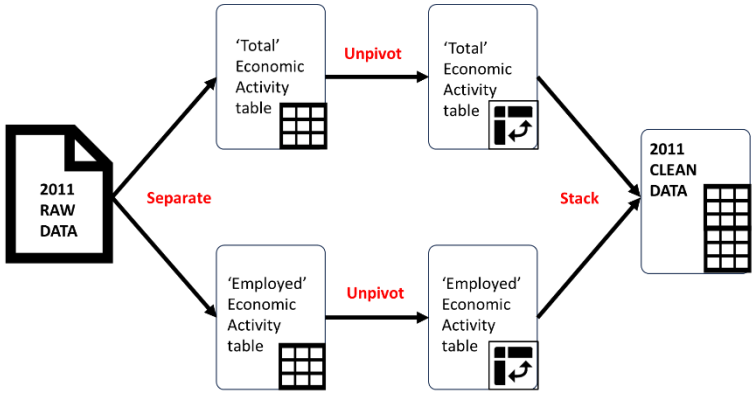

Figure 2.1: Visual summary of the data preparation steps on 2011 raw data.

The data acquired for the 2021 data required minimal preparation since it came in an unpivoted form. The 2021 data also needed an additional column to specify the year. Given that the 2021 data had an additional age range of 'Aged 15 years and under', this had to be removed to make it consistent with the 2011 data.

To merge the 2011 and 2021 data into a single CSV file, both tables were 'inner' joined on the column containing MSOA code. The resulting table was split into the respective years and stacked one over the other to ensure it was in its long format in Tableau. Figure 2.2 visually summarises the aforementioned process and Table 2.1 shows the data abstraction for all three dashboards.

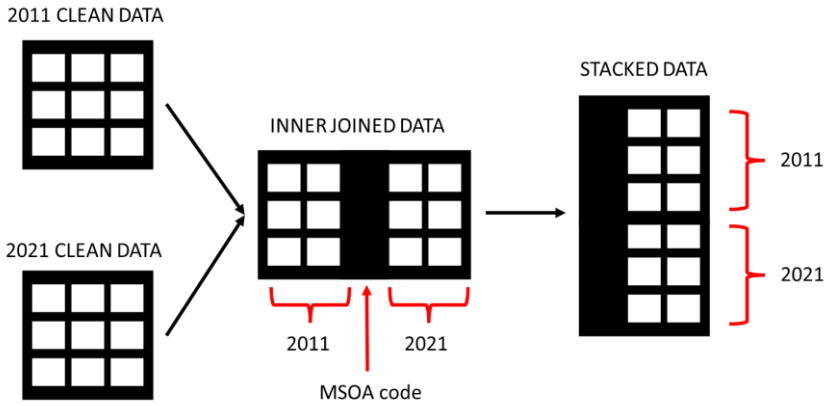

Figure 2.2: Visual summary of the data preparation to merge 2011 and 2021 data in long format table.

Table 2.1: Attributes (and types) used in the dataset for each of the three dashboards.

Attribute	Year	Middle Layer Super Output Areas Code	Middle Layer Super Output Areas	Age	Sex	Economic Activity	Highest level of qualification	Distance travelled to work	Hours worked	Observation
Attribute Type	Ordinal	Nominal	Nominal	Interval	Nominal	Nominal	Nominal	Nominal	Interval	Ratio
Preliminary Analysis	✓	✓	✓	✓		✓				✓
Sub-question 1		✓	✓	✓		✓	✓			✓
Sub-question 2		✓	✓		✓			✓	✓	✓

To generate the choropleth map in the first dashboard, three calculated fields were generated: 2011 Employed Ratio (ER), 2021 Employed Ratio (ER), and Percentage Delta (%Δ). The relationship between these three fields is described in equation 2.1. As seen in the dashboard the map has 2 null values. These null values are as results of non-identifiable MSOA based on Tableau's geocoding package. It is also observable that there are more than 2 empty regions in the graphs with no data and this is because the original data does not cover the entirety of England and Wales.

