Advanced Data Analytics Coursework

This report contains all the answers to the Summative Assessment for Advanced Data Analytics Coursework.

1 Abstract

This report will explore and analyse the industry demographics of employed people aged 16 to 74 in England and Wales in 2011. In particular, the study focuses on finding links between people working in various industries and their links with other socio-economic factors such as age, sex, economic activity, and occupation. Tableau (a data visualisation software) and data collected from the 2011 census in England and Wales were used to conduct this study. Dimensionality reduction algorithms such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were used to make projections and find clusters in the raw data. An interactive dashboard has been created as part of this report to provide insights for policymakers and ordinary people.

2 Introduction

Socio-economic factors are the factors that determine the overall development of the country in terms of the society's economic and health well-being. Employment, education, income, wealth and where a person lives are some of the socio-economic factors. One of the essential socio-economic factors taken into consideration is employment. Employment refers to a person's job, i.e., what a person does for a living. The report focuses on the industry demographics of employed people in England and Wales in 2011. Some of the research questions this study hopes to answer are:

Task 1: How are the employed people distributed across England and Wales? Which areas are relatively better to look for employment?
Task 2: What kind of industries dictate the England and Wales job market and by how much?
Task 3: What is the gender distribution of working employees in various industries? Are there any industries dominated by a single gender?
Task 4: What age groups are most common in a particular industry?
Task 5: Was the start-up/self-employment culture prevalent in 2011?
Task 6: What kind of occupations are common amongst various industries? Does any occupation dominate every industry?

Dimensionality reduction is a process of reducing the number of features of a large data set. Two dimensionality reduction algorithms were tested by making projections on raw data.

- **Principal Component Analysis (PCA):** It is a linear dimensionality reduction technique used for extracting a small number of uncorrelated variables known as principal components from a large raw data set.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE):** It is a non-linear, unsupervised dimensionality reduction technique used for visualising and exploring high dimensional datasets. It works by calculating similarity measures in high and low dimensional space and later tries to optimise them by using a cost function.

3 Data preparation and abstraction

3.1 Data Source

For this report, the data from the 2011 census related to the industry demographics of employed people in England and Wales has been taken through the Nomis Website. The links to the taken datasets and their descriptions are listed below:

- Industry by Age
- Industry by Sex
- Industry by Economic Activity
- Industry by Occupation

3.2 Dataset Availability

The entire datasets with the local authorities (districts) information were taken for the analysis. The datasets were static files (fixed data that do not change frequently) last updated in 2014.

3.3 Dataset Type

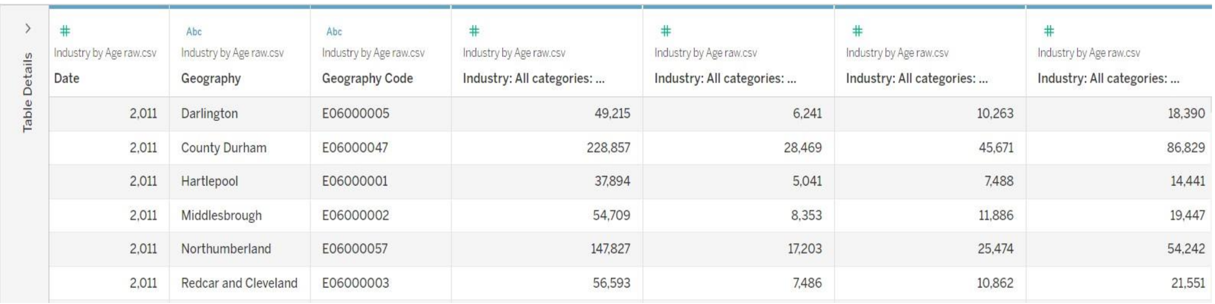
The downloaded datasets were in the form of tables that contained multiple rows and columns. Each row represents an item, and each column represents an attribute. The datasets also contain geographical data in the form of local authorities (districts). Tableau natively does not support the visualisation of geographical data of the United Kingdom (UK). Hence, to display the local authorities, geocoding package for the UK has been installed in the tableau repository.

