Abstract

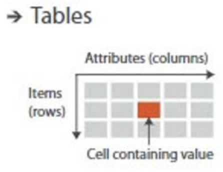
The project deals with datasets from the 2011 census in England and Wales. The aim is to design, construct, and evaluate an exploratory data analysis of census data using suitable visualisation techniques and data projection. The study strives to explore the aspect of employability and socio-economic life in England and Wales and their relationship with industry and the public's qualifications. This analysis clarifies various tasks related to this topic and presents it to the end-user in the most effective way to perceive it according to the vis principles.

Introduction

The study analyses the economic activities in England and Wales and their relationship with the industry and the public’s qualifications.
It is necessary to explore these variables and how they relate to each other to get an overall picture of the economic status in England and Wales.
In this study, census data in England and Wales of 2011 is explored, which revolve around three aspects: economic activity, qualifications and Industry.
Performing an exploratory data analysis on these datasets can help to gain useful insights to help in making better-informed decisions.
The economic activity dataset represents information that classifies usual residents by economic activity and additional information about the unemployed population (including full-time students), with estimates of the number of long-term unemployed, those who have never worked, and the classification of unemployed by key age groups. The qualification dataset outlines the information that classifies usual residents by their highest level of qualification and the industry dataset describe information that classifies usual residents by the industry in which they work.
The study aims to analyze and examine this data to discover useful findings and present it smoothly and efficiently for end-users to interact with it.
The outline approach is to design, create, and evaluate an exploratory analysis of the census dataset using different visualisation techniques and data projection.
The rest of the report is organized as follows: Data Preparation and Abstraction is presented in Section 2. In Section 3, I describe the Task Definition. The visualisation Justification that was based on to derive and present the plots are described in Section 4. Evaluation of the work is described in Section 5, conclusions in Section 6.

Data Preparation and Abstraction

The data has been collected from the National Offender Management Information System (NOMIS) website using its query search, which provides various selection of tables for census 2011.



Dataset Structure

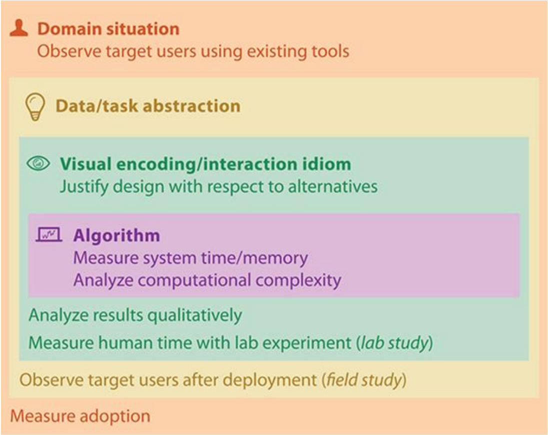
The dataset consists of various static tables. The variables/attributes types are nominal such as country and gender, ordinal such as qualification levels and discrete measures.

- **Dimensionality Reduction (PCA)**

Here I Included a feature dimensional analysis for the economic activity per county. The economic activity dataset has high dimensionality features and I reduced them into low dimensionality features with two principal components. Then, I projected the data into a two-dimensional space and clustered the data points with the county as the index label. It seems to give good information with only two features and does not have any extra knowledge.

Evaluation

Validation is very critical for the visualisation design because the vis design space is large, and most designs are ineffective. Therefore, it is a continuous process to improve the vis appearance to make it more interactive and user friendly. Here I assessed my design process based on the four-level of design as shown in the graph below.


Domain situation
Observe target users using existing tools
Data/task abstraction
Visual encoding/interaction idiom
Justify design with respect to alternatives
Algorithm
Measure system time/memory
Analyze computational complexity
Analyze results qualitatively
Measure human time with lab experiment (*lab study*)
Observe target users after deployment (*field study*)
Measure adoption

Four levels of design

In the domain situation, I validated the group of target users, their domain, their questions and their data. Users have a certain vocabulary to describe the problem domain. Hence, I addressed the tasks of my visualisations to my classmate's evaluators by writing a brief description of the main tasks and sub-tasks. Then, I asked them in the questionnaire form questions related to the domain of the study to answer them on a scale-out of 5 where 5 is the highest confidence. An example question is whether the visualisations techniques clearly illustrate the main goal of the study to the user and fully display the information that needs to be explored. Moreover, validating if these visualisations help the audience to clearly understand the result of the research.

Regarding the data/task abstraction, I seek to reduce the cognitive load for the user by making sure that the visual representation addresses the task problem. The questions I asked the users in this validation section including whether the goals of the tasks clear and well defined. Similarly, If the data for both tasks well abstracted and used, that is, the tasks are conveyed to a generic description. I also asked the user to give their feedback on the definitions of the tasks presented such as Economic activity in each county.

In the visual encoding/interaction idiom, I validated to make sure that
the design of the visualisation is guided by the abstract tasks and measure my idiom design choices. So, I asked the user in the questionnaire form to reflect on my visualisations design for good interactivity. Besides, assessing my visual graphs in giving the user a clear and

straightforward perception. The colours are another aspect here whether they signify both genders in the study clearly.

Finally, for algorithm evaluation, the goal is to measure the computational complexity of the algorithm.
I asked the user to evaluate the design of the PCA dimension reduction algorithm and the K-means clustering algorithm.

Conclusion

Overall, regarding the study task, there is a close relationship between the employment statistics of the population of England and Wales and the level of education of the public, which reflects this in the distribution of different jobs in the industry. As is evident in the analysis that I conducted in this study, there is a significant bias in employment, with most of them centred in England more than Wales and especially in the southwest region. As for the visual analysis presentation, I learned that there are many forms of displaying information, but it requires a lot of effort, thinking and experiments to choose the most appropriate visualisation technique that makes access to information by the user easier and more effective.