Visual Analytics Coursework

Which industries have seen a rise in employment, and are these AI-related jobs?

Abstract:

Abstract:
This report deals with the employment shifts across industries in England and Wales between 2011 and 2021, with a particular focus on identifying IT sectors and whether they are related to artificial intelligence (AI). Using census data from both years, I have done data cleaning, Bayesian modelling for missing value imputation, and dimensionality reduction techniques (PCA, t-SNE, and UMAP) and finally tableau to identify patterns in employment distribution. The analysis shows important growth in the technology sector, scientific research and a dip in low tech sector with evidence showing co relation between growth areas, educational qualification and AI- related occupation. Educational qualifications emerge as a important socioeconomic indicator, with distinct relationships between qualification levels and access to growth employment sectors. Interactive visualizations show regional variations in employment transformation, highlighting a widening socioeconomic divide between areas with high concentrations of knowledge intensive jobs and those dominated by low tech sectors. The findings indicate that while AI related sectors show a substantial growth, access to these opportunities remains unequally distributed across educational groups and geographic regions which highlight the existing socioeconomic inequalities.

Introduction:

The advancement in the development of artificial intelligence AI is mind boggling and over the last couple of years, there has been a huge debate about the impact of AI on employment and socio-economic trends and structures. As AI is spreading its wings, certain jobs face an automation and new roles are coming up to manage these technologies. This has shifted the landscape of the labour market and also has had socio-economic impact in the form of job losses and education quantification patterns change.

The time between 2011 and 2021 censuses shows a critical decade in AI development, with important advances in machine learning, natural language processing and computer vision. I have done this analysis keeping in mind the coming future generation who can predict on the basis of this report, which industrial sectors have more scope in the future and what educational qualifications are more demanding as time advances. Also, with the development of AI, UK economy especially after financial crisis, Brexit and the COVID-19 pandemic. We need to understand the employment shifts during this period as it provides important insights into how technological change interacts with economic trends to change the labour market dynamics.

My Research question and Socio-economic importance:

This report deals with insights into a primary question “**Which industries have seen a rise in employment, and are these AI-related jobs?** It has objectives:

1. Identify industries that experienced significant employment growth between 2011 and 2021.

2. Determine whether these growth sectors correlate with AI-related fields.
3. Notice the regional variations in the employment patterns.
4. Assess the socioeconomic implications of the employment shifts by looking at the education qualification changes.

The socioeconomic importance of this topic extends beyond simple labour market statistics. Changes in the industrial employment patterns can:

- Drive income inequality as high skilled AI-related positions often demand premium wages.
- Create geographic disparities between regions and attract new industries.
- Reshape educational requirements, and exclude workers without specific qualifications.
- Affect social mobility by creating different opportunity landscapes across regions.
- Transform community structures as employment patterns influence housing, servicing, and local economics.

Target Users and applications:

This report is made in a way that it addresses policymakers, economic analysts, and education planners who need to understand changing employment landscapes to make appropriate changes. Key user groups and specific needs include:

- Government agencies: geographic distribution of growth sectors to target interventions, analysis of qualification requirements to design education initiatives.
- Educational institutions: they need to adapt the curricula to emerging employment opportunities and understand regional employment patterns to develop local course offerings.
- Career advisors: guide individuals through changing job markets and get a clear visualisation of qualification requirements for growth sectors.
- Businesses: in making strategic decisions about workforce development, they have here, a regional analysis of skills.

The above-mentioned stakeholders require clear visualisations that allow then to identify the trends in educational requirements, compare regional opportunities, and explore the relationship between industry and growth and potential AI influence, all this while they consider the socioeconomic changes.

Data Preparation and Abstraction:

2.1 Data sources and Selection criteria:

The analysis is mainly from 2011 and 2021 census in England and Wales, mostly focusing on tables selected for their importance to industry trends, educational qualifications and the socioeconomic impact.

