WOMEN IN TECHNOLOGY –––––––––––––––––––––––Colleges

Origin of data, work methodology and visualizations

CS and Women in Scotland

Ester Giménez, 1 September 2021

Table of Contents

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

1. Introduction

This report is part of the project “Women in Tech”, final 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.

Following up with the investigative work on the presence of women in technology in Scotland, we took a look at girls attending courses in software development and in digital design and web development in colleges.

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 the use of our conclusions at their lease.

2. Research of the websites

The datasets are a compilation of queries using the website of INFACT, Scottish Founding Council, https://stats.sfc.ac.uk/infact/QueryBuilder/Basic from 2005 to 2019.

**Values**

This website does not show any values, because it is a direct link to the database of the Scottish Founding Council, the organization in charge of helping with the funding of colleges in Scotland.

The website of the Scottish Founding Council, <http://www.sfc.ac.uk/funding/funding.aspx>, explains their values and purpose, however the INFACT is a website apart from the SFC.

**Purpose**

The purpose is not stated in their website; however, it is just to offer data in several formats (pdf, chart, in columns/rows, etc).

**Their business**

This is not specified in the website, however, it is assumed that the data is made available as per FOI requests from the public and Government requirements.

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

The data comes from the colleges, who inform the council, in order to receive the appropriate funding. INFACT publicizes the anonymized data per gender, age, areas of depravation, courses attended, qualifications, etc.

It is in the best interest of colleges to provide accurate data, in order to avoid misunderstandings regarding funding requests to the Council. Due to this, we considered at the beginning of this study, that the data available in INFACT was reliable and accurate. Some discrepancies will be pointed out in this report in the following sections.

4. Business intelligence and data-driven decision making

One of the objectives of Code Division, as stated in the Introduction, is to improve the reach 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 (Computer Science) in schools and in the life of work. It has been highlighted, 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 will help visualize differences in the number of girls vs number of boys in CS in colleges for two main subjects: “Software Development” and “Digital Design and Web Development”. The data will include only numbers for HND qualifications, which are the ones that allow students to enter directly in the World of Work without having to complete a degree at the University. These visualizations can inform about:

* + Which areas of Scotland have availability of these subjects. This can help explain differences in girls attending CS in different regions of the country and if some areas need improvement.
  + The grade of inequality in gender distribution within these subjects, by area, by college and by year.

The completion of the above points depends on the quality of the data provided and can help with funding decisions.

5. Domain knowledge and the business context

The website of the Scottish Funding Council, http://www.sfc.ac.uk/funding/funding.aspx, uses the data in INFACT to deliver investment as per the requirements of each college (and also universities). As per their information:

“We invest around £1.9 billion a year in Scotland's [19 universities](http://www.sfc.ac.uk/funding/universities-we-fund.aspx) and [26 colleges (within 13 college regions)](http://www.sfc.ac.uk/funding/colleges-we-fund.aspx) for learning and teaching, skills development, research and innovation, staff, buildings and equipment.Each college and university sets out in its [Outcome Agreement](http://www.sfc.ac.uk/funding/outcome-agreements/outcome-agreements.aspx) what it plans to deliver (in line with Ministerial priorities and SFC’s [strategic framework](http://www.sfc.ac.uk/about-sfc/strategic-framework/strategic-framework.aspx)) in return for SFC funding. “

In order to perform the necessary calculations, the colleges provide data on number of students they had in the past, so they can organize and plan the necessary number of teachers, material resources, etc for the following academic year. The Council depends entirely on data provided by the Colleges. We have not been able to find any other information as to how the council obtains their data.

6. Business processes involved and data flows

At 2 July 2021, there are 27 colleges in Scotland, 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. In order to obtain this information, the INFACT website provides data on gender distribution by regions and colleges, age distribution, courses, etc.

