Table of contents

List of figures 4
List of tables 5
Abstract 6
Acknowledgements 7
1. Introduction 8
1.1 Background 8
1.2 Aims 9
1.3 Objectives 9
1.4 Motivation 9
1.5 Investigation Report Summary 10
1.6 Project Specifications 11
1.6.1 Functional Requirements 11
1.6.2 Non-Functional Requirements 12
1.7 Report Structure 12
2. Design 13
2.1 Data Collection 13
2.1.1 Skills Development Scotland (SDS) RSA Data Matrix 13
2.1.2 Higher Education Statistics Agency (HESA) 14
2.1.3 Scottish Funding Council (SFC) Data 15
2.1.4 Scottish Index of Multiple Deprivation (SIMD) Data 16
2.2 Data Preparation 17
2.2.1 Data preparation using Excel 17
2.2.2 Data preparation using R Studio 18
2.3 Choice of Software Tool 23
2.4 Choice Visualisations 24
3. Implementation 26
3.1 Correlation matrix 26
3.1.1 Attendance vs Enrolments 27
3.1.2 Attainment vs Enrolments 28
3.1.3 No qualifications vs Enrolments 28
3.1.4 Working age population vs Enrolments 29
3.1.4 Employment rate vs Enrolments 29
3.1.5 Income rate rate vs Enrolments 30

3.2 Dashboard 30
3.2.1 Overall Trend 31
3.2.2 STEM enrolments by UHI College 31
3.2.3 Subject enrolments 32
3.2.4 Enrolments by Subjects by UHI College 33
3.2.5 UHI Region Map 34
3.3 Challenges faced during Implementation 34
4. Testing and Evaluation 36
5. Conclusion 39
5.1 Summary and Achievements 39
5.2 Reflections 39
5.2.1 A good preparation is essential 39
5.2.2 Regular supervision meetings 39
5.2.3 Adaptability to overcome hurdles 39
5.3 Review of Ethical, Social, Legal, Commercial and Professional Issues 40
5.4 Improvements and Future Work 40
6. References 42
Appendix A: Project Plan 44
Appendix B: Correlation Plots 45

List of figures

Figure 1 CRISP-DM Process Flow Diagram (Manasson, 2019) 11
Figure 2: SIMD Indicator Descriptions("Scottish Index of Multiple Deprivation 2020 - Gov.Scot" 2022) 17
Figure 3: Data structure in R Studio 19
Figure 4: Converting data types of variables 19
Figure 5: Checking for missing values 19
Figure 6 : Filtering instances with "*" 20
Figure 7: Mean of all variables for year 2016 20
Figure 8: Mean of all variables for year 2020 21
Figure 9: Combining averaged variables data for year 2016 and 2020 21
Figure 10: Exporting SIMD 2016-2020 merged file 21
Figure 11: Loading STEM data 22
Figure 12: Converting datatypes in STEM data 22
Figure 13: Creating new columns Date and enrolments 22
Figure 14: Converting datatypes for variables Date and enrolments 23
Figure 15: Merging SIMD and STEM dataset by Year and Council area 23
Figure 16: Exporting the SIMD and STEM merged dataset 23
Figure 17: Types of Charts ("Cheatsheet For Charts: Pair The Right Chart With The Right Set Of Data" 2022) 24
Figure 18: Variable selection for Correlation Matrix 26
Figure 19: Plotting Correlation Matrix 26
Figure 20: Correlation Matrix 27
Figure 21: Attandance vs Enrolment 27
Figure 22: Attainment vs Enrolments 28
Figure 23: No qualifications vs Enrolments 28
Figure 24: Working age population vs Enrolments 29
Figure 25: Employment rate vs Enrolments 29
Figure 26: Income rate vs Enrolments 30
Figure 27: Overall Trend 31
Figure 28: Tool tip description 31
Figure 29: STEM Enrolments by UHI College 32
Figure 30: Enrolments by Subjec 32
Figure 31: Enrolments by Subject and College 33
Figure 32: UHI Region Map 34
Figure 33: Stacked bar for STEM subject dispersion 36
Figure 34: Area chart showing STEM subject dispersion 37

List of tables

Table 1: MoSCoW Analysis of Functional Requirements 12
Table 2: MoSCoW Analysis of Non - Functional Requirements 12
Table 3: Compliance to functional requirements 38
Table 4: Compliance to non functional requirements 38

Abstract

The objective of this project report is to thoroughly record the design, implementation, testing, and evaluation of the HISP data dashboard.

The project starts with a discussion on the findings from investigation phase with a focus on the design and implementation needs as agreed upon with the HISP team. Data acquisition, data preparation, data visualisation and analysis were all incorporated during the design stages.

The implementation phase will look at how the correlations between the SIMD and the STEM data were plotted using R language commands in R Studio. We then look at creating the data dashboards using Tableau desktop. The various chart formats and colour schemes used were thoroughly described. Each dashboard includes a detailed explanation as well as a trend analysis.

The next section examines the dashboards' testing and evaluation. The evaluation method includes regular feedback from the HISP team and a thorough description of how developmental testing was used. All necessary requirements, including functional and non-functional, will be examined for compliance or non-compliance.

Finally, lessons learned from the entire project will be reviewed. Previously considered risks will also be revisited to determine compliance with the ethics of carrying out such an undertaking. Meanwhile, efforts to further raise the system's level of service quality are encouraged.

Acknowledgements

I would like to recognise and express my gratitude to my supervisor, Pam Johnston. Her guidance and wisdom guided me through every stage of writing my project. I would also want to thank my placement supervisor Dawne Bloodworth, the chair of the UHI HISP programme, Su Bryan, and The Data Lab for providing the opportunity to work on this project.

In addition, I would like to thank my husband, Kamran Quraishi, my children, Adeel Quraishi and Aairah Quraishi, my mother, and my siblings for their constant support and patience while I conducted research and wrote my project. All your prayers have sustained me this far.

Finally, I would like to thank God for helping me to overcome every difficulty. You made it possible for me to earn my master's. I will continue to have faith in you for my future.

1. Introduction

This chapter starts with a background of the project, describes the project's motivation, aims, and objectives. It summarises the research done to meet the project's deliverables, briefly reviews the project specifications, and then describes the structure of the report.

1.1 Background

The report meets two of the requirements, the MSc in Data Science programme and is also a record of the work done during the placement at the University of Highlands and Islands. The goal of this project is to make a data visualisation of data about STEM student enrolments in the Highlands and Islands, Moray, and Perthshire. This will help the Highlands and Islands STEM Partnership (HISP) understand what students want to study, how the population is changing, plan activities, and predict what will happen in the future.

Data visualisation is the graphical representation of information or data. By utilising visualisations like charts, graphs, and maps, data visualisation tools make it simple to discover and comprehend trends, outliers, and patterns in data. Data visualisation tools and technologies are crucial for analysing massive volumes of data and making data-driven decisions in the age of big data (Tableau 2022). A data dashboard is a tool used by businesses to monitor, analyse, and show data, generally to better understand the overall health of the firm, a particular department, or even a particular process. Dashboards enable businesses to harvest crucial data from many sources and present it in a user-friendly way by connecting all types of metrics, data sources, APIs, and services in the background. Performance monitoring, data transparency and accessibility, agility, and forecasting are some advantages of using dashboards (Microsoft 2022).

STEM stands for science, technology, engineering, and mathematics. The Highlands and Islands STEM collaboration seeks a data dashboard, which is the goal of this project. As part of a larger economic strategy, the Scottish Government adopted the STEM (Science, Technology, Engineering, and Maths) Education and Training Strategy in 2017. Its objectives are to enhance the relationship between STEM education and training and the labour market, close equity gaps in participation and accomplishment, inspire young people and adults to pursue STEM jobs, and increase capacity to deliver great STEM learning. The aim of this strategy is to boost education and training in order to meet the rising demand for STEM skills and to make sure that the supply of these talents can keep up with the demand and support growth. The Minister for Further Education, Higher Education, and Science is in charge of the national oversight of the programme through the STEM Strategy Implementation Group. The 13 regional leads are joined by representatives from important national STEM organisations in a group called STEM regional leads. This organisation is coordinated and led by the Energy Skills Partnership (ESP).

The University of the Highlands and Islands (UHI) oversees and serves as chair of the Highlands and Islands STEM Partnership (HISP). The HISP region is distinctive in Scotland because of its vastness and rural setting. In contrast to the other 12 STEM zones, UHI has 11 active FE colleges participating in STEM and centrally coordinates STEM involvement throughout UHI.

1.2 Aims

The purpose of this project phase is to design, implement, and evaluate a prototype self-service dashboard to help STEM providers in the Highlands and Islands STEM Partnership plan activities, understand enrolments in STEM subject entrants at college and forecast future trends.

1.3 Objectives

1. Investigate the availability & viability of data sources for inclusion in the visualisations.
2. Review state of the art in data visualisation dashboards including available software tools.
3. Create an efficient and usable data visualization.
4. Evaluate the effectiveness in terms of usability & accuracy.

The first objective of the project was accomplished during the investigation phase, establishing a foundation for the completion of the other deliverables.

1.4 Motivation

STEM refers to science, technology, engineering, and mathematics. Through project-based learning, STEM-related education and training aims to help young people build their ability to collaborate across disciplines in addition to their skills and knowledge in each specific area. Through these methods, young people can learn about how STEM knowledge and skills are applied in the workplace. Individuals and groups with different specialties and competencies would collaborate in the workplace and in industries to produce novel insights, concepts, and products ("Supporting Science, Technologies, Engineering and Mathematics (STEM) At Home | Learning At Home | Parent Zone" 2022).

Numerous employers, particularly those in sectors unrelated to STEM, place a high value on the abilities acquired through STEM. Young people in Scotland have several employment options in fields related to STEM. Most of Scotland's economic sectors depend on STEM such as life sciences, energy, food and drink, financial and business, universities, tourism and creative industries (Npfs 2022).

For instance, WEEE Scotland limited, which specialises in the repair, refurbishment, and remanufacturing of various commodities, and Walker Precision, a manufacturer of space, aerospace, and defence parts, have offered placements to recent STEM graduates who are unemployed or struggling to find meaningful work due to the effects of Brexit and the COVID pandemic ("STEM Graduates Wanted For 50 Fully Paid Work Placements" 2022).

Scotland, like other nations around the world, provides a variety of thrilling and lucrative STEM vocations. Many of Scotland's major economic sectors, including energy, the creative industries (including digital), food and beverage, and life sciences, rely on STEM.

It's a common misconception that these occupations are only open to people with a university degree in a STEM-related field, but there are also plenty of options for modern apprentices and technician-level positions (Npfs 2022).

The University of the Highlands and Islands (UHI) provides a wide variety of courses in subjects like business management and administration, computing and ICT, construction, engineering, hospitality and tourism, land-based industries, nautical studies, and science across the Highlands and Islands, Moray, and Perthshire at various locations on campuses at UHI Argyll and Bute, UHI Inverness, UHI Outer Hebrides, UHI Moray, UHI Orkney, UHI Perth, and UHI Shetland as well as online for students who wish to study. Degree-level, HNs-level, postgraduate-level, professional-level, and vocational-level courses are all provided (UHI 2022).

The abilities acquired through STEM can lead to professions in a variety of other fields. For instance, transferrable skills like the ability for problem-solving, precise prediction, and reliable conclusion are helpful in a variety of occupations. As a result, businesses in both STEM and non-STEM sectors value STEM expertise (Npfs 2022).

My main motivation in doing this project is that the academic and non-academic workers at the university will have a focal point, allowing them to focus their efforts in the appropriate areas, for example, recruitment of students. By examining student data and employability rates through this initiative, more students will select STEM majors, helping to close skills gaps and encouraging more STEM-related research that will spur innovation and be good for society as a whole.

In doing this project I will also be learning a new tech stack, and this will help me enhance my knowledge and prepare me to confidently face any challenges that I may come across in my career as a future data scientist.

1.5 Investigation Report Summary

In the initial investigation report, the definition of STEM was examined in and the prospective open data sets, such as the RSA data matrix, HESA data, Scottish Funding Council data, and Scottish Index of Multiple Deprivation 2020 data, were obtained and explored. All of the aforementioned bodies' data was examined to see what it had to say, how to use it most effectively for this project, or, if it couldn't be used, why it couldn't be used. The Scottish funding council STEM enrolment proprietary data and the Scottish Index of Multiple Deprivation data are the datasets being used for the design of this project because the geographic locations in both datasets could be used as a reference points. We also examined data visualisation, categories and types. With the line charts, users can see how enrolments in STEM subjects have changed over time, while the bar charts show how many students have been enrolled in UHI STEM providers over time. Since UHI campuses are spread out over a large area, the HISP STEM team would like to include maps in the final visualisation to learn more about rurality and poverty. The idea behind the visualisation is to use colour blind palettes to make it easy for everyone to use. One or two members of the HISP STEM team will look at the visualisation to make sure it is clear. This will make it easier to get feedback and make any changes that are needed before showing it to the whole team.

Software tools RShiny, Power BI and Tableau were explored by laying out their pros and cons. It was concluded that Microsoft Power BI will be the tool of choice for the implementation of this project because it does not cost anything to licence and it works well with Microsoft 365, which is a widely used app that UHI also uses. This meant that the UHI didn't have to pay for a new licence for another tool or piece of software.

Some past work that was similar to the goal of this project, which is to make a data visualisation of student data was also looked at. This also highlighted how HE sis emerging to be a competitive market and its importance to the universities, colleges, politics, job markets and labour markets.

Then the methodology, software tools and the functional and non-functional requirements were examined. An agile approach to the CRISP-DM process has been used in this project as there were regular meetings with the placement supervisor and the HISP project team to gain feedback. Also, the functional and non-functional requirements were assessed.

The report concluded with an evaluation of the project's legal, ethical, social and professional issues in undertaking this project.

1.6 Project Specifications

The project investigation report included more information about the suggested prototype dashboard's specifications. The official name of this will be HISP- Dashboard.

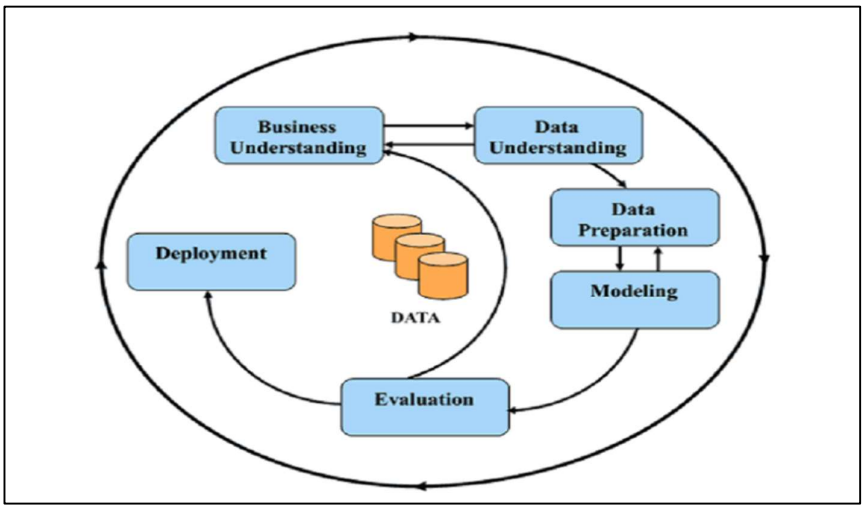


Figure 1 CRISP-DM Process Flow Diagram (Manasson, 2019)

The development of the dashboard followed a professional agile approach to the CRISP-DM process (Manasson, 2019) as can be seen in Figure 1.

The "plan" phase of the cycle is comprised of its functional and non-functional requirements, which were established during the project investigation phase. The agile methodology's adaptability made it possible to communicate progress and obtain feedback in order to make any necessary adjustments. The MoSCoW prioritisation technique was used to manage the requirements and identify the implementation-related musts, shoulds, coulds, and woulds. Table 1 and Table 2 reiterate this information.

1.6.1 Functional Requirements

Below are the functional requirements for the project:

1. What information is available and what it reveals must be determined by the investigation.
2. The data must be specific to the Highlands and Islands, Moray and Perthshire region.
3. The data must be specific to STEM subjects and STEM student enrolments.
4. The investigation should be able to identify gaps in data.
5. The data visualisation should have a geographic map.
6. Higher Education STEM data could also be helpful.
7. The HISP would like statistics that show preferred student pathways.

In Table 1 Table 2: MoSCoW Analysis of Non - Functional Requirements can be seen

Requirement no.	MoSCoW Analysis (Priority ranking)
1	Must
2	Must
3	Must
4	Should
5	Should
6	Could
7	Would

Table 1: MoSCoW Analysis of Functional Requirements

1.6.2 Non-Functional Requirements

The below have been identified as non-functional requirements (CEIAS 2022):

1. Accessibility - The dashboard must be made accessible to all its users.
2. Integrity – The information provided by the dashboard should be accurate. Measures can be enforced to follow strict guidelines and any data that does not adhere to these should not be accepted.
3. Reliability - Whether the dashboard can be accessed to write or read data is a measure of reliability. This will ensure that the data saved is accurate.
4. Recoverability – It is the capacity to get back to a working state in the event that something goes wrong. This can be accomplished by periodically making backups.

In Table 2 Table 2: MoSCoW Analysis of Non - Functional Requirements can be seen.

Requirement no.	MoSCoW Analysis (Priority ranking)
1	Must
2	Must
3	Must
4	Should

Table 2: MoSCoW Analysis of Non - Functional Requirements

1.7 Report Structure

The remaining of the report contains:

Chapter 2 Design. This chapter discusses the design concepts as well as the specifics of the design, beginning with the stage of data collection, data preparation, types of charts or visuals utilised, and tool selection.
Chapter 3 Implementation. This chapter outlines the technique for implementation, the functionality that was used, and both specific and broad design choices. issues that arose during the project and their solutions.
Chapter 4. Testing and evaluation. This explains the kind and specifics of testing the dashboard to verify the implementation of the solutions.
Chapter 5 Conclusion. This chapter gives a brief overview of the project, along with observations and lessons learned. It ends with a description of the remaining tasks, the project's scope moving forward, and any potential improvements.