$$\begin{aligned} \%\Delta &= (2021\ ER - 2011\ ER) \times 100 \\ &= \left(\frac{\sum Employed\ Count\ 2021}{\sum Total\ Count\ 2021} - \frac{\sum Employed\ Count\ 2011}{\sum Total\ Count\ 2011} \right) \times 100 \end{aligned} \tag{2.1}$$

2.2 Sub-questions 1 & 2

For sub-questions 1 and 2, the original data required minimal cleaning which mainly involved removing any 'Does not apply' fields from attributes or removing 'Aged 15 years or under' counts. The map produced in the second dashboard has 184 null values for the same reasons as the map in the first dashboard. The number of null values is significantly larger in the second dashboard because the majority of the null values from the 2021 dataset were removed when it was inner-joined with the 2011 dataset for the first dashboard.

Data preparation was done on the cleaned dataset for each of the sub-questions to be able to apply dimensionality reduction. As seen in Table 2.1, the majority of attributes were categorical which cannot be used to produce clusters and therefore the tables need to be pivoted (summarised in Table 2.2). t-SNE and UMAP techniques were selected since both can be used for clustering. However, sub-question 2 uses UMAP instead of t-SNE because UMAP is faster with larger datasets (McInnes et al., 2018). The dataset for sub-question 2 has approximately 146,000 more data points than the dataset for sub-question 1. In both reductions, it was decided to cluster the data using the MSOA and the Highest level of qualification to be able to identify regions which are anomalies.

Table 2.2: Summary of the pivot table structure used for the dimensionality reductions.

	Dimensionality Reduction	**Pivot Table Format**		
		Index	**Columns (i.e. dimensions reduced)**	**Aggregate Values**
Sub-question 1	t-SNE	Middle Layer Super Output Areas, Highest level of qualification	Sex, Age, Economic Activity	**Sum** of Observation
Sub-question 2	UMAP		Sex, Hours worked, Distance travelled to work	

3 Task Definition

For each of the dashboards, the task definitions are described as an 'Action-Target' pair, where keywords are italicised, as suggested by Munzer (2014). All dashboards are designed so that they can also be used as standalone. Therefore, any categorical labels which are not self-explanatory are accompanied by further description and legends are appropriately used.

3.1 Preliminary Analysis

The primary task definition of the first dashboard was to *present* the *dependency* between Economic Activity on Age by *comparing* the change in the *distribution* of Age between 2011 and 2021. Dependencies are presented by having a choropleth map showing a change in the percentage of employed population and a 100% horizontal stacked bar which shows the age distribution for each of the years. Note that the values in the views are purposely represented as percentages only because it was identified that the number of data points in 2021 was less than in 2011 even though the population was estimated to have increased by 6.5% for England and 1.5% for Wales in the decade (Park, 2022). This suggested a smaller proportion of the

population was captured in the 2021 census so having absolute numbers would distract the user from the original task definition.

A secondary task definition was to *explore* the data for any *outliers*. This is achieved by allowing users to select one or more regions on the map and view the age distributions. Moreover, the extremes of the distributions are captured in the form of a table that represents the 'Top 10' and the 'Bottom 10' values which in turn would also help in *identifying outliers*.

3.2 Sub-question 1

The purpose of the second dashboard was to follow up on the preliminary analysis to understand more about the demographic characteristics of the 2021 population for economic activity and qualification level. As such the primary task definition would translate to *discovering* the age/gender demographic *relationship* between Economic Activity and the Highest level of qualification. This is achieved by having a doughnut chart of age ranges which can act as a filter for the bar charts showing information on economic activity and qualification level. The labels in the doughnut chart are in the form of percentages as proportions are the most useful aspect of the view, but absolute values are also encoded in the tooltip should more information be required. The categories in each of the bar charts can also act as filters for the remaining views except for the t-SNE data projection. A sub-task of this primary task was to *compare* the *distribution* of the sexes for different categories of economic activity and qualification level so a grouped bar chart was used to be able to split the information into the respective genders. Bar charts only show absolute values because the information is derived from one data set (2021) but also because percentages with grouped bar charts make it difficult for readers to understand. To overcome this issue a choropleth map is used again to present the distribution of the percentage of employed population which summarises the information in the Economic Activity bar chart. The secondary purpose of the map was to act as a filter for location should the user want to explore individual areas.

