Investigation into Factors influencing Smoking and General Health in Ireland

Kaden Armstrong

N00222605

Supervisor: Cyril Connolly

Second Reader: John Montayne

Year 4 2025/26

DL836 BSc (Hons) in Creative Computing

Abstract

This project explore smoking prevalence and health outcomes across Ireland through the use of spatial data analysis and data visualisation techniques. The project was developed using spatial datasets along with tools like R programming language and the Rshiny framework. Interactive maps, charts, and statistical models were created to examine geospatial patterns and relationships between smoking behaviour and socioeconomic factors.

Acknowledgments

I would like to express my sincere gratitude to supervisor, Cyril Connelly, for his guidance, support, and feedback throughout this project. Cyril's expertise in data visualisation, including Tableau and R, and advice and feedback helped this project succeed.

I would also like to thank my lecturers for their assistance during my time at IADT, as well for the knowledge and skills they have helped me develop throughout my studies.

Lastly, I would like to thank my family and friends for their continuous support and encouragement during the last four years, helping me stay motivated and focused throughout my studies.

The incorporation of material without formal and proper acknowledgement (even with no deliberate intent to cheat) can constitute plagiarism.

If you have received significant help with a solution from one or more colleagues, you should document this in your submitted work and if you have any doubt as to what level of discussion/collaboration is acceptable, you should consult your lecturer or the Course Director.

WARNING: Take care when discarding program listings lest they be copied by someone else, which may well bring you under suspicion. Do not to leave copies of your own files on a hard disk where they can be accessed by other. Be aware that removable media, used to transfer work, may also be removed and/or copied by others if left unattended.

Plagiarism is considered to be an act of fraudulence and an offence against Institute discipline.

Alleged plagiarism will be investigated and dealt with appropriately by the Institute. Please refer to the Institute Handbook for further details of penalties.

The following is an extract from the B.Sc. in Creative Computing (Hons) course handbook. Please read carefully and sign the declaration below

Collusion may be defined as more than one person working on an individual assessment. This would include jointly developed solutions as well as one individual giving a solution to another who then makes some changes and hands it up as their own work.

Declaration:

I am aware of the Institute's policy on plagiarism and certify that my thesis is my own work.

Signed: Kaden Armstrong

Failure to complete and submit this form may lead to an investigation into your work.

Contents

1 Introduction and Project Context

1.1 Introduction

Smoking is one of the most significant preventable causes of illness and premature death worldwide, creating major challenges for healthcare systems and policymakers. From cardiovascular disease to cancer, tobacco use has been strongly associated with a range of serious health conditions. Despite increased awareness of these risks, smoking continues to affect a major proportion of the global population.

According to the World Health Organization (WHO), approximately 1.2 billion people used tobacco in 2024. In addition, the use of electronic cigarettes has grown, with an estimated 86 million adults and 15 million adolescents ages 13-15 reported worldwide. The WHO also identified Europe as the region with the highest prevalence of adult tobacco usage globally, with 24.1% of adults using tobacco in 2024 (World Health Organization, 6 October 2025).

A number of strategies have been shown to reduce smoking prevalence and improve public health outcomes. These include proper surveillance to monitor tobacco use, bans on tobacco advertising, promotion and sponsorship, and policies to increase tax on tobacco products. Among these measures, tobacco taxation has consistently been one of the most effective measures for reducing tobacco consumption, particularly with younger populations (World Health Organization, 25 June 2025).

Given the health burden associated with smoking, there is significant value in analysing smoking patterns at regional and local levels. This project focuses on the visualisation of smoking statistics in Ireland through the use of spatial data analysis and interactive dashboards. By presenting smoking data geographically, the project aims to improve understanding of regional trends and support evidence-based discussion surrounding public health.

1.2 Project Aim & Objectives

The overall aim of this project is to develop an interactive R Shiny dashboard that visualises smoking and health statistics across Ireland through the use of spatial data analysis. The dashboard was designed to present complex public health data in a clear and accessible format, allowing users to explore regional patters, compare areas, and gain a better understanding of factors influencing smoking and health outcomes.

By combining geographic visualisation techniques with statistical analysis, this project aims to improve the interpretation of data relating to health and smoking.

