WOMEN IN TECHNOLOGY –––––––––––––––––––––––ReGIONAL DIFFERENCES

Origin of data, work methodology and visualizations

CS and Women

Ester Giménez, 26 June 2021

Table of Contents

[**Introduction ..**…………………………………………………………………………… 3](#_Toc65783358)

**Planning** ……………………………………………………………...………………….. 7

**Cleaning Data** ………………………………………………………………………...………………….. 7

**Cleaning Data** ………………………………………………………………………...………………….. 7

**Cleaning Data** ………………………………………………………………………...………………….. 7

**Cleaning Data** ………………………………………………………………………...………………….. 7

**Cleaning Data** ………………………………………………………………………...………………….. 7

**INTRODUCTION**

Our customer is Code Division, a charity dedicated to provide high quality computer science courses to people that had not much chance of preparing themselves in this subject, when they were attending school. Usually, students belong to underrepresented minorities, women (girls are not motivated enough to study computer science (CS) at school and the unemployed. The aim of the charity is to promote CS across the population of Scotland and to increase the number of underrepresented parts of the population in the CS working pool.

This necessity has been more noticeable during the pandemic provoked by the COVID-19 virus. Health boards needed people to collect data of the development of the infection, in order to make models and temporal projections for future care. The reduced number of professionals dedicated to data analysis was made apparent, when added to this necessity, a high number of companies realized the necessity of collecting data for their own use and began looking for data analysts as well.

This report is part of the project of “Women in Tech”, end project of the PDA in Data Analysis, Level 8 (first cohort, July 2021). This PDA, provided by Code Division, will help mitigate the urgent need for Data Analysts in Scotland.

The project has been divided in 5 parts, which will be taken by each one of our 5 team members. I will be the acting project manager/coordinator in this case. The divisions are:

* **University:** subjects teach at universities in Scotland in CS, comparisons of men/women that attend those courses, entry requirements, etc.
* **SQA board:** differences in qualifications between boys/girls across all levels of primary, secondary, vocational studies, etc.
* **Regional differences:** differences between schools across Scotland and which schools have CS as a subject. If possible, to compare boys/girls on those schools that take CS across the country.
* **Pipeline to tech:** comparison in the curriculums of CS between Scotland vs England and number of girls studying CS due to changes. If possible, to compare them with other countries.
* **Role of charities / out of the pipeline organisations:** what is the impact of charities in CS in Scotland and specially in girls.
* **Role of colleges:** to add college numbers of CS subjects and girls studying CS across the country.
* **Further research ideas:** there is a lot to develop and study in this project. Our ideas will be added at the end of the final report.

The aim of this project is not to advise on procedures or to apply changes, but only to highlight points of interest that appear in this study. We leave Code Division to use our conclusions at their lease.

**PLANNING**

My study will focus in the “Regional Differences”.

* **Business intelligence and data-driven decision making** 
  + Business intelligence focuses on descriptive analytics, which provides a summary of historical and present data to show what has happened or what is currently happening.
  + One of the objectives of Code Division, as stated in the Introduction, is to improve the number and/or level of their courses. In the last years, there have been a number of articles and studies (please see the reference section), explaining the status of CS in schools and in the life of work. It has been pointed out several times, that women have a very reduced presence in CS in general. Code Division decided to help increase the number of women that go into CS, and in order to do this, they need to know where are the main gaps.
  + This report, based solely in the “Regional Differences” area, will help visualize differences of number of girls vs number of boys in the following:
    - Comparison by type of school (primary, secondary, independent, etc).
    - Comparison by location of the school in Scotland (regions, cities, rural and urban).
    - Schools with/without computer science subjects (current and historical).
    - Comparison of schools with CS and enough teachers.
  + These comparisons and visualizations will help identify areas in Scotland with CS deficiencies, if some types of schools need more assistance in this regard and how big is the gap between girls and boys taking CS studies.
* **Domain knowledge and the business context** 
  + The datasets come from the Scottish Government website [https://www.gov.scot/publications/](https://www2.gov.scot/Topics/Statistics/Browse/School-Education/Datasets/contactdetails)
  + The Scottish Government receives information directly from schools, colleges and universities in Scotland on a yearly basis: contact details of the premises, demographic data of teachers and pupils, information about subjects given, etc.
  + These data help the Government organize the education needs across the country, in terms of budget, moving pupils of schools, changing the curriculum, etc.
  + The Government also requests for specific studies if there is an area in which some data is lacking. This is why in the lasts years it was detected an increase in external workforce being hired by companies. Some surveys were able to highlight an important decrease in pupils taking CS in schools in Scotland, which leads to a lack of adult professionals to cope with the demand of the companies (private and public).
  + Furthermore, it was assessed that very few girls take this career path in comparison with boys. The Government believes that having a more balanced genre numbers in CS will be an important advantage for Scotland. On the other hand, having more young people in CS will benefit the future of computer science in Scotland, because the country will be less dependent on external resources.