A secondary task definition was to *identify* any *outliers* in the data. This is achieved by a t-SNE data projection reducing dimensions of Sex, Age and Economic Activity so that each point is a combination of MSOA and Highest level of qualification. To serve the purpose of this task definition, the scatter plot is encoded so that it can filter information on the remaining views but none of the other graphs can filter information on the t-SNE chart.

3.3 Sub-question 2

The purpose of the third dashboard was to understand the working conditions of those employed in England and Wales. As such the primary task definition was to *discover* any *relationship* between qualification level, hours worked and distance travelled to work. This is achieved by allowing the users to select one or more bars in each chart to act as a filter in the remaining views except for the UMAP reduction. The filters can be layered such that the bars can be selected from one chart and then to another. Note for personal *enjoyment* of *data*, users can also add additional filters such as geographic regions or sex.

Similar to the previous two dashboards, the secondary task definition was to *identify* any *outliers* in the data. This is achieved by a UMAP data projection reducing dimensions of Sex, Hours worked and Distance travelled to work. As with the t-SNE, the UMAP projection has a one-way filter action which allows it to filter the other views but not vice versa.

4 Visualisation Justification

The majority of the storyboard consists of easy-to-interpret visualisations. Therefore, there are many bar charts of different types since most people are taught to read a bar chart from a very young age (Berkeley Library, 2024). What makes bar charts easy to read is the fact the human eye can distinguish the differences in length much more accurately (Munzer, 2014). On the first dashboard, the 100% horizontal stacked bar chart representing age groups has percentage labels also included since the lengths of categories are very

similar which can make it harder to differentiate the proportions. The only reason why age ranges were presented as a doughnut chart in the second dashboard is because its primary purpose was to act as a filter so compactness was prioritised over comparing proportions of categories. As compensation, percentage labels were used to give a quick idea of proportions to the user. Since age ranges are represented across two dashboards, the colours for the categories were kept consistent so that the user could understand the same information without having to look at the legend. This was applied to any attributes that are displayed in more than one dashboard. A blue-orange colourblind colour palette was chosen for the age categories so the visualisation is interpretable by a colour-blind person (Munzer, 2014). The colours were strategically assigned to different age ranges such that adjacent categories were different in either hue or saturation or both (Munzer, 2014). On the second dashboard, grouped bar charts were used instead of a stacked bar chart to have sex categories side-by-side which makes it easier to compare proportions (Munzer, 2014). The colours for gender were chosen as blue for males and pink for females since those hues are often associated with genders (Scholarly Community Encyclopedia, 2022) which makes it intuitive for the user to understand gender without looking at the legend.

The second type of graph used was a choropleth map, as seen in the first two dashboards. The choropleth map in the first dashboard has a diverging blue-orange colour scheme for not only colour-blind users but also to distinguish a net increase (blue) or decrease (orange) (Munzer, 2014). The colour legend is stepped instead of having a smooth gradient so it is easier for users to identify the shade of the colour for comparison (Munzer, 2014). To select a suitable number of steps, the maximum absolute value was rounded up to the nearest 10^{th} and then divided to have an even number of steps. In the case of the first choropleth map, the maximum absolute value was 16.36% which was rounded to 20% and then decided to increment in 5% to have 4 steps in each colour. The choropleth map in the second dashboard also has a diverging blue-orange colour scheme to allow for more steps in the legend as opposed to when having just a single colour. The legend has 10 steps in increments of 10% from 0% to 100%, with the midpoint being 50% where the colours change hue.

The final type of graph used was a scatter plot for the data projections, as seen in the second and third dashboards. Given there were many data points in the projections and that some clusters partially overlapped, the two deciding factors were the size of the scatter points and the colour. The size of the points was found by trial and error such that the outlier points were not overlayed with other points. Similarly, the lighter colours were ordered to be always in front of darker colours.

Both the choropleth graphs and scatter plots were produced with Shneiderman's mantra of overview first, zoom and filter, details on demand (Munzer, 2014). Both types of graphs are encoded with a colour scheme and legend to give an overview of the big picture followed by the ability to zoom in on individual geographic regions or outlier scatter points. By hovering, further details are provided on demand and selecting the point would allow other graphs to filter to provide specific details.