1.2.1 Specific Objectives

1. Identify, collect and prepare relevant Irish datasets relating to smoking, health, population, and geographic boundaries for use within the dashboard.
2. To develop interactive Leaflet maps that display smoking rates across different geographic areas, including Small Areas and Local Electoral Areas, using colour-coded boundaries and legends.
3. To create additional visualisations, including GeoFacet charts and data tables, to present health statistics in a clear and comparative format across Irish counties.
4. To implement regression modelling, correlation analysis, and supporting charts to explore relationships between smoking prevalence and selected socioeconomic or health variables.
5. To integrate all visualisations into a user-friendly R Shiny dashboard that enables users to navigate between maps, charts and tables using tabs.

1.3 Success Criteria & Scope

1.3.1 Success Criteria

The success of this project was measured against functionality and usability criteria. These metrics were used to evaluate whether the final dashboard met its intended aims and objectives.

1. Successful development of a working R Shiny dashboard in which all maps, charts and tables displayed correctly.
2. Users were easily able to switch between dashboard tabs, allowing access to maps, GeoFacet visualisations, data tables, and regression analysis output without errors.
3. Leaflet maps were successfully updated when different geographic layers were selected.
4. Statistical datasets were correctly joined with geographic boundary files, ensuring smoking and health data were accurately represented within mapped regions.
5. Visualisations used colour scales, appropriate legends and labels to ensure complex health and smoking data could be interpreted easily by users.
6. The dashboard successfully incorporates regression modelling, correlation plots, and comparative charts to support deeper analysis pf smoking and health trends.
7. The completed application ran successfully within a local R Studio environment without major performance or rendering issues.

1.3.2 Scope

The scope of this project focused on the design and development of a locally hosted interactive dashboard for analysing smoking and health statistics in Ireland. The project included data collection, preprocessing spatial mapping, statistical modelling, and visualisation within the R Shiny environment.

The dashboard was designed to be analytical rather than a large scale commercial system. As a result, the project scope did not include any live real-time data or online deployment. Instead, it focuses on producing an accurate, functional, and user-friendly tool capable of demonstrating how spatial data analysis can be applied to Irish health data.

2 Research and Background

2.1 Literature Review

2.1.1 Visualisation techniques for health data

Visualisation techniques play a crucial role in making huge and complex health and medical data accessible and interpretable for clinicians, researchers, doctors and policy makers. In healthcare, data can vary widely in type and scale, from individual patient records to population-level census data, effective visualisations can help reveal patterns, trends and anomalies that numbers alone struggle to convey. Given the complexity and scale of modern healthcare datasets, selecting appropriate visualisation methods is essential for revealing patterns, trends, and relationships that inform decision-making (Ofori et al., 2025).

Basic visualisation techniques remain the foundation in healthcare analytics due to their simplicity, clarity and familiarity to everyone, not just those in data fields. In their scoping review of public health visual communication, Ofori et al. (2025) identified 25 different visualisation techniques used across health-related studies. These included common charts and graphs such as bar charts, line charts, pie charts, bubble charts, box plots, and scatter plots, which are frequently used to display trends, comparisons, and relationships in univariate and bivariate data. Additionally, map-based visualisations such as choropleth maps, hotspot maps, and heatmaps were widely used to represent spatial patterns and geographic differences in health outcomes.

Beyond these better-known forms, the review also highlights the use of more specialised visualisations, including sunburst diagrams, alluvial plots, upset plots, and circos diagrams, being useful for exploring multidimensional and complex health datasets. These visualisations were implemented using a variety of programming languages, statistical tools and platforms, including R, Python, Power BI, Tableau, ArcGIS, and custom web-based applications (Ofori et al., 2025). Altogether, these techniques support key analytical tasks like improving data accessibility, identifying patterns and trends, examining similarities between variables, and exploring multidimensional health data.

In addition to documenting techniques, broader reviews of healthcare visualisations highlight the key role of visualisation analysis for making sense of complex, multidimensional healthcare data. Large surveys highlight that visual analytic approaches, which combine visual encoding with algorithmic methods, help reduce data complexity and uncover hidden associations, enhancing healthcare professionals' ability to interpret and act on diverse types of clinical and population data (Tan et al., 2024). These combined perspectives demonstrate that basic visualisation techniques are crucial to achieving efficiency, insight, and understanding in healthcare analytics.