2. Design

The design procedure that was used for the project is described in this chapter. Building and delivering a set of functionalities that satisfy the needs of the stakeholders, both users and non-users, is the goal of every software implementation project. To accomplish this purpose, it is necessary to comprehend all requirements and create an effective design that will enable them to be implemented into a practical solution.

2.1 Data Collection

During the investigation phase three potential open datasets from the HISP project specifications were reviewed. As seen in the investigation report:

2.1.1 Skills Development Scotland (SDS) RSA Data Matrix

The data matrix covers:

i. Skills supply- the supply of people within the labour market.
ii. Skills demand - the demand for skills within the labour market.
iii. Skills mismatches- where there is a gap between the demand for skills and the supply of skills within the labour market. ("Regional Skills Assessments" 2022)

This dataset was examined to determine if any data regarding the chosen pathways and desired career paths of young people could be found. The HISP team might use this to gain understanding of current patterns in STEM subject enrolments as well as predict and forecast them. SDS provide funding for foundation apprentices to study through local authorities (councils). The number of FAs (Foundation Apprentices) and GAs (Graduate Apprentices) attending college and universities is also monitored by SDS.

The data from this source cannot be used for the following reasons:

- The specific Colleges within Highlands and Islands region, such as UHI Perth and UHI Argyll and Bute, individually have information relating to STEM applicants and enrolments. The course title, level, and other information would be included in that specific college report, but Skills Development Scotland is unable to give this information to me directly.
- As the data contained personal level information, we required to understand Skills Development Scotland's policy on data security, these specific datasets relating to the desired careers/preferred paths, which were of particular importance to this research, could not be released. Web scraping was an option to look at for this dataset, but I chose against using the web scraping because:
    i. As discussed in the investigation report section 2.2.5 that SFC and SDS do not count the same elements. SFC funds and counts students attending colleges and universities. Apprentices who obtain funds from other sources, such as those who study at local authorities, are not subject to SFC monitoring. Whereas SDS offers funding to foundation apprentices who study via local authorities (councils). It keeps track of how many FAs and GAs enrol in colleges and universities, but those students are not funded by just SFC. Therefore, even though scraping would have provided the data hidden behind the RSA data matrix, it could not be utilised for this project.
    ii. Of the project's time constraints.

2.1.2 Higher Education Statistics Agency (HESA)

The Higher Education Statistics Agency, which is the official statistical organisation, is the source for data on higher education in the UK.

HESA releases information on every aspect of the UK higher education sector. The information is updated for each academic year and is available for download as a comma separated file. T he HE student data pages, which are grouped into five different web pages is looked at, as our project's main focus is students. These include information on:

i. Who is studying in HE? This shows information about student numbers and their personal characteristics which is further broken down by sex, personal characteristics (age group, disability, ethnicity) by HE provider, personal characteristics by subject of study (Medicine and dentistry, computing etc.), detailed disability and ethnicity breakdowns.
ii. Where do the HE students come from? Information about the origins(domicile) of UK higher education students may be found in the tables and charts on this page. It contains a detailed breakdown on where UK HE students come from like HE student enrolments by domicile and region of HE provider, UK domiciled HE students by HE provider and domicile. It also shows changes over time for example, HE student enrolments by domicile for academic years 2016/2017 to 2020/2021, first year non-UK domiciled students by domicile for academic years 2006/2007 to 2020/2021 etc. There are also details about non-UK HE students by HE provider and country of domicile, information about HE student enrolments by level of study, mode of study and domicile. It also provides details of transnational education i.e., students who pursue degrees or courses at UK universities while studying abroad and avoiding travel to the UK. These students are counted separately from the HESA Student record population and have data returned via HESA's Aggregate offshore record.
iii. Where do HE students study? Information about where in the UK HE students study is shown on this page. Students by HE provider, location of study and student accommodation details can be found on this page.
iv. What do HE students study? This page details a complete subject breakdown, breakdown by personal characteristics, changes over time and breakdown by subjects by HE provider.
v. What are HE student's progression rates and qualifications? This page details number of qualifications i.e., number of undergraduates and number of postgraduate students, details of classifications they have achieved i.e., first class honours, upper second-class honours, second class honours etc ("HESA - Experts in Higher Education Data and Analysis" 2022).

The data from this source could not be used in this specific project for the below reasons:

- Firstly, how are science subjects defined by HESA.
  The Common Aggregation Hierarchy (CAH) provides a standardised hierarchical grouping of subject codes and terms developed from the Higher Education Classification of Subjects (HECoS) codes and terms that is suitable for the majority of applications. The grouping of science courses has been created by HESA. The science grouping is an amalgamation of CAH level 1 codes CAH01 through to CAH13 and CAH26 (Geography), with the exception of CAH26-01-03(Human geography), is included in the non-science category known as "Geographic and environmental studies (social sciences)". The science category titled "Geographic and environmental studies (natural sciences)" is where all other CAH level 3 codes within CAH26 are presented.

The second question that arises is, are the science subject areas same as STEM.
In section 2.1 in the investigation report STEM has been looked at in detail, its definition, what it stands for and how it plays an important role in education, training and developing skills. However, according to HESA, there is no distinct definition of what constitutes a STEM subject (science, technology, engineering, and maths). The HESA science category includes fields including agriculture, nursing, and medicine that may not be covered by other STEM definitions ("What Do HE Students Study? | HESA" 2022).

- HESA's open data offers very insightful information about students, employees, graduates, finances, UK performance measures, etc. The level of granularity in terms of country of HE provider is England, Scotland, Northern Ireland, and Wales, and by region of HE provider for Scotland, is limited to Scotland as a region. There are no statistics available around the level for this project's focus areas i.e., the Highlands and Islands, Moray, and Perthshire.
- Lastly, STEM applicants and enrolments data is returned to the UHI institutions by way of the regional Colleges, such as UHI Perth and UHI Argyll and Bute. This particular college report would contain information about the course's title, level, etc., but HESA does not provide this information specific to STEM.

2.1.3 Scottish Funding Council (SFC) Data

The Scottish Funding Council (SFC) is working to make Scotland the world's top location for learning, research, and innovation. Over half a million people can receive opportunities that will change their lives thanks to the £1.9 billion in public funds invested annually by the SFC.

Every one of Scotland's 19 universities is able to conduct internationally renowned research because of SFC funding for university research. Exciting collaborations between industry and university research are taking place thanks to their investment of more than £120 million in innovation centres. Their efforts to increase access are bringing colleges and universities closer together and giving more people access to greater learning and skill-building opportunities ("Scottish Funding Council Home Page" 2022).

SFC fund and count students going through colleges and universities. SFC does not monitor apprentices who receive funding from other sources, such as those who study at local authorities. Additionally, SFC funds all other university and college provisions.

I haven't been able to access the URL listed as a potential open data source in the HISP STEM data project description which is Infact (sfc.ac.uk) because it has been offline for maintenance with no estimated return date. I was obliged to make do with the proprietary data that the SFC had to offer. One of the SFC representatives confirmed that this information is collected annually and tends to be available in January for the preceding academic year. So, in January this year i.e., 2022 the data was available for 2020-21.

This dataset, which is an excel file, contains statistics on STEM FE (further education) enrolments by subject and STEM providers unique to the Highlands and Islands, Moray, and Perthshire areas. The dataset consists of the following information:

i. STEM subjects, which list the STEM subjects which are Business Management and Administration, Computing and ICT, Construction, Engineering, Hospitality and Tourism, Nautical Studies, and Science.
ii. STEM colleges. The Further Education (FE) colleges that come under the UHI are UHI Argyll and Bute, UHI Inverness, UHI Outer Hebrides, UHI Moray, UHI Orkney, UHI Perth, UHI Shetland, UHI West Highland and UHI North Highland. This dataset however does not have information about the UHI West Highland and UHI North Highland because as per the SFC

representative who provided this proprietary data, the searches did not return any FE STEM provision under these two colleges.

iii. Number of enrolments for academic years 2014-2015 to 2020-2021for each of the STEM providers by subject.

The SFC team has been extremely busy and does not currently have the capacity to deliver the Higher Education (HE) STEM data, which was another prerequisite for this project. However, they have assured that that they will be able to do so by the end of this month i.e., August 2022.

Hence, the STEM FE enrolments data will be used to create a visual representation.

2.1.4 Scottish Index of Multiple Deprivation (SIMD) Data

In the investigation report we have seen in depth information regarding what SIMD is,how it was created and the SIMD tool.

A comparative indicator of deprivation across 6,976 local regions is the Scottish Index of Multiple Deprivation (called data zones). If a place is labelled as "deprived," this may refer to people having poor incomes, but it may also indicate that there are less resources or possibilities available. Income, employment, education, health, access to services, crime, and housing are the seven categories that SIMD examines to determine the degree of deprivation in a given location.

Since 2004, the Scottish Government has been developing the Scottish Index of Multiple Deprivation internally. The latest revision was released in 2020. The collection, processing, and quality assurance of the data as well as the creation of the accompanying resources and documents all take roughly 18 months. It took around six months to process the data for the access to services domain.

Alongside the core team within the Scottish Government, SIMD entails processing data for approximately 30 indicators from a variety of data suppliers:

- Scottish Government Education Analytical Services
- Department for Work and Pensions
- Her Majesty’s Revenue and Customs
- National Records of Scotland
- NHS Scotland Information Services Division
- Scottish Qualifications Authority
- Higher Education Statistics Agency
- Skills Development Scotland
- Police Scotland
- Ofcom

The SIMD update effort started in late 2017 by going over a thorough assessment completed in 2013 and 2014 to guide the creation of SIMD 2016. The Measuring Deprivation Advisory Group (MDAG), which included both users and data providers of the Scottish Index of Multiple Deprivation, was in charge of overseeing this. The group offered guidance on matters such user demands, development priorities, methodological possibilities, output quality, dissemination, and guidance on the use of outputs.

Data zones in rural areas sometimes span a vast geographic area and reflect a more diverse population with varying degrees of disadvantage. In contrast to the greater pockets of deprivation prevalent in metropolitan areas, this means that SIMD is less effective at recognising the smaller pockets of deprivation found in more rural areas. If analysed separately from urban data zones or supplemented with other data, SIMD domain indicators can still be helpful in rural areas.

Previous SIMDs were published in 2004, 2006, 2009, 2012 and 2016.The latest i.e., SIMD 2020 was published on 28 January 2020.Figure 2 shows a description of the indicators in this dataset ("Scottish Index of Multiple Deprivation 2020 - Gov.Scot" 2022).

.

	Column	Indicator type	Description
Geography	Data_Zone	Code	2011 Data Zone
	Intermediate_Zone	Name	2011 Intermediate Zone name
	Council_area	Name	Council area name
Population	Total_population	Count	2017 NRS small area population estimates
	Working_age_population	Count	2017 NRS small area population estimates and state pension age
Income	Income_rate	Percentage	Percentage of people who are income deprived
	Income_count	Count	Number of people who are income deprived
Employment	Employment_rate	Percentage	Percentage of people who are employment deprived
	Employment_count	Count	Number of people who are employment deprived
Health	CIF	Standardised ratio	Comparative Illness Factor: standardised ratio
	ALCOHOL	Standardised ratio	Hospital stays related to alcohol use: standardised ratio
	DRUG	Standardised ratio	Hospital stays related to drug use: standardised ratio
	SMR	Standardised ratio	Standardised mortality ratio
	DEPRESS	Percentage	Proportion of population being prescribed drugs for anxiety, depression or psychosis
	LBWT	Percentage	Proportion of live singleton births of low birth weight
	EMERG	Standardised ratio	Emergency stays in hospital: standardised ratio
Education, Skills and Training	Attendance	Percentage	School pupil attendance
	Attainment	Score	Attainment of school leavers
	no_qualifications	Standardised ratio	Working age people with no qualifications: standardised ratio
	not_participating	Percentage	Proportion of people aged 16-19 not participating in education, employment or training
	University	Percentage	Proportion of 17-21 year olds entering university
Geographic Access to Services	drive_petrol	Time (minutes)	Average drive time to a petrol station in minutes
	drive_GP	Time (minutes)	Average drive time to a GP surgery in minutes
	drive_PO	Time (minutes)	Average drive time to a post office in minutes
	drive_primary	Time (minutes)	Average drive time to a primary school in minutes
	drive_retail	Time (minutes)	Average drive time to a retail centre in minutes
	drive_secondary	Time (minutes)	Average drive time to a secondary school in minutes
	PT_GP	Time (minutes)	Public transport travel time to a GP surgery in minutes
	PT_Post	Time (minutes)	Public transport travel time to a post office in minutes
	PT_retail	Time (minutes)	Public transport travel time to a retail centre in minutes
	broadband	Percentage	Percentage of premises without access to superfast broadband (at least 30Mb/s download speed)
Crime	crime_count	Count	Number of recorded crimes of violence, sexual offences, domestic housebreaking, vandalism, drugs offences, and common assault
	crime_rate	Rate per 10,000 population	Recorded crimes of violence, sexual offences, domestic housebreaking, vandalism, drugs offences, and common assault per 10,000 people
Housing	overcrowded_count	Count	Number of people in households that are overcrowded
	nocentralheat_count	Count	Number of people in households without central heating
	overcrowded_rate	Percentage	Percentage of people in households that are overcrowded
	nocentralheat_rate	Percentage	Percentage of people in households without central heating

Missing and suppressed data
In some data zones, the population in the considered age range is zero for some years. In these cases, a rate cannot be determined. This is denoted by '*'. Data have been suppressed for some indicators for reasons of disclosure control where small numbers are involved. This is also indicated by '*'.

Standardised ratios
A value of 100 is the Scotland average for a population with the same age and sex profile.

Figure 2: SIMD Indicator Descriptions("Scottish Index of Multiple Deprivation 2020 - Gov.Scot" 2022)

The data is available to download as an excel and comma separated file from the website. For the purpose of this project SIMD 2016 and SIMD 2020 data will be used to compare and find any positive or negative correlations between any of the SIMD indicators and the STEM FE enrolments. In addition to the other characteristics, this study may be affected by the education, skills, and training data in particular from this source. The council area of this dataset within the geography parameter is also helpful for the project.

2.2 Data Preparation

2.2.1 Data preparation using Excel

Currently, three data sets are available: STEM FE enrolment data, SIMD 2016 data, and SIMD 2020 data. Combining SIMD 2016 and SIMD 2020 was the first step followed by combining the resulting dataset with the STEM dataset, in determining whether there were any relationships between UHI STEM enrolments and the two SIMD datasets. So, in establishing common traits between them and given that all the three files were Microsoft Excel spreadsheets, some basic preparation was done manually to save time on writing a code as follows:

i. It was observed that SIMD 2016 and SIMD 2020 had a few characteristics that were unique to each, it was necessary to identify common traits. Characteristics of NEET and HESA were discovered to be unique to the 2016 SIMD dataset. While SIMD2020v2 _Rank, SIMD_2020v2_Percentile, SIMD2020v2_Vigintile, SIMD2020v2_Decile, SIMD2020v2_Quintile, SIMD2020v2_Income_Domain_Rank, SIMD2020_Employment_Domain_Rank, SIMD2020_Health_Domain_Rank,

SIMD2020_Education_Domain_Rank, SIMD2020_Access_Domain_Rank, SIMD2020_Crime_Domain_Rank, SIMD2020_Housing_Domain_Rank, not participating, University and broadband were identified as being unique to the SIMD 2020. These characteristic values were manually deleted from the datasets respectively.

ii. A new column called "SIMD year" was introduced, with the values 2016 and 2020 for the SIMD 2016 data and SIMD 2020, respectively. Later, a manual Excel merge of the two datasets was performed and a unified "SIMD 2016-2020" dataset was obtained.

iii. This dataset was then filtered by the council areas of the UHI colleges i.e., Argyll and Bute for UHI Argyll and Bute, Highland for UHI Inverness, Na h-Eileanan an Iar for UHI Outer Hebrides, Moray for UHI Moray, Orkney Islands for UHI Orkney, Perth and Kinross for UHI Perth, Shetland Islands for UHI Shetland to create a new dataset titled "SIMD 2016-2020 by UHI region council areas".

iv. The proprietary dataset also required some manual pre-processing. The STEM dataset now includes extra columns for Council area, Latitude, and Longitude. The names of the council areas for each UHI college were added. Geocoding was used in Google Sheets to calculate the latitude and longitude data. This action was done in consideration of the HISP project's demand for a location map. Additionally, a new worksheet was made containing the name of UHI College, address, latitude, longitude, and council area so that a map could be constructed using this information.

v. As section 2.1.3 noted that searches for UHI West Highland and UHI North Highland did not turn up any FE STEM provisions, the corresponding council areas had to be left off the enrolment data sheet because including just the college's name and the council area would not be useful. However, the values of the columns in the address worksheet were not altered.

As a result of this pre-processing, we have two datasets "SIMD 2016-2020 by UHI region and council areas" and STEM enrolment dataset which have common attributes which are council area and UHI college. The next logical step would be to merge these in order to achieve any correlations.

2.2.2 Data preparation using R Studio

The remaining pre-processing was completed in R Studio. The new data set i.e., "SIMD 2016-2020 by UHI region council areas" required to be averaged by each characteristic and by each council area for "SIMD year" characteristic values 2016 and 2020. In order to do so R Studio data science software solution utilising R commands was the choice due to the ease and flexibility of performing the tasks. The actions taken within R Studio are listed below:

i. As seen in Figure 3, the dataset was loaded in R studio. There were 35 characteristics and 1688 records. Figure 4 shows how the variables had to be converted in order to guarantee that all the datatypes were accurate. This demonstrated that every attribute was of the appropriate datatype.