5 Evaluation

A key aspect of Munzer's method is to validate the visualisation (Munzer, 2014). As per Munzer, there are four levels of validation: domain situation, data/task abstraction, visual encoding and algorithm (Munzer, 2014) and so the storyboard was evaluated by four peers. Considering the user requirements, as mentioned in the introduction, and the four levels of validation, a questionnaire was produced. The questionnaire mainly focused on assessing data/task abstraction, visual encoding and algorithm. Table 5.1 summarizes the quantifiable results of the questionnaire. Additionally, based on the written feedback from the peers, it was evident that the hardest views to interpret were the dimensionality reductions. Moreover, there were some comments on the choropleth map being too small or in the wrong orientation and also being too granular in terms of the regions.

Based on the results, it was evident that most visualisations in the storyboard were easy to interpret which validated that the construction of idioms was to a good standard. It is evident with the long loading times

of the workbook that the algorithm was not very quick to execute. This was believed to be due to the large CSV files that were used but also because the number of items in each table was proportional to the granularity of geographic areas chosen (i.e. MSOA). Therefore, if a less granular level of detail was chosen, not only would it make the maps easier to interpret but also reduce the number of unique area codes. This problem highlights that the data needs to be transformed further by grouping the geographic areas for better data abstraction. The dimensionality reductions were hardest to interpret because the original projections (Figures 5.1 and 5.2) reduced categorical data from the unpivoted tables which would not form proper clusters. This problem was identified and rectified as in the current visualisation. However, due to the high level of granularity of geographic regions, the number of points on the projections is still too high so it can be critiqued as the root cause of poor idiom construction. Additionally, the choropleth map in the second dashboard was changed from landscape to portrait so it is easy to view (see Figure 5.3).

Table 5.1: Summary of quantifiable results of the questionnaire.

Question Topic	Level of Validation	Average Result
Ease of interpretation	Data/task abstraction	8/10
Visual appealingness	Visual encoding	9.25/10
Choice of colours	Visual encoding	9/10
Time taken to open the workbook	Algorithm	8 seconds
Total time taken to load individual dashboards	Algorithm	14 seconds

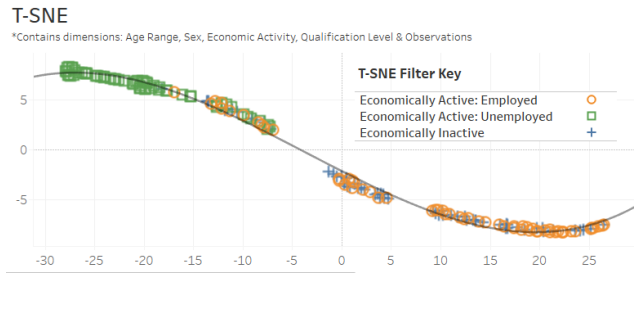

Figure 5.1: Original t-SNE data projection.

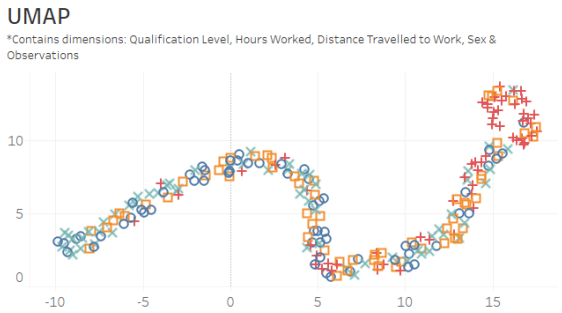

Figure 5.2: Original UMAP data projection.