2.2 Similar applications

1. U.S. Centers for Disease Control (CDC) Places

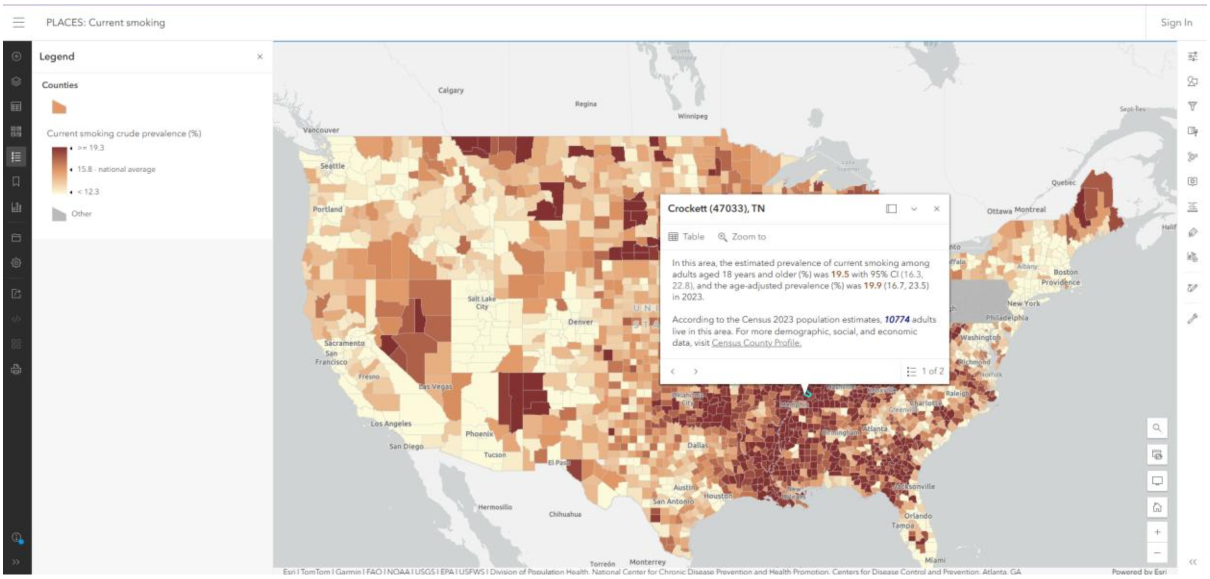

Figure 1 CDC Places Dashboard

Features:

- Ability to add, toggles, and manage multiple geographic data layers for comparative analysis.
- Pop-up feature displaying attribute information when selecting regions.

Pros:

- Highly interactive and visually intuitive interface for exploring geographic data.
- Strong layer management.
- Pop-up attribute to enhance accessibility.

Cons:

- No direct functionality for downloading or exporting datasets for external analysis.