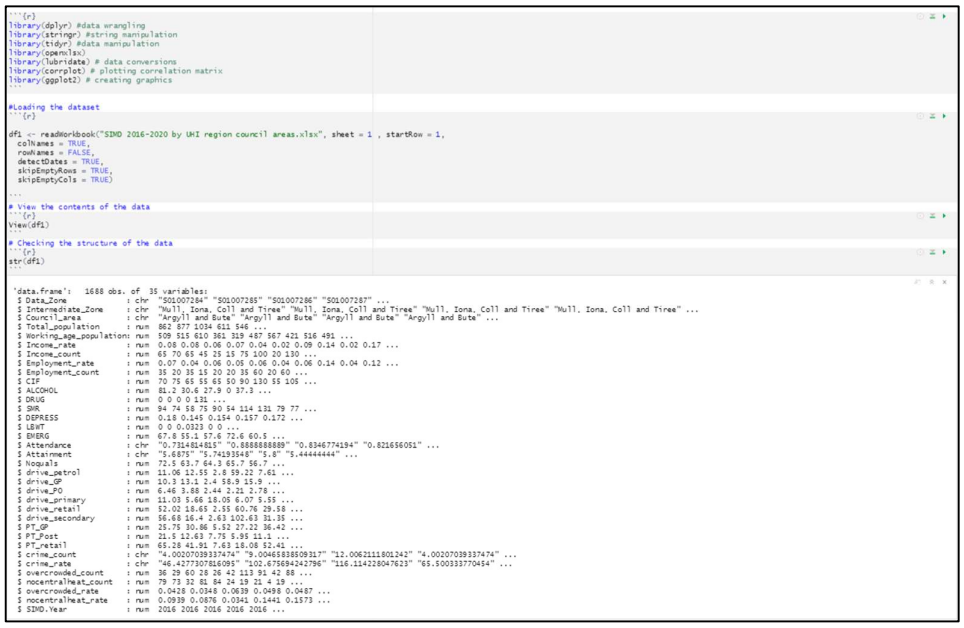


Figure 3: Data structure in R Studio

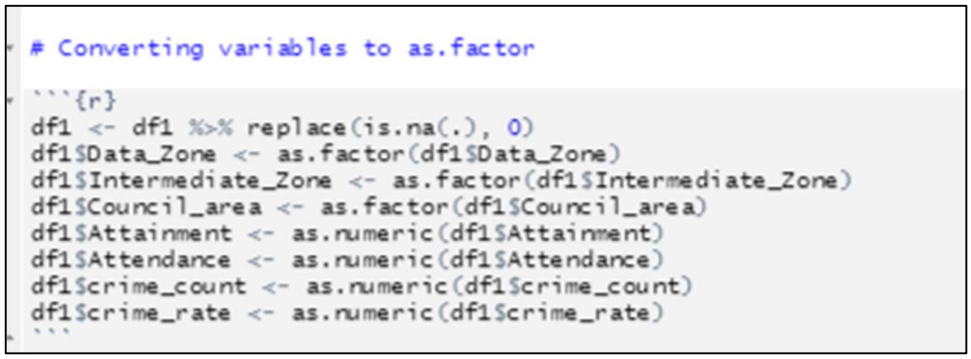


Figure 4: Converting data types of variables

ii. When looking at the SIMD dataset, it was pretty clear that there were certain cases with "*" values or missing values that needed to be handled. This was aso verifed using R commands as seen in Figure 5.

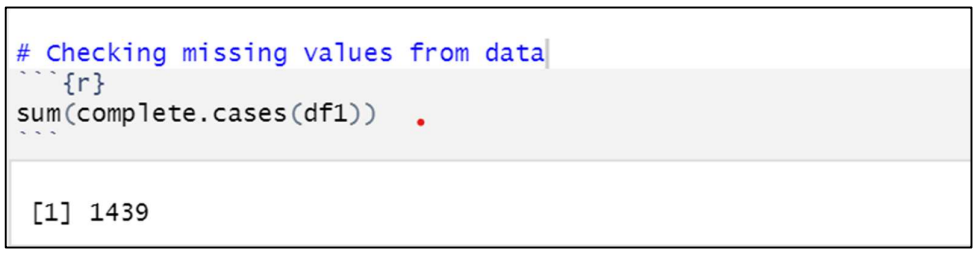


Figure 5: Checking for missing values

As per the SIMD indicator description ("Scottish Index of Multiple Deprivation 2020 - Gov.Scot" 2022) the population in the considered age range is zero in some data zones for a

period of time. Under these circumstances a rate cannot be established. This is indicated with a "*." It was therefore decided to remove these instances.

It was determined that the variables Attainment, Attendance, Crime rate, and Crime count contained the values of "*". To remove of them, the code seen in Figure 6 was applied.

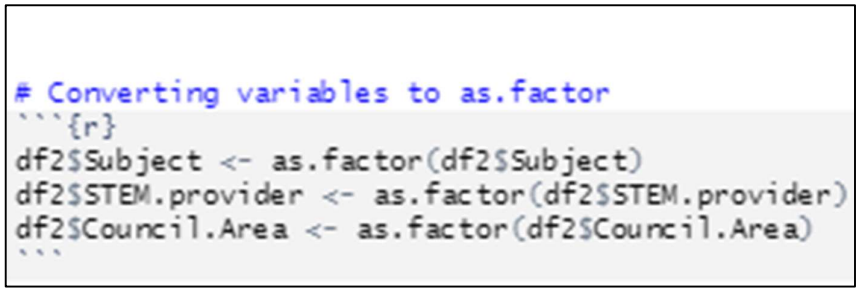
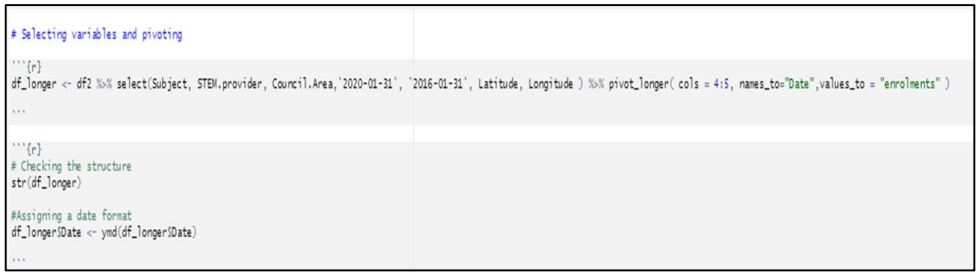
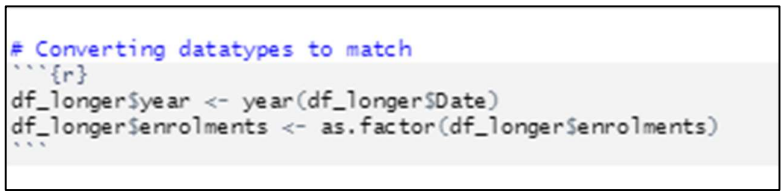
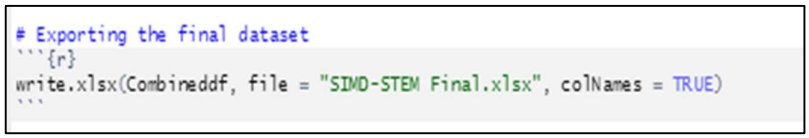
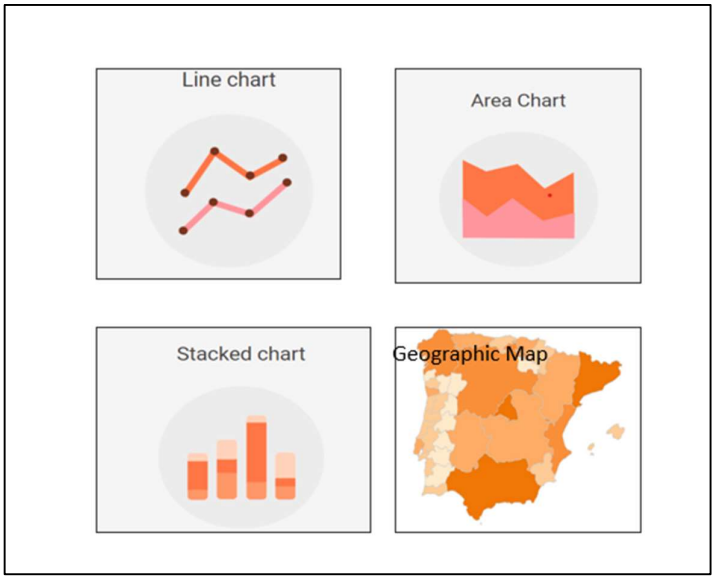
```
# Removing instances with "*" as values from the Attainment, Attendance, crime_rate and crime_count variables

```{r}
df1 <- df1 %>% filter(!Attainment == "*") %>% filter(!Attendance == "*") %>% filter(!crime_count == "*")%>% filter(!crime_rate == "*")
```
```

Figure 6 : Filtering instances with "*"