The 9 colleges listed in the Highlands and Islands region were promoted to university campuses of the University of the Highlands and Islands (UHI) and they do not appear in the database of the INFACT:

Argyll College UHI www.argyll.uhi.ac.uk

Inverness College UHI www.inverness.uhi.ac.uk

Lews Castle College UHI www.lews.uhi.ac.uk

Moray College UHI www.moray.ac.uk

North Highland College UHI www.northhighland.uhi.ac.uk

Orkney College UHI www.orkney.uhi.ac.uk

Perth College UHI www.perth.uhi.ac.uk

Shetland College UHI www.shetland.uhi.ac.uk

West Highland College UHI www.whc.uhi.ac.uk

The website of the UHI, <https://www.uhi.ac.uk/en/>, lists 13 main campuses and 12 appear in their map of the country. The UHI has 70 local learning centres with an extensive offer of online courses. For example, Argyll College has 9 local centres. We are not going to talk about UHI in this study and we will focus solely in those centres that have been considered as colleges since 2005, year of initiation of the INFACT database, which would be 18 in total.

The data can be obtained through 3 types of queries in INFACT. We used the BASIC and the COMPLEX. As examples:

**a) Basic:**

Year 2019-20

Variable College names (actual for each year)

Statistic Number of enrolments

As Count

Show data where Course title contains

"web"

AND

Course title contains

"development"

AND

Qualification aim of course

= Higher National Diploma or equivalent

**b) Complex:**

Year 2019-20

Row variable 1 College Region (Glasgow, Highlands, etc)

Row variable 2 College names (actual for each year)

Row variable 3

Column variable Gender of student (male, female, other)

Statistic Number of enrolments

As Count

Show data where Course title contains

"web"

AND

Course title contains

"development"

AND

Qualification aim of course

= Higher National Diploma or equivalent

The complex query options would be shown at the end of the page, Like for example:

“Number of enrolments **BY** College Region (Glasgow, Highlands, etc) **BY** College names (actual for each year) **BY** Gender

of student (male, female, other) **for year** 2019-20 **Where :-**Course title contains "software" **and** Course title contains

"development" **and** Qualification aim of course = Higher National Diploma or equivalent”

The two queries were used for every academic year, from 2005-2006 to 2019-2020, for every college and every geographical area. Only qualifications of HND were sorted, as those are the ones that companies request for students entering the world of work. The basic query provided the total of students by year and course and the complex, the same data by gender distribution. This was done because the totals from both queries did not match, when they should.

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

This point will be discussed in the section of “Dataset Discussion”, because it is related to the origin of the data and how the data is obtained from the INFACT website.

8. Review of Data Tools

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

* **Excel:**
* The final dataset is composed of 89 lines for web development and 168 lines for software development. The datasets are well within the limits of Excel and this program will be used for the cleaning of the dataset.
* **Python (Jupiter Notebooks and Pandas):**
  + Jupiter Notebooks and Matplotlib libraries has been very useful for showing some visual graphs of demographic comparisons of pupils versus years, colleges and geographical areas.

**9. Dataset discussion - Cleaning**

There data downloaded from the INFACT website was copied to an Excel spreadsheet, every year beneath the previous one.

1. **HND-g-WebDev-Data.xlsx**: contains the downloaded data from the queries from INFACT and a sheet with explanations of what is in every column.
2. **HND-g-WebDev2.csv**: the same data as in the previous file but prepared for Jupiter Notebooks. Formulas have been copied and pasted as values, subtotals have been deleted and column titles have been deleted to leave just the titles on row 1. The years have been added in the inserted column A.

The xlsx file includes the following columns:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| College Region (Glasgow, Highlands, etc) | College names (actual for each year) | Male | Female | Total not real | Total real | Difference | Female real |

The ones in purple come from the complex query, the ones in white from the basic query. The overalls from the complex query were **higher or the same** as the overalls of the basic query. The reason seems to be an unusual high amount of “5” in the columns of “Females” and “Other/not disclosed” (not shown here) in the complex query. The differences between the two overalls were calculated in the column “Difference”.

In order to differentiate both overalls, it was decided that the totals from the basic query were closer to the real numbers and the column was renamed “Total real”, while the other was “Total not real”. As a result, we have a column of “Female” from the complex query and the number of females after taking into consideration the differences, called “Female real”.