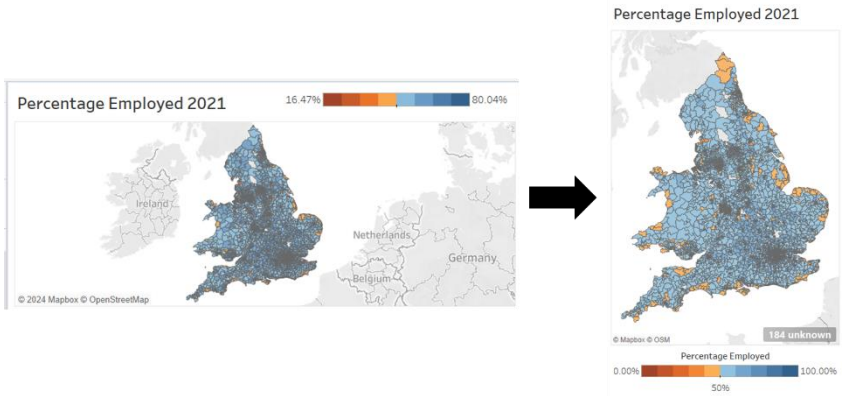

Figure 5.3: Change in orientation of the choropleth map in the second dashboard.

6 Conclusion

The work in this report summarised the use of Munzer's task taxonomy to develop a Tableau visualisation to answer the question: what is the relationship between economic activity and education in England and Wales in 2021?

Below are some key findings of the preliminary analysis and the question above:

- Overall, there is a decrease in the percentage of the employed population between 2011 and 2021 mainly due to an aging population.
- Areas such as Leeds 111, South Staffordshire 006, Lancaster 014 and along with some more regions from the 'Bottom 10' table on the first dashboard suggest that the ageing population is not the only reason for a decrease in employment, which formed the basis of the original question.
- More men are employed than women, even at older age ranges (50+); however, more women are economically inactive than men which is due to a higher life expectancy of women and therefore there is a larger population of women above the age of 65.
- There is a smaller proportion of level 4 qualified (degree level) in relation to other qualification levels with the older generation (50+) than the younger generation (under 50). This could be due to advancements in technologies which require employees to be more knowledgeable and skilled.
- More women are level 4 qualified than men.
- Most people in England and Wales work between 31 to 48 hours per week regardless of their qualification level.
- As people's qualification level increases from 1 to 4, the proportion of people working from home increases, suggesting that highly skilled jobs are less location-dependent.
- People who travel less than 10 kilometres but do not work from home are least likely to work longer (49+) hours. On the other hand, people who travel more than 30 kilometres are most likely to work longer hours, which could be because they prioritize personal time less when they are far from family.

On a more reflective level, the three key lessons learnt from conducting a large-scale data visualisation using Munzer's task taxonomy were:

1. Data/task abstraction is crucial and has a cascading effect on how well the algorithm performs but also how easy it is to interpret the visualisation. This was learned as a result of choosing an incorrect level of granularity for the geographic data which made maps harder to interpret but also made the visualisations slower.
2. A key challenge of any exploratory data analysis is finding the right data in the right format. Therefore, data cleaning is extremely important even though in most cases it consumes the most time in the project.
3. Developing a visual dashboard is an iterative process which requires continual feedback from the user to ensure that the requirements set are fulfilled to a satisfactory level.

References

Bayliss, J., & Sly, F. (2010). Ageing across the UK. *Regional Trends*, *42*, 2-28.

Berkeley Library. (2024). *Library Guides: Data Visualization: Choosing a Chart Type*. University of California Berkeley. Retrieved 17/05/2024 from https://guides.lib.berkeley.edu/data-visualization/type#:~:text=Inc.%2C%20forthcoming.-,Bar%20Chart,good%20for%20showing%20exact%20values.

Levin, H. M. (1987). Work and education. In *Economics of education* (pp. 146-157). Elsevier.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Munzer, T. (2014). *Visualization Analysis & Design*. Taylor & Francis Group.

Office for National Statistics. (2011). *2011 Census*. Retrieved 12/05/2024 from https://www.nomisweb.co.uk/sources/census_2011

Office for National Statistics. (2021). *2021 Census - Office for National Statistics*. Retrieved 12/05/2024 from https://www.ons.gov.uk/census

Park, N. (2022). *Population estimates for the UK, England, Wales, Scotland and Northern Ireland*. Office for National Statistics. Retrieved 15/05/2024 from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2021

Scholarly Community Encyclopedia. (2022). *Gendered Associations of Pink and Blue*. Scholarly Community Encyclopedia. Retrieved 17/05/2024 from https://encyclopedia.pub/entry/30019