* **Business processes and data flows** 
  + The Government publish some datasets with information compiled from the data received by the centres in their webpage of the Education Directorate.
  + The number of schools (primary, secondary and independent) can vary from year to year from 2400 to 2600; 357 are secondary schools. Some of the schools join efforts, while others close and yet others open. The datasets would not be very big in this regard.
  + From the website of the SCIS, it is possible to obtain a list of **independent schools**, which are members of the SCIS (71 of the 102 in total in Scotland, <https://www.scis.org.uk/find-a-school/>). Another register of independent schools can be found at https://www.gov.scot/publications/independent-schools-in-scotland-register/.
    - I compiled manually the **independent schools** that are members of the SCIS. The website does not offer this type of list and clients need to browse for every single school, in order to obtain the data. At the moment is not completed, because it is unknown which ones offer CS courses and how many girls attend those courses. Hence it may be used in the future or maybe not. The compiled file is called “8—Sec-Ind-Schools.xlsx”.
  + For a **comprehensive list of all schools** in Scotland (primary, secondary, independent, special) is better to take the list of the Government website, https://www.gov.scot/publications/pupil-census-supplementary-statistics/
  + **Colleges:** there are 27, listed in the website [www.collegesscotland.ac.uk](http://www.collegesscotland.ac.uk). The information provided does not specify CS courses, or if they have teachers, or how many pupils or girls take CS. For this, every college needs to be consulted separately. Another member of my team will take them.
  + Apart from the information above, I will use **other datasets from the Education Directorate:**
    - School level summary statistics and summary statistics schools bulleting tables
    - School contacts list (mentioned above)
    - Summary statistics follow up leaver destinations
    - A dataset called “tables-charts” when downloaded from the website, which in reality is the “Summary statistics for attainment and initial leaver destinations”.
  + From the **SQA website**, https://www.sqa.org.uk/sqa/91419.html, I will take some historical data up until 2019, because the data for the years 2020/2021 is not yet available. The datasets were done separately for each grade and each file contains the subjects given by the schools and the number of boys and girls that took the examinations by year.
* **Data quality/Data Bias** 
  + **Data Quality**: after giving a look at the datasets, I believe that all of them are of high quality. The Government provides corrections of data when needed and gives explanations and rationale for the changes. This is very welcome and gives credibility to the data.
  + Regarding **data bias**, I have detected that the datasets provide information about girls/boys, location of schools, teachers information and CS subjects but it is not possible to connect them. For example:
    - If there is information about school’ s location and pupils by sex/age, the subjects taught are not included.
    - If there is information about subjects and pupils by sex, the schools are not included. The age can be obtained by the certification year they are in, using the SQA datasets for every SQCF qualification.
    - However, **the school information to compare with the subject information is lacking**. Taking into account that some schools, mostly secondary schools and independent ones, teach the English system or even external ones not related to the Scottish SQCF, the information about which subjects they teach, especially CS, becomes more important.
    - I would not like to say that there is a bias here, but certainly there is some information that has not been compiled properly or that it is not showing in the Government database and/or the websites of the colleges, independent schools, etc. That is which schools offer CS (now and in the past) and how many boys/girls attended those courses.
* **Stages in the data analysis process** 
  + **Choosing the datasets:**
    - This was done in the previous sections. Due to the fact that the biggest dataset has less than 2700 lines, I thought about choosing the columns I need from the sheets and files and do a “manual join” within a file.
  + ***Cleaning the datasets:***
* **Descriptive, diagnostic, predictive and prescriptive analysis**
  + *I will provide a descriptive analysis of the dataset (median, mode, mean, std deviation, etc) of cost of projects and number of projects per area/county.*
* **Data management including security** 
  + All data will be treated as per the GDPR conditions in the respective websites.
  + *The final report and the datasets used will be forwarded to Code Division for further consideration.*
* **Ethical implications of business requirements** 
  + No ethical implications will be discussed here, as the pupil’s data is confidential and does not show in the datasets of the website of the government. Only demographic information will be used for comparison purposes.
  + The school information is public.
* **Tools for data analysis** 
  + In this analysis, I will use Jupiter Notebooks, with some specialized libraries for mapping data and graphics.
  + The information will be compiled and prepared in Excel, because the maximum number of rows is well below 40,000 (I have, approximately, 2600 rows maximum).