2021 Census Tables:

ts060-msoa

ts067-msoa

ts063-msoa

2011 Census Tables:

QS605EW

QS501EW

DC6604EW

These tables were selected because:

- Initially I selected 10-11 potentially valid tables as I needed to look deeper into them.
- Did the exploratory analysis to find out information content and quality.
- Final selection was based on completeness, comparability, and relevance to the research question, with particular attention to educational qualification data.

2.2 Data Cleaning and Preparation:

1. Data frames: first I converted all the tables into data frames, so that I can easily work with them.
2. Standardized form: then I made the tables in a standard form where the tables from 2011 match the ones from 2021, as some extra rows were present.
3. Null Value Identification: initially I checked whether the data had any null values so that I could remove or predict the values accordingly
4. Geographic Alignment: the 2021 census contained 533 additional geographic areas not present in the 2011 data due to boundary changes and new developments. I identified these by data frame merging with a left join operation on geographic identifiers, these were then stored separately.
5. Bayesian imputation: rather than applying simple mean/mode imputation for missing values, I implemented Bayesian Linear Regression to generate estimates. This approach was applied as it provided several advantages:
    - Uncertainty Qualification: this method provides not just point estimates but entire probability distributions, which allows for better understanding of potential values.
    - Incorporation of Prior Knowledge: the model used known relationships on the basis of the tabular data available.
    - Robustness to small sample sizes: in areas with limited data, Bayesian approaches are less prone to overfitting that traditional imputation methods.
    - Improved accuracy: this approach reduced error by a good amount.
6. Then the final comparison: here I made sure before performing PCA and t-SNE, the tables should have the same values, so renamed the columns that they could match and added a few columns before merging.

Task Definition:

Following Munzer's task taxonomy, I identified the following analytical tasks for the visualization, organizing them in hierarchical order:

1. **Primary Tasks**: to identify trends and patterns:
    a. Discover: identify trends and patterns
        i. Search: for industries showing major employment growth
        ii. Identify: Investigate the correlation between industry sectors and the qualification levels of their workforce to understand skill transitions.

        iii. Locate: regions with high concentrations of potentially AI- related employment
    b. Compare: Check the differences between time periods and regions:
        i. Specific changes between 2011 and 2021 employment data
        ii. Differentiate between regions with varying qualification profiles
        iii. Correlate: qualification levels with changes in industrial composition.
    c. Summarize: provide an overview of:
        i. Aggregate employment shift.
        ii. Relationships between education and employment.
        iii. Analyse evidence for connection between qualification requirements and AI development.

2. **Secondary Tasks:**
    a. Explore: enable open-ended investigation:
        i. Browse: through different geographic regions.
        ii. Filter data by qualification level, industry type, or region.
        iii. Refine queries based on emerging patterns.
    b. Present: communicate findings:
        i. Arrange views to emphasize key relationships
        ii. Annotate important patterns and outliers
        iii. Change visualisation parameters to highlight different aspects of data.
3. **User Scenarios and Task Alignment:** to ensure the visualization meets user needs I developed specific scenarios aligned with the defined tasks:
    a. Scenario 1: Regional Economic Development Officer:
        i. Primary task: is to discover
    b. Scenario 2: Education Curriculum Planner:
    c. Scenario 3: Socioeconomic Researcher:

Visualisation Justification.

Dimensionality Reduction Approach: for the complex multidimensional census data, dimensional reduction methods are very important and reveal the underlying patterns while keeping the information intact. I employed two methods, each with its own foundation and strengths

4.1.1 Principal Component Analysis (PCA): I selected this as a primary dimensional reduction technique for various reasons:[1]

1. Linearity reduction: PCA creates orthogonal components that represent maximum variance in the dataset, which allows for clear interpretation of primary factors driving employment differences.
2. Preservation of global structure: PCA very nicely preserves distances between dissimilar points, which makes it great for identifying broad patterns in educational qualifications and industrial transformation.
3. Computational efficiency: PCA scales well the high dimensional census data, which enables rapid interactive visualisation necessary for the exploratory tasks in section

Analysis of the results:

- **Temporal Shift Pattern:** there is a clear temporal progression in the data points which indicates that the yellow points (2021) are denser and more dispersed than the blue point (2011), indicating a greater diversification in the employment patterns in 2021.An outward expansion along the PC1 is seen, this suggests an increase in the new employment sectors in the 2021 data

- **Regional Dispersion:** the 2021(yellow) data points show greater dispersion along both principal components compared to 2011 data, showing a regional differentiation in the industry profiles.
- **Outlier Regions:** several points appear as outliers, especially in the positive range of PC1(right side), mostly representing the regions with exceptionally high concentrations of highly qualified workers, they are in 2021 data.

