DE5 Group Dashboard Project

Health in Scotland

Project Description Outline

Group 1: David Currie, Mark Donaldson, Calum Sey, Geraldine Smith

December 2020

Team Roles and Responsibilities

David focused on part of the overview data, looking at life expectancy as the chosen indicator. David used his skills in spatial data to create an interactive map showing life expectancy by health board. He also incorporated SIMD quintile data to a plot to show any potential trends in life expectancy versus deprivation. David also assisted the other team members with his debugging skills.

Mark explored data related to smoking and its impact on health in Scotland. He created interactive plots to show trends in smoking by council area. Mark used his skills in data wrangling and cleaning to provide insights into the temporal, geographic and demographic trends in smoking in Scotland.

Calum's main task in the project was data exploration on the themes of drinking and drug behaviours in Scotland, in both a temporal and geographical sense. He cleaned and wrangled the required data to give insights into how these behaviours have been affecting key NHS services across the country, such as hospital admissions. Calum also managed the team's main project management tool, Trello.

Geraldine focused on another part of the Scottish health overview data, looking at life satisfaction as the chosen indicator of quality of life. She cleaned and wrangled the data to allow comparison of the indicator by sex and region(NHS board and country-wide). She also took on the task of creating the project report.

Everyone in the team took ownership of drafting the project presentation, and the final set up of the dashboard after all features were implemented.

Dashboard Topic

The team created a dashboard containing two separate themes. The first was an overview of general health in Scotland. This section provides data on two main indicators of general health:

- Life expectancy
- Life satisfaction

The dashboard is split into four main tabs:

- Health Overview
- Smoking
- Alcohol
- Drugs

Health Overview
This section provides insight into two key indicators of general health - life expectancy and life satisfaction. Life expectancy data is displayed geographically in the form of a map and temporally. SIMD quintile data is also displayed and can be viewed in conjunction with life expectancy. Life satisfaction data displays data geographically and by sex demographic. The user can compare data by NHS health board and also across Scotland as a whole.

Smoking
This dashboard section is in a separate tab and allows the user to explore the percentages of people who smoke in Scotland. Trends in the data can be determined over different demographics such as age range, gender and council area. One issue encountered in selecting specific genders/age range was that no data was plotted. Extension work for this would be to display a text message on screen to inform the user that insufficient data was available to display this combination of parameters.

Alcohol
This section displays data related to alcohol related conditions treated in Scottish hospitals. Trends in the data can be determined over different demographics such as age range, gender and council area. Data is available for the last 10 years, allowing temporal trends to be explored.

Drugs
This dashboard tab displays data relating to drug related discharges from Scottish hospitals over the last 10 years. Data can be compared across different geographic regions (council areas) and compared temporally.

Project Stages

Team Dynamic and Project Plan
The first stage of our project involved the team exploring effective ways of working together using the learning support material provided earlier in the course. We established ground rules and a basic plan of how each day should run. We also discussed which health topics everyone was interested in. The project brief was reviewed and key requirements recorded in a MoSCoW structure. A general timeline was also proposed thus:

- Thurs 3rd Dec: Team dynamics, team objectives, data topics exploration
- Fri 4th Dec: Decide on chosen data topics, best practice for cleaning data, basic dashboard structure
- Weekend: Data Exploration of chosen topics
- Mon 7th Dec: Team discussion on most suitable data sets, further cleaning/wrangling
- Tue 8th Dec: Create cleaning scripts and basic shiny layout for each section
- Wed 9th Dec: Continue with shiny layout and plots required, begin merging of features
- Thur 10th Dec: Debugging completed dashboard, work on report and presentation
- Fri 11th Dec: Group presentations

The team used the project management tool Trello to formalise our strategy:

