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

Origin of data, work methodology and visualizations

CS and Women

Ester Giménez, 26 June 2021

Table of Contents

1. [**Introduction ..……………………………………………………………………………… 3**](#_Toc65783358)
2. **Research of the websites……………………………………...………………………... 7**
3. **Review the data sources the organization uses to gather the data …………….. 7**
4. **Business intelligence and data driven decision making ..…………………..…….. 7**
5. **Domain knowledge and the business context …………...………………………….. 7**
6. **Business processes involved and data flows……………...………………….…….. 7**
7. **Review of data quality and data bias used in data decision making process ….. 7**
8. **Review of data tools ………………………………………………………………..………**
9. **Dataset discusision - Cleaning……………………………………………………………..**
10. **Descriptive, diagnostic, predictive and prescriptive analysis……………………….**
11. **Data management including security…………………………………..……………**
12. **Ethical implications of business requirements……………………………………**
13. **Create a data model (Star Schema)…………………………………………………..**
14. **Document the results of the analysis…………………………………………………….**
15. **Document business intelligence gained…………………………………………………**
16. **Document additional research……….…………………………………………………….**
17. **References……………………………………………………………………………………..**

1. 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 can be useful for future decision making. We leave Code Division to use our conclusions at their lease.

My study will be a pre-study in “Regional Differences”, awaiting the reception of data regarding pupils that study CS in schools. This information was requested to the Scottish Government at the end of June 2021.

2. Research of the websites

The datasets that I will use were downloaded from the website of the Scottish Government, hence my review will be of this website alone.

My report also mentions SQA website and others. Those will be reviewed by other team members, because they will use more data from those websites than from the Scottish government website.

**Values**

The website of the Education directorate, <https://www.gov.scot/about/how-government-is-run/directorates/learning/> lists what type of information is available (list of policies in the right-hand side of the webpage). It does not need to state their “values” as such. Those are usually specified in each document and report.

**Purpose**

The purpose as per their webpage states that they “ensure success for Scotland's learners through an effective school system and wider learning environment”. This gives an idea of what the directorate is striving for.

**Their business**

The same webpage lists the items that the directorate is responsible for. This also included the issue of statistics related to education.

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

The law requires all schools to be registered in a database every year. The schools have to inform if they are active, descriptive statistics of their teachers and pupils, what subjects they teach, etc.

Local authorities, like councils, collect the data, which goes through a validating software called ProcXed.Net. This information is used to make reports and statistics for stablishing budgets, or looking after specific needs that are detected. Parents can also use this information to look for schools for their children at <https://education.gov.scot/parentzone/find-a-school>.

The data is as good as the information given by the schools. Reviewing the datasets (please see the section on choosing the datasets for this report) it was shown that some schools do not provide their contact details, or simply do not say the grade level of their pupils, just to mention a couple of examples. Some of those points can be corrected manually, looking in internet or calling the schools directly. Due to the short period given for the completion of this study, schools will not be contacted.

4. Business intelligence and data-driven decision making

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.

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. Unfortunately, instead of showing comparisons between schools and CS, a descriptive analysis of schools in Scotland will be given, because there is no data available in the websites of the Scottish Government or the SQA. Further below there is more information about this point.

5. Domain knowledge and the business context

The datasets used in this report 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.

6. Business processes involved 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 special) 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.
  + This list is not complete, because every school needs to be contacted to obtain information about their CS courses and how many girls attend them. Because of this, this contact list can be useful for future studies. 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 be in charge of 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”.

7. Review of data quality and data bias used in data decision making process

**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 all the information. 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 systems 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 would be, which schools offer CS (now and in the past) and how many boys/girls attended those courses and for which grades.

8. Review of Data Tools

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

* **Excel:**
  + If the data collected does not exceed 30.000 rows.
    - The school contact dataset contains less than 2600 rows, making Excel very appropriate for the initial manipulation of the data.
    - Excel can also manage the column size, because there are only 64 columns.
  + Excel can be used to join data from different datasets and if the case, for descriptive analysis, bar graphs and pivot tables (that show number of pupils by grade and region, or by grade and SIMD classification, urban vs. rural, etc).
  + For joining of some datasets, please see the Planning section and the discussion on datasets.
* **Python (Jupiter Notebooks and Pandas, Folium, Geopandas libraries):**
  + For the cleaning of the school contact dataset, I prefer the use of Excel in this case, because the data came initially from 2 different datasets and additions had to be manually done. Please see the section on dataset discussion for more information.
  + For my report, I would prefer to use the worked dataset with Pandas for extracting description data and for showing some bar graphs on rural/urban classifications, for example. This dataset is also good for showing a map of Scotland and the positioning of the schools. The lack of data regarding how many boys/girls per school study CS can be added later in another study, using this same dataset and the prepared Jupiter notebook.

**9. Dataset discussion - Cleaning**

**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 has some discrepancies, as some schools offer primary and secondary tuition, however this is not showing in the dataset. This can be caused by the absence of the independent schools, which offer both. 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. On the contrary, the SQA website has data on the subjects taught, but no information about regions/areas.

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 information for the academic year 2020/2021 was posted the first week of July 2021. All data are divided by the type of certification and/or level obtained as per the Curriculum for Excellence. Due to this, the data for only CS / Information Technology and similar subjects, needs to be manually extracted from each one of those files. There are 38 files for each year. The tables needed are Table 3 and Table 4 of each file. The numbers are in percentage, so they need to be transformed back into totals, 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). The totals for Scotland can be extracted from the Scottish Government website “Pupils in Scotland 2020” census. This information, however, does not tell us anything about regional differences.