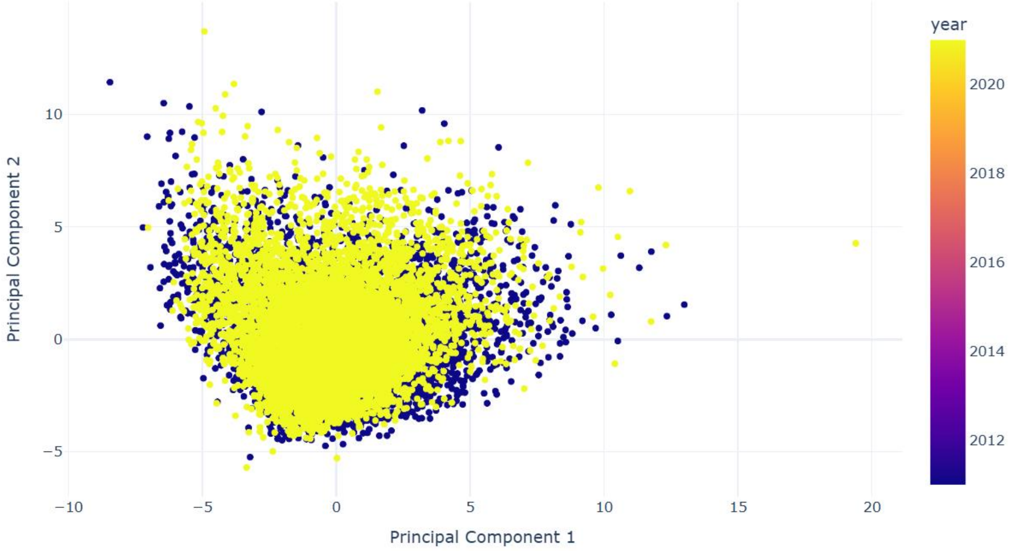

Fig1: Industry data

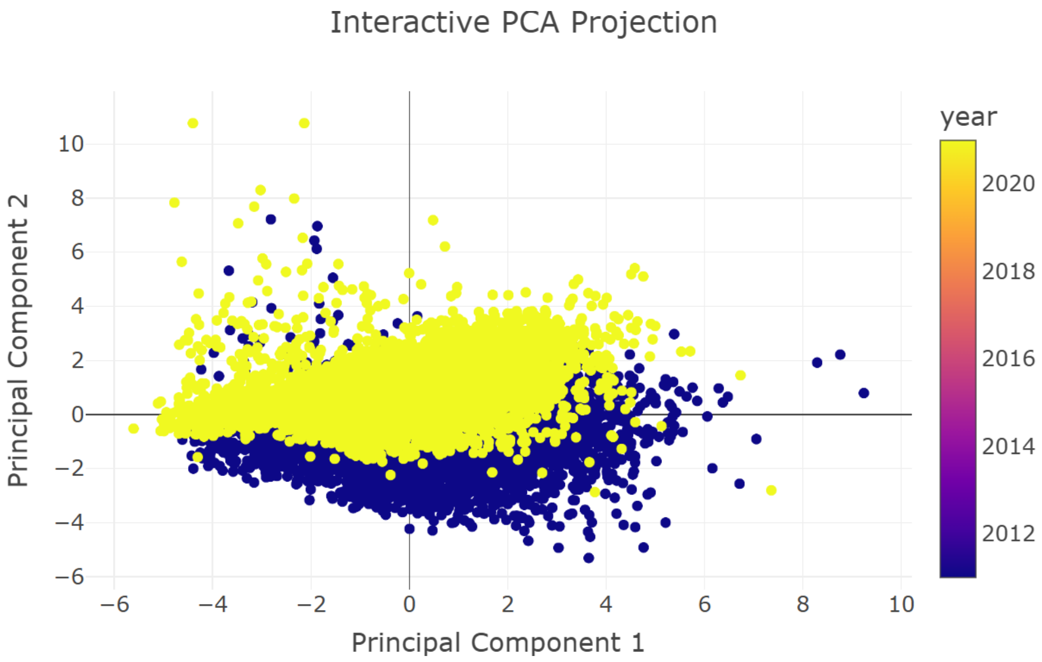

Fig 2 : Qualification data

4.1.2 Distributed Stochastic Neighbor Embedding (t-SNE): while PCA provides a global view, t-SNE captures non-linear relationships: [2]

1. Preservation of local relationships: it excels at preserving cluster structures and local similarities, which makes it particularly valuable for identifying regions with similar employment patterns.
2. Non-Linear pattern detection: t-SNE can reveal complex nonlinear relationships between the past and new year's that might be missed by PCA, especially important for socioeconomic data where relationships are often non-linear.
3. Enhanced cluster visualization: this technique is mostly effective at separating distinct qualification-employment pattern clusters, allowing for identification of regional types.

Analysis of the results:

- It shows fine-grained patterns and clusters and preserves local structures, which helps visualize niche AI-related growth.
- It shows more scattered clusters, yellow points (2021) dominate, indicating that new employment clusters have emerged post-2011. Smaller and denser groupings suggest growth, particularly in sectors such as **IT sector,** and other areas.
- It is ideal for highlighting sector clusters with similar behaviour across years.