3.4 Data Abstraction:

Before using the data to make visualisations in Tableau, it was cleaned and simplified. The raw data consisted of unnecessary data such as the year and total/all categories information. The redundant information was filtered/deleted, and the required information was neatly organised. Detailed steps taken are as follows:

The year column was not required as the entire data consisted of information from only a single year (2011). Hence, it was deleted.

The downloaded census data contained information in separate columns as shown below:

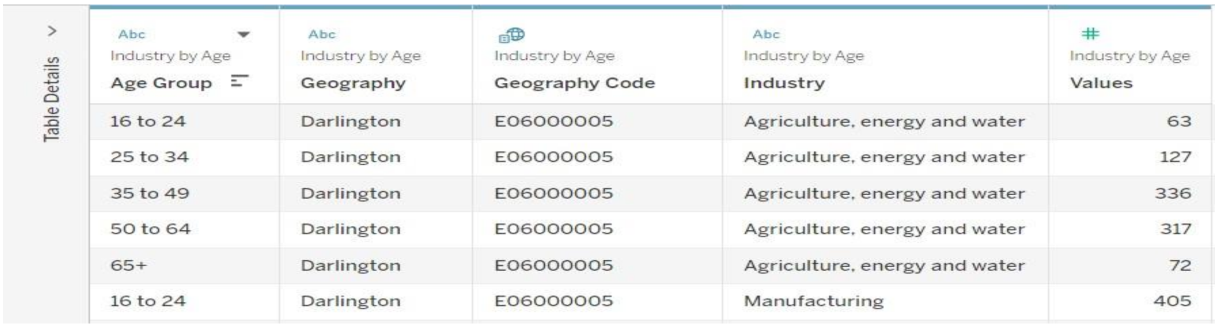


#	Abc	Abc	#	#	#	#
Industry by Age raw.csv	Industry by Age raw.csv	Industry by Age raw.csv	Industry by Age raw.csv	Industry by Age raw.csv	Industry by Age raw.csv	Industry by Age raw.csv
Date	**Geography**	**Geography Code**	**Industry: All categories: ...**	**Industry: All categories: ...**	**Industry: All categories: ...**	**Industry: All categories: ...**
2,011	Darlington	E06000005	49,215	6,241	10,263	18,390
2,011	County Durham	E06000047	228,857	28,469	45,671	86,829
2,011	Hartlepool	E06000001	37,894	5,041	7,488	14,441
2,011	Middlesbrough	E06000002	54,709	8,353	11,886	19,447
2,011	Northumberland	E06000057	147,827	17,203	25,474	54,242
2,011	Redcar and Cleveland	E06000003	56,593	7,486	10,862	21,551

As a result, the data set is broad, and data in this format can be complex for Tableau to visualise. For this reason, all the data except geographical information was pivoted to make it easier to work with. Pivoting data is a data shaping technique that can be used to transform a broad dataset into a tall dataset and vice versa. Apart from the geography and geography code columns, the remaining attributes are pivoted to increase the simplicity.

Tableau is advanced visualisation software that can calculate the total values of all attributes on its own. Therefore, there was no need for total/all categories data. Such kind of data was filtered out to reduce dimensions and complexity.

Once the data was pivoted, it was cleaned further by splitting the columns. An example of one of the cleaned datasets is shown below:



Industry by Age	Industry by Age	Industry by Age	Industry by Age	Industry by Age
Age Group	Geography	Geography Code	Industry	Values
16 to 24	Darlington	E06000005	Agriculture, energy and water	63
25 to 34	Darlington	E06000005	Agriculture, energy and water	127
35 to 49	Darlington	E06000005	Agriculture, energy and water	336
50 to 64	Darlington	E06000005	Agriculture, energy and water	317
65+	Darlington	E06000005	Agriculture, energy and water	72
16 to 24	Darlington	E06000005	Manufacturing	405

4 Task definition:

Task analysis is essential for creating good visualisations. Tamara Munzer's taxonomy has been used as a reference for the task descriptions carried out in this report. Each task has been represented in the form of actions and targets. The task descriptions for the earlier mentioned research questions are given below:

- Task 1:

  Action: Present a map of England and Wales that contains information on employed people at the local authority (district) granularity. Add interactivity to increase curiosity and make the user discover new insights regarding an industry.