iii. Calculating the averages of all characteristics for the years 2016 and 2020, leaving each UHI college and council region with one value for each variable for each year i.e., 2016 and 2020. This can be seen in Figure 7 and Figure 8.

```
# Rounding off few variable values, filter by SIMD year 2016 and calculating mean of all the variables.

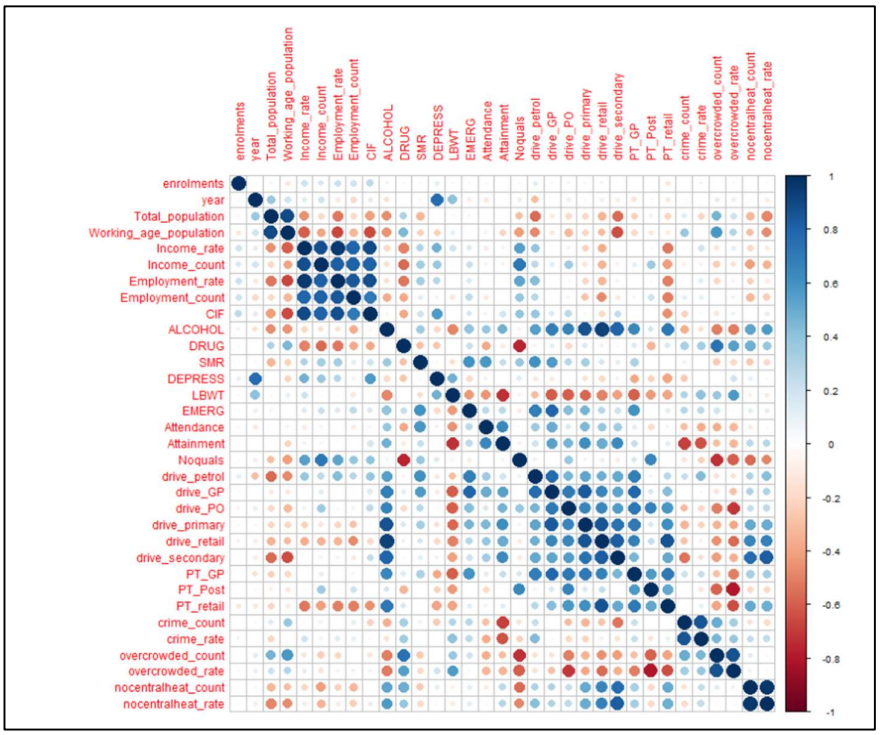
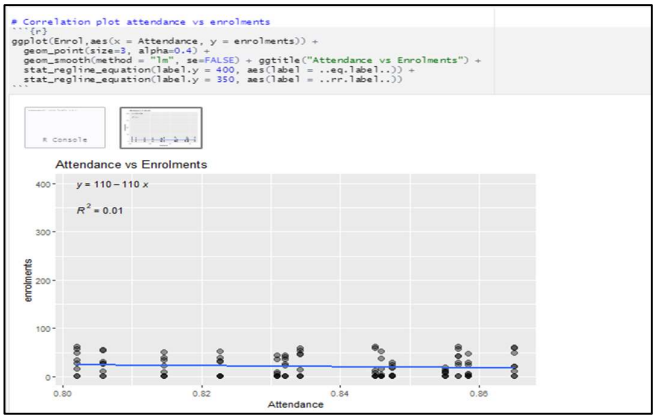
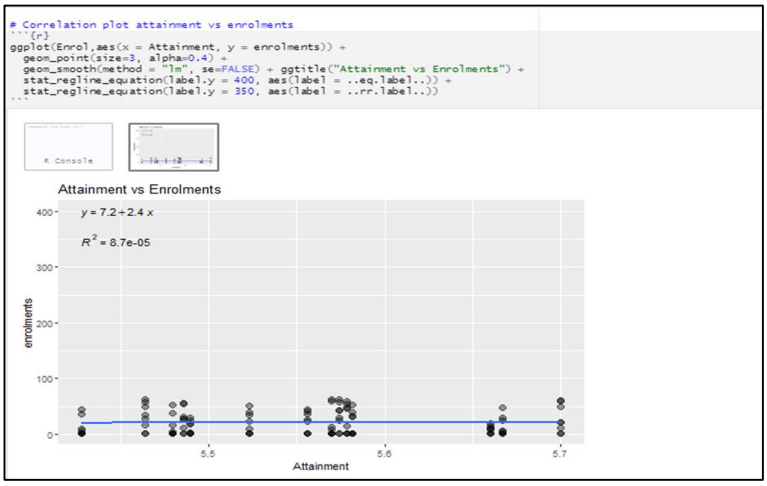
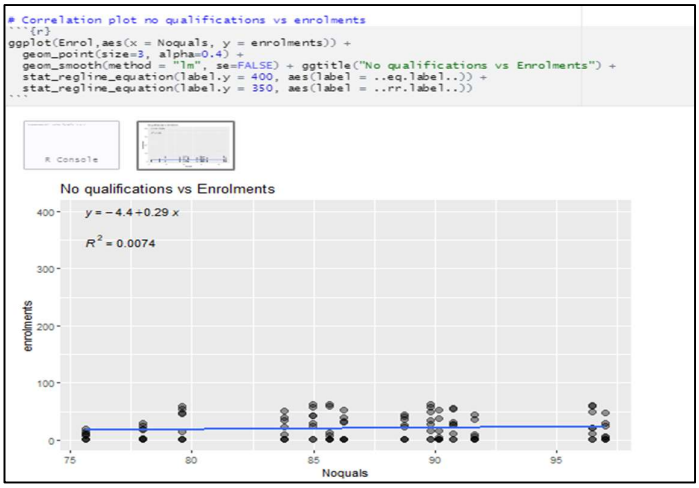
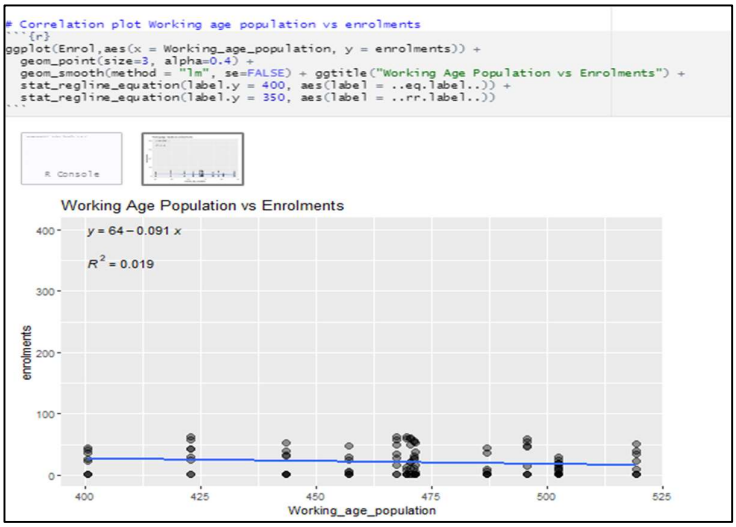
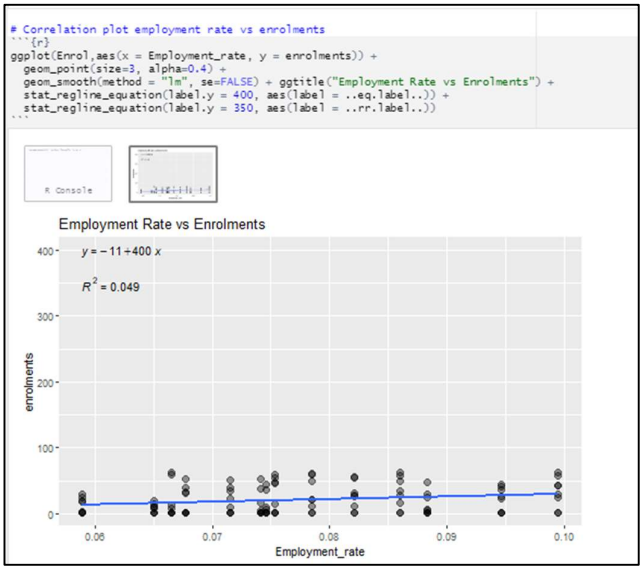
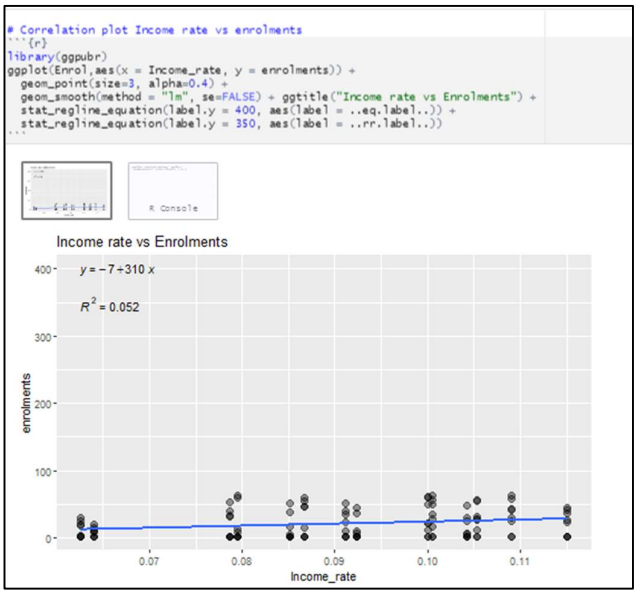
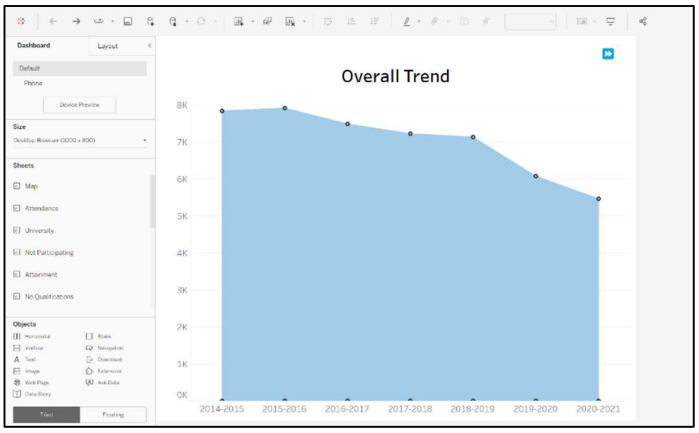
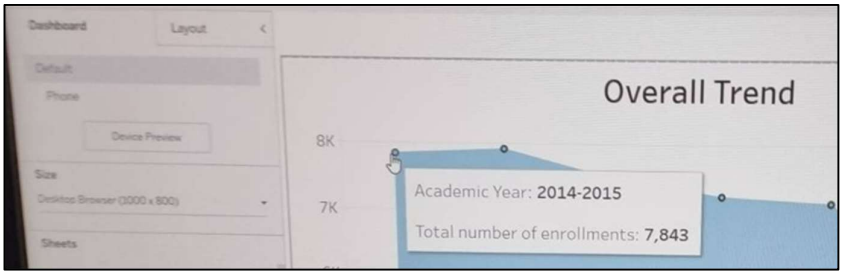
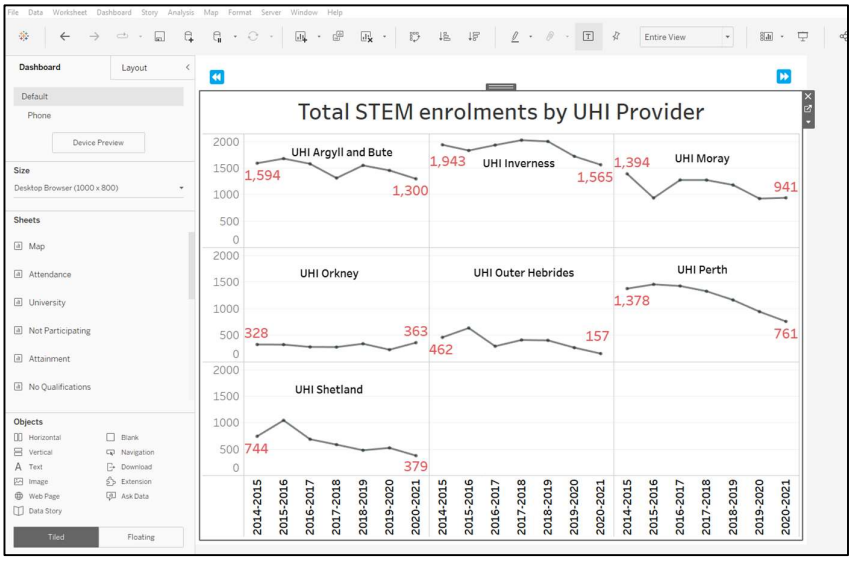
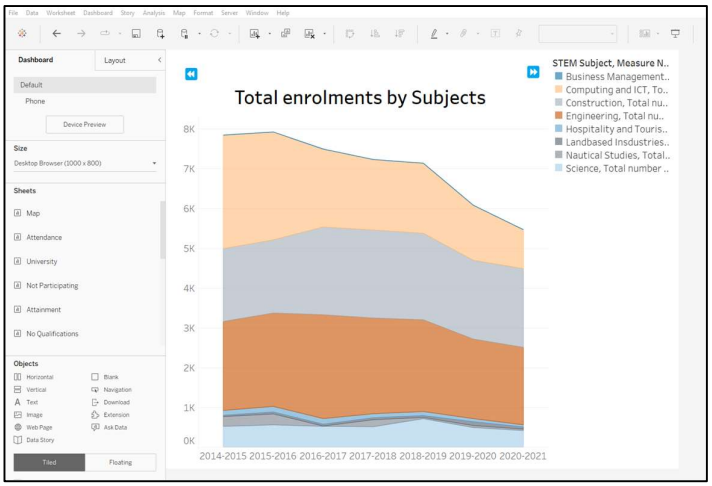
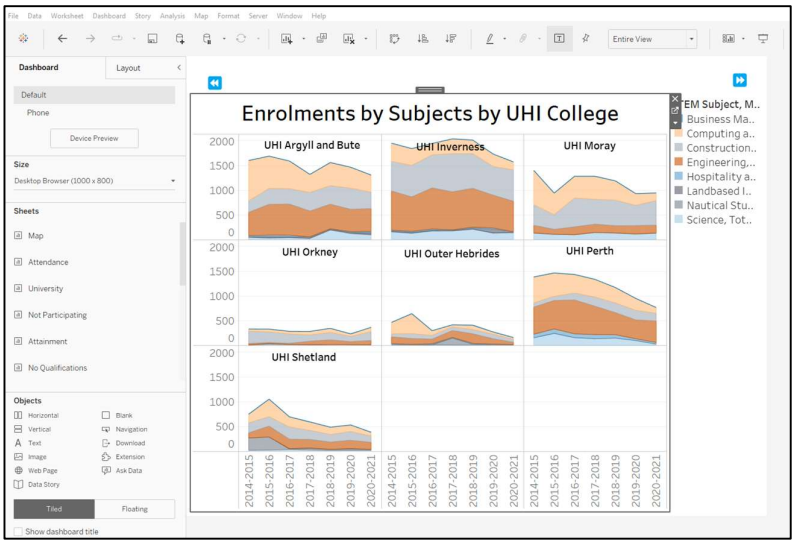
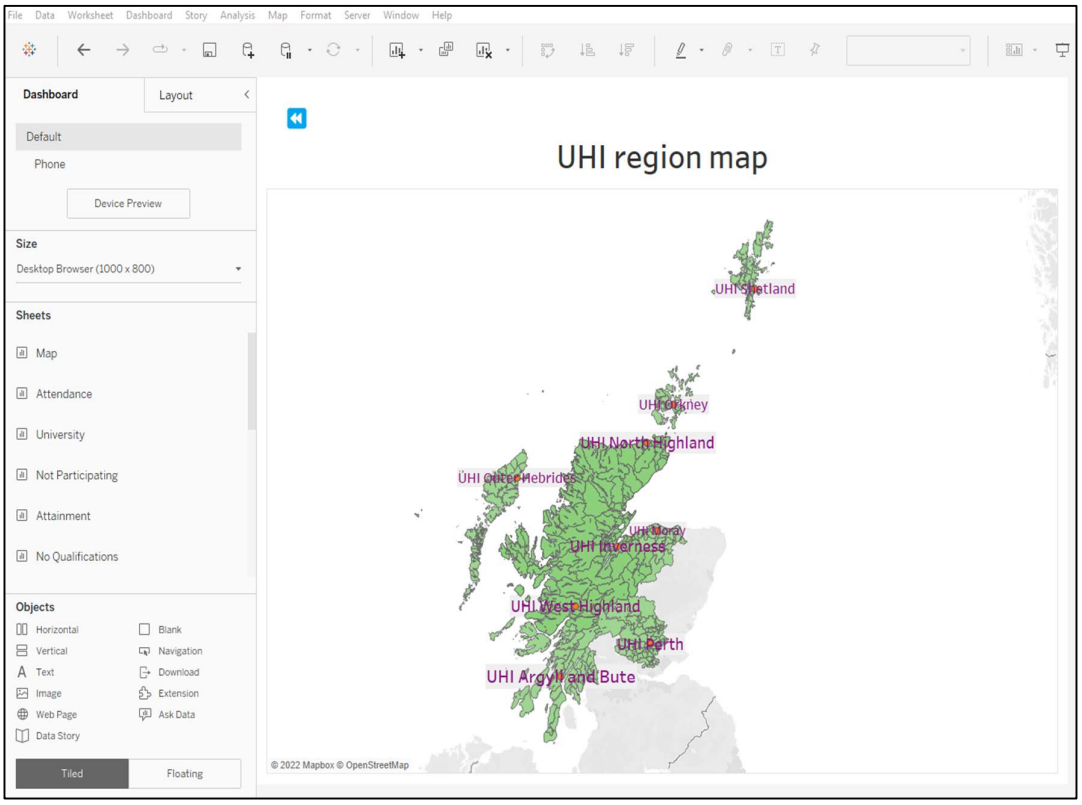
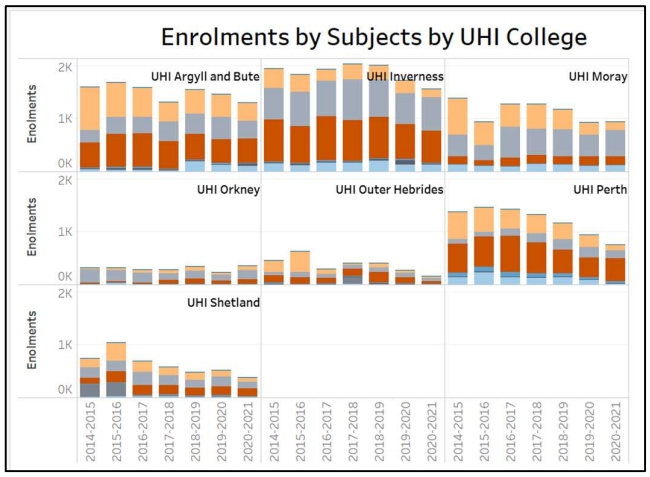
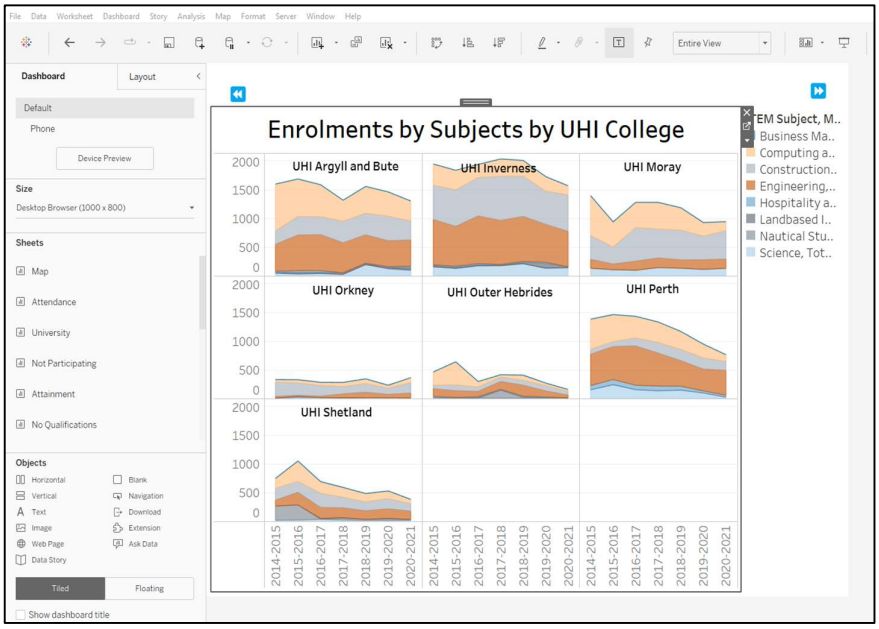
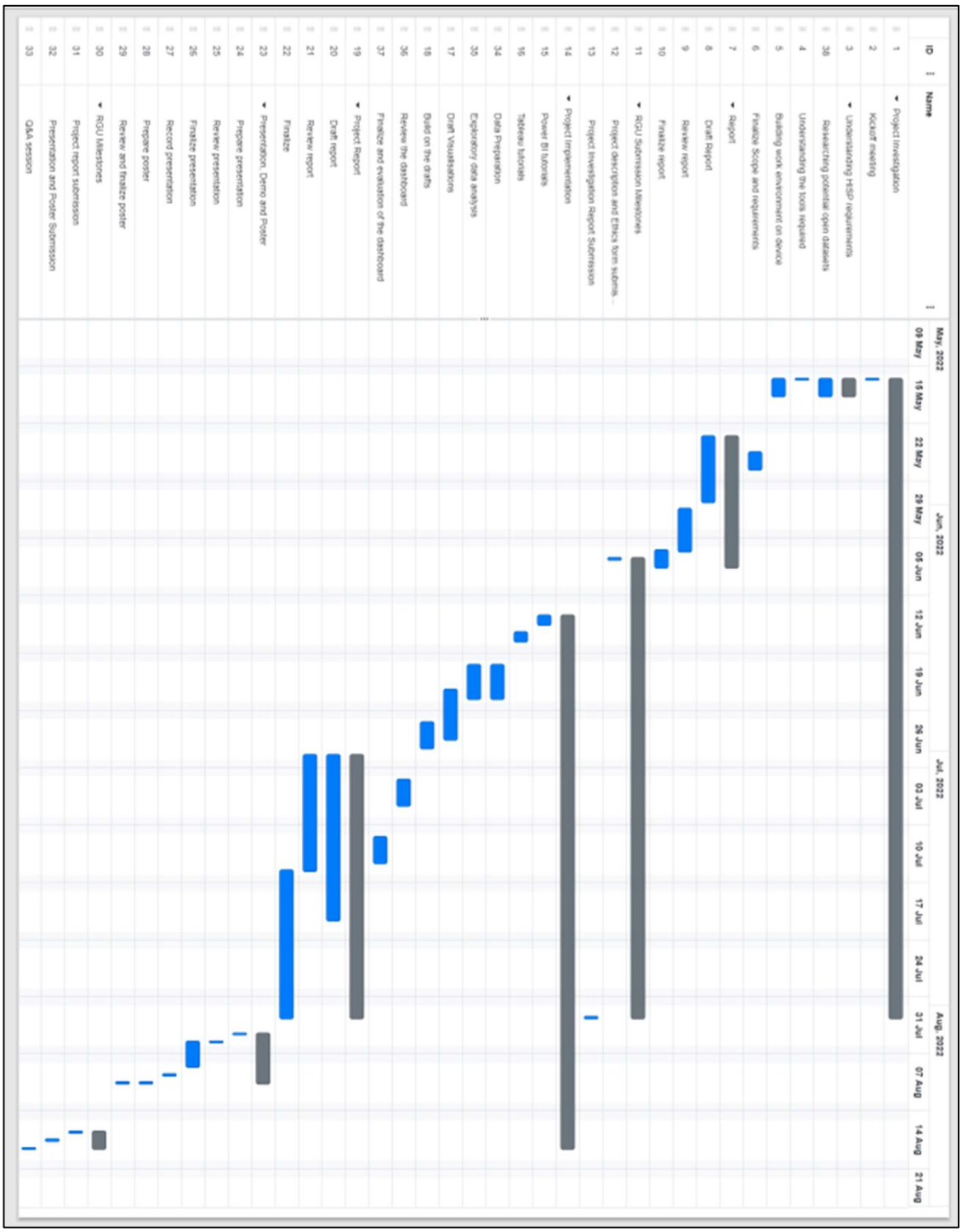
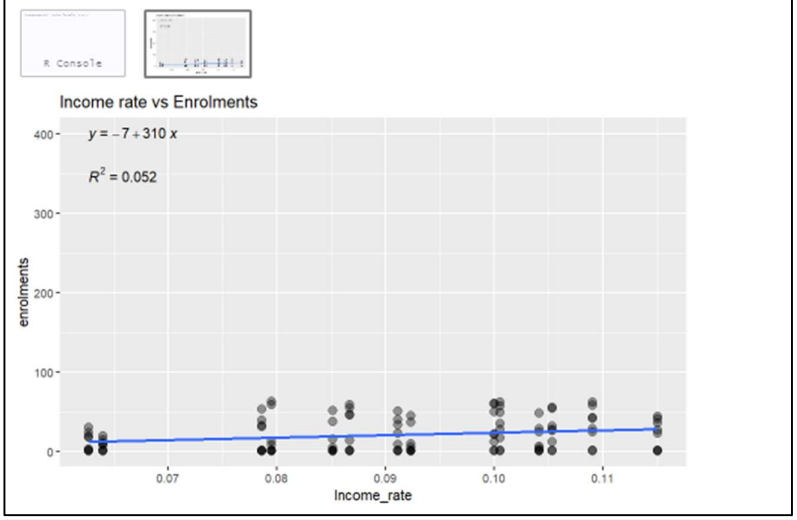
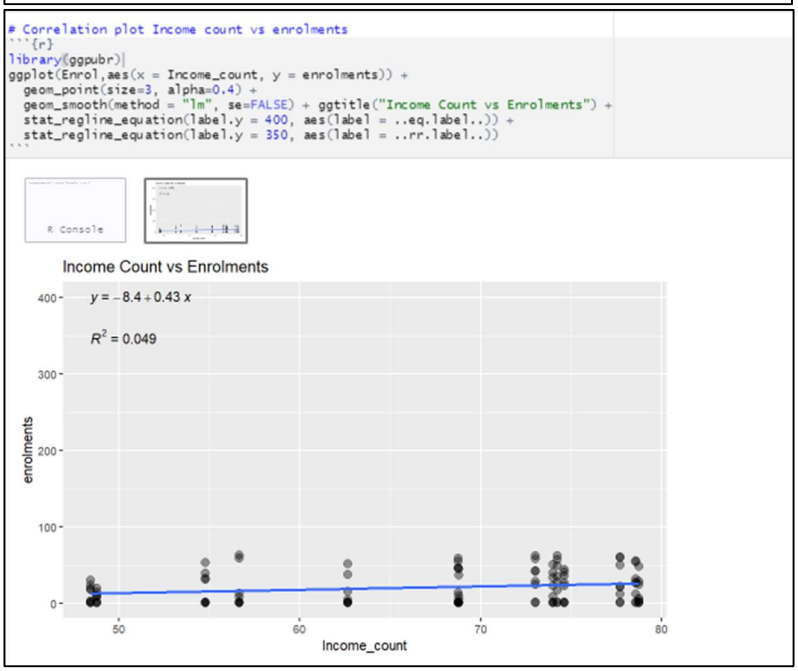
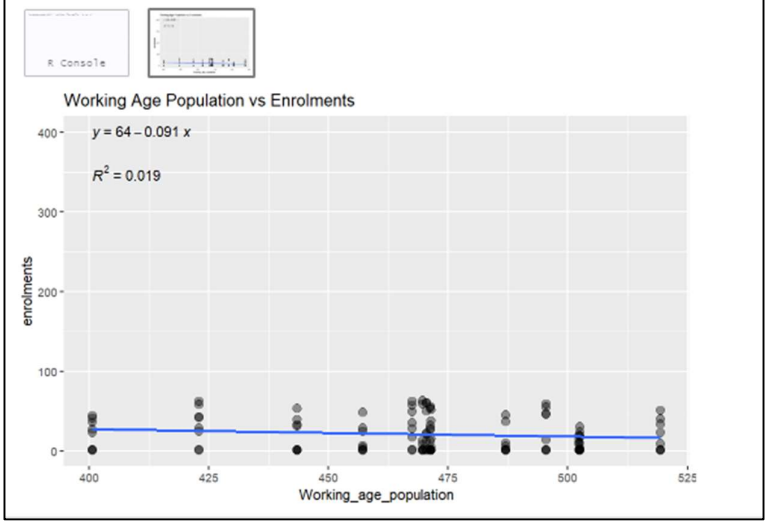
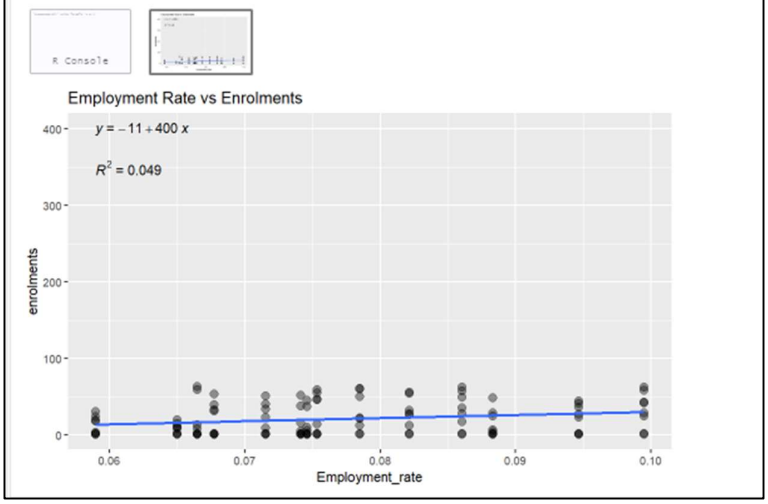
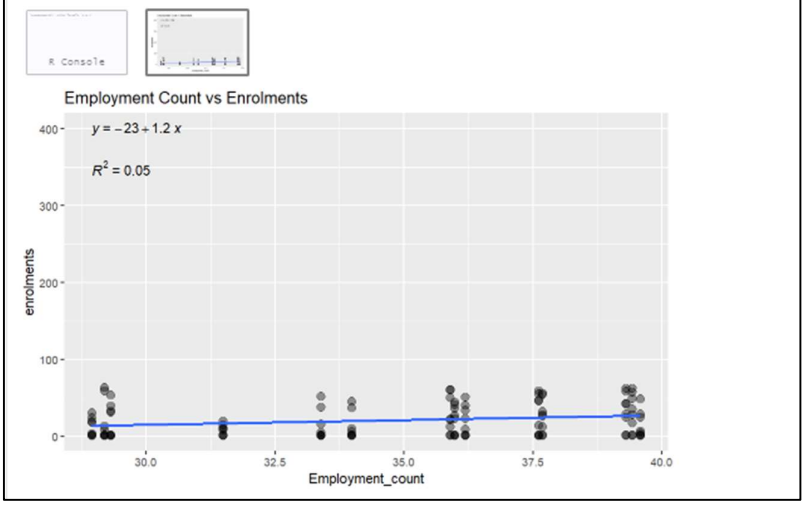
```{r}
SIMD2016 <- df1 %>% replace(is.na(.), 0) %>%
  mutate(Attendance=round(Attendance,digits=1),
         Attainment = round(Attainment,digits=1),
         crime_count = round(crime_count,digits=1),
         crime_rate = round(crime_rate,digits=1)) %>%
  filter(SIMD.Year=="2016") %>%
  group_by(Council_area) %>%
  summarise(
    Year= max(SIMD.Year),
    Total_population = mean(Total_population),
    Working_age_population = mean(Working_age_population),
    Income_rate = mean(Income_rate),
    Income_count = mean(Income_count),
    Employment_rate = mean(Employment_rate),
    Employment_count = mean(Employment_count),
    CIF = mean(CIF),
    ALCOHOL = mean(ALCOHOL),
    DRUG = mean(DRUG),
    SMR = mean(SMR),
    DEPRESS = mean(DEPRESS),
    LBWT = mean(LBWT),
    EMERG = mean(EMERG),
    Attendance = mean(Attendance),
    Attainment = mean(Attainment),
    Noquals = mean(Noquals),
    drive_petrol = mean(drive_petrol),
    drive_GP = mean(drive_GP),
    drive_PO = mean(drive_PO),
    drive_primary = mean(drive_primary),
    drive_retail = mean(drive_retail),
    drive_secondary = mean(drive_secondary),
    PT_GP = mean(PT_GP), PT_Post = mean(PT_Post),
    PT_retail = mean(PT_retail),
    crime_count = mean(crime_count),
    crime_rate = mean(crime_rate),
    overcrowded_count = mean(overcrowded_count),
    overcrowded_rate = mean(overcrowded_rate),
    nocentralheat_count = mean(nocentralheat_count),
    nocentralheat_rate = mean(nocentralheat_rate) )
```
```

