Advanced Data Analytics Coursework Report

1. Abstract

The report discusses the procedures and methods of creating a visual analytics project. The project's goal is to focus on the demographics of occupation in England and Wales. Occupations in a workplace population with multiple socio-economic aspects have been considered for this problem. Firstly, Data relevant to the topic is taken from the NOMIS website. The data taken was collected in 2011. Data preparation techniques are used to clean the data to be visualized in tableau. Secondly, two interactive dashboards have been created to gain relevant insights. Finally, two data projection algorithms are applied to the dataset and recorded their results.

2. Introduction

The topic I have chosen to explore is occupation-based on different socio-economic factors in the workplace. The socio-economic factors I considered to link with the occupation are age, qualification, industry, distance travelled to work and method of travel to work. All these datasets are taken from the NOMIS website, and they all are workplace population datasets in local authorities of England and Wales. Workplace population is the estimate of a population working in that area. It includes all the individuals who live and work in the local authority and individuals who work in the local authority but commute from a home elsewhere within England and Wales. I have considered five datasets to approach this topic:

- Occupation by Age: This dataset consists of nine occupations with eleven age groups
- Occupation by Industry: This dataset consists of nine occupations with eight industries
- Occupation by Distance Travelled to work: This dataset consists of nine occupations with seven ranges of distances travelled to work
- Occupation by Qualification: This dataset consists of nine occupations with six qualifications
- Distance Travelled to work by Method of Travel to work: This dataset contains seven ranges of distances travelled to work with six methods of travel.

All these datasets contain local authorities (geography, geography code) of each district/region. The main aim is to create an interactive dashboard visualisation where the user can easily navigate and filter the dashboard and gain valuable insights regarding the demographics of occupation in England and Wales. For example: if the user wants to know the occupations in Bristol (workplace occupation), a simple click on the region of Bristol must bring the user the distribution of occupations. Furthermore, the user should be able to see what kind of age groups are in different occupations, what industries are popular in Bristol relating to an occupation, what kind of qualifications relating to that occupation are popular in Bristol, and how far are the people coming from to work in Bristol and what is their method of travel. The user should also be able to filter the factors accordingly. All this information should be accessible without confusing the user.

3. Data Preparation and Abstraction

All the five data sets from NOMIS were differently formatted and were harder to process in tableau for a couple of reasons. Firstly, the data had many columns in which, some of them contained totals of other columns. These types of columns could be removed from the data. Secondly, the column names were formatted with special character separators (: ;). This made it harder to read and understand the purpose of each column.

Occupation: All categories: Occupation; Highest Level of Qualification: Level 3 qualifications; measures: Value									
C	D	E	F	G	H	I	J	K	L
raphy co	Occupation:	Occupation:	Occupation:	Occupation:	Occupation:	Occupation:	Occupation:	Occupation:	Occupation:
00005	52096	5317	7697	9840	9335	15742	4165	4687	416
00047	196872	23840	30599	37778	33318	55059	16278	17453	1614
00001	34197	4492	5048	6436	6047	9042	3132	2779	298
00002	62839	7096	8662	11032	10352	20761	4936	4761	488
00057	123686	12985	19612	24400	20468	35967	10254	12403	1048
00003	45844	5606	6987	8614	8069	12141	4427	3894	386

I used tableau data manipulation commands to clean the data for analysis. All five datasets contained the above mentioned problems. Firstly, I removed all the columns containing totals of other columns. I did this using the hide command in tableau. Now the data had columns with separators like (: ;). I selected all the columns and pivoted them into rows using the pivot command. After this, I split the data with (;) in each row using a custom split command. This splits the pivoted data into separate columns. For example: taking the above image as the reference, the occupations and qualifications are separated by a semicolon (;). Splitting it with the semicolon using custom split divided the single column of occupations and qualifications into two individual columns. The data is mostly cleaned with some minor challenges like the misread column names and strings by tableau. I used the rename and aliases commands in tableau to resolve such issues. Finally, the data is clean and ready for analysis. I cleaned all five datasets using the same procedure.

