KING'S College LONDON

University of London

Population visualization tool for Lambeth Council

Final Project Report

Author: Jan Aldous Torres

Supervisor: Dr Simon Miles

Student ID: 1323454

April 15, 2017

Abstract

(Will be filled after the completion of the report). The abstract is a very brief summary of the report's contents. It should be about half-a-page long. Somebody unfamiliar with your project should have a good idea of what your work is about by reading the abstract alone.

Originality Avowal

I verify that I am the sole author of this report, except where explicitly stated to the contrary. I grant the right to King's College London to make paper and electronic copies of the submitted work for purposes of marking, plagiarism detection and archival, and to upload a copy of the work to Turnitin or another trusted plagiarism detection service. I confirm this report does not exceed 25,000 words.

Jan Aldous Torres

April 15, 2017

Acknowledgements

I would like to thank my supervisor Dr. Simon Miles for his support in this project. It is much appreciated.

I thank Lambeth Council for their support in the requirements and evaluation stages of the project. Taking time off their busy schedules was much appreciated.

Contents

1 Introduction **4**
1.1 Report Structure 5

2 Background **6**
2.1 Data analysis in government 6
2.2 Customer segmentation 6
2.3 Related applications 7
2.4 Data visualization and exploration 8

3 Requirements **9**
3.1 Interview with Lambeth Council 9
3.2 Objective 10
3.3 The users 10
3.4 Description of data used 10
3.5 Functional Requirements 11
3.6 Method of clustering 12
3.7 Data visualization 12
3.8 Functional Requirements 13
3.9 Non-functional requirements 13

4 Specification & Design **14**
4.1 System architecture 14
4.2 User interface 15

5 Report Body **16**
5.1 Section Heading 16

6 Implementation **17**
6.1 Section Heading 17

7 Professional and Ethical Issues **18**
7.1 Section Heading 18

8 Results/Evaluation **19**
8.1 Software Testing 19
8.2 Section Heading 19

9 Conclusion and Future Work **20**
Bibliography 21

A Extra Information **22**
A.1 Tables, proofs, graphs, test cases, ... 22

B User Guide **23**
B.1 Instructions 23

C Source Code **24**
C.1 Instructions 24

List of Figures

4.1 Pie chart visualization 14

Chapter 1

Introduction

[**?**] [5] [6] [7] [**?**] [9] [1] [2] [8] [4] [3]

Local government councils are challenged with the recent cuts in local government funding and a diverse population with different needs. Therefore, there is a need to efficiently allocate resources. The usual approach is to tackle issues raised by individual social services. However, there has been a shift in local government, which is promoted by the national government, to utilize the concept of customer segmentation. This will let them understand the needs of the population better thus letting them utilize and allocate resources in the most effective way based on those needs [4]. This approach would require the council to identify and therefore focus resources on vulnerable population groups. For example, instead of focusing on issues on education, they would focus on the children in care who are most susceptible to lower educational, social and employment outcomes. Some social services deal with the same population groups so this approach would require an integration of data from the different services. This gives an overview of all data from services provided by the councils. This is unlike the former approach which requires individual social services to conduct data analysis using their own data.[TJA1]

Lambeth Council in London wants to implement the new approach. However, current methods to analyze and visualize data take a long time before the analyses and visualizations would reach the commissioners, the people in the council who formulate strategies to tackle issues.

There are off-the-shelf software packages available which supports users in customer segmentation from clustering to the visualization of their data. In government, using data analysis in policy decision making is becoming more popular(new york guy video). The Local Government

Agency has created a guideline to support councils in tailor-made clustering and visualization [4]. Moreover, there have been projects by other local councils to visualizations of data on prominent population groups.

Lambeth Council would like to have a bespoke population segmentation tool that would be customized to their own needs and available data. This project aims to implement a prototype of such a tool as a Django web application.

The tool will focus on the situation in which population groups are in through visualizations. The situations they face will be defined by the data. Visualizations of this data will also help to compare characteristics of the group to the wider population and subgroups. This comparison may lead to observations on which problems a group or subgroup is facing. This may lead to the conclusion that the council may want to increase resources spent on a group or service. The tool basically aid the commissioner to apply customer segmentation on their population and identify which problems segments of the population are facing thus help make decision on the allocation of their resources.

Clustering will provide a more complete perspective into the different situations groups face instead of viewing a group as one entity and defines variables which distinguish the group from the rest of the population using clustering algorithms. Data integration from open government datasets will create a more in depth view of groups. And possible relations between geographical locationsâĂŹ characteristics ie facilities (GP), situation (crime) in a ward and whether they are on the London living wage. Since the user may not necessarily know what they are looking for when they read the data, the visualizations will be in the form that will let the user explore more aspects of the data.