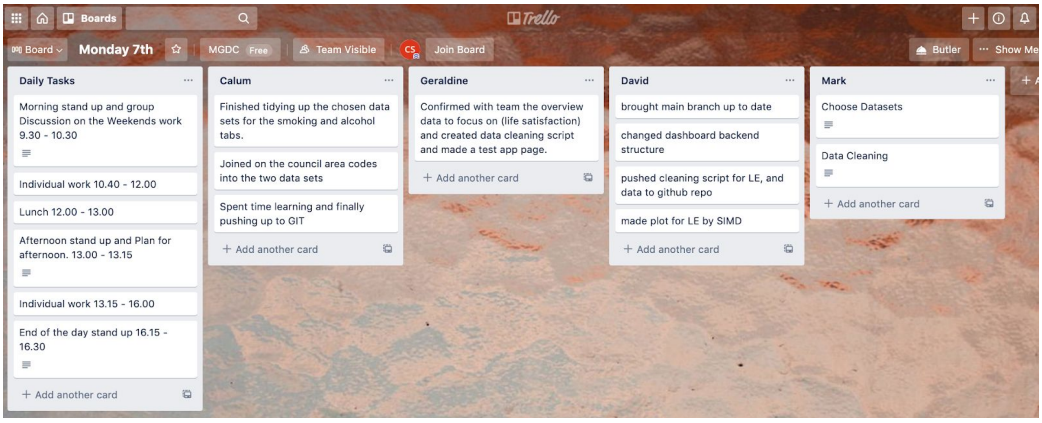

Figure 1 - Team Trello project management board

Data Exploration and Dashboard Concept

The second stage of the project involved data exploration of our chosen topics with a view to narrowing down our selections to allow us to create a dashboard wireframe. The team decided to focus on two topics for general overview - life expectancy and life satisfaction. Our chosen topic of focus was decided through a team review of the Scottish government publication "Public Health Priorities for Scotland" [1]. It was decided the specific health topic to focus on would be Scottish Government Priority 4 - *"A Scotland where we reduce the use of and harm from alcohol, tobacco and other drugs"*. Therefore data sets were chosen relating to drug, alcohol and smoking behaviours in Scotland. Having chosen the topics, the dashboard wireframe could then be constructed:

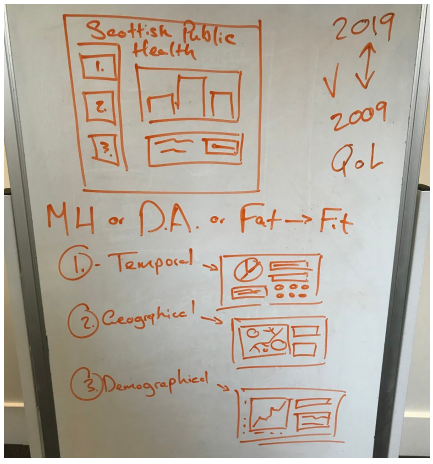

Figure 2 - initial dashboard wireframe

Data Cleaning and Wrangling

The third stage of the project required the chosen data sets to be wrangled/cleaned so that they were in a suitable structure for use in plotting and the dashboard. Data insights were also discussed and noted for coverage in the final presentation. At this point, each member of the team created feature branches on Git to allow commits to be pushed for each section. Each

team member also tested their data set on a remote dashboard to ensure the code was working prior to beginning the merge process.

Feature Merging and Consolidation
The penultimate stage of the project involved merging each feature, which in this case was a seperate tab for each topic area and an additional plot on the “General Overview” tab. This was carried out and any conflicts were discussed and fixed by the team as a whole, and the app was tested after each feature addition to ensure it was still working.

Final Dashboard Review and Presentation
The final stage of the project involved a mob programming session to ensure any kinks in the dashboard were corrected, and the team discussed the insights drawn from each of the data sets. The group presentation was also created and the team finalised the plan for presenting.

Project Management Tools

The team used a number of tools to help the project run smoothly over the week. The main communication tool used was Zoom, which allowed remote team members to communicate easily with the other team members and instructors. Twice-daily stands ups were held, and this allowed everyone to understand the team objectives for the day. Zoom and screen-sharing were also used for any mob programming/debugging sessions. Slack was an additional communication tool used, which was helpful for sharing code or debugging errors in code.

Trello was used as a project planning tool. It allowed the team to record their progress each day, and also kept a record of any planned tasks that still had to be completed. Trello was also used as a hub to store the team ground rules.

Git/GitHub were used primarily as tools for version control, so that each team member could work on an individual feature and then merge these features together. It was also used as a back-up system, so that if any team member lost data they would be able to retrieve it.

Synthesising Project Requirements

The team reviewed the project brief regularly whilst exploring the available data sets. As mentioned previously, a MoSCoW strategy was used to decide on the features that were critical to satisfying the brief, whilst including “nice-to-have” feature options if there was available time towards the end of the project.