2. GeoHive Map Viewer

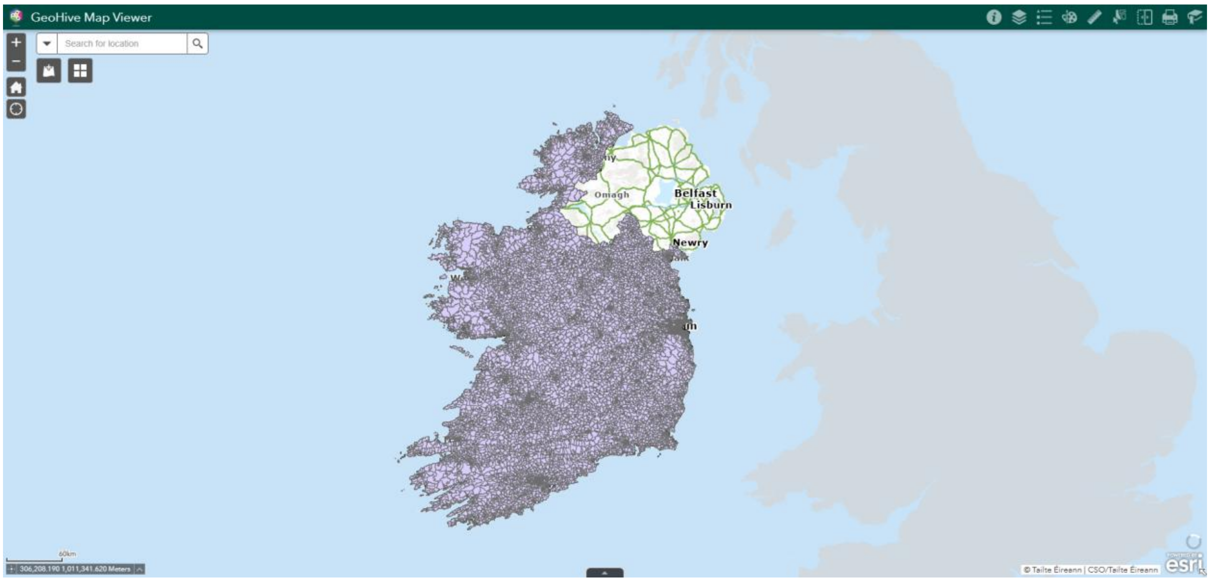

Figure 2 GeoHive Map Viewer

Features:

- Interactive map viewer allowing users search, display and layer multiple datasets within a single interface.
- Ability to visualise both current and past geographic data across a variety of themes.
- Integration of multiple government datasets in a single platform.

Pros:

- Strong dataset coverage across multiple domains, including demographic, environmental, and health data.
- Promotes user engagement, as layers must be manually added.

Cons:

- Interface is less intuitive compared to modern dashboards, particularly for new users.
- Limited analytical functionality beyond spatial visualisations.
- Pop-up windows display numerical values, which may not be meaningful to all users without additional context.

2.3 Technical Research

2.3.1 Datasets

- CSO Census data
- Tailte Eireann MapGenie

CSO Census data was necessary for providing the statistical information used in this project, including smoking rates, health indicators, and other socioeconomic variables. This data was the basis for both the visualisations and analysis. spatial boundary files enabled this data to be mapped geographically, allowing values to be displayed across Small Areas and Local Electoral Areas.

2.3.2 Libraries and packages

- Tidyverse/ dplyr
- Sf
- Leaflet
- Ggplot2
- GeoFacet
- DT
- Ggcorrplot and GGally
- R (lm)
- R Shiny

The Tidyverse and dplyr libraries were necessary for cleaning, transforming, and merging datasets, while the sf package handled spatial data and geographic boundary files. Leaflet supported the creation of interactive maps to display smoking rates, and ggplot2 with GeoFacet was used to produce a GeoFacet grid map to display health data. The DT packages was used to create interactive tables with filters, search, and export functionality. Statistical analysis was carried out using lm() for regression, with Ggcorrplot and GGally for correlation and pairwise visualisations. R Shiny enabled the integration of all visualisations into an interactive dashboard.

2.3.3 Platforms and Applications

- Rstudio
- Tableau

Rstudio was used as the main development environment for coding and data processing and allowed for the creation of interactive maps and charts by downloading libraries and other packages. Tableau was used to test and explore datasets in early development and validated before implementation.

3 Requirements Analysis

3.1 Requirements Gathering

To help design the system, CDC Places and GeoHive were reviewed. Both provided interactive mapping, layer management, and pop-ups to display data which highlights the importance of clear and accessible spatial visualisations. Limitations were identified in these two applications, with lacking export functionality in CDC dashboard and GeoHive having a less intuitive interface.