1.1 Report Structure

The remainder of this paper is organized as follows. Chapter 2 introduces related work to the tool and data exploration, and Chapter 3 discusses the requirements of the tool. Chapter 4 outlines the specification and design of the tool. Chapter 5

Chapter 2

Background

This chapter includes work done on formulating better ways of with data exploration. It will later describe in more detail the work done on the access and visualization of open government data to non-technical people who require access to this type of data. The chapter ends with related application to the prototype of this project.

incorporates all relevant issues

main issues are analysed and developed

2.1 Data analysis in government

Give overview of using data in policy decision making

There has been work to make government data more accessible to people which do not have the technical abilities to merge and manipulate data such as journalists, data journalists, etc. [10]. This is especially true in open government data platforms where datasets take non-standard forms and whose data may be encoded in a way that is not immediately understandable. There have been efforts to evaluate government platforms and to enumerate features that would effectively let users make sense of the data [8].

2.2 Customer segmentation

Market segmentation or customer segmentation is concept used in business regarding the division of a homogenous population into segments which have similar attributes, wants, needs or demands. âĂIJIts objective is to design a marketing mix that precisely matches the expectations of customers in the targeted segment. Few companies are big enough to supply the needs

of an entire market; most must breakdown the total demand into segments and choose those that the company is best equipped to handle.âĂİ (businessdictionary.com). âĂIJThe four basic market segmentation-strategies are based on: behavioral, demographic, psychographic, and geographical differencesâĂİ (businessdictionary.com). Much of the need to have this is due to the lack of resources needed for the whole market. For many of the local government units, it is the constraint of the budget that would require an efficient allocation of resources by choosing specific groups. The national government and LGA promote the application of this concept usually used in the private sector into local government.

Clustering algorithms could be used to divide the population.

According to the LGA [], the collation of demographic data and the use of clustering algorithms such as k-means could be used to divide the population into segments. Choosing which

2.3 Related applications

There have been applications developed similar to the tool this project aims to implement. These include more general purpose customer segmentation tools which allows the user to analyze and then segment data using algorithms, to the visualization and use of that data. There have also been local governments which are following the concept of customer segmentation to also spend their resources more effectively.

Customer segmentation tool like Mosaic by Experian (mosaic) and Acorn by CACI (acorn) are tools both to allow population segmentation through customer profiling based on set of demographic and other indicators.

Customer classification is a concept promoted by the government and the Local Government Association (LGA) has created a guidance document for any council that wishes to implement such a tool (LGA) (smart cities).

Kent and Medway has tool which segments its population by social class and aims to highlight the key features which make each Group distinctive, to help you visualise the segmentation data and understand the essence of each Group (Kent Medway). It lets the user compare the values between all groups. This tool is bespoke to Kent MedwayâĂŹs data and the tool cannot be reused for other councilâĂŹs data.

There have been similarities between tools and recommendations to differentiate and describe each group. (Kent) (LGA) suggested to include maps showing the concentration of that group in each ward and in addition to a textual description, they include pictures to describe

each group. (Barnet) has a report on the customer segmentation of its population, and includes the percentage and textual description of its population. (Kent) includes graphs to visualize data but also word maps of key characteristics. (LGA) has suggested that the use of spider diagrams which detail the variables compared to the town and district average. Similarly (Acorn) (Kent), they graph pieces of data with an index of the group compared to the overall population. This visualizes both the average value and the groups relation to that average displays whether they are higher or lower than that average however each graph is created for each variable.

LGA has suggested data integration to enable other uses of data. (smart cities) defined use of explicit, âĂIJrecords of who has or is using a serviceâĂİ and implicit âĂIJknowledge of staff on customers using a serviceâĂİ customer data in an effort to know their customers in great detail.

2.4 Data visualization and exploration

Numerous applications exist to allow users to interact with data. Data exploration involves different types of interaction and functionality and there has been work to create a framework for data exploration [9].

There has been work done on enumerating the problems with data exploration through visualizations [10].

There has also been work done on the visualization of clusters through graphs, which show how closely related the clusters are to each other [7].

Show thinking of why some features of related applications are used here Use cases/features of Kent and Medway: - See group data, more specifically see pictures and phrases which describe the group, distribution between wards, data about services used, benefits used and demographics of the group as an index value (0-200) and a map of the groupâĂŹs population density in the councilâĂŹs area Features of neighbourhood.statistics.gov.uk - The user selects the data to be displayed on the map and chart. Map and geographical unit breakdown chart interacts with each other where hovering over one will highlight the area on both map and chart. The map is divided into geographical units (i.e. ONSâĂŹ lower output areas) which are colored according to the band which the area is in. The chart is according to the userâĂŹs setting.