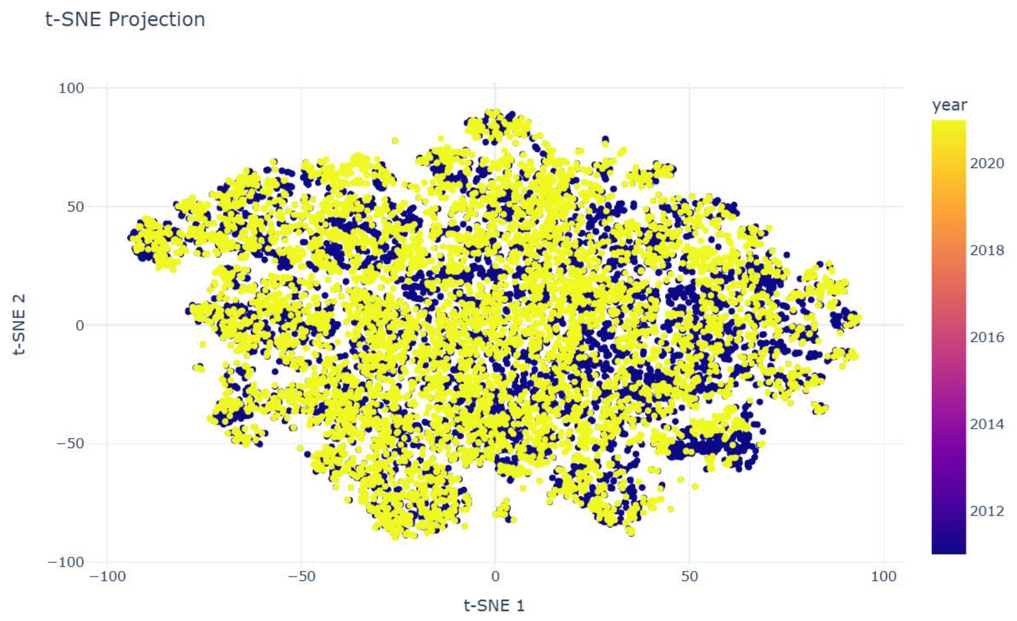

Fig 3 Industry t-SNE

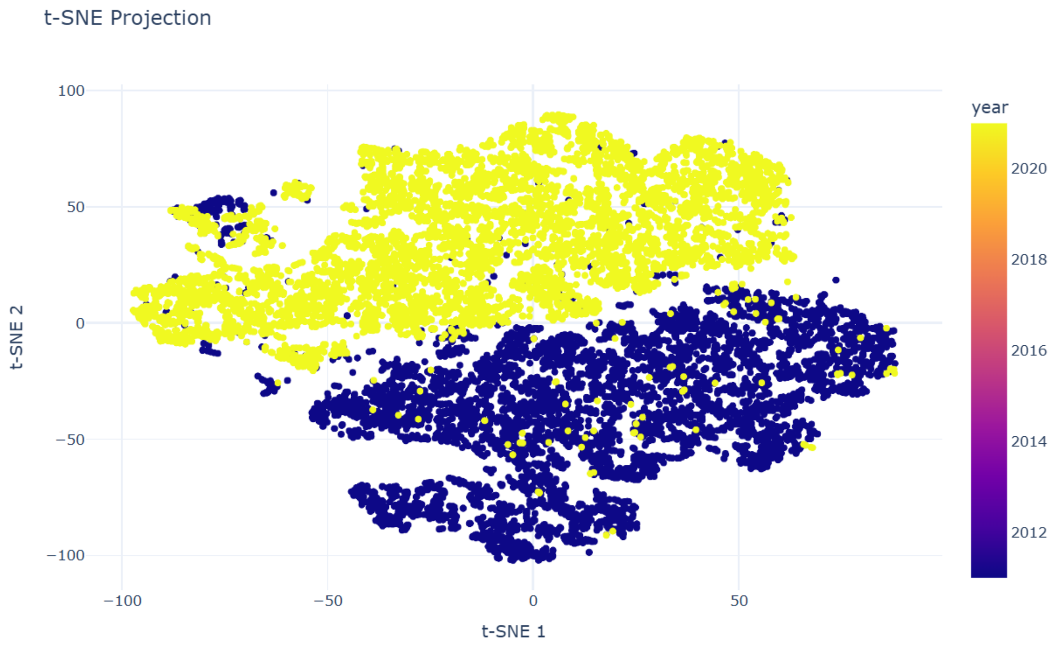

Fig 4: Qualification t-SNE

4.1.3 (UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction): this is another reductional analysis technique that I found to be more distributed and gave me a clear view of how the data is distributed, it can be observed here that data from

2021 is very well distributed and spread out more.[3]