**Data Cleaning**

**1.- Dataset’s discussion**

**a) School Level Summary Statistics 2020.xlsx**

This file contains information about pupils attending at each level in every school in Scotland.

There is specific information of both genre boys/girls for the total of each school but not for grades.

Column D of school type is not entirely correct, as some schools offer primary and secondary tuition.

The number of schools listed is 2477.

**b) Schools contacts list.xlsx**

This file does not contain the independent schools. The number of schools listed is 2467.

My idea was to join both files, however, due to the difference in the number of lines affected, which does not match between both files, I decided to make a copy of the addresses into the file of the summary statistics using Excel – vlookup function.

Firstly, I moved the column of the Seed Code in the Summary Statistics\2020 file to column A (the same as the contact list file) from column C. I moved the comments written at the end of the table in the Summary Statistics file to the sheet of the Background notes.

Secondly, I inserted the necessary columns in the file of the Summary Statistics after the School name column.

Third, I used a vlookup function to look for the values of the same Seed Code (with the name of the school does not work) and to copy the address in the file of Summary Statistics:

**=VLOOKUP(A3,'[2--School+Contacts+List+April+2021.xlsx]Open Schools'!$A$1:$S$2468,5,FALSE)**

Column 5 is the first address line in the file of contacts list. This formula can be used with the rest of the columns, modifying the number 5 for 6, 7, 8… until 15.

Checking upon the Sort function for this new created column with addresses, there are 3 cells with =N/A. Their seed codes are:

8244022

8240922

8108625

These 3 schools lacked the contact details. In order to find them, I looked for the name of the school in internet. For the UPRN numbers, I went to <https://uprn.uk> and used the postcode. Barshare Primary School has 2 indistinguishable numbers, while the others were easily found. The email addresses were not found and in one case, not even the website. I left them highlighted in blue, just in case in the future I needed further working on them.

In total I obtained data on 2477 schools in Scotland, without independent schools and colleges.

**c) Denominations of the schools**

If this database is used in the future for further comparisons that are out of scope of this project, I verified the denominations of the schools used in the file of “Schools Contact List” and the “Summary Statistics 2020”. There are discrepancies, highlighted by a comparison made in column BL. The “false” rows are in pink. I believe that the best column to use for denominations should be the one in the contact list file, added in column R.

**d) Computer subjects in the datasets**

None of the datasets that I checked in the website of the Scottish Government contain data on the subjects taught in schools. There is a pool of information in teachers and pupils, but the numbers are added for the whole country and not specified by school.