Chapter 3

Requirements

This section outlines the requirements of the prototype based on correspondence with the user. The first subsection outlines the objectives. The following subsections describe the operations on data by Lambeth Council which is followed by the functional and non-functional requirements.

The user has expressed their intention of following an iterative implementation of the final product, thus this is only the requirements for the first iteration of the prototype.

3.1 Interview with Lambeth Council

These requirements were based on what I have learned in communication with, Noel Hatch, a line manager at Lambeth Council who communicates with the commissioners in the council. He suggested a tool that would allow the commissioners to create groups through the method of grouping described below (see Method of grouping) using Kent and MedwayâĂŹs application as inspiration. The tool will allow the user to navigate between groups and visualize the data. Should I have time, allowing the commissioner to analyze the data further is a secondary main objective. He also described the tool should be user-friendly in that the IT skills and data analysis skills of the commissioners are limited.

I had limited time with the council such that after the initial requirements, I was not able to communicate until the evaluation stage of the project. I continued the project with the information I had and used this and inspiration from related applications to come up with features that could potentially support the users. Thus the requirements, design and implementation therefore may not be what the users have envisioned.

3.2 Objective

The purpose of this application is to create a piece of software which will allow local government councils to explore their residential data through graphical and map visualizations. The intent of which is to support them to identify which segments of the population or parts of Lambeth geographically and socially do resources require the most.

3.3 The users

The primary users for the program are the commissioners, those responsible for formulating strategies to allocate resources for a council based on the presentation of data analysis given to them. They do not necessarily have technical or statistical analysis skills in terms of being able to operate applications and interpret data.

3.4 Description of data used

A combination of local and national data provided by Lambeth and data collected in past censuses will be used:

- **CSV of Lambeth's 2016 Residential Survey**: a survey of a sample of LambethâĂŹs population consisting of 1024 people about their quality of life, what they thought of LambethâĂŹs services and about the respondents themselves. The data is in the form of CSV text files where each row contains a personâĂŹs entries as categorical data in the form of code (see below Survey Code Translation). Each single answer question has its own column and entries are in code as an integer between 1-100. For the questions which have multiple answers, each answer is regarded as a sub question in the form of âĂIJQ5AâĂİ meaning question âĂIJ5âĂİ, choice code âĂIJAâĂİ (makes it look like a string but its supposed to be an integer) (see below Survey Code Translation). Therefore, a sub question has its own column and entries are in code as a âĂIJ1âĂİ or âĂIJ0âĂİ. There are also other fields which have been added to the original survey results such as group, subgroup, quintile, which is a result of previous data analysis.
- **LambethâĂŹs 2016 Residential Survey Code Translation**: a Microsoft Word document of the original survey with the original questions and code used in the data associated with the question. Under each question is the choice code and English meaning of the choice. The choice code is either a number for single answer questions (e.g. 1. Male,

2. Female) or letters for multiple answer questions (e.g. A. Access to nature, B. Activities for teenagers, etc.).

- **LambethâĂŹs Open Data**: from LambethâĂŹs website, it is a compilation of mainly geographical data such as the locations of public amenities and geographical specifications of each ward. These are in the form of CSV or GeoJSON or files.
- **Office of National Statistics data**: The ONS has a vast number of open data sets. They include datasets from the census, such as employment, housing tenure, etc. They are also in the form of JSON files

3.5 Functional Requirements

Functional requirements with the words textbfshould or textbfmust is a required feature. Requirements with the word textbfmay, is a desirable feature which may or may not be implemented should there not be enough development time.

1. Data storage
2. The system may be able to save access (through URLs to CSV and JSON files) or actual files of the datasets (CSV or JSON files themselves). The URLs must be kept in persistent storage. If the input is in the form of a file, the file must be stored.
3. The text in the entries may be of any length.