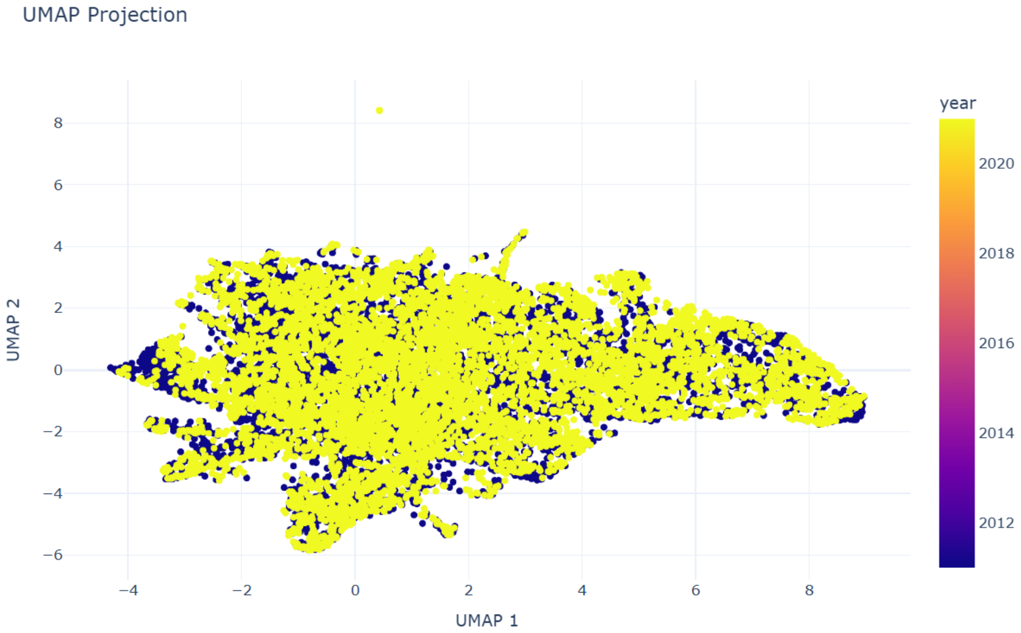

Fig 5: Industry UMAP

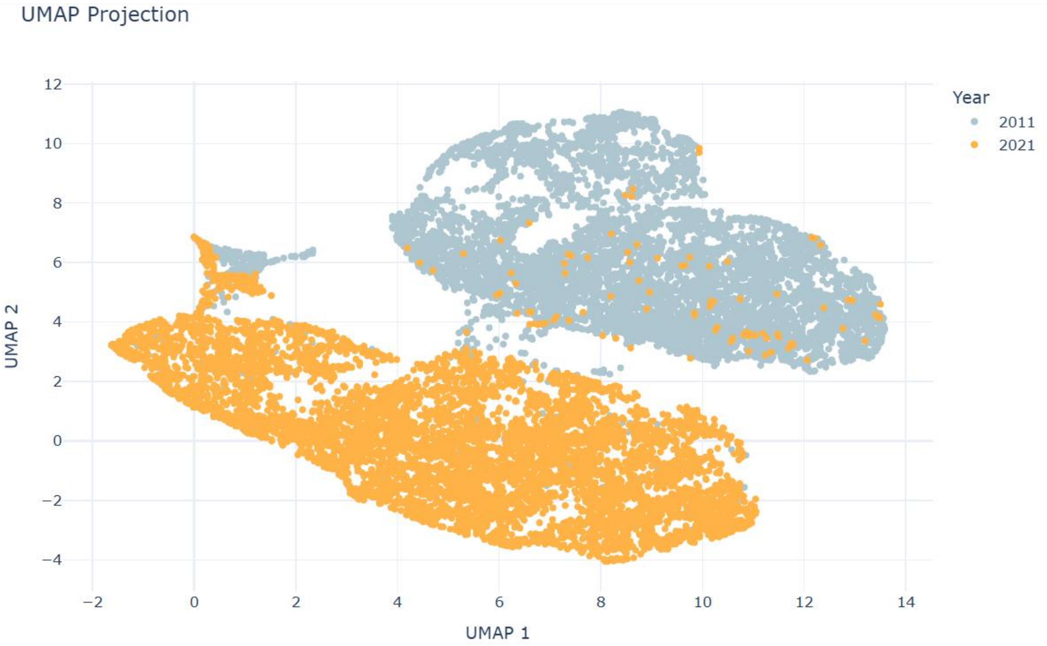

Fig 6: Qualification UMAP

Integration of both methods:

1. PCA gives a clear, interpretable overview of global structural changes.
2. t-SNE shows detailed local groupings and uncovers hidden non-linear relationships.
3. U map adds up everything well and gives and overall idea about the varied data distribution.
4. Together, they provide a good understanding, PCA for broad macro trends, t-SNE for micro-sectoral insights.

Tableau:

File 1:

Dashboard 1: Compares the difference between 2011 and 2021

Socio economic: Regional Disparity:

- It shows darker green regions have higher relative presence of the IT/Information and communication industry and vice versa.
- Urban clustering: denser IT presence in metropolitan areas like London, Manchester, Birmingham, etc.
- Rural and urban divide: Less IT concentration in rural or peripheral areas.
-

Dashboard 2: Compares the difference between IT and manufacturing (Low tech)

Socio economic: Regional variation:

- The map has yellow to dark orange gradient.
- High concentration areas in London, the Thames Valley, Birmingham, Leeds etc show they are tech hubs.
- These regions have universities, strong connectivity and are better economies.
- Low concentration areas are rural areas like South West, Midlands etc.
- They lack digital infrastructure or educational institutes.
- Bar chart shows high percentage in IT and low in low tech.
- Thus, clearly showing UK has shifted towards IT.
- From the policy point of view, must invest in digital infrastructure

File 2:

Dashboard 3: Answers the sub question about the share of AI (Artificial Intelligence) in the growth of industries and employment.