Figure 7: Mean of all variables for year 2016

```
# Rounding off few variable values, filter by SIMD year 2020 and calculating mean of all the variables.

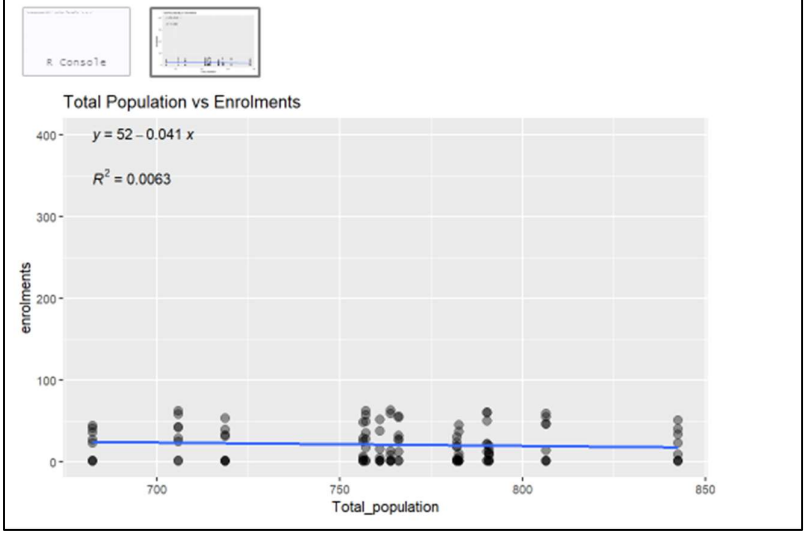
```{r}
SIMD2020 <- df1 %>% replace(is.na(.), 0) %>%
  mutate(Attendance=round(Attendance,digits=1),
         Attainment = round(Attainment,digits=1),
         crime_count = round(crime_count,digits=1),
         crime_rate = round(crime_rate,digits=1)) %>%
  filter(SIMD.Year=="2020") %>%
  group_by(Council_area) %>%
  summarise(
    Year= max(SIMD.Year),
    Total_population = mean(Total_population),
    Working_age_population = mean(Working_age_population),
    Income_rate = mean(Income_rate), Income_count = mean(Income_count),
    Employment_rate = mean(Employment_rate),
    Employment_count = mean(Employment_count),
    CIF = mean(CIF), ALCOHOL = mean(ALCOHOL),
    DRUG = mean(DRUG),SMR = mean(SMR), DEPRESS = mean(DEPRESS),
    LBWT = mean(LBWT), EMERG = mean(EMERG),
    Attendance = mean(Attendance),
    Attainment = mean(Attainment),
    Noquals = mean(Noquals), drive_petrol = mean(drive_petrol),
    drive_GP = mean(drive_GP),
    drive_PO = mean(drive_PO),
    drive_primary = mean(drive_primary),
    drive_retail = mean(drive_retail),
    drive_secondary = mean(drive_secondary),
    PT_GP = mean(PT_GP), PT_Post = mean(PT_Post),
    PT_retail = mean(PT_retail),
    crime_count = mean(crime_count),
    crime_rate = mean(crime_rate),
    overcrowded_count = mean(overcrowded_count),
    overcrowded_rate = mean(overcrowded_rate),
    nocentralheat_count = mean(nocentralheat_count),
    nocentralheat_rate = mean(nocentralheat_rate) )
```
```

Figure 8: Mean of all variables for year 2020

These two datasets were then combined as seen in Figure 9.

```
# Combining averaged SIMD 2016 and SIMD 2020

```{r}
dfFinal<- rbind(SIMD2016,SIMD2020)
```
```

Figure 9: Combining averaged variables data for year 2016 and 2020

iv. Exporting the merged file as seen in Figure 10

```
# Exporting the dataframe as an Excel file

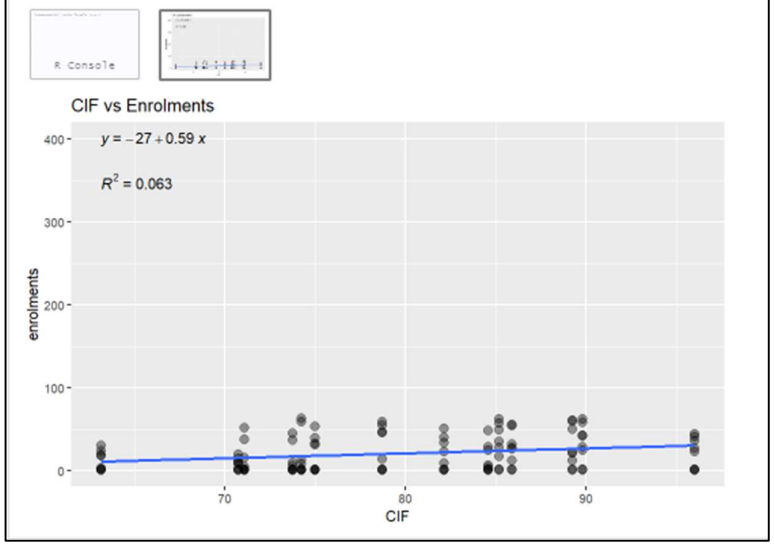
```{r}
write.xlsx(dfFinal, file = "Averaged SIMD 2016-2020 by UHI region Council areas .xlsx", colNames = TRUE)
```
```

Figure 10: Exporting SIMD 2016-2020 merged file

v. Loading the STEM dataset and preparing it to be merged with the SIMD dataset. A similar approach to the above was taken to prepare this data.

Figure 11: Loading STEM data

Figure 11 shows code for loading the dataset, viewing the contents of the data and checking the structure of data.



Figure 12: Converting datatypes in STEM data

Figure 12 shows code used to convert the datatypes of a few variables for further processing.



Figure 13: Creating new columns Date and enrolments

Figure 13 shows that two new columns were created namely Date which has a date of 31/01/2016 for a value 2016 in the year column and 31/01/2020 for a value 2020 in the year column. And the new enrolments column takes the number of enrolments corresponding to those years.

Figure 14 shows conversion of datatypes for columns Date and enrolments.



Figure 14: Converting datatypes for variables Date and enrolments

Figure 15 shows merging of the SIMD and STEM datasets by the year and council area.

Figure 15: Merging SIMD and STEM dataset by Year and Council area

Figure 16 shows the exporting of the merged dataset.



Figure 16: Exporting the SIMD and STEM merged dataset

This dataset will be used to plot any correlations.

2.3 Choice of Software Tool

As seen in the investigation report, the initial choice for implementation was Microsoft Power BI. As Microsoft Power BI requires no licencing fees and integrates effortlessly with Microsoft 365, a popular programme that UHI also uses. As a result, the UHI do not have to pay money for a new licence for a different piece of software.

But after studying the technology and comparing it to Tableau, I realised that Tableau was a better choice for this project. The Gartner Magic Quadrant for Analytics and Business Intelligence Platforms, released in March 2022, positioned Tableau too as a Leader (Gartner 2022).

Users of all data literacy levels can use analytics and generate visualisations because to Tableau software's straightforward design. New users and data analysts alike can quickly create visualisations and dashboards with Tableau software's drag-and-drop features, simple drill-down capabilities, and natural language querying. This allows users to obtain insight practically immediately. Analysts can respond to complicated inquiries and produce detailed visualisations without sacrificing detail or data quality, while beginner users may easily read dashboards without getting lost or overwhelmed. For example, if a user wants an axis to represent specific fields or parameters, they can simply import the data source and drag the required field onto the axis (Brockman 2022). One of the simplest examples, is selection of multiple worksheets or dashboards, Tableau allows this multiple selection, whereas it was not possible to do in Power BI. Also, the spatial/map features are better in Tableau.

Tableau automatically creates beautiful, interactive maps with 16 levels of zoom using the location data and information already present. Users may also create custom geocodes to map the information that is important to their requirements. There are built-in census-based population, income, and other common demographic datasets. Tableau allows for importing any spatial files or custom geocoded to make geographic data more readily available, interactive and shareable via Tableau Cloud, Tableau Public, and Tableau Server (Tableau 2022).

This project will not see any use of machine learning libraries or algorithms because:

- The data is comparatively clear and uncomplicated.
- In addition, there aren't enough data points in the main dataset, which is the STEM data from the SFC, to use machine learning ("How Much Data Is Required for Machine Learning? - Postindustria" 2022).

2.4 Choice Visualisations

Figure 17 shows the selection of charts that are considered for the data visualisation.



Figure 17: Types of Charts ("Cheatsheet For Charts: Pair The Right Chart With The Right Set Of Data" 2022)

- **Line charts and area charts:** Predictive analytics typically employs line charts and area charts to illustrate change in one or more quantities by charting a series of data points across time. While area charts connect data points with line segments, stack variables on top of one another, and use colour to differentiate between variables, line graphs use lines to show these changes.
- **Stacked bar charts:** Each section of these graphs represents a component of the whole. They offer an easy method for compiling data and evaluating the relative sizes of various components.
- **Geographic Maps:** The purpose of this type of visualisation is to analyse and present geographically related data in the form of maps. This type of data expression is more transparent and intuitive. We can see the proportion or distribution of data in each region ("Top 10 Map Types in Data Visualization" 2022).

By examining all the data visualisations and relating it to this project, a conclusion can be drawn that line charts, stacked bar charts, area charts and geographical maps are most relevant to this project. The line charts can be used to see an overall trend, trend in STEM enrolments by UHI colleges, while the stacked bar charts or area charts can be used to see the spread of STEM subjects. As the UHI campuses are dispersed over a large geographical area, the HISP team is interested in incorporating maps into the final visualisation to gain insights regarding rurality and deprivation. Thus, using geographic maps is the preferred method. Furthermore, an effort has been made to leverage the built-in colour-blind palette in Tableau to increase the accessibility of the visualisations.

Correlation plots and a corelation matrix have been produced using R language in R studio. The libraries used to do so are ggplot2 and corrplot. However, the variables with strong relationships only will be displayed in the report. While it made sense with the SIMD to use the 2016 and 2020 data, however, the STEM data is available for academic years i.e., 2015–2016, 2016–2017, 2019–2020, and 2020–2021. Some data for the year 2016 will be available in the academic year 2015–2016 and 2016–2017, the same for year 2020. Choosing STEM enrolment data was unclear in order to look for correlations.

As discussed in section 2.1.3, the information is collected annually and tends to be available in January for the preceding academic year. So, in January this year i.e., 2022 the data was available for academic year 2020-21. An assumption was made to select January 2016 and January 2020 STEM enrolments for correlation plots. But instead of being used in the dashboard, the plots will be shown separately giving the flexibility to take it forward when a decision is made on what enrolment data can be used to find precise relationships.

3. Implementation

This section explains how the visualisations were created before briefly addressing any interpretations or key takeaways.