Grouping requirements 2. The user should be able to create groups based on the parameters to the 5 factors (see Method of grouping subsection). The groupsâĂŹ parameters should be kept in the database. 3. The user should be able to view data about a group through graphs. a. Single code questions or questions which requires only one answer should be visualized as stacked bar charts or pie charts b. Multi code questions or questions which requires more than one answer should be visualized as bar charts c. Selecting data on a graph should display that data on the map (see functional requirement 5) 4. The user should be able to compare a data variable between all groups such that groups should be compared to the average value of the variable (e.g. compare the percentage of disabled people who are male to the percent of overall population who are male). 5. The system should display data based on the respondentâĂŹs residence on a map. The map may display data within the councilâĂŹs wards. Clustering 6. The system should divide the group using a clustering algorithm. 7. The system should display

how the group has been clustered and any statistical information about the differences between the clusters. 8. The user should be able to compare information between the groupâĂŹs data and each of the clusterâĂŹs data. 9. The main data used in the visualizations should be the CSV of LambethâĂŹs Residential Survey and its Code Translation. Data integration 10. The system may integrate data from other sources (i.e. governmentâĂŹs open data, councilâĂŹs open data) to a ward (or post code if possible). 11. The system may display integrated data on a map. The user may be able to compare a map about a field in the integrated data to a map about a field in the residential data.

3.6 Method of clustering

There are 5 factors in which a user can create a cluster:

1. Disability/illness
2. Whether they are on benefit support
3. Educational/employment activity
4. Whether they are on the London Living Wage or not
5. Housing tenure (i.e. council tenant, private owner, etc.)

A cluster is based on the set of answers to the 5 factors the user inputs. A resident must have the exact set of answers to be part of that cluster.

3.7 Data visualization

After a cluster has been created, the user must be able to see the answers to the following survey questions (question number in brackets denotes the column name in the SPSS file):

1. What matters most to them most (Q5)
2. What their last contact with the council was (Q26)
3. How they use the website (Q27)
4. What services they have used (Q39)
5. How they access the internet (Q50),

6. How well the changes have benefited them (Q11)
7. What they value in terms of community cohesion (Q13)
8. Gender (QGEN)
9. Age (QAGE)
10. Ethnicity (QETH)

The answers to the questions above will be visualized under each of the clusters.

3.8 Functional Requirements

The following enumerates the functional requirements for the first iteration of implementation. Another set of requirements will be added for the following iterations.

1. The user should be able to input an SPSS file into the system.
2. The user should be able to create clusters by setting parameters to the 5 factors 3.6. The setting for each parameter should be saved for future use.
3. The user should be able to navigate between clusters easily.
4. The system should visualize data as described in 3.7 for each cluster.

3.9 Non-functional requirements

Usability: One of the main aims of the system is to interface data to the user in an effective way. Since the users are not technical, the ease of use is imperative to the design of the user interface.

Maintainability: This version of the system will be added on to the future with future data integration and other features, therefore there will be a need for maintainability of the code.

Security: The data being used is anonymous therefore there will be no need for security measures for the data.

Chapter 4

Specification & Design

This section describes a high-level design of the system to be implemented. As mentioned in the Requirements section, this will also be the first version of the design documentation. It will be added on to in the future.

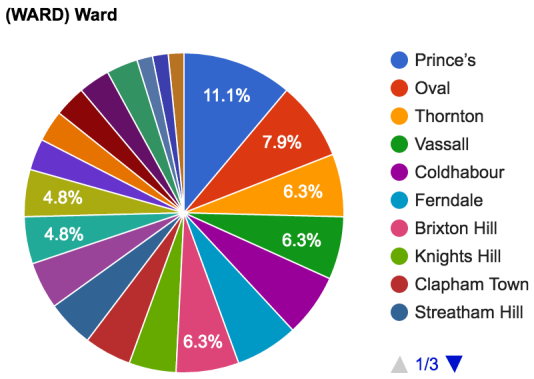

Figure 4.1: Pie chart visualization

4.1 System architecture

Due to the iterative nature of implementation, the architecture chosen will make it easy to add more components and even more data sources. The system will follow the three-layer architecture. The data layer will consist of the survey data and cluster criteria, parameters (based on the 5 clustering factors) for clusters set by the user. The logic layer will consist of creating clusters and extracting clusters from survey data. The presentation layer includes the

form to create clusters and the navigation between clusters and visualization of cluster data.

4.2 User interface

The systemâĂŹs main goal is to present data and let the user interact with the user in an effective way. Thus the design of the user interface is integral to the success of the system.

The cluster creation UI will enable the user to name and enter the parameters for a cluster. The choices for the parameters will be constrained to the fields in the survey and the fieldâĂŹs answers.

To enable the easily navigate between each cluster, there will be a navigation pane in the form of tabs for each cluster. Once a tab is selected, the graphs of data listed in section (data visualization) will be displayed. To enable comparison between data on each cluster, inspired by (kent and medway), each graph will be shown as indexed horizontal bar charts (figure). This will visualize the comparison of the values of the cluster to average value for that field in the whole population.