The Q&A session with the client was used to show them our wireframe proposal for the dashboard, and discuss the chosen general health indicators, as well as deep-dive topics. We also confirmed with them the temporal, geographical and demographic features that would be included in each section and it was agreed the team were on the right track.

Chosen Data Sets - Motivation

The motivation for choice of the data sets is outlined below, along with any potential reasons for bias:

Life expectancy [2] - the team viewed this indicator as a suitable, simple measure of the overall health of the nation. It is used globally as an indicator of general health. The data set chosen also contained variables relating to a number of demographics such as age and gender. Data over a number of years was available, so this was viewed as useful in assessing temporal trends. This data set also related directly to SIMD quintiles and could therefore be used for extension work if required.

Life satisfaction [3] - the team wanted to include a measure of quality of life, as this was viewed as an important factor in wellbeing. Life satisfaction was chosen as the indicator, because this is a self-assessed parameter and as such takes into account the individual's mental and physical wellbeing. The chosen data set allowed comparisons across gender and health board areas. The data was also pre-aggregated so allowed comparisons at a country versus health board level.

Smoking [4] - the data set chosen was from Scottish Survey Core Questions and provided percentages of those who smoked. Comparisons were possible across temporal, geographical and various demographics so potentially allowed a number of insights into factors which affect smoking prevalence.

Drugs [5] - the data set chosen to assess drug behaviours in Scotland was Drug Related Hospital Discharge. This was chosen as it gives an understanding of the impact drug behaviour has on hospital resources across Scotland. Comparison was also possible geographically with this data set.

Alcohol [6] - Alcohol Related Hospital Statistics was chosen as the data set to assess the impact alcohol abuse has on the Scottish health care system. Insights into how this data varied by region and temporally were possible with this data set.

Data Quality, Potential Bias and Cleaning

The data quality, potential bias and cleaning methods for each data set are discussed below. It was noted that each individual team member carried out their own cleaning method and created an individual cleaning script for their feature branch on GitHub, however it was felt the project structure would be more efficient with one single cleaning script that cleaned all required data sets. The individual scripts were therefore merged into one after all merging had been completed:

Life expectancy - in terms of data quality, it must be recognised that life expectancy at birth is an estimate. The metric does not account for future changes to health and social care

standards. The data set provides confidence intervals and these were included to give a higher standard of data quality. It is unlikely that bias exists in this data set, because it is a calculated measure based on mortality rates and these would be recorded in absolute terms, so there should be no sampling bias. This data set was cleaned by filtering extraneous data, such as life expectancy measured at ages beyond birth. Variables such as urban rural classification were also removed. The data was joined to an additional data set, which defined health board regions by name, so this could be paired with the feature code in the original data set.

Life satisfaction - the data quality of this data set is dependent on the honesty of the respondents in the survey. It is a representative sample of the general population living in private households. Therefore, it does not take into account individuals in institutions such as care homes or long-term health facilities. This would introduce bias into the data set, as it is likely those individuals would perhaps view their quality of life on the lower end of the scale. It was also not possible to infer temporal trends in this data set as it was already aggregated into 4 year groupings. This data set was cleaned to remove all the indicators that were not of interest at this time, such as self-assessed obesity, fruit and vegetable consumption and various others. Filtering was used to allow comparisons by NHS health board level, as well as Scotland-wide. Joining was also required to match the health board name with feature code, as feature code also included council area. Minimal recoding was required to classify the life satisfaction variable into a cleaner format.

Smoking - The quality of this data set was deemed satisfactory, as the sample size given in the metadata was 20,000, which potentially allowed analysis to be carried out across many demographics such as age, gender, education level and household type. Confidence intervals were also included in this data set, but the team chose to focus on the point measurement which was deemed satisfactory to show general trends in smoking behaviours. The metadata also highlights that care should be taken if comparing by electoral ward with other more recent data sets as the ward boundaries may have changed. The dataset was cleaned by removing all the indicators that were not of interest at this time, such as tenure type and household type. Filtering was used to allow comparisons by council area, gender and age. A join with a separate csv data set was also required as this was used to match the council area name with feature code.