  Target: Visually encode the attributes and enable users to search for areas popular in an industry.

  How: Use a map to display the encoded information regarding employment in an industry. Map data with nonspatial visual channels such as colour (luminance) to make the user easily understand the information.

- Task 2, 6:

  Action: Present the user with employment information for each industry/occupation.

  Target: Visually encode the attribute by sorting the information. Let the user compare information amongst different industries/occupations.

  How: Align the necessary data into descending order. Add colour to make it easier for the user to carry out the task.

- Task 3, 5:

  Action: Analyse the information and present the user with knowledge regarding each industry's gender distribution/economic activity.

  Target: Encode the attributes to visualise comparison. Enable the user to understand and compare the data.

  How: Arrange the data in an understandable manner. Add colour to make it easier for the user to carry out the task.

- Task 4:

  Action: Depict the age group information to the user.

  Target: Encode the attributes to make it easier for the user to identify common age groups.

  How: Arrange the data in an ordinal manner. Add colour to simplify the process of carrying out the task for the user.

5 Visual Justification:

All the visualisations made were intended for the general public and policymakers as the target audience. Visual justification for each visualisation has been given below.

5.1 Employment in England and Wales

The worksheet contains the map of England and Wales. This visualisation aimed to show the density of the number of employed people in various local authorities (districts). Colour was added to the measure values of employed people to make the user easily identify the denser areas quickly. A filter has been added to the values of employed people for interactivity and to enable the user to uncover new insights. The tooltip contains information regarding the local authority, number, and percentage of employed people in that local authority.

Insights:

The top 3 Local Authorities with the most significant number of employed people across all industries are:
1. Birmingham
2. Leeds
3. Cornwall
People actively looking for employment opportunities can extend their search to these locations.

5.2 Top Industries

The worksheet visualises the industries present in England and Wales. This visualisation aimed to enable the user to identify top industries and easily compare them. A bar chart has been used to visualise the quantitative value attribute (% of people) and the categorical key attribute (industry). The vertical axis contains different industries, and each bar represents an industry. The chart is encoded by the bar heights and percentage of people in the industry. This was done to make the user quickly identify the top industries. Additionally, the bars were also colour coded (luminance) to increase the ease of carrying out the task. The tooltip contains the industry name, number and percentage of people employed in that industry.

Insights:

The top 3 industries in England and Wales are:
1. Public Administration, education, and health
2. Distribution, hotels, and restaurants
3. Financial, Real Estate, Professional and Administrative activities
These three industries amount to 67% of the job market in England and Wales. This indicates that there is no shortage of employees in these industries. Agriculture, energy and water industries have a severe workforce shortage.

5.3 Sex Distribution by Industry

The worksheet visualises the sex distribution of employed people in their respective industries. This visualisation aimed to make the user understand the part to the whole relationship of sex distribution in each industry and identify which industries are dominated by men and women. A 100% stacked bar chart has been used to display the relationship between one quantitative value attribute and two categorical attributes. The key used to distribute the bars along the axis is the industry, and the key used to colour and stack each bar is gender. This type of bar chart was chosen to quickly show the user the relative differences between males and females in each industry.

Insights:

- Male-Dominated Industries:
  1. Construction
  2. Agriculture, energy, and water
  3. Manufacturing
  4. Transport and communication
- Female-Dominated Industries:
  1. Public administration, education, and health

It can be noted that most of the industries that males dominate in require hard manual labour (blue-collar workers), whereas the females dominate the services-based industries (white-collar workers). It is to note that women lead the top industry (Public administration, education, and health) by a margin of 70-30 ratio.