3.2 Requirements Modelling

3.2.1 Personas

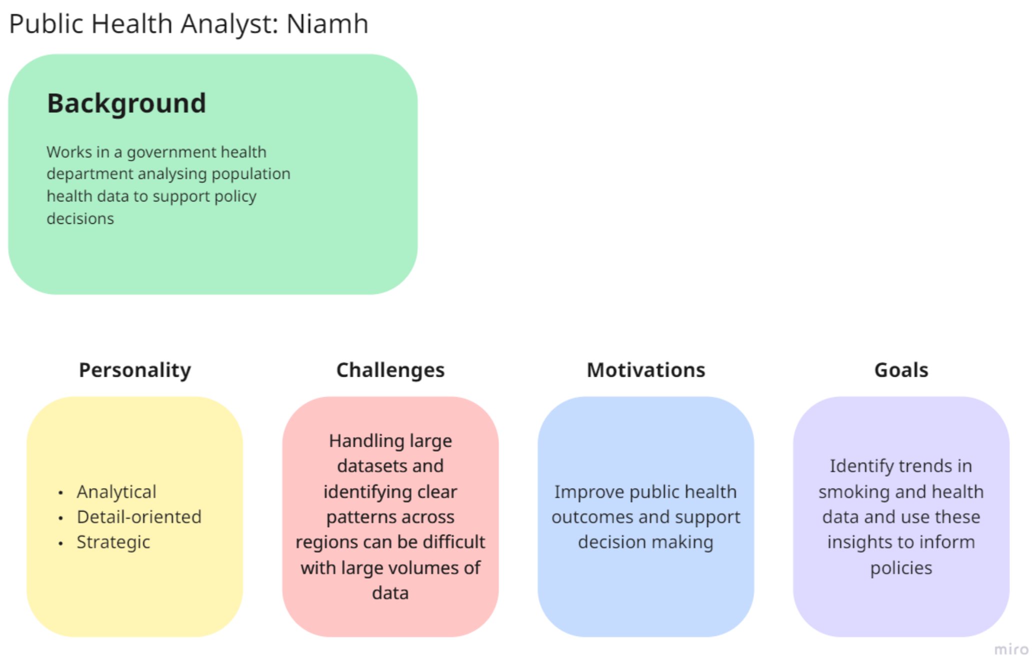

Figure 3 Persona 1

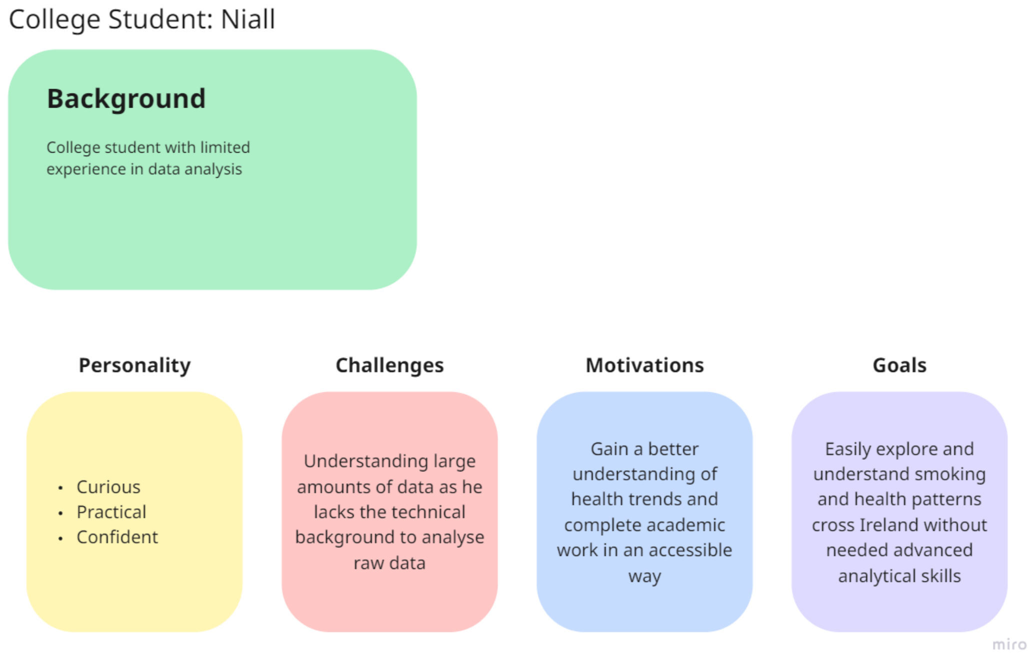

Figure 4 Persona 2

3.2.2 Functional requirements

The dashboard should be able to display smoking and health data through interactive visualisations, allowing users to explore the information across different geographic regions in Ireland. This includes rendering Leaflet maps to show smoking rates in different regions across multiple levels, with colour-coding and legends to help users understand the data being shown.