Abc Occupation By Qualification Geography (Occupation ...	Abc Occupation By Qualification Geography Code (Occu...	Abc Occupation By Qualification Occupation (Occupatio...	Abc Occupation By Qualification Qualification	# Occupation By Qualification Count (Occupation By Q...
Darlington	E06000005	Managers, directors and senio...	Apprenticeships and other qu...	295
Darlington	E06000005	Managers, directors and senio...	Level 1	647
Darlington	E06000005	Managers, directors and senio...	Level 2	726
Darlington	E06000005	Managers, directors and senio...	Level 3	684
Darlington	E06000005	Managers, directors and senio...	Level 4 and above	1,919
Darlington	E06000005	Managers, directors and senio...	No qualifications	416
Darlington	E06000005	Professional occupations	Apprenticeships and other qu...	145
Darlington	E06000005	Professional occupations	Level 1	160

After cleaning the data, I exported each datasheet to excel format and merged all the datasheets into a single excel file. This made loading all the datasheets in tableau at a time easier. I combined datasheets using relationships with common columns like occupation, geography, and geography code in the logical layer of tableau. These are the datasheets with relationships in the logical layer.

4. Task Definition

I considered these three typologies while creating the visualisations. They are:

- Why is this task performed?
- How is the task performed?
- What does the task pertain to?

These are some of the questions focused on the problem statement:

- What are the top occupations in a region?
- How are the age groups distributed in a particular occupation of a region?
- Which industry do people work in a particular occupation of a region?
- What is the highest qualification in a particular occupation of a region?
- What per cent of people are travelling more than 10kms to their work and what occupation do they belong to?
- What is the popular method of travel of people travelling 30kms or more and what occupation do they belong to?

To answer these questions implementing interactivity is important. For instance, to answer the top occupations in a region, there needs to be a geographic map of England and wales to let the user hover and select the desired region. With filters, the selected region displays the occupations in that region. Sorting the information in descending order displays the top occupations in a region. The task is completed by answering the question of the user.

The dashboard visualisations created aim to show the demographics of occupations in a workplace region. These occupations are linked to five socioeconomic aspects of England and Wales, occupations, their age distributions, demographics of occupations in industries, highest levels of qualification in occupation, distance travelled to work and their occupations and the method of travel. So, rather than having a set of defined questions and answering them to the user. For example: visualising a static bar graph of the top 5 occupations answers a single question. I want the users to interact with the dashboard and form their questions and find answers through the dashboard. The dashboard acts as a gateway between the user's question and the available information. However, the dashboard is limited to a pertinent topic: demographics of occupation in a workplace linked to five socioeconomic factors. So, for example, if a user wants to know the top occupations in Bristol with qualifications above level 3, between the age of 25-29 in the financial industry, the user can filter this by interacting with the dashboard.

5. Visualisation Justification

From the data obtained, there is only one feature with numeric values in each datasheet. Due to this, bivariate numeric plots like scatter plots are not possible.

For the first visualisation (Occupations), I created a bar graph visualisation of occupations (categorical) and their counts (numerical). I arranged the graph in descending order to sort the graph from the largest occupation to the least based on the counts. This makes it easy for the user to identify the top n or bottom n occupations quickly. I colour coded the graph in blue to make it visually stand out from the background. The primary issue with this graph was that the occupation names were not completely visible, so I included the occupation in the tooltip alongside the counts. I labelled each bar graph with the per cent of the total counts to display the user the per cent of each bar (occupation) compared with others. The reason for considering the bar graph is that I wanted to show the different occupations with their counts. From the graph, we can infer that in all of England and Wales, Professional occupations constituted the largest occupation which was 17.4%, and process plant and machine operatives constituted the lowest occupation, which was 7.2%.

For the second visualisation (Age), I created an area plot visualisation which showed the age distribution (categorical) with counts (numerical). I colour coded the area plot in green to differentiate it from the previous visualisation. I only labelled the highest and lowest counts on the plot. I did this because there are eleven categories of age groups and displaying the count of all the categories makes the plot messy and harder for the user to understand. I added all the counts and the names of the age groups in the tooltip so that the users can see the relevant information by hovering over the plot. I chose

an area plot for this visualisation because I wanted to show the range of ages over the counts, and this plot looked like a proper structure to represent the age category. From the plot, we can infer that people in the age range of 40 years to 49 years are the most while 65 to 74 years are the least in England and Wales.