Chapter 5

Report Body

The central part of the report usually consists of three or four chapters detailing the technical work undertaken during the project. **The structure of these chapters is highly project dependent**. They can reflect the chronological development of the project, e.g. design, implementation, experimentation, optimisation, evaluation, etc (although this is not always the best approach). However you choose to structure this part of the report, you should make it clear how you arrived at your chosen approach in preference to other alternatives. In terms of the software that you produce, you should describe and justify the design of your programs at some high level, e.g. using OMT, Z, VDL, etc., and you should document any interesting problems with, or features of, your implementation. Integration and testing are also important to discuss in some cases. You may include fragments of your source code in the main body of the report to illustrate points; the full source code is included in an appendix to your written report.

5.1 Section Heading

5.1.1 Subsection Heading

Chapter 6

Implementation

6.1 Section Heading

Chapter 7

Professional and Ethical Issues

Either in a seperate section or throughout the report demonstrate that you are aware of the **Code of Conduct & Code of Good Practice** issued by the British Computer Society and have applied their principles, where appropriate, as you carried out your project.

7.1 Section Heading

Chapter 8

Results/Evaluation

8.1 Software Testing

8.2 Section Heading

Chapter 9

Conclusion and Future Work

The project's conclusions should list the key things that have been learnt as a consequence of engaging in your project work. For example, "The use of overloading in C++ provides a very elegant mechanism for transparent parallelisation of sequential programs", or "The overheads of linear-time n-body algorithms makes them computationally less efficient than $O(n \log n)$ algorithms for systems with less than 100000 particles". Avoid tedious personal reflections like "I learned a lot about C++ programming...", or "Simulating colliding galaxies can be real fun...". It is common to finish the report by listing ways in which the project can be taken further. This might, for example, be a plan for turning a piece of software or hardware into a marketable product, or a set of ideas for possibly turning your project into an MPhil or PhD.

References

[1] 3. customer segmentation and profiling.

[2] Customer segmentation.

[3] Customer segmentation.

[4] Developing a customer classfication tool: Guidance document for local authorities.

[5] Kent & medway interactive guide.

[6] Brigitte Boden, Roman Haag, and Thomas Seidl. Detecting and exploring clusters in attributed graphs: a plugin for the gephi platform. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 2505–2508. ACM.

[7] Drew Conway. Data science through the lens of social science. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1520–1520. ACM.

[8] Anna Gogolou, Marialena Kyriakidi, and Yannis Ioannidis. Data exploration: A roll call of all user-data interaction functionality. In *Proceedings of the Third International Workshop on Exploratory Search in Databases and the Web*, ExploreDB '16, pages 31–33. ACM.

[9] Alvaro Graves and James Hendler. Visualization tools for open government data. In *Proceedings of the 14th Annual International Conference on Digital Government Research*, dg.o '13, pages 136–145. ACM.

Appendix A

Extra Information

A.1 Tables, proofs, graphs, test cases, ...

The appendices contain information that is peripheral to the main body of the report. Information typically included in the Appendix are things like tables, proofs, graphs, test cases or any other material that would break up the theme of the text if it appeared in the body of the report. It is necessary to include your source code listings in an appendix that is separate from the body of your written report (see the information on Program Listings below).

Appendix B

User Guide

B.1 Instructions

You must provide an adequate user guide for your software. The guide should provide easily understood instructions on how to use your software. A particularly useful approach is to treat the user guide as a walk-through of a typical session, or set of sessions, which collectively display all of the features of your package. Technical details of how the package works are rarely required. Keep the guide concise and simple. The extensive use of diagrams, illustrating the package in action, can often be particularly helpful. The user guide is sometimes included as a chapter in the main body of the report, but is often better included in an appendix to the main report.

Appendix C

Source Code

C.1 Instructions

Complete source code listings must be submitted as an appendix to the report. The project source codes are usually spread out over several files/units. You should try to help the reader to navigate through your source code by providing a "table of contents" (titles of these files/units and one line descriptions). The first page of the program listings folder must contain the following statement certifying the work as your own: "I verify that I am the sole author of the programs contained in this folder, except where explicitly stated to the contrary". Your (typed) signature and the date should follow this statement.

All work on programs must stop once the code is submitted to KEATS. You are required to keep safely several copies of this version of the program and you must use one of these copies in the project examination. Your examiners may ask to see the last-modified dates of your program files, and may ask you to demonstrate that the program files you use in the project examination are identical to the program files you have uploaded to KEATS. Any attempt to demonstrate code that is not included in your submitted source listings is an attempt to cheat; any such attempt will be reported to the KCL Misconduct Committee.

You may find it easier to firstly generate a PDF of your source code using a text editor and then merge it to the end of your report. There are many free tools available that allow you to merge PDF files.