3.1 Correlation matrix

All the categorical variables were filtered to be able to proceed with the correlation matrix as seen in Figure 18.

```
# Variable selection to do a correlatin matrix

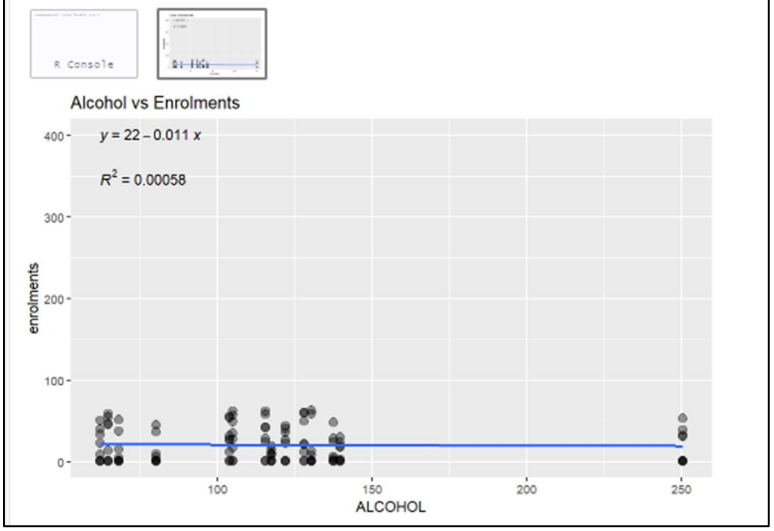
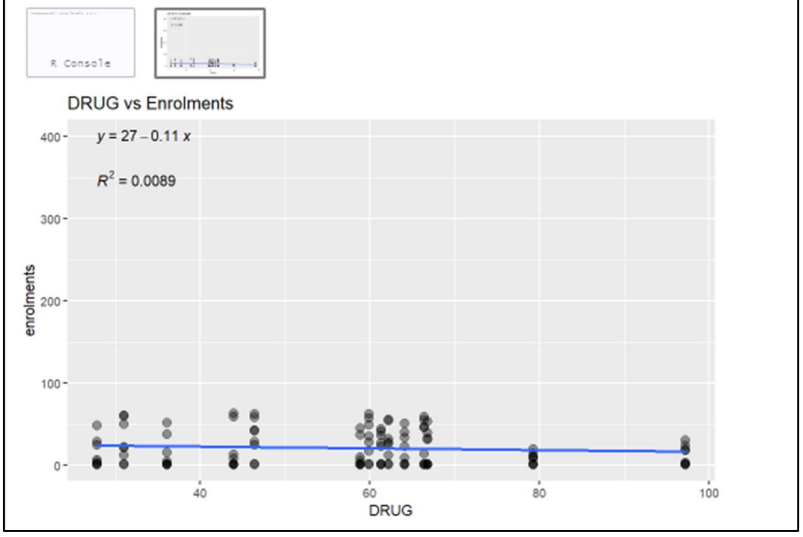
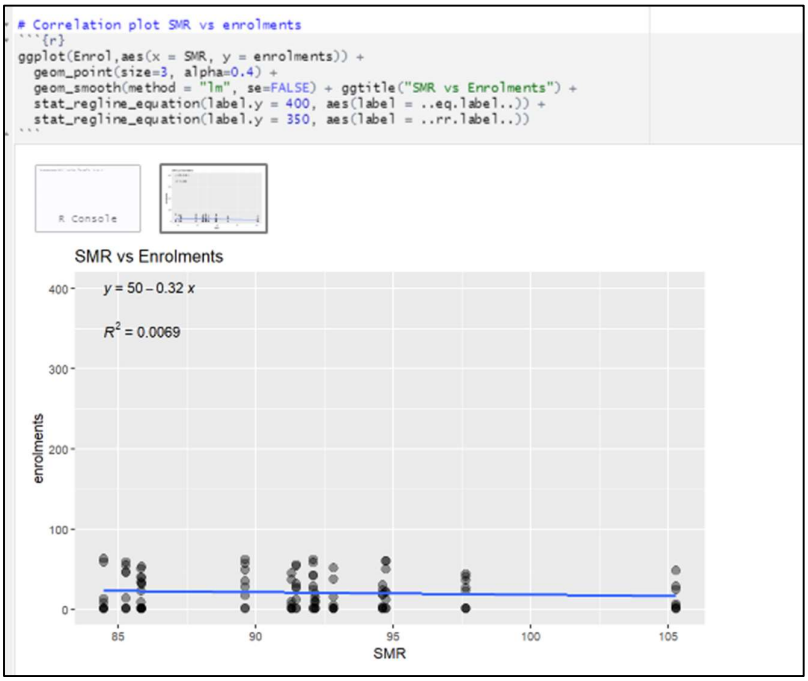
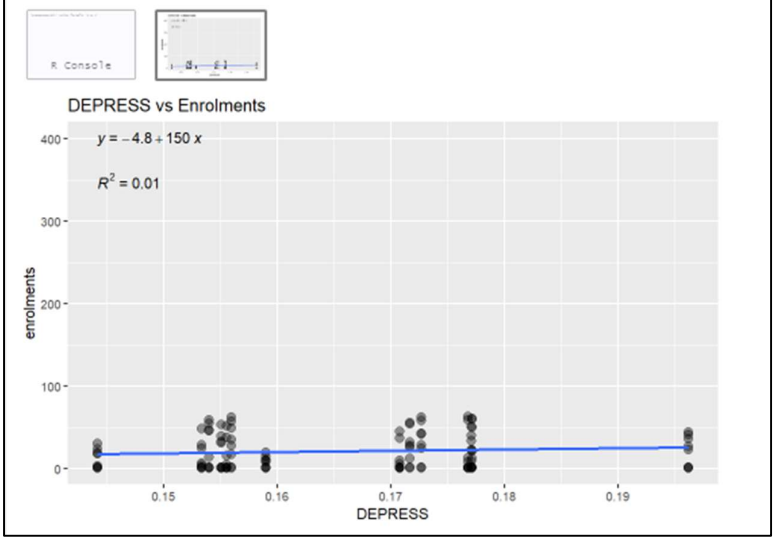
```{r}
Enrol <- subset(Combineddf, select = -c(Subject,STEM.provider,Council.Area, Latitude,Longitude,Date))
```
```

Figure 18: Variable selection for Correlation Matrix

A correlation matrix was plotted in R studio using R commands as can be seen in Figure 19.

```
# Correlation matrix

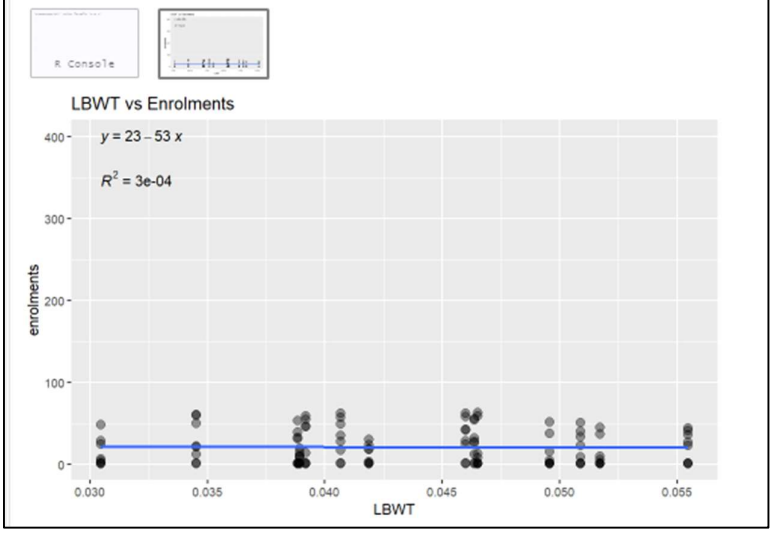
```{r}
round(cor(Enrol, method = "spearman"), 2)
corrplot(cor(Enrol, method = "spearman"))
```
```

Figure 19: Plotting Correlation Matrix

The resulting matrix can be seen in Figure 20.



Figure 20: Correlation Matrix

Although there are many positive and negative correlations, we only pay attention to our target variable, which is the number of enrolments. The matrix clearly shows that enrolments are correlated with some of the other factors, including working age population, employment rate, and income count. Each of the variables was plotted against the variable enrolments in order to closely assess how positively or negatively each of the variables is associated. The Appendix B: Correlation Plots has a list of all the plots, but we only address the ones that stand out and are particularly pertinent to this project.

3.1.1 Attendance vs Enrolments

Attendance and enrollments appear to be negatively correlated, as seen in Figure 21. The poor attendance could be interpreted as a sign that the students lack the prerequisites for continuing their study, which would explain the negative trend.



Figure 21: Attandance vs Enrolment

3.1.2 Attainment vs Enrolments

As seen in Figure 22, there is a positive association between enrollments and attainment. The increased tendency could be explained by the fact that pupils who have higher academic achievement are more inclined to continue their education.



Figure 22: Attainment vs Enrolments

3.1.3 No qualifications vs Enrolments

No-qualifications and enrollments have a positive correlation, as can be seen in Figure 23
The growing trend could be explained by the fact that those without qualifications are also more likely to return to education in order to obtain certifications that could help them land a job or receive a salary raise or promotion by enrolling in additional courses to acquire the necessary skills.



Figure 23: No qualifications vs Enrolments

3.1.4 Working age population vs Enrolments

This indicator in the SIMD is described based on 2017 NRS a small area population estimate and state pension age ("Scottish Index of Multiple Deprivation 2020 - Gov.Scot" 2022). Figure 24 illustrates how negatively correlated working age population and enrolment are. We can infer that people are less likely to enrol in any education the older they get, especially if they're getting close to retirement age.



Figure 24: Working age population vs Enrolments

3.1.4 Employment rate vs Enrolments

Employment count and enrolments have a positive correlation, as can be seen in Figure 25.
The fact that the rate of employability rises with education may help to explain the upward trend.



Figure 25: Employment rate vs Enrolments

3.1.5 Income rate rate vs Enrolments

The salary you can make is also influenced by your education.This is evident from Figure 26 showing a positive correlation between income rate and enrolments.



Figure 26: Income rate vs Enrolments

3.2 Dashboard

The main objective of the project was to create a data visualisation that can help the HISP team to plan activities, understand enrolments in STEM subject entrants at college and forecast future trends.

The proprietary STEM FE dataset consists of the following information:

i. STEM FE subjects, which list the STEM subjects which are Business Management and Administration, Computing and ICT, Construction, Engineering, Hospitality and Tourism, Nautical Studies, and Science.
ii. UHI FE colleges. The colleges that come under the UHI are UHI Argyll and Bute, UHI Inverness, UHI Outer Hebrides, UHI Moray, UHI Orkney, UHI Perth, UHI Shetland, UHI West Highland and UHI North Highland. This dataset however does not have information about the UHI West Highland and UHI North Highland because as per the SFC representative who provided this proprietary data, the searches did not return any FE STEM provision under these two colleges.
iii. Number of enrolments for academic years 2014-2015 to 2020-2021for each of the STEM providers by subject.

With the use of this information the below dashboards were built using Tableau desktop.

3.2.1 Overall Trend

To view the overall enrollment for all UHI colleges and all subjects, the visualization below in Figure 27 was created. The X-axis represents the academic years 2014–2015, 2015–2016, 2016–2017, 2017–2018, 2018–2019, and 2020–2021, while the Y-axis represents the total number of enrollments for each academic year.

The dots that can been seen represent each academic year and the total enrolments for that particular academic year. In the tableau workbook when you hover over these dots it displays the same as can bee seen in Figure 28.



Figure 27: Overall Trend



Figure 28: Tool tip description

A line chart was the most appropriate form to use as the goal of this plot was to identify a general trend.

3.2.2 STEM enrolments by UHI College

The enrollment data for each UHI college was further broken down to look for any patterns. Small multiples were used create this visualization. A data visualisation known as a small multiple is made up of several charts that are placed in a grid. This makes it simple to compare all the data. They are also known as panel, grid, trellis, and lattice charts (Bock 2022). Given that the purpose of this plot was to identify a trend in all the UHI colleges, a line chart was the most suitable form to utilise.



Figure 29: STEM Enrolments by UHI College

The decline in enrolment at all of the colleges is a general trend that is apparent in Figure 29. Although the overall number of students enrolments has decreased, this does not necessarily indicate that teaching activity has reduced.

As per the College Statistics 2020-2021 for Scotland, published by the Scottish Funding Council in January 2022, Full-time FTEs (across all funding sources) in Further Education (FE) dropped by 5.8% to 50,087 in 2020–21, although part-time FTEs increased by 3.4% to 38,033 during the same year ("Scottish Funding Council" 2022).

In order to quantify college student activity, additional criteria including demographics, course enrolment, credits, and full-time equivalents (FTEs) must also be taken into account.

3.2.3 Subject enrolments

Figure 30 shows an overall dispersion of Subjects. We can observe that Engineering, Construction, Computing, and ICT are the fields with the most students enrolled.



Figure 30: Enrolments by Subjec

3.2.4 Enrolments by Subjects by UHI College

We break down the subject distribution to each UHI college in further detail. We can see that engineering has the highest enrollment at UHI Argyll and Bute, UHI Inverness, and UHI Perth, followed by construction Figure 31.



Figure 31: Enrolments by Subject and College

3.2.5 UHI Region Map

As UHI campuses are dispersed over a large geographical area, the HISP STEM team is interested in incorporating maps into the final visualisation to gain insights regarding rurality and deprivation.

The map shown in Figure 32 was created using the data that was available at the time of this research. The UHI campuses are highlighted on the map, which also includes council areas with the total population figures for all of the UHI regions.



Figure 32: UHI Region Map

3.3 Challenges faced during Implementation

The HISP data project implementation process is not an exception to the rule that challenges are a natural part of the software implementation experience. However, the emphasis was on finding appropriate workarounds to cope with them and offer a good solution within the strict time limits.

The biggest challenge was gaining access to the data which caused a major delay in the implementation phase.Also, due to some of the new software tools that had to be learned, the actual building of dashboards was tiresome.

Another challenge was selection of data to plot correlations. While it made sense with the SIMD to use the 2016 and 2020 data, however, the STEM data is available for academic years i.e., 2015–2016, 2016–2017, 2019–2020, and 2020–2021. Some data for the year 2016 will be available in the academic year 2015–2016 and 2016–2017, the same for year 2020. Choosing STEM enrolment data was unclear in order to look for correlations.

As discussed in section 2.1.3, the information is collected annually and tends to be available in January for the preceding academic year. So, in January this year i.e., 2022 the data was available for academic year 2020-21. An assumption was made to select January 2016 and January 2020 STEM enrolments for correlation plots.

4. Testing and Evaluation

Despite the fact that this project's nature prevented the use of the conventional testing and evaluation techniques, a developmental testing approach was implemented. The dashboard was tested at the time building to ensure:

i. Choosing the appropriate charts to depict various dashboards. For example, line charts and area charts are used to illustrate change in one or more quantities by charting a series of data points across time. While area charts connect data points with line segments, stack variables on top of one another, and use colour to differentiate between variables, line graphs use lines to show these changes. Whereas in stacked bar charts each section of these graphs represents a component of the whole. They offer an easy method for compiling data and evaluating the relative sizes of various components.
In order to visualise patterns, including the overall trend for enrolments over the academic years and the sum of enrolments for all UHI colleges throughout the academic years, a line chart was used.
Initially, a stacked bar was selected to demonstrate the subject dispersion for the colleges. However, as can be seen in. Figure 33, the bars were too compressed, were challenging to read and looked cluttered.



Figure 33: Stacked bar for STEM subject dispersion

The area charts I tested next provided a much better visualisation, as can be seen in Figure 34



Figure 34: Area chart showing STEM subject dispersion

ii. To ensure that the numbers, or enrolments, displayed in the dashboards were an accurate depiction of the STEM proprietary data, they were carefully cross verified in Excel.

Throughout the implementation process, there were regular/weekly meetings with the placement supervisor and fortnightly with the HISP team to present progress and get input to ensure the dashboard's clarity. Before presenting the visualisation to the complete team, two members of the HISP STEM team examined it toward the project's final stage. This made it easier to get feedback and make any additional improvements that were required.

Given the nature of the project, the dashboard is solely provided for demonstration and validation purposes; rigorous usability testing of the dashboard is not necessary at this stage.

Compliance to functional and non-functional requirements can be seen in Table 3 and Table 4.

Requirement no.	Requirement	MoSCoW Analysis (Priority ranking)	Compliant (Yes/No)
1	What information is available and what it reveals must be determined.	Must	Yes
2	The data must be specific to the Highlands and Islands, Moray and Perthshire region.	Must	Yes
3	The data must be specific to STEM subjects and STEM student enrolments.	Must	Yes

4	The investigation should be able to identify gaps in data	Should	Yes
5	The data visualisation should have a geographic map	Should	Yes
6	Higher Education STEM data could be helpful	Could	No
7	HISP team would like statistics that show preferred pathways	Would	No

Table 3: Compliance to functional requirements

Requirement no.	Requirement	MoSCoW Analysis (Priority ranking)	Compliant (Yes/No)
1	The dashboard must be made accessible to all its users	Must	Yes
2	The information provided by the dashboard should be accurate.	Must	Yes
3	Whether the dashboard can be accessed to write or read data.	Must	Yes
4	The capacity to get back the dashboard to a working state in the event that something goes wrong.	Should	Yes

Table 4: Compliance to non functional requirements

5. Conclusion

The work completed on the project is summarised in this chapter, which completes the project report. It also considers some of the lessons learned along the course of the project. It then goes through the ethical principles followed and makes suggestions for how HISP dashboard might be improved.

5.1 Summary and Achievements

This objective of this phase was to thoroughly record the design, implementation, testing, and evaluation of the HISP data dashboard.

The project started with a discussion on the findings from investigation phase with a focus on the design and implementation needs as agreed upon with the HISP team. Data acquisition, data preparation, data visualisation and analysis were all incorporated during the design stages.

The implementation phase looked at how the correlations between the SIMD and the STEM data were plotted using R language commands. We then looked at creating the data dashboards using Tableau desktop. The various chart formats and colour schemes used were thoroughly described. Each dashboard included a detailed explanation as well as a trend analysis.

The next section examined the dashboards' testing and evaluation. The evaluation method included regular feedback from the HISP team and a thorough description of how developmental testing was used. All necessary requirements, including functional and non-functional, were examined for compliance or non-compliance.