Users need to be able to interact with the application through controls such as dropdown menus, allowing them to switch between different map layers, and switch to different maps and graphs using tabs. Additional functionality should include viewing GeoFacet grids displaying health data, accessing an interactive data table with search and filtering capabilities, and downloading data in formats like CSV or Excel. The application also includes statistical analysis features, including regression modelling, correlation plots, and pairwise comparisons.

3.2.3 Non-functional requirements

The dashboard should be easy to navigate and understandable for both technical and non-technical users. This involves a clear layout design, tabs to navigate across different pages, legends, labels and colour schemes to improve readability.

4 Design

4.1 System Architecture

The system architecture of the application is based on a reactive web application model implemented using the R Shiny framework. The user interface is responsible for displaying the visual components such as maps, charts, and tables, while the server handles data processing, analysis, and output rendering.

The server loads and processes the datasets ready for output. Spatial data is joined with census data before being visualised using Leaflet and ggplot2.

4.2 Interface Design

The interface was designed using R Shiny with a focus on clarity and ease of use. The dashboard layout uses tabs to help users navigate between maps, charts, data tables, and statistical models and charts without cluttering the interface.

Interactive elements like dropdown menus were used to allow users to switch between different map layers, including Small Area and Local Electoral Area views. Leaflet maps were used as the main spatial visualisation tool, with colour coded polygons and legends to represent smoking rates across different regions.

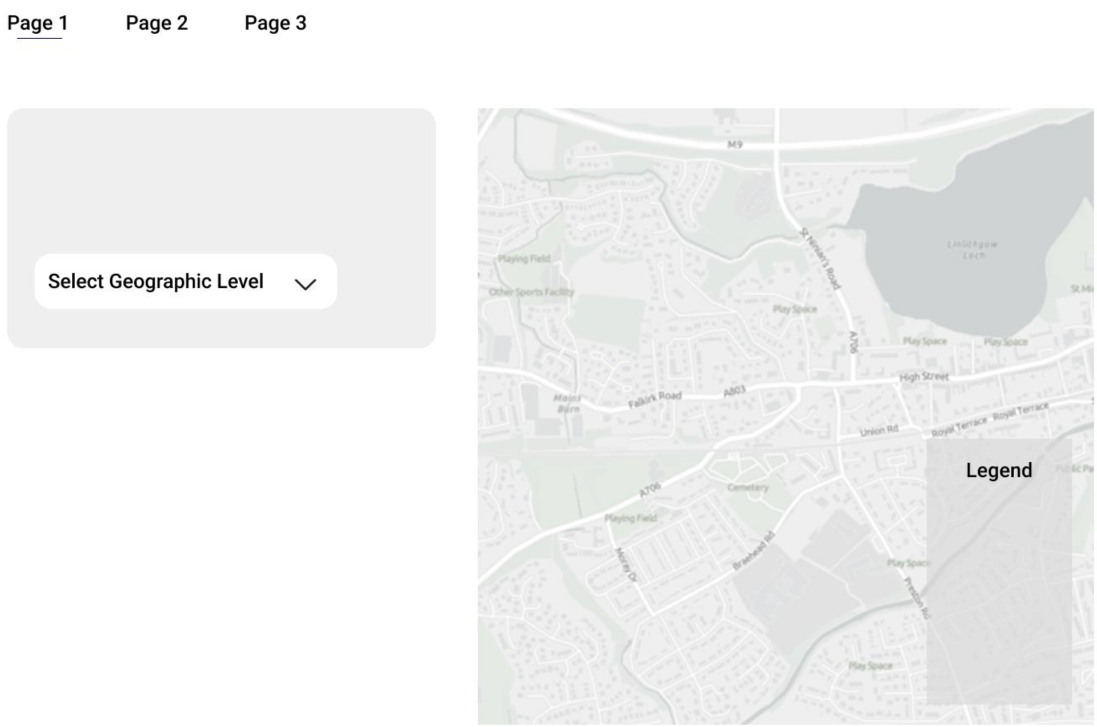

Figure 5 Early Prototype Design

4.3 Process Design

Multiple libraries were used and each had different functionality. The Leaflet library was used for interactive spatial mapping, while sf handled spatial data processing and geographic boundary files. Data manipulation was carried out using dplyr and Tidyverse, which allowed for efficient filtering, joining, and transformation of datasets. The GeoFacet grid was implemented using the GeoFacet and ggplot2 libraries, and DT was used to create an interactive data table of this data and included filtering and exporting.