Drugs - the quality of this data set was deemed satisfactory as it records drug related hospital admissions in absolute terms, i.e. is not a representative sample. The data quality is checked by ISD and compared to previous years/expected trends to ensure no obvious anomalies are detected. The potential bias that exists in this data set could be related to drug treatments which are carried out outwith a hospital setting. The cleaning process involved reading the data through janitor to standardise the data format. Then the necessary columns were selected and renamed appropriately. The date codes were ordered through mutate and some strings were removed or edited within the chosen columns. Afterwards, the council area codes were read in on a .csv and merged with the data along the feature_code variable, which produced the clean data sheet that was used within the project.

Alcohol - the quality of this data set is similar to the drug data set in that it records absolute acute health incidents related to alcohol consumption. Therefore, the data set does not account for any effects of alcohol which cannot be directly related to excessive consumption. Again, the data quality is checked by ISD and compared to previous years/expected trends to ensure no obvious anomalies are detected. The potential bias that exists in this data set could be related to alcohol treatments which are carried out outwith a hospital setting. This data was cleaned in a similar fashion to the drugs data set.

Data Structure

The data on statistics.gov.scot is stored as linked data. This allows multiple data sets to be linked. The datastore is made up of millions of "triples" - that is a subject - predicate - object arrangement. This is shown in the diagram below:

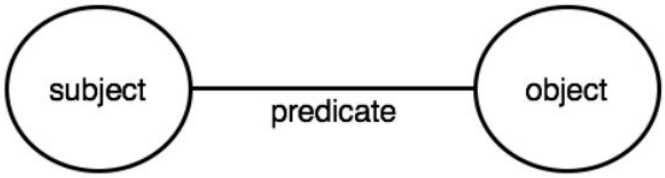

Figure 3 - from https://guides.statistics.gov.scot/article/34-understanding-the-data-structure

An example of what this means in "real" data terms is shown below:

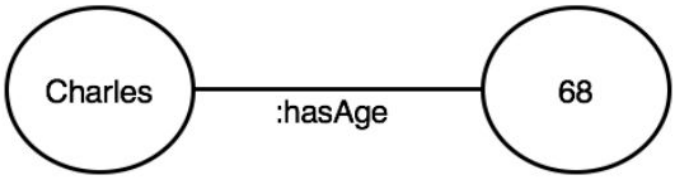

Figure 4 - from https://guides.statistics.gov.scot/article/34-understanding-the-data-structure

This means that multiple "triples" can then be linked to one subject, for example data could also be stored on Charles' geographical location, family members, health conditions etc. in other triples. Essentially the structure of the database is therefore stored in the data itself, and not in separate database schema. Advantages to storing data in this way are:

- It allows the data model to be changed easily
- No schema need to be defined in advance of data being stored
- ID's are stored as url's so are unique which minimises potential issues i.e. merging data sets

Ethical and Legal Considerations

All of the data used in this project are from statistics.gov.scot. The data sets are covered by an Open Government License for public sector information. This means that the data is free to be copied, published and distributed by the user. Additionally, the information can be exploited for personal or commercial use, but in doing so the user must reference the information source and link to the license if possible.

There are unlikely to be any ethical considerations with the data used in this project because it is aggregated to a high enough level that no individuals can be identified through it. Also, it has been assumed that each data source has confirmed consent with every data provider. This can be assumed with a high degree of confidence because the data sets come from NHS health boards(with their own data ethics procedures) and surveys in which the participants are volunteers.

References

1. https://www.gov.scot/binaries/content/documents/govscot/publications/corporate-report/2018/06/scotlands-public-health-priorities/documents/00536757-pdf/00536757-pdf/govscot%3Adocument/00536757.pdf)
2. https://statistics.gov.scot/resource?uri=http%3A%2F%2Fstatistics.gov.scot%2Fdata%2FLife-Expectancy
3. https://statistics.gov.scot/resource?uri=http%3A%2F%2Fstatistics.gov.scot%2Fdata%2Fscottish-health-survey-local-area-level-data
4. https://statistics.gov.scot/resource?uri=http%3A%2F%2Fstatistics.gov.scot%2Fdata%2Fsmoking-sscq
5. https://statistics.gov.scot/resource?uri=http%3A%2F%2Fstatistics.gov.scot%2Fdata%2Fdrug-related-discharge
6. https://statistics.gov.scot/resource?uri=http%3A%2F%2Fstatistics.gov.scot%2Fdata%2Falcohol-related-hospital-statistics