5.2 Reflections

Every such project ends, perhaps, with nostalgic views about the process and the lessons learned from it. At the end of the experience, the following lessons were learned:

5.2.1 A good preparation is essential

The preliminary framework of the investigation report, which took into account the time invested in conducting meaningful research into the problem area, showed good judgement in the pre-selection of the crucial IT tools and methodologies to carry out the project's objectives for the following phase. That was a good mental preparation technique for the project's more challenging elements.

5.2.2 Regular supervision meetings

This project's success can be attributed to the great management of its constituents. I was encouraged to enhance my software development, writing, and research skills during weekly meetings that monitored progress and helped the project's goals be accomplished within the stipulated time frame.

5.2.3 Adaptability to overcome hurdles

Overcoming some of the difficulties encountered during the implementation process required flexibility to respond to unforeseen circumstances. Due to some of the new abilities that had to be learned, the actual building of dashboards was tiresome. The biggest obstacle was gaining access to the information sources, and waiting for responses to my queries made me feel frustrated and bewildered. This caused a major delay in my implementation phase.However,this encounter boosted my overall confidence and I utilised this waiting time in learning the tools

5.3 Review of Ethical, Social, Legal, Commercial and Professional Issues

During the investigation stage of this project, ethical and related issues were taken into account. But it is crucial to demonstrate at the project's conclusion that it was carried out in accordance with the ethical standards originally thought of.

From a legal aspect there are no issues that have been identified. In order to be in compliance with the laws for intellectual property (GDPR) ("Information Commissioner's Office (ICO)" 2022) and with the best practises for publicly available material, all sources of information, including any photographs, books, articles, datasets, and so forth, have been clearly cited and the authors have been given the appropriate credit.

From an ethical point of view, no individual is directly involved in the creation of this work, which is morally justifiable. No personal information is needed. All the data is publicly available or has been made anonymous. By conducting this project, I will not be adding any new ethical concerns.

There are no social issues identified at this given time. Instead, it can be viewed as a direct contribution to the research that attempts to explore ways to make STEM education more accessible to even more students so they can develop their skills.

All the open-source third-party software libraries utilised for the application's development have been duly acknowledged. No official licensing was required. The HISP project team has been advised to use Tableau Reader, a free tool, in order to make the project accessible; but, if this is to proceed, they will need to purchase a Tableau Licence as they do not currently possess one.

From a professional perspective, this project is undertaken to fulfil two requirements, my MSc in Data Science programme and as a summary of the work carried out during the Highlands and Islands STEM data project as a summer internship for the University of Highlands and Islands. The research was conducted in a professional way, making sure all software is legal and that the results are as useful as possible to the HISP STEM project team. This report can be used to determine future directions in data collection and sharing, access, and improved planning for the promotion of STEM professions.

5.4 Improvements and Future Work

Although the overall requirements for HISP data dashboard were fairly implemented in relation to the goals outlined in the project brief, there is still potential for improvement, making it far from a perfect system.

- SDS, HESA and SFC were not able to be put to use together for this project because, despite the fact that each of them, in its own way provides very insightful information, there is no commonality between them that could be used to gain insights. These could be utilised in future projects if common grounds can be identified.
- UHI North Highland and UHI West Highland data can be added once available.
- HE data once available can be used in a similar way to gain insights.
- At this point in time there are not enough data points to perform predictive analysis, hence no machine learning algorithms could be used. This could be added to the future scope of work.
- Effect of COVID-19 on the enrolments could be identified.
- User testing/usability – Although developmental testing has been carried out when building the dashboards, user testing was not performed.

To summarise we can say that this is a exploratory Data Science project and the outcome of this project can be used to determine future directions in data collection and sharing, access, and improved planning for the promotion of STEM professions.

6. References

- Ali, J., 2022. MSc. Project Investigation Report. Robert Gordon University.
- BOCK, T., 2022. *What are Small Multiples?*. [online]. Displayr. Available from: https://www.displayr.com/what-are-small-multiples/ [Accessed 22 August 2022].
- British Education Research Association Journals,2022. [online]. Available from: https://berajournals.onlinelibrary.wiley.com/doi/full/10.1111/bjet.12849?saml_referrer [Accessed 18/06/2022].
- Brockman, A., 2022. *8 Features that Set Tableau Software Apart from the Rest*. [online]. Blog.csgsolutions.com. Available from: https://blog.csgsolutions.com/8-features-that-set-tableau-software-apart-from-the-rest [Accessed 21 August 2022].
- Busch, L. (2017). *Knowledge for sale: The neoliberal takeover of higher education*. London, UK: MIT Press.
- CEIAS,2022. [online]. Available from: https://www.ceias.nau.edu/capstone/projects/CS/2014/TheAviators/assets/TheAviators_Requirements.pdf [Accessed 26/07/2022].
- *Courses -UHI*, 2022. [online]. Available from: https://www.uhi.ac.uk/en/courses/ [Accessed 02/06/ 2022].
- EDUCATION, IBM., 2022. *What is Data Visualization?*. [online]. Ibm.com. Available from: https://www.ibm.com/cloud/learn/data-visualization [Accessed 01/07/2022].
- Espeland, W. N., & Sauder, M. (2016). *Engines of anxiety: Academic rankings, reputation, and accountability*.New York, NY: Russell Sage Foundation.
- Gartner, 2022. [online]. Available from: https://www.gartner.com/doc/reprints?id=1-2955ETOT&ct=220215&st=sb. [Accessed 25/06/2022].
- Gulson, K., & Sellar, S. (2018). Emerging data infrastructures and the new topologies of education policy. *Environment and Planning D: Society and Space*, *37*(2), 350–366. https://doi.org/10.1177/0263775818813144.
- *HESA - Experts in higher education data and analysis*, 2022. [online]. Available from: https://www.hesa.ac.uk/ [Accessed 10/06/2022].
- HESA,2022. [online]. Available from: https://www.hesa.ac.uk/files/HESA_Corporate_Strategy_2016-2021.pdf [Accessed 18/07/2022 August 2022].
- *How Much Data Is Required for Machine Learning? - PostIndustria*, 2022. [online]. Available from: https://postindustria.com/how-much-data-is-required-for-machine-learning/ [Accessed 28/07/2022].
- *Information Commissioner's Office (ICO)*, 2022. [online]. Available from: https://ico.org.uk [Accessed 26/05/2022].
- KPMG. (2015). *The blueprint for a new HE data landscape: Final report*. Author.
- Manasson, A., 2019. *CRISP-DM: Manage your Data Science Projects*. [online] Towards Data Science. Available from: https://towardsdatascience.com/why-using-crisp-dm-will-make-you-a-better-data-scientist-66efe5b72686 [Accessed 27 July 2022].
- Microsoft, 2022. *What is a data dashboard*. [online]. United Kingdom: Microsoft. Available from: https://powerbi.microsoft.com/en-us/data-dashboards [Accessed 25/05/2022].
- Microsoft,2022. [online]. Available from: https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview. [Accessed 20/06/2022].
- Muller, J. Z. (2018). *The tyranny of metrics*. Oxford, UK: Princeton University Press.
- Npfs, 2022. [online]. United Kingdom. Available from: https://www.npfs.org.uk/wp-content/uploads/edd/2020/06/NPFS_STEM.pdf [Accessed 02/06/2022].
- Peck, J., & Theodore, N. (2015). *Fast policy: Experimental statecraft at the thresholds of neoliberalism*. London, UK: University of Minnesota Press.
- *Power BI Desktop—Interactive Reports | Microsoft Power BI*, 2022. [online]. Available from: https://powerbi.microsoft.com/en-us/desktop/#:~:text=The%20new%20AI%20capabilities—pioneered,drive%20more%20strategic%20business%20outcomes. [Accessed 16/07/2022].

- ScaleFresh. 2022. *Cheatsheet for Charts: Pair the Right Chart with the Right Set of Data*. [online] Available at: https://scalefresh.com/cheatsheet-for-charts/ [Accessed 20/07/2022].
- *Scottish Employer Skills Survey 2020*, 2022. [online]. Available from: https://www.gov.scot/publications/scottish-employer-skills-survey-2020-2/pages/5/ [Accessed 10/06/2022].
- *Scottish Funding Council home page*, 2022. [online]. Available from: https://www.sfc.ac.uk/ [Accessed 15/06/2022].
- *Scottish Index of Multiple Deprivation 2020 - gov.scot*, 2022. [online]. Available from: https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/ [Accessed 31/06/ 2022].
- *Shiny*, 2022. [online]. Available from: https://shiny.rstudio.com [Accessed 25/06/2022].
- Slota, S. C., & Bowker, G. C. (2017). How infrastructures matter. In U. Felt, R. Fouche, C. A. Miller, & L. Smith-Doerr (Eds.), *The handbook of science and technology studies* (4th ed., pp. 529–554). London, UK: MIT Press.
- *STEM graduates wanted for 50 fully paid work placements*, 2022. [online]. Available from: https://www.glasgowworld.com/business/stem-graduates-wanted-for-50-fully-paid-work-placements-3559923 [Accessed 02/06/2022].
- *Supporting science, technologies, engineering and mathematics (STEM) at home | Learning at home | Parent Zone*, 2022. [online]. Available from: https://education.gov.scot/parentzone/learning-at-home/supporting-science-at-home/ [Accessed 02/06/2022].
- Tableau, 2022. *What is data visualization*. [online]. Seattle: Tableau. Available from: https://www.tableau.com/en-gb/learn/articles/data-visualization [Accessed 26/05/2022].
- Tableau 2022. *Maps.* [online]. Seattle: Tableau. Available from: https://www.tableau.com/solutions/maps [Accessed 21 August 2022].
- *Top 10 Map Types in Data Visualization*, 2022. [online]. Available from: https://towardsdatascience.com/top-10-map-types-in-data-visualization-b3a80898ea70 [Accessed 20/07/2022].
- *UHI*, 2022. [online]. Available from: https://www.uhi.ac.uk/en/ [Accessed 02/06/2022].
- *Visualizations That Really Work*, 2022. [online]. Available from: https://hbr.org/2016/06/visualizations-that-really-work [Accessed 01/07/2022].
- *What is CRISP DM? - Data Science Process Alliance*, 2022. [online]. Available from: https://www.datascience-pm.com/crisp-dm-2/ [Accessed 25/07/2022].
- *What do HE students study? | HESA*, 2022. [online]. Available from: https://www.hesa.ac.uk/data-and-analysis/students/what-study [Accessed 10/06/2022].
- *Who's studying in HE? | HESA*, 2022. [online]. Available from: https://www.hesa.ac.uk/data-and-analysis/students/whos-in-he [Accessed 10/06/2022].

Appendix A: Project Plan



Appendix B: Correlation Plots

```
# Correlation plot Income rate vs enrolments
```{r}
library(ggpubr)
ggplot(Enrol,aes(x = Income_rate, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Income rate vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot Income count vs enrolments
```{r}
library(ggpubr)
ggplot(Enrol,aes(x = Income_count, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Income Count vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot Working age population vs enrolments
```{r}
ggplot(Enrol,aes(x = Working_age_population, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Working Age Population vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot employment rate vs enrolments
```{r}
ggplot(Enrol,aes(x = Employment_rate, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Employment Rate vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot Employment count vs enrolments
```{r}
ggplot(Enrol,aes(x = Employment_count, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Employment Count vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot Total population vs enrolments
```{r}
ggplot(Enrol,aes(x = Total_population, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Total Population vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



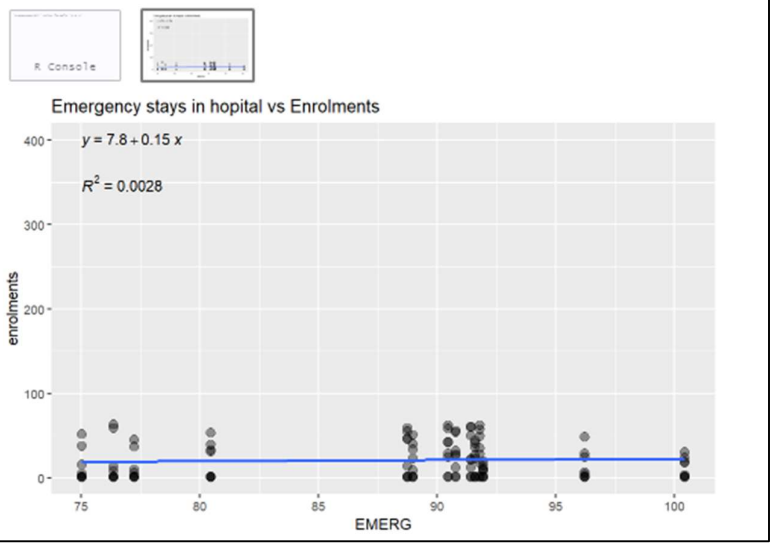
```r
# Correlation plot CIF vs enrolments
```{r}
ggplot(Enrol,aes(x = CIF, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("CIF vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```

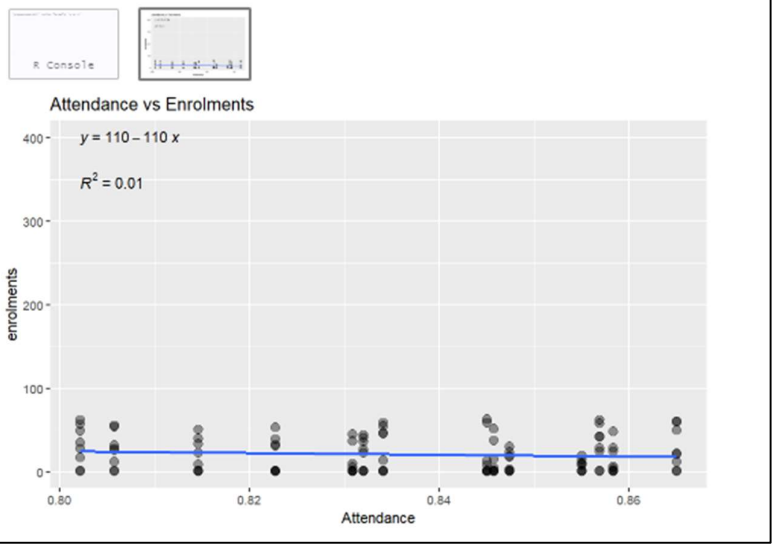
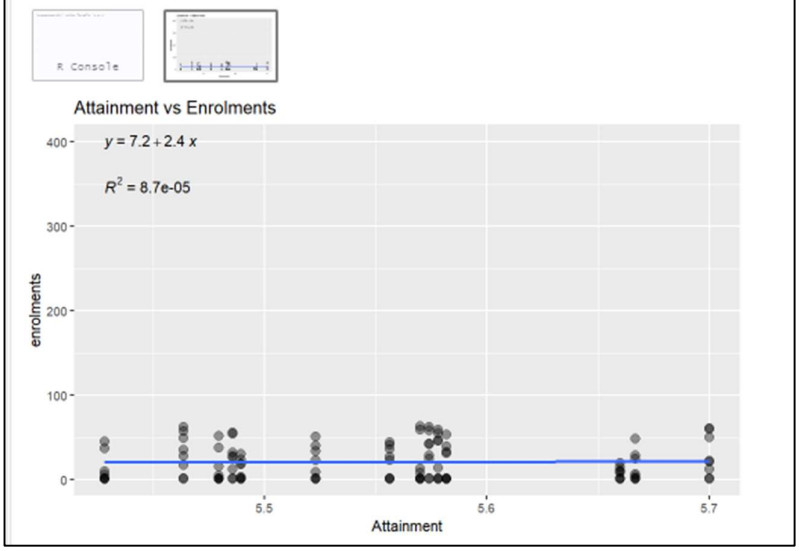
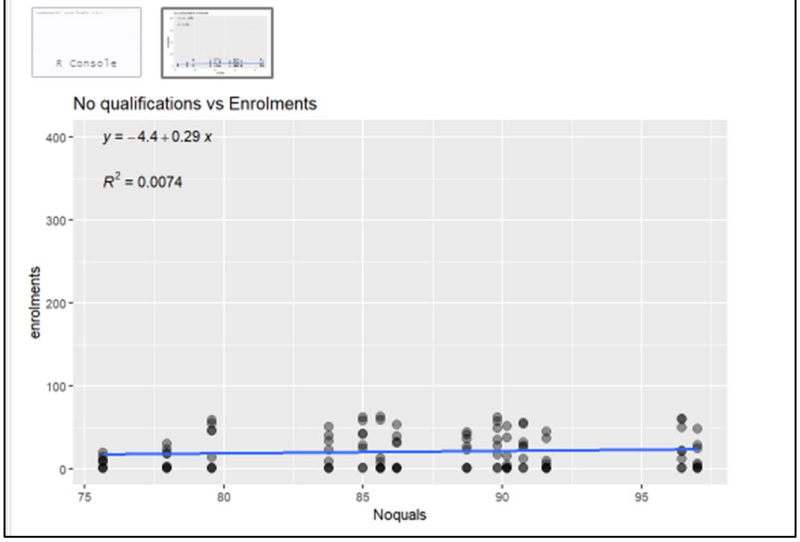


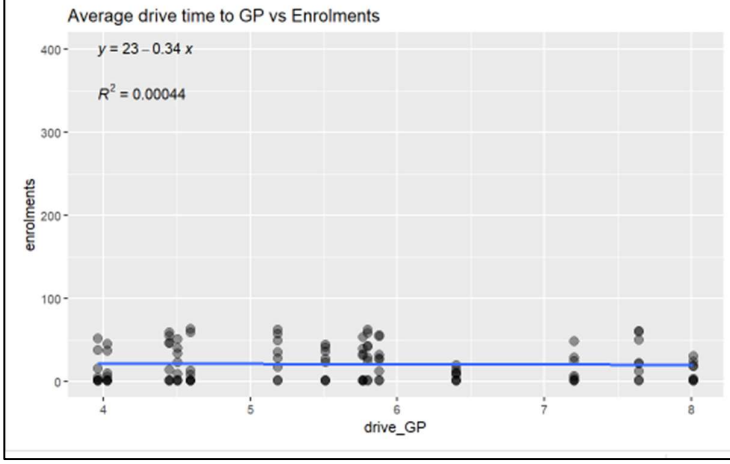