5.4 Age Groups

The worksheet shows various working-class age groups by industry. This visualisation aimed to visualise the distribution of employed people across various age groups. A discrete area chart has been used to envision this relationship between the quantitative value attribute (% of employed people) and the ordered attribute key attribute (age groups). A line chart has been filled in visually to allow the user to easily identify general trends and typical age ranges of employed people. An industry filter has been added to increase the user's curiosity and enable them to discover new insights.

Insights:

35-49 was the most common age group across all the industries. The younger age group (16-24) has the most opportunities in distribution, hotels & restaurants, and other industries. The oldest age group (65+) was most active in agriculture, energy, and water industries.

5.5 Economic Activity by Industry

The worksheet displays the economic activity of employed people. This visualisation aimed to show the user the distribution of self-employed people and people employed in an organisation. A single 100% stacked bar has been used to visualise this information. This bar graph lets the user notice the relative differences between employees and self-employed people.

Insights:

The percentage of people working part-time as an employee (24.56%) alone is greater than that of self-employed people (15.35%) across all industries. Most people self-employed are in male-focused industries.

5.6 Occupation by Industry:

The worksheet displays various occupations by industry. This visualisation aimed to make the user easily identify occupations within a selected industry. A lollipop chart has been used to visualise the quantitative value attribute (% of people) and the categorical key attribute (occupation). The horizontal axis contains different occupations, and each bar represents an occupation. The chart is encoded by the bar heights and percentage of people in the occupation. This was done to make the user quickly identify the top occupations and compare them. Occupations were also colour coded to make the user more easily carry out the task. Instead of a simple bar graph, this chart type has been chosen to add variety to the dashboard.

Insights:

The Top 3 occupations across all industries are:
1. Professional Occupations
2. Associate professional and technical occupations
3. Skilled trades occupations
These three occupations amount to 41.54% of the occupations in England and Wales. No occupation dominates every industry.

5.7 Data Projection Methods

PCA and TSNE have been used to test and perform dimensionality reduction on the raw dataset - Industry by Occupation. Initially, the dataset had 93 features. Categorical columns were dropped, and the remaining numerical columns were scaled using StandardScaler() before performing dimensionality reduction. The results after reducing the dimensions were visualised on Tableau.

For PCA, eigenvalues and eigenvectors were calculated. A threshold plot was plotted to find the optimal number of principal components. 11 eigenvectors were required for the chosen dataset for the proportion of variance explained to exceed 0.95 (threshold). Therefore, 11 components were used to fit the scaled data. The results obtained were stored in a data frame and visualised in Tableau. Eight different clusters were observed. These clusters represent the local authority (districts). All the local authorities' information from the raw data can be grouped into eight different clusters.

t-SNE, by default, uses only two components to fit the scaled data. Upon visualisation, 14 different clusters were observed. This implies that the local authority (districts) can be grouped into 14 different clusters in the chosen industry by occupation dataset.

6 Conclusion

Extensive visualisations and analysis have been done to find solutions to the socio-economic research problems stated at the report's beginning. In 2011, Birmingham and Leeds were the top economic centres that drew much attention amongst people from various industries regardless of gender and age. There was a lot of gender disparity in many industries. Especially the industries that required physical work were dominated heavily by men. The start-up/self-employment culture in 2011 was not as prevalent as it is now in 2022. 35-49 was the most common age group across all the industries. Professional and associate professional occupations were most popular back in 2011. An interactive dashboard has been created for the end-user to increase curiosity and find new insights. Projections were made using dimensionality reduction algorithms on the "Industry by Occupation" dataset. The work done in this report can help a typical person be aware/understand the industry demographics and aid policymakers in making better decisions regarding the socio-economic life in England and Wales.

References

- Course lectures and Lab Worksheets
- Tamara Munzner. Visualisation Analysis and Design, CRC Press, 2014.