For the third visualisation (Industries), I created a bar graph visualisation of industries (categorical) with counts (numerical). I also arranged this graph in descending order to sort the graph from the largest industry to the least, making it easy for the user to quickly identify the top n and bottom n industries. I coloured the bar graph in mustard yellow to differentiate it from the other visualisations. Like the first visualisation, due to the longer industry names, the graph had an issue displaying the complete names of the industries. To resolve this, I included the names of the industries in the tooltip alongside the values. I labelled the bar graph using the per cent of the total counts to display the percentage of each bar (industry) compared with others. I considered a bar graph in creating this visualisation because I want the user to be able to summarise a large set of categories in visual form quickly. From the graph, we can infer that public administration, education, and health are the largest industries constituting 28.4% of the total industries in England and Wales.

For the fourth visualisation (qualifications), I created an area plot like the second visualisation to visualise the highest level of qualifications (categorical) with counts (numerical). I colour coded the plot in orange to make it different from the other visualisations. I displayed the minimum and maximum values to avoid the plot from being messy. I added all the counts and names of the qualifications in the tooltip to let users see the information while hovering over it. I used this visualisation because an area plot looked more structured and related to the categories of qualifications. From the plot, the highest level of qualification for the majority of people in England and wales is level 4 and above.

For the fifth visualisation (method of travel), I created a packed bubbles visualisation where each bubble consisted of a category in the method of travel to work (categorical) and the size of the bubble is determined by the counts (numerical). I colour coded the plot where each bubble is a different colour, and each colour is a category. I showed the legend to ensure that the user understood the colour and its respective category. I also labelled the category's name for better clarity in the visualisation. The bubbles are larger when the value counts are higher and vice versa. I chose this plot because the data does not need to be ordered, and the bubbles are compact and easy to understand. From the visualisation, In England and Wales, 14.4 million people commuted to their work using a car or a van, followed by public transport by 4.2 million.

For the sixth visualisation (distance travelled to work), I created a treemap visualisation where each tile consists of different distances travelled to work (categorical), and the size of the tile represents the counts (numerical). The higher the value, the larger the tile and vice versa. I colour coded the treemap according to the counts. The higher the count, the darker the hue of blue and vice versa. I displayed the colour legend to ensure that the user understood the colour hue. I labelled the names of the category along with the per cent of the total counts on the tile. I added the value of the counts in the tooltip so the user can see it while hovering over the tile. I chose this plot because the categories are in a hierarchical order, and a treemap is intended to visualise hierarchical data. From the visualisation, In England and Wales, 25.7% of the people travel less than 5 kilometres to reach their workplace whereas 4.6% of the people travel more than 60 kilometres to reach their workplace.

For the seventh visualisation (Map), I assigned geographic roles to the local authority regions to create a map of England and Wales. I colour coded the map to a lighter red to make the map stand out from the surrounding countries. I added the names of the regions on the tooltip to let the users see the names of the regions while hovering on the map. I used this visualisation because it forms the dashboard's centre, where the users can filter the regions by clicking an area on the map.

The above visualisations constitute to individual analysis. The actual aim of the report is resolved in creating a dashboard visualisation where the occupation is linked to every other socio-economic aspect. This can be achieved because the data consists of occupation in every datasheet apart from the method of travel datasheet.

I included occupation visualisation in the first dashboard with three different socio-economic visualisations: age groups, the highest level of qualifications, and industries. The reason for considering every visualisation in a different colour is to ensure that the user can differentiate all the factors. This dashboard is one of the two dashboards created. It has two rows consisting of five visualisations and a

separate filter column. I have placed the map of England and Wales in the dashboard's centre to let the user easily navigate and click on the workplace region of their choice. I added annotations on the map and occupations visualisations to make it easy for the user to understand the dashboard action filters. The filters column on the far right provides additional filtering options to the user for every visualisation. I added a custom region name at the end of each visualisation title so that if a user clicks on a particular region, the visualisations get filtered according to the region and the same region is reflected in all the titles. For example, if a user clicks on Bristol, the occupation visualisation title changes from occupations in all to those in Bristol. This happens to all the visualisations in both dashboards.