```r
# Correlation plot Alcohol vs enrolments
```{r}
ggplot(Enrol,aes(x = ALCOHOL, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Alcohol vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot drug vs enrolments
```{r}
ggplot(Enrol,aes(x = DRUG, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("DRUG vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot SMR vs enrolments
```{r}
ggplot(Enrol,aes(x = SMR, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("SMR vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot Depression vs enrolments
```{r}
ggplot(Enrol,aes(x = DEPRESS, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("DEPRESS vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot LBWT vs enrolments
```{r}
ggplot(Enrol,aes(x = LBWT, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("LBWT vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



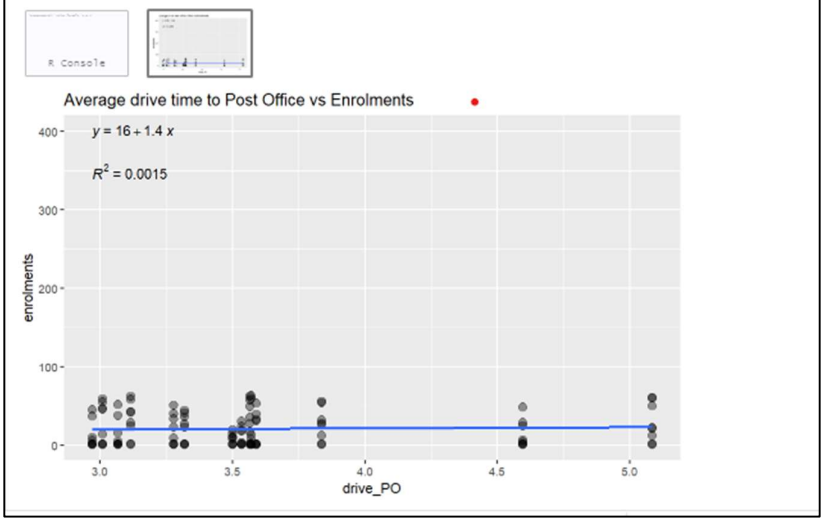
```r
# Correlation plot Emerg vs enrolments
```{r}
ggplot(Enrol,aes(x = EMERG, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Emergency stays in hopital vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



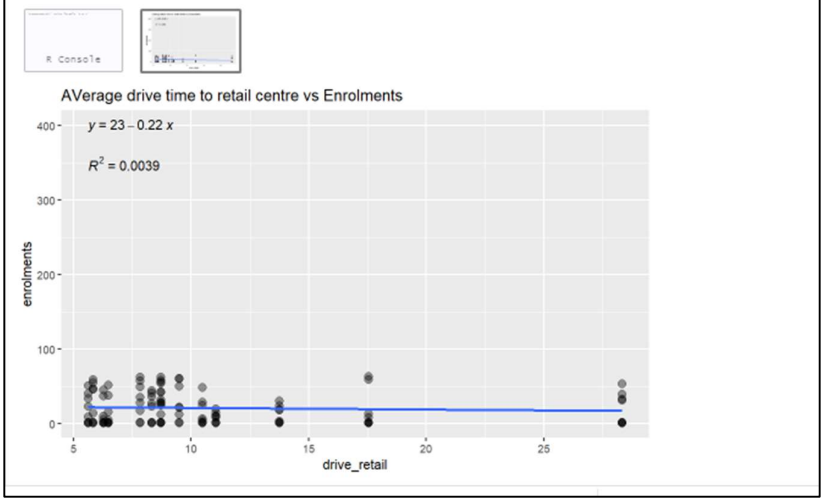

```r
# Correlation plot attendance vs enrolments
```{r}
ggplot(Enrol,aes(x = Attendance, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Attendance vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot attainment vs enrolments
```{r}
ggplot(Enrol,aes(x = Attainment, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Attainment vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot no qualifications vs enrolments
```{r}
ggplot(Enrol,aes(x = Noquals, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("No qualifications vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot drive_petrol vs enrolments
```{r}
ggplot(Enrol,aes(x = drive_petrol, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Average drive time to Petrol station vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```


```
# Correlation plot drive to GP vs enrolments
```{r}
ggplot(Enrol,aes(x = drive_GP, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Average drive time to GP vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



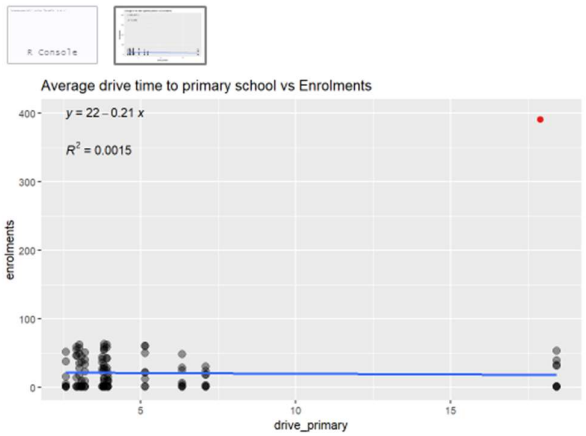
```r
# Correlation plot drive to PO vs enrolments
```{r}
ggplot(Enrol,aes(x = drive_PO, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Average drive time to Post Office vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```

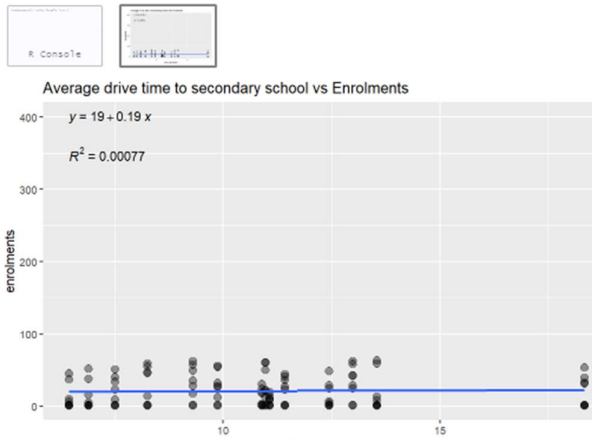
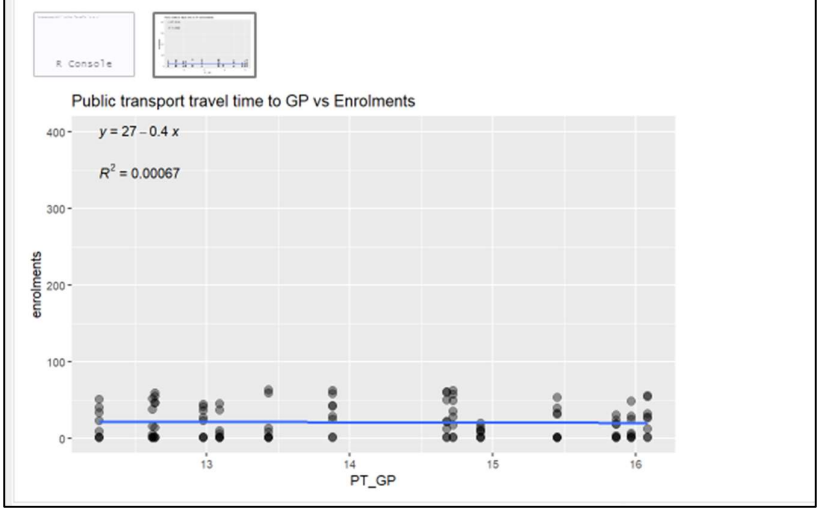
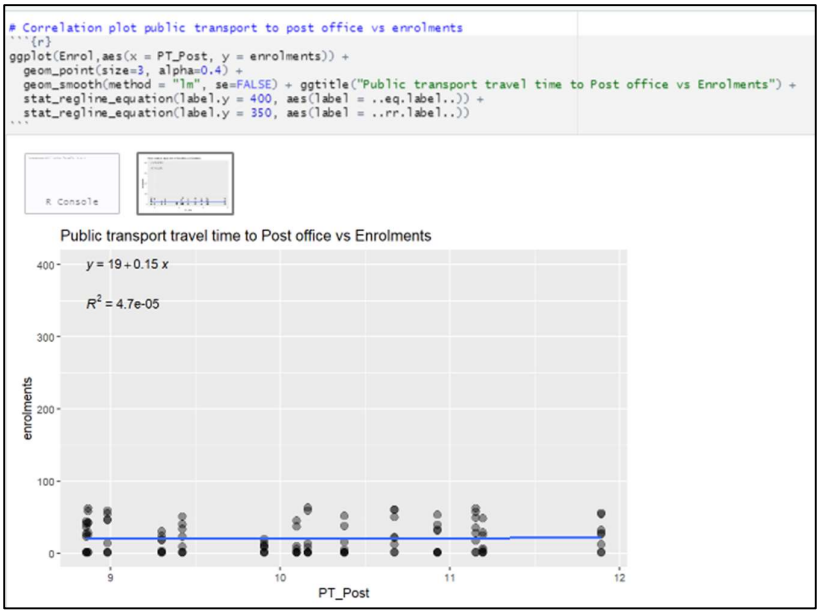
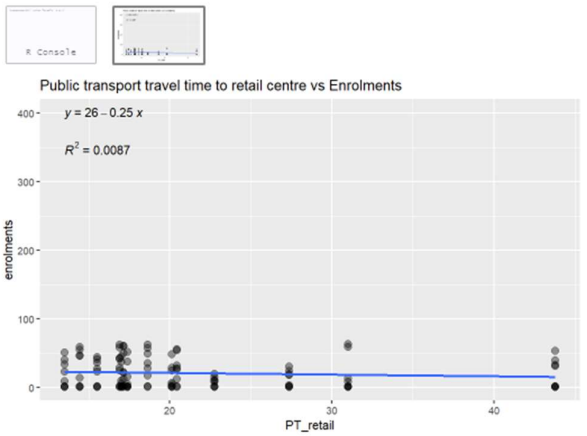
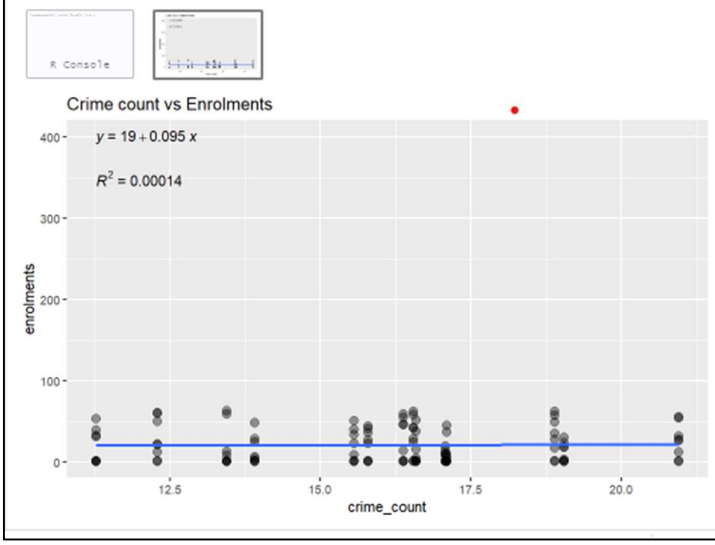
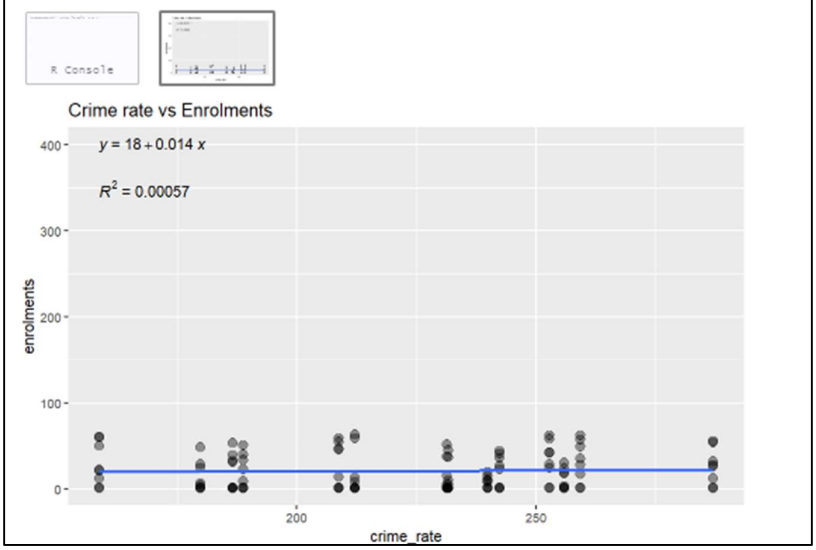
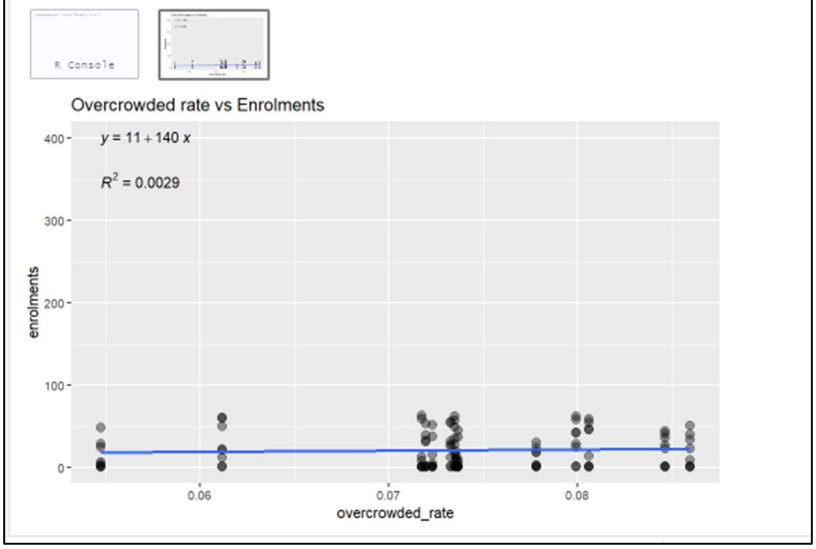
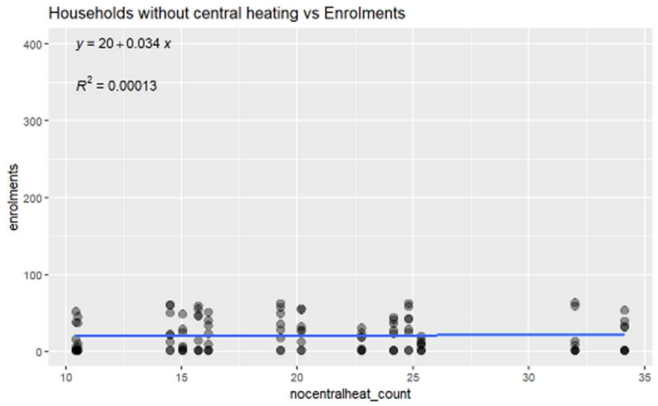
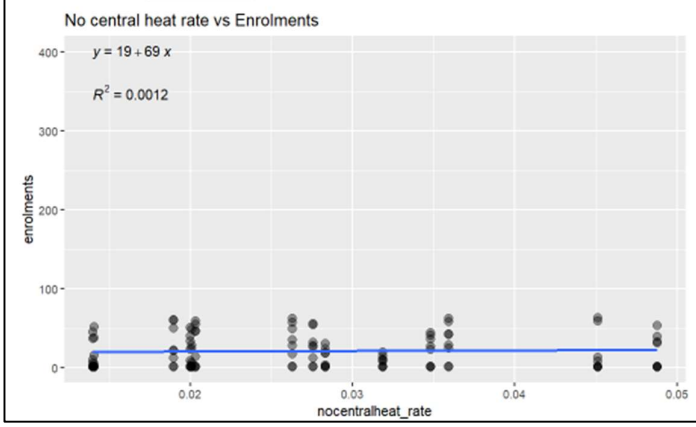



```r
# Correlation plot drive to retail vs enrolments
```{r}
ggplot(Enrol,aes(x = drive_retail, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("AVerage drive time to retail centre vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```r
# Correlation plot drive to primary school vs enrolments
```{r}
ggplot(Enrol,aes(x = drive_primary, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Average drive time to primary school vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```r
# Correlation plot drive to secondary school vs enrolments
```{r}
ggplot(Enrol,aes(x = drive_secondary, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Average drive time to secondary school vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot public transport to GP vs enrolments
```{r}
ggplot(Enrol,aes(x = PT_GP, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Public transport travel time to GP vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot public transport to post office vs enrolments
```{r}
ggplot(Enrol,aes(x = PT_Post, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Public transport travel time to Post office vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot public transport to retail vs enrolments
```{r}
ggplot(Enrol,aes(x = PT_retail, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Public transport travel time to retail centre vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot crime count vs enrolments
```{r}
ggplot(Enrol,aes(x = crime_count, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Crime count vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot crime rate vs enrolments
```{r}
ggplot(Enrol,aes(x = crime_rate, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Crime rate vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot overcrowded rate vs enrolments
```{r}
ggplot(Enrol,aes(x = overcrowded_rate, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Overcrowded rate vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot no central heating count vs enrolments
```{r}
ggplot(Enrol,aes(x = nocentralheat_count, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Households without central heating vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```




```
# Correlation plot no central heating rate vs enrolments
```{r}
ggplot(Enrol,aes(x = nocentralheat_rate, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("No central heat rate vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```



```
# Correlation plot overcrowded count vs enrolments
```{r}
ggplot(Enrol,aes(x = overcrowded_count, y = enrolments)) +
  geom_point(size=3, alpha=0.4) +
  geom_smooth(method = "lm", se=FALSE) + ggtitle("Overcrowded households vs Enrolments") +
  stat_regline_equation(label.y = 400, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 350, aes(label = ..rr.label..))
```
```