**e) STEM website datasets**

Lastly, the STEM Strategy for Education and Training in Scotland reports (last one issued in March 2020), available from the [www.gov.scot](http://www.gov.scot) website, shows comparisons by gender and apprenticeships, festival attendance, science centres support, etc. KPI IIIa shows a gender comparison by gender, however it has no relation with regions or subjects. KPI IIIb shows a gender and subject comparison, but no regions.

The “KPI Ic by LA” shows which regions in Scotland offer IT as IT hardware system support or software development within the STEM Foundation Apprenticeship framework. “KPI Ic” gives me the total number of pupils that completed the apprenticeship in the last 3 cohorts by subject. Joining the data of both sheets would be very challenging and it would be better to ask the Government for the proper datasets if we are going to show comparison of this specific apprenticeship.

I believe this is out of scope for this project, as it does not relate to schools, so this report will not be included in this study.

**10. Descriptive, diagnostic, predictive and prescriptive analysis**

With the data available, I can provide a descriptive analysis of the school’s dataset (median, mode, mean, std deviation, etc) of the number of boys-girls per region.

A map can be drawn of the geographic distribution of schools in Scotland.

The results and the maps could be used in the future to add data of boys/girls attending CS per school.

**11. 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.

**12. 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 government. Only demographic information will be used for comparison purposes.

The school information is public.

13. Create a data model (Star Schema)

(redo)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.



14. Document the results of the analysis

(redo)The OSPHD data shows cumulative data for 2 weeks on 2 weeks (the data is updated every 2 weeks). For example, the pivot 2 that shows the number of projects by county, does not really have a great total of 13565 projects for the 2 years 2013 and 2014 but only the number of projects in December of both years should be taken into account.

Due to this property of the data available, the pivot tables present data for the 31 Dec 2013 for cost and number of projects and the same for the 24 Dec 2014. Each pivot has its own year, which is better when looking at the totals. I formatted the results as numbers with no decimals points (it would be illogical to have 1.29 number of projects, it should be 1 or 2 depending on how it is rounded up).

**8.1 STATISTICS**

The basic statistic data obtained from the years 2013 and 2014 are the following (tab “statistics”):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **COST OF PROJECTS** | | **NUMBER OF PROJECTS** | |
|  | **2013** | **2014** | **2013** | **2014** |
| **highest value (the maximum of the data)** | 3,662,365,402.10 | 3,202,260,043.00 | 2,580.00 | 2,544.00 |
| **lowest value (the minimum of the data)** | 1,887,262.07 | 288,401.00 | 11.00 | 4.00 |
| **mean value** | 371,151,064.12 | 344,819,907.57 | 165.33 | 159.88 |
| **median value** | 125,551,545.50 | 103,313,928.86 | 63.50 | 64.00 |
| **standard deviation** | 718,755,824.57 | 671,141,181.74 | 351.00 | 348.23 |

**8.2 REGRESSIONS**

I ran two linear regressions to see if the cost of projects increased with the number of projects.

* The main number of projects are below 100.
* The average cost is well below $100Mi.
* The regression coefficients of cost in relation to number of projects are 0.79 for 2013 and 0.72 for 2014. The coefficients are higher than expected, and because of that, I will rely on the results.
* The slopes are positive, which indicates that, in general, greater number of projects higher the cost.
  + This is logical, as it would be expected that with more projects, more money is needed to cover them.
* The slope of 2014 is smaller than the 2013 one (1,394,091.95 to 1,690,085.69), which indicates that there is a decrease in the number of projects in 2014 in relation to 2013. Due to this, the overall cost in 2014 is less than in 2013.

**8.3 GRAPHS**

In the tab “graphs”, the bar charts **for cost** show some outliers, xxxxxxxxxxxxxxxxxx

The bar chart **for the number of projects** shows xxxxxxxxxxxxxxxxxxxxxxx

**8.4 Z-TESTS**

Both tests, gave the same result of no statistical difference between the means of cost of projects between 2013 and 2014:

µ mean of our column of data in 2014

µ0 mean of our column of data in 2013

1. **First test:**

H0: µ <= µ0 the mean of the cost is less in 2014 than in 2013 with α = 0.05 significative level

H1: µ > µ0

left one-tail (α = -0.025)

p > α 0.5 > -0.025

H0 is accepted and the means of both years are statistically similar

There is no statistical difference between the data of 2013 and 2014.

1. **Second test:**

H0: µ = µ0 The means of the costs for both years are the same, there are no differences

H1: µ ≠ µ0 We take the grand total for cost of projects for 2013 and 2014

two tails (α = 0.05).

p > α 0.838 > 0.05

Same as before, no statistical difference.

These results could mean that, even if a slight decrease in the cost was observed in the lineal regressions (slope was slightly smaller for 2014, this decrease is not statistically significative.

15. Document business intelligence gained

Given the results of the graphs and the linear regressions, the California area runs a very high number of projects with costs that go into the hundreds of millions of dollars. The majority of projects stay below $100 Mi and from time to time, there are a few big projects with the same value of hundreds of small ones.

It is my opinion that this is what has happened during

16. Document additional research

All links to websites used for the elaboration of this report are located in their respective sections.

**17.** **References**

Scottish Index of Multiple Deprivation 2020

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

[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/>

SIMD webpages:

[www.gov.scot/SIMD](http://www.gov.scot/SIMD)

Statistics Scotland:

<http://statistics.gov.scot/>