In the second dashboard, I included the remaining two visualisations: method of travel and distance travelled to work. This dashboard also contained two rows of visualisations and a filter column. I added the same map from the first dashboard and integrated both the maps with dashboard actions. So, when a user clicks on a region of the map in the first dashboard, the filter of the region is applied to both the dashboards. If the users want to filter the region again in the second dashboard, they can click on the map in the second dashboard without going back to the first dashboard. I added annotations to both the map and the distance travelled visualisation. The reason for adding an annotation to the distance travelled visualisation is that there is no direct relationship between occupation and the method of travel. Distance travelled is the common feature in both the datasheets. When a user filters the required occupation by clicking on the occupation visualisation, the filter is applied to all the visualisations, excluding the method of travel. To include the method of travel, a user needs to filter from the distance travelled visualisation. This way, an indirect relationship is formed between occupation and method of travel. The filters column on the far right provides additional filtering options to the user for every visualisation. To navigate between both the dashboards, I created two buttons:

- Further info: The button is placed beside the title of the first dashboard. Upon clicking, it sends the user to the second dashboard.
- Back: The button is placed at the bottom of the filter column in the second dashboard. Upon clicking, it sends the user to the first dashboard.

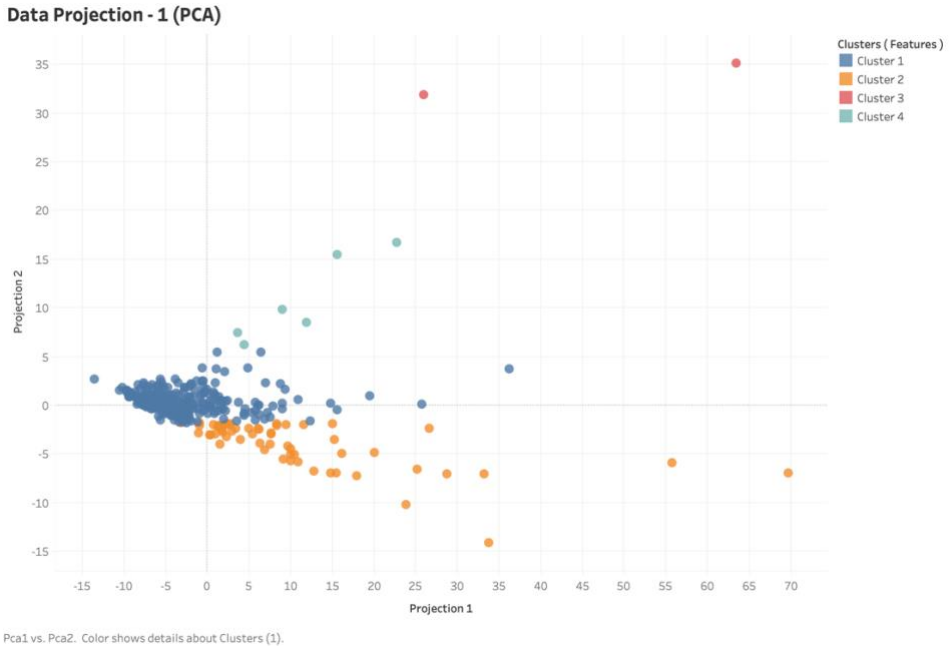
Example Task:
The user wants to get an insight into the most popular occupation in the workplace of Bristol and its demographics.
A user interacting with the dashboard:
The user first clicks on the Bristol region on the map. Due to this, a geography filter is applied on both dashboards. The user notes down the popular occupation in Bristol. Then the user clicks on the most popular occupation in Bristol in the occupations visualisation. As a result, the occupation filter is applied on both dashboards. The user notes down the results and navigates to the second dashboard by clicking the further info button. In the second dashboard, the user filters the distance travelled by most people to their work by clicking on the tile. The distance travelled is filtered across the dashboard. The user notes down the observations.
Observations:
21.7% (51.2k) of the people working in Bristol are employed in professional occupations making it the popular occupation, in which 53.3% (27.2k) of the people are from the public administration, education, and health industries. In these professional occupations, a majority (8.6k) of the people are in the age group of 30 to 34 years and have the highest qualifications of level 4 and above. 38.2% (13.2k) of the people employed in professional occupations travel less than 5 kilometres to their work, and the majority (42.9k) commute to their work using a car or van.