A linear regression model was implemented using lm(). Ggcorrplot, and ggPairs were used to display a correlation plot and pairwise comparisons of this model.

User inputs triggered reactive updates in the server, which then automatically refreshed the corresponding visual, such as switching map layers. this ensures dynamic dashboard response.

5 Implementation

5.1 Development Process and Methodology

The development process of this project followed an agile sprint methodology, which allowed the application to be developed iteratively. The implementation process was divided into a series of sprints, each focusing on one aspect of the application. This ensured that functionality could be tested throughout development.

The first stage of implementation focused on data acquisition and processing. Irish Census data and shapefiles were collected, cleaned, and structured. This involved standardising variables and transforming raw data into percentages to enable comparison across regions.

Once the datasets were prepared, the census data was joined to shape files using common identifiers. This allowed for spatial visualisation of smoking and health indicators across Ireland. Tableau was also used to ensure valid data and if the datasets were suitable for mapping within R.

Leaflet maps, GeoFacet grids, regression models, and correlation analysis were then implemented in R. Each feature was initially developed and tested independently before being integrated into the R Shiny environment. Refinement of individual components was quicker and less complex.

The last stages of implementation was dashboard integration in R Shiny. This involved taking all previously developed components and combining them into a single interactive application. This also included adding tabs for user navigation and ensuring all visuals updated correctly based on their input.

5.2 Challenges and Solutions

The first major challenge faced during the development of this app was data compatibility. Initially, this project aimed to include Small Area and County level shapefiles. After implementing the County shapefile, the Leaflet map would fail to render the County polygons and prevented other components in the application from loading. The solution for this was to replace the County data with Local Electoral Area shapefiles, which provided similar levels of geographic visualisation to the County level, while also maintaining performance and usability.

Another challenge was managing multiple map layers within the dashboard. Initially, switching between layers caused the entire map to reload, which slowed down the application as each layer being loaded had to be re-rendered. The solution for this was to implement the leafletProxy() function. This allowed each layer of the map to be dynamically shown or hidden depending on which input was selected from the dropdown menu, without requiring the map to be fully reloaded. This significantly improved performance when selecting a new map.

The final major challenge faced during development was the GeoFacet data preparation. The GeoFacet visualisation required county data; however, census data for cities with larger populations was split into multiple administrative regions. Since the GeoFacet grid used in this application can only support one value per county. The solution for this was to manually gather the data from each administrative region within a county and combine them to get accurate county totals. The data then had to be restructured into long format to be properly interpreted by the GeoFacet library. This was necessary to ensure compatibility with the grid layout and produce an accurate visualisation.

5.3 Development Environment Tools

The majority of the development of this project was carried out within the Rstudio environment, and was the main workspace for coding, testing, and debugging the application. Rstudio provided the interface for managing scripts, visualising and analysing data, and was a suitable environment for development of spatial and statistical components.

This project was developed using the R programming language, which is derived from the S programming language. Rshiny was used to build the interactive dashboard and enabled the combination of interface components and server logic, which allowed the application to respond dynamically to user inputs such as dropdown selectors and tab navigations.

During development, visualisations and analytical components were created and tested outside the Shiny environment within Rstudio. This include building Leaflet maps, GeoFacet visualisations, regression models, and correlation plots separately. Once these components were working correctly, they were integrated into the Shiny dashboard to complete the application.

Other tools such as Tableau were used during early stages of this project for feasibility testing of the datasets and spatial visualisations. This validated the data and ensured it was suitable form mapping before implementing in R. Excel was used for preprocessing the GeoFacet and regression model data to ensure compatibility with their respective libraries.

Within the Rstudio environment multiple libraries were used to support development, including libraries from spatial analysis such as Leaflet and sf, data manipulation using dplyr and Tidyverse, statistical modelling through base R, and data visualisation using ggplot2 and GeoFacet.

