Codeclan Health Dashboard

CodeClan group project - Scottish health dashboard

Repository Description

This repository contains the data and R files used in the CodeClan group project to create a dashboard and presentation using Scottish health statistics.

Group Members

- Jonathan D'Costa
- Emily Jaco
- John Wood

Roles & responsibilities of each member

Jonathan D'Costa worked on the design, creating and entering data into the dashboad. During the design phase various health related dashboards were looked at and some of the best features were noted so that they could be entered into the teams dashboard. During this process ways to simplify creating the dashboard were also selected, such as using Shiny dashboard design. This saved considerable time during the creation process as we had only a week to complete the dashboard. The data was enetered in collaboration with the team to ensure the message and analysis they wanted to get across was portrayed the way they intended.

Emily Jaco worked on part 1: the Scottish Health Survey Local Data. She began by reviewing the different survey categories and selecting 4 categories to look at in greater detail: Smokers, Alcohol drinkers, E-cigarette users and obesity levels. She used spatial data that was divided by the Health Boards in Scotland and created geographical ggplot spreads with a colour range from green and red based on low to high percentages.

John Wood worked on data relating to pregnancy and infant health: post-natal smoking, age of first-time mothers and the number of births classed as having low birthweight. Various visualisations of this data over time and relations between the data sets were presented.

Everyone worked on the overall dashboard concept, review and debugging of code and the structure of the final presentation

Brief description of dashboard topic

Our dashboard contains selected metrics from the Scotland Health Survey and public health statistics related to pregnancy and birth. The data used is published by Public Health Scotland.

Our health topics are Scotland Health Survey and Birth Statistics.

The first tab of our dashboard outlines our health survey in terms of alcoholism, obesity, cigarette and e-cigarette usage for Scotland as a whole and individual councils.

The second tab outlines the factors effecting baby birthweight for Scotland as a whole.

The Third tab outlines the factors effecting baby birthwight for individual councils and compares them to Scotland as a whole.

The forth tab outlines the statistics for the factors effecting baby birthweight to compare the effects of smoking and Mother's age.

Which tools were used in the project

Zoom (daily stand-ups and occasional mob programming) Git/GitHub (collaboration & version control) R Studio (with the below libraries)

Libraries used

1. tidyverse
2. janitor
3. here
4. shiny
5. shinydashboard
6. ggplot2
7. skimr
8. DT
9. sf
10. leaflet
11. rmapshaper

Structure

The project is structured as follows:

1. data_raw

This folder contains the raw data downloaded from the Scottish Government statistics website statistics.gov.home

2. data_ cleaning_scripts

This folder contains .R scripts used to clean and transform the data into Tidy format.

3. data_clean

This folder contains the clean data produced by the .R scripts.

4. notebooks

This folder contains the .RMD notebooks which perform the analsyis and produce the results.

5. health_app

This folder contains all code required for the Shiny app.

6. output

This folder contains all outputs.

Stages of the project

- Researched existing dashboards and Scotland Public Health data sets
- Created Github repo and relevant folders to organise
- Developed a wireframe markup
- Reviewed data as a team and selected points of interest
- Team members split up to work on different tasks of the project i.e. data handling and app development
- Creation of UI along with simultaneous data handling
- Entered data into the app and made changes to the UI as required

How did you gather and synthesise requirements for the project?

Requirements for the project were set by CodeClan, with an overall brief set to produce a

dashboard. The requirements also specified the source data to be used.

We each individually reviewed the requirements before coming together to discuss and agree our plans.

As time was limited we agreed to prioritise a small number of metrics and to limit the initial scope of the dashboard, however we did leave scope to extend and expand our analysis and dashboard if time permitted.

Motivations for using the data you have chosen

For part 1 the team chose to use the dataset on the Scottish Health Survey Local Data because they found it was an interesting overview into health levels across Scotland. Visualising this data gave a good insight into the 14 different Health Boards across Scotland broken down by sex (All, Female, Male) and year (2015-2018 and 2016-2019). It also would be a great data set to add future years to as you could compare with previous years to see any key changes.

Part 2 looked at pregnancy and infant health statistics. Early-years development is one of the Public Health Scotland strategic priorities, and there are specific mentions of the importance of pregnancy and infant health. The datasets available enabled us to look at changes in key metrics over time and to visualise any relationships between them.

Data quality and potential bias, including a brief summary of data cleaning and transformations

As the data was pre-cleaned there wasn't a huge amount of cleaning to be done, however, the team spent time reviewing the data to see what would be the most beneficial to the project. Following that they spent time wrangling the data to best present the data in visualisations. The main transformations were related to linking health board/council area identifiers to datasets and joining national and local level datasets to allow comparisons.

Some of the data relies on self-reported responses (for example on alcohol use, smoking) rather than directly observable information (such as birthweight). This introduces potential bias in results as individuals may not be able to accurately assess, or may choose not to accurately report, their own responses.

How is the data on statistics.gov.scot stored and structured

The data on statistics.gov.scot is in the form of Linked Data - any data point contained in a Linked Dataset should have a unique identiifer which is a URL. The datastore on statistics.gov.scot is in the form of millions of statements, or 'triples' which is a combination of three bits of information that form a subject - predicate - object arrangement.

This means that the data point can be browsed using a web browser and that the data point can be refernced by other Linked Data sets.

Benefits of storing the data like this are that the structure of the database is contained within the data so there is no need to have complex databases.

Ethical and legal considerations of the data

The dataset has a high level of anonymisation from the data i.e. there are at least 3-5 degrees that separate an individual from being identified.

Individuals responding to the Health Survey should have been made aware of how their data would be collected stored and used, and how it would be appropriately anonymised. Similarly new mothers should also have been made aware of how their data, and that of their child, would be used.

The datasets are covered by the Open Government License, which means individuals or businesses in the public sector are allowed to license the use and re-use of the National Archives information under a common open licence.