The differences affect the quality of the data, as it is not possible to know the real gender distribution. It also gives concern about the origin of the data (information provided by colleges) and/or the treatment of the data by INFACT, because the data shown by the queries is not reliable.

The differences in the two overalls bring some bias issues to the data in INFACT, because it is not possible to know from which column, “Female” or “Not disclosed” we need to subtract the excess. For example, in the Web Development results in 2019:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | College Region (Glasgow, Highlands, etc) | College names (actual for each year) | Male | Female | Not disclosed | Total not real | Total real | Difference | Female real |
| 2019 | Aberdeen and Aberdeenshire | NE Scotland Col | 25 | 6 | 0 | **31** | 31 | 0 | 6 |
| 2019 | Edinburgh and Lothians | Edinburgh College | 33 | **5** | **5** | **43** | 38 | 5 | 0 |
| 2019 | Fife | Fife College | 9 | **5** | 0 | **14** | 12 | 2 | 3 |
| 2019 | Glasgow | City of Glasg Coll | 41 | 14 | 0 | **55** | 55 | 0 | 14 |
| 2019 | Glasgow | Glasgow Clyde Col | 13 | **5** | 0 | **18** | 9 | 9 | **-4** |
| 2019 | Lanarkshire | New Col Lanarkshire | 9 | **5** | 0 | **14** | 11 | 3 | 2 |
| 2019 | **OVERALL** |  | **130** | **40** | **5** | **175** | **156** | 19 | 21 |

* **NE Scotland College**: there is no difference between the two totals, so we assume that 6 women attended this course in 2019.
* **Edinburgh College**: 5 individuals appear as a difference. There is a bias issue in this line, because we need to delete one of the “5”, or the females or the not disclosed, or half each (3 one, 2 the other or vice versa). In this case, it was decided to delete the not disclosed, because this column appears only in 2018 and 2019 with lots of “5”.
* **Fife**: the difference is 2, so we deleted 2 females from the 5 in the column.
* **Glasgow Clyde College**: the difference is 9 (half of the total of the complex query). We realized that all 5 females had to go, as well as 4 boys. The assumption could have been to take out 9 of the 13 boys and leave all the girls. This provokes bias issues, however, the most common scenario is to decrease the number of females.

The bias issues appear in all years in both web and software development. In order to present the assumptions more clearly, the graphs show both the number of females with the complex query (red) and the number with the basic query after the modifications (blue).

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

With the data available, it is possible to provide a descriptive analysis of the gender distribution in colleges (median, mode, mean, standard deviation, etc.).

A map of the location of the colleges can be done in the future with the library Folium.

The results and the maps can be used in the future to add data of CS for the year 2020 and onwards.

**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**

Due to the discrepancies identified in the totals shown in the INFACT database, there are some ethical considerations that are derived from these results. The differences in totals should not exist and they can affect the reputation of INFACT and the Scottish Funding Council when dealing with the public, and to the colleges, whey they request funding.

13. Document the results of the analysis

The graphs show the gender distribution of pupils by years, colleges and geographical regions. As mentioned before, the colleges of the Highlands and Islands were joined in the UHI and do not appear here. The colleges are located in the Central Belt, and from Dumfries and the Borders, to Aberdeenshire and Fife.

There is a total of 15 years of data, from the academic year 2005 -2006 (posted as 2005) to 2019-2020 (posted as 2019).

The proportions of girls vs boys are taken from the columns “Male” and “Female real”, being the second one the number of females after updates from the discrepancies.

All data was filtered by HND – High National Diploma.

**13.1 STATISTICS**

**13.1.1 SOFTWARE DEVELOPMENT**

* This dataset has 168 lines of data, each corresponding to a year and a college, giving an average of 11.2 colleges a year that were offering this subject.
* From 2005 to 2019, there have been 32.7 boys to 3.47 girls in average.
* Boys had a range of 5 (minimum) to 152 (maximum), while girls moved from 0 to 20.