6 Testing and Evaluation

6.1 Test Plan

6.1.1 Usability Testing

Test Number	Description	Input	Expected Output	Actual Output	Comments
1	Users can navigate between tabs	User selects each tab	Correct content is displayed for each tab	All tabs displayed correctly	Navigation is clear and easy to use
2	Users can switch between map layers	User selects small area or local electoral area from dropdown	Map updates to selected layer with respective boundaries and legend	Map correctly updates with boundaries and legend	Dropdown was easily understood and used
3	Users understand the colour legend	User views selected map and legend	User can identify the colour scale	Users were able to correctly interpret the colour coding	Legend was clear and easily understood
4	User can use and understand data table	Users scroll and use filters in the data table	Table updates based on filters and is easily read	Table functioned correctly and filters worked	Some users couldn’t understand some column headings

6.1.2 User Testing

Test number	Description	Input	Expected Output	Actual Output	Comments
1	Users can identify areas with high smoking rates	Users were asked to find a region with high smoking percentage	User can identify that darker region have higher smoking rates	Users were correctly able to identify high smoking areas using the legend	Map colours effectively highlighted smoking patterns
2	Users can identify in which counties women have better health than men	Users were asked to find which counties have more women in very good health over men	Users are able to identify which counties had more women in very good health	Users could correctly identify the counties with more women in very good health than men	GeoFacet grid charts helped users compare health data being shown
3	Users can download data from the data table	Users were asked to download the data from the data table in Excel form	Users successfully downloaded the file	File downloaded and opened correctly	Feature worked correctly
4	User can interpret regression plot and other model charts	User views correlation plot and other charts	Users are able to understand factors influencing smoking rates	Users could understand correlation plot using the legend but not the other charts	Visuals were helpful, but more knowledge with data is necessary to be fully understood

6.1.3 Feature Testing

Test number	Description	Input	Expected output	Actual output	Comments
1	Dropdown menu switches map layers	Select small area/ local	Selected map is displayed	Map layer switched correctly	Functionality works as intended

		electoral area			
2	Legend updated with selected map layer	Select small area/ local electoral area	Correct legend is displayed for selected map	Legend updated correctly	No issues when selecting different map
3	Tabs display correct content	Click each tab	Correct content is displayed across each tab	All tabs displayed the correct output	Navigation functioning correctly
4	Data table export buttons work	Click csv/ excel download	Files download with correct data	Files were successfully downloaded	Export function works as intended

6.2 Testing and Feedback

Usability testing was carried out to evaluate how easy the application was to use and how well the information was communicated to users. This assessed the clarity of the interface. This included the layout of tabs, dropdown menus, and map legends. Feedback from this testing indicated that the layout was clear and easy to use, however some users found that some column headings of the data table were difficult to understand, so clearer labelling could improve usability.

User testing was conducted by allowing users to interact with the application and complete simple tasks. This included identifying regions with high smoking rates or which counties displayed better health in women. The results showed that users were able to understand and navigate the dashboard and correctly interpret most visualisations, however while some users could understand the correlation plot using the legend, users had a harder time understanding the more detailed statistical outputs.

Feature testing verified which components of the application functioned correctly. This included testing dropdown menus, tab navigation, map legend, and export function. Each component was tested to ensure it responded correctly to input and produced the expected output. All features were found to be functional, with no errors observed.

6.3 Error Handling

6.3.1 Debugging techniques

Functions such as View() and print() were the main debugging techniques used during the development of this application. These functions were used to examine datasets when validating joins between census data and shapefiles to correctly determine the common variable to create the join. View() displays the whole dataset and allowed for further analysis of the data. Print() displays the dataset in the console, showing each column name and their data type.

6.3.2 External Resources

External resources such as Stack Overflow and official R package documentation were used throughout the development of this project. These resources were used for solutions for common issues, guidance on implementing specific functions, and tips for working with spatial data, visualisations, and shiny applications.

6.3.3 Unresolved Bugs

Some issues within the application remain unresolved. One issue is in the Leaflet map displaying Local Electoral Areas, certain polygons did not display colour and appeared as blank spaces, however when hovered over, the correct percentage was still displayed. The issue for this could be related to colour scaling or rendering, with the value of the area falling just outside