In order to find how many pupils studied CS, I consulted the website of the SQA, <https://www.sqa.org.uk/sqa/91419.html>. The latest information available is for the year 2018/2019 (this academic year is not yet finished). All data are divided by the type of certification and/or level obtained as per the Curriculum for Excellence. Due to this, I am obliged to collate the data for only CS / Information Technology and similar subjects that appear in each of those files and also for the previous years, in order to obtain historical data. The tables that I need are Table 3 and Table 4 of each file. The numbers are in percentage, so I will need to transform them back to integers, in order to represent them (percentages can be the same for different initial amounts. For example, 5000 students one year can yield 50% women and 50% men, and the next year, 500 students can have the same percentage, however the initial data shows a big decrease from 5000 to 500).

*This information is not specified by school, so it would not be included in the maps. Moreover, this can overlap the work of another colleague that is dealing with changes in curriculum during several years. She needs to know how many boys/girls have taken CS every academic year (every year equals to a certification file in the SQA website) and relate it to those changes. In my case, I need that information divided by school, in order to add it to the visualization maps.*

**References**

[Scottish Index of Multiple Deprivation 2020 - gov.scot (www.gov.scot)](https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/?utm_source=redirect&utm_medium=shorturl&utm_campaign=SIMD)

<https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/?utm_source=redirect&utm_medium=shorturl&utm_campaign=SIMD>

 Consulted on 26/06/2021.

[Statistics 2019 - SQA](https://www.sqa.org.uk/sqa/91419.html)

<https://www.sqa.org.uk/sqa/91419.html>

<https://education.gov.scot/parentzone/find-a-school>

<https://www.gov.scot/publications/foi-17-01802/>

<https://www.gov.uk/government/statistics/provisional-entries-for-gcse-as-and-a-level-summer-2021-exam-series/provisional-entries-for-gcse-as-and-a-level-summer-2021-exam-series>

<https://www.thescottishsun.co.uk/news/4020266/scotland-secondary-schools-results-chart-worst-best-glasgow-edinburgh/>

<https://www.gov.scot/publications/foi-17-01802/>

===========================================================================================================================================

1. Research of OSHPD website

**Values**

**Purpose**

**Their business**

2. Review possible Data Tools

In order to manipulate this data, there are several tools that can be used:

* **Excel:**
* **Python (Jupiter Notebooks and Pandas library):**

3. Review the data quality and data bias used in data decision making process

4. Review business processes involved

5. Review the data sources the organization uses to gather the data

6. Plan to analyze the data

The plan will be divided in 8 phases:

**6.1. Ask the right questions - To understand the business problem.**

**6.2. Explore and collect data - From database, web logs, customer feedback, etc.**

**6.3. Extract the data.**

**6.4. Clean the data - Remove erroneous values from the data.**

**6.5. Find/Replace missing values - Check for missing values and replace them with a suitable value.**

**6.6. Normalize data - Scale the values in a practical range. Transform the data to a standardized format.**

**6.7. Analyze data - find patterns and make future predictions.**

**6.8. Represent the result - Present the result with useful insights in a way the "company" can understand.**

7. Create a data model (Star Schema)

The data model for this analysis will be divided in a **fact table**, containing the quantitative data (cost and number of projects) and around it, several **tables of dimensions**, for location, time and project status. All of them can be linked by a key (Id) than can be the *order number* of each line of information. This key is not present in the datasets but can be inserted if the data analysis will be done with SQL methodologies. The same can be said about the dimension tables, which can be moved to new sheets or files to be used with the same methodologies.



8. Document the results of the analysis

**8.1 STATISTICS**

The basic statistic data obtained from …

**8.2 REGRESSIONS**

**8.3 GRAPHS**

**8.4 Z-TESTS**

9. Document business intelligence gained

10. Document additional research

.