- The map shows green to blue gradient; dark blue shows a high count of programming-related jobs.
- Dense regions are seen in London, Bristol, Manchester.
- Less in rural areas in North West, Cornwall etc.
- This again suggests a rural urban digital divide.
- From the bar chart such as Swindon, Warwick where AI is approaching.
- This maybe due to R&D hubs and shows regional specialisation.

File 3:

Dashboard 4: Qualification and employment:

- The darker shade represents higher qualification and the lighter shade lower level of qualification.
- Also, they co relate with manufacturing or heavy industry which do not require formal education.
- Such as Cambridge and Oxford clearly from the map show a higher percentage of education and hence industrial growth as well.
- Top right shows higher qualification while as the bottom right shows the no qualification bar.

Peer Evaluation:

As a part of the evaluation process, 4 team members from a group of 7 were asked to evaluate my dashboard by performing analytical tasks and responding to questions. These tasks were visual appeal, usability, comprehension, and alignment with perceptual principles.

Visual Appeal:

- 3 of my peers rated this 10/10, while 1 gave 8/10.
- The consistent and appealing colour gradient was praised by all for clarity, especially in the qualification one.
- One peer said that the map was immediately readable and well aligned.

Interactivity:

- All 4 said the layout easy to navigate, especially how I had used different patterns.
- The bar charts and the tree maps were helping with the comparison tasks,
- One suggestion was to change the colour of the tree map.

Legand and colour:

- 2 of the participants said that the background of one of the maps was too dark.
- Then they also said there was a lack of an informative legend initially made the heatmap harder to interpret.
- A small portion showed difficulty differentiating lighter shades.
- In response I changed both the things later.

Data Segmentation and Clarity:

- Everyone liked my data was split into highest qualification and also how I took the percentage and differences and how I managed to go for a different industry.
- The visual hierarchy was said to be clear, with headings and layout as well.
- The story was very smooth and well written.
- The noticed the text boxes and chart vales and appreciated the time and effort I put in.

Overall, the peer feedback validated my overall design as engaging, informative, and visually balanced.

Conclusion:

• What I learned about the socio-economic problem:

The visualisation and the techniques highlight important socio-economic issues: the regional digital divide and unequal access to high value technological employment, especially within the AI related fields of programming. Here:

- Regional Disparities: the programming related jobs are highly concentrated in the urban areas and lesser in the rural ones where opportunities, income potential and innovations are less.
- The socioeconomic challenge: the rural and post-industrial areas may fall behind in the digital transition, this shifts the social cohesion.
- Emergence of AI: some small towns like Warwick and Swindon show high AI jobs that shows niche strengths in AI might be because of universities.

- Inequality: due to the divide the inequality may grow in certain regions can keep getting higher and a growing socio-economic fault line. Here workers in the low-tech regions may lack digital skills for transitioning into high-growth sectors and hence skills-based inequality.
- National Growth: this shows that the growth is not balanced. Policy makes must ensure that AI and programming related opportunities are equally accessible.

• What I have learned about information visualisation:

By the work that I have done, I'm amazed how much I've grown. Tableau has a lot to teach:

- It is great for understanding socio economic topics, by using geographic, industrial and educational datasets, I learned how it connects raw data and human interpretation.
- Different maps I represented the distribution of technology related employment across England.
- Bar charts were interactive and helped me focus on the overall trend and the individual city trend such as IT vs low tech.
- Dashboards: these are so insightful and fun to work with and they make data reading and interpretation very easy and user friendly.

References:

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html
3. McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv preprint arXiv:1802.03426. Retrieved from https://umap-learn.readthedocs.io/en/latest/