**13.1.2 DIGITAL DESIGN AND WEB DEVELOPMENT**

* This dataset has 89 lines of data, each corresponding to a year and a college, giving an average of 6 colleges a year that were offering this subject. This shows that software development was more popular than web development from 2005 to 2019.
* From 2005 to 2019, there have been 19.3 boys to 4.7 girls in average.
* Boys had a range of 5 (minimum) to 69 (maximum), while girls moved from 0 to 16.
* There were far less number of boys attending web development, while girls had approximately the same numbers as with software development.

**13.2 GENDER DISTRIBUTION BY REGION**

**13.2.1 SOFTWARE DEVELOPMENT**

The main hub for learning software development is Glasgow, followed by Aberdeen, Dundee and Edinburgh. Aberdeen has been increasing its importance as a hub in this subject in the latest years.

The proportion of girls vs boys by region shows:

* The worst region, Forth Valley, with 1 girl for every 15 boys
* The best region is Glasgow with 1 girl for every 8.3 boys.

West Lothian had the smallest number of boys during 2005 to 2015 (only 30) but one of the best with the girls vs boys ratio: 8.37 (they had 11 girls during the same period). This could be explained, not because more girls attended the courses, but less boys were present.

Ayrshire had the least number of girls during these years, only 9, but also a small number of boys (35), so the ratio is not much different from other regions (11.55).

**Region Proportion boys/girls real**

Aberdeen and Aberdeenshire 9.06

Ayrshire 11.55

Dundee and Angus 9.93

Edinburgh and Lothians 10.41

Fife 10.94

Forth Valley 14.87

Glasgow 8.3

Lanarkshire 10.72

West 8.86

West Lothian 8.37

**13.2.2 DIGITAL DESIGN AND WEB DEVELOPMENT**

The main regions and from high to small number of students are:

* Glasgow
* Edinburgh
* Aberdeen
* Ayrshire

Ayrshire is unexpectedly included in the first positions, while Dundee is far down the list. It looks like software development is more sought after in Dundee than web development, while web development is more popular in the west of Scotland.

The proportion of girls vs boys by region shows:

* The worst region, West, with 1 girl for every 5.7 boys.
* The best region is Forth Valley, with 1 girl for every 2.7 boys. This could be explained, not because more girls attended the courses, but because this college has the smallest number of boys in all the colleges in 15 years, just 61.
* Fife had the least number of girls during these years, only 18, but also a very big number of boys (85), making it one of the worst areas for the number of girls vs boys (4.72).

**Region Proportion boys/girls real**

Aberdeen and Aberdeenshire 3.5

Ayrshire 3.44

Dundee and Angus 2.77

Edinburgh and Lothians 4.45

Fife 4.72

Forth Valley 2.65

Glasgow 4.25

Lanarkshire 5

West 5.66

**13.3 GENDER DISTRIBUTION BY YEAR**

**13.3.1 SOFTWARE DEVELOPMENT**

It looks like the number of boys attending software development was quite constant until 2009, when the classes began to be more populated, reaching maximums from 2014 to 2017 and slightly decreasing after that. The tendency could be to decrease a little bit more in the future, depending on the job market and prospects.

Regarding the number of girls, it looks like it has been more or less constant during these 15 years, with a slight increase from 2014 to 2016, remaining more or less stable since then.

* Boys went from a minimum of **127** in **2007** to a maximum of **590** in **2015**.
* Girls went from **17** in **2008** (in average, 1 female per college) to **72** in **2016**.

In this case, it looks like when boys reached a minimum, girls did the same the following year, and the same with the maximums. These are coincidences, because gender distribution followed a wave, with up and downs since 2005, however those are more clearly visible with boys than with girls, due to the difference in numbers.

**13.3.2 DIGITAL DESIGN AND WEB DEVELOPMENT**

Boys and girls follow the same curve (when there are more boys, there are more girls and when there are less, the number of girls decreases as well). From 2010 to 2012 there is a big decrease, and only **8** women studied web development in 2010. The maximum of **44** was reached in 2015.