Two data projection algorithms have been used to reduce the dimensionality of the datasheets. Initially, the raw data consisted of a vast number of features. To reduce the number of features while retaining most of the information in the data is dimensionality reduction. I implemented two data projections using the algorithms in python and visualised the results in tableau. These are the two data projection algorithms used:

- **Principal Component Analysis (PCA)**: a statistical procedure that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

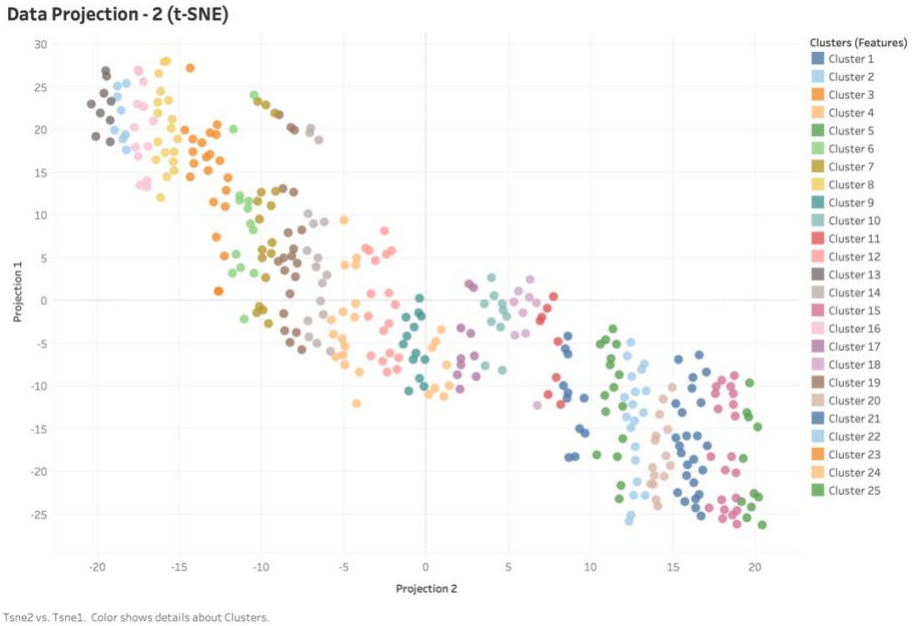
- **t – distributed Stochastic Neighbor Embedding (t-SNE)**: a non-linear dimensionality reduction algorithm used for exploring high-dimensional data. It maps multi-dimensional data to two or more dimensions suitable for human observation.

These two data projection algorithms are implemented on the occupation by age dataset. This dataset consisted of 348 rows and 123 columns. I removed all the categorical columns of the dataset. The data is not scaled to fit into the algorithms, so I scaled the data using a standard scalar. After scaling the data, I computed the eigenvalues and eigenvectors. I used the eigenvalues to compute a threshold plot to determine the optimal number of components having a threshold variance above 95%. From this plot, four components met the criteria of the threshold variance. I fit transformed the data using PCA with parameter n_components set to 4. I plotted the results and saved the dimensionally reduced data into a data frame. Only two components were considered as we cannot plot greater than two dimensions in tableau. In addition to that, I visualised the data in tableau using a scatter plot and clustered it to obtain the results.



The above visualisation represents data projection on the occupation by age using PCA in tableau. PCA reduced the dimensionality of the dataset from 119 features to 4 features (from the threshold plot). The clusters represent the number of groups the local authority regions can be formed given the occupation by age. We can group the local authority regions into four groups/clusters given the occupation by age data.

t-SNE is a probabilistic technique rather than a mathematical one like PCA. I followed a similar procedure as PCA to implement t-SNE but did not visualise a threshold plot as t-SNE considers the parameter n_components to 2 by default. I used random state to keep the same results.



The above visualisation represents data projection on the occupation by age using t-SNE in tableau. t-SNE reduced the dimensionality of the dataset from 119 features to 2 features. The clusters represent the number of groups the local authority regions can be formed, similar to the count of clusters given the occupation by age. We can group the local authority regions into 25 groups/clusters given the occupation by age data.

6. Conclusion

Successfully carried out a complete visual analytics project which consisted of various phases from data preparation to data projections. Created interactive dashboards by considering Munzner's task taxonomy that help the user to view, understand and visualise the demographic of occupations in a workplace region of England and Wales, which forms the basis of the visualisation. Overall in England and Wales, Professional occupations were the highest in which the majority of the people belonged to public administration, health and education industries. The majority of people in this occupation belonged to an age group between 30 to 34 years, with most of them having level 4 and above qualifications. More than 25% of the people in this occupation travelled less than 5 kilometres to their workplace in which most of them commuted using a car or a van. This is only an overall insight regarding the demographics of occupation in England and wales. The user is given more freedom to find more insights region-wise using the interactive dashboards. In addition to this, using two data projection algorithms reduced the complexity of the data by implementing dimensionality reduction and explained the clusters formed in tableau.

Thank You