For men, the minimum is 41 in 2010 and 169 in 2015.

**13.4 GENDER DISTRIBUTION BY COLLEGE**

**13.4.1 SOFTWARE DEVELOPMENT**

In total, 29 campuses (not colleges) offered Software Development in Scotland from 2005 to 2019. The 5 most successful ones with boys were, from big to small:

* City of Glasg Coll (779 boys in 15 years)
* Glasgow Clyde Col (733)
* NE Scotland Col (464)
* Forth Valley (446)
* Dundee College (355)

For girls, we have:

* Glasgow Clyde Col (97)
* City of Glasg Coll (85)
* NE Scotland Col (46)
* Aberdeen College (35)
* Edinburgh College (34)

Even with such a high proportion of girls vs boys in all campuses, the Glasgow area continues to be a hub for Software Development, with the NE and Dundee, while girls prefer Aberdeen and Edinburgh. For example, Aberdeen College has a proportion of **1:8.9**.

Glasgow had offers in several campuses:

* Glasgow Kelvin College closed their offer in 2015.
* Anniesland was only active in 2013.
* Stow College was active from2007 to 2013.
* Cardonald, Langside had been offering software classes since 2005 and closed them in 2013.
* The only 2 campuses still active for Software Dev are Glasgow Clyde College and City of Glasgow College.

It would be interesting to investigate how the different campuses have developed with the years, in order to identify possible reasons of girls not attending, if colleges in the area do not offer the subject.

**13.4.2 DIGITAL DESIGN AND WEB DEVELOPMENT**

In total, 20 campuses (not colleges) offered Web Development in Scotland from 2005 to 2019 (9 campuses less that offered Software Dev.). The 5 most successful ones with boys were, from big to small:

* City of Glasg Coll (337 boys in 15 years, not even half of attendees to Soft. Dev.)
* Edinburgh College (284, not comparable with only 34 in Soft. Dev.)
* James Watt College (224; this college did not show as an important one for Soft. Dev.)
* NE Scotland Col (187; it looks like this college has a better offer in Soft. Dev than in Web. Dev. for men)
* Jewel & Esk Valley Coll (121; this college did not show as an important one for Soft. Dev.)

For girls, we have:

* City of Glasg Coll (82, *exactly the same number* as with Soft. Dev.)
* Edinburgh College (66, almost double than the girls in Soft. Dev.)
* James Watt College (51; this college did not show as an important one for Soft. Dev.)
* NE Scotland Col (53; a little below the number of 46 that attended Soft. Dev.)
* Jewel & Esk Valley Coll (26; this college did not show as an important one for Soft. Dev.)

Like in Software Development, the proportion of girls vs boys is very high in all campuses (**1:4** in the campuses listed above). Glasgow is very important for Web. Dev., however the numbers are more evenly distributed through all the areas. For other colleges, like Glasgow Clyde College, the proportion is **1:7**, while in Forth Valley, the proportion is **1:2.65**.

14. Document business intelligence gained

Given the results of the graphs and the calculations obtained after the corrections were made, I believe that:

* Very few women study Software or Web Development in colleges (in comparison to men).
* More women study Software than Web Development. This goes against the assumption that it would be the other way round, because Web Development has an important component in design and creativity.
* Not all colleges have offered HNDs in both subjects during the last 15 years. This needs further study, in order to ascertain if this is due to lack of mobility of students, pricing, subject content, etc, that reduces the number of pupils continuing their formation in IT in colleges.

15. Document additional research

In the near future, maps will be provided of the location of the colleges in Scotland, together with percentages of girls/boys attending Software and Web Development. This will help to visualize the locations and how available are in terms of commute, etc. It will also give a visual for the publicum.

**16.** **References**

Below there is a list of links, with information and datasets that were used or consulted in the elaboration of this report.

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

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

Consulted in August 2021, to obtain the titles of the HND awards offered.

Scottish Funding Council, INFACT Database

<https://stats.sfc.ac.uk/infact/QueryResults/Basic>

Scottish Funding Council

<http://www.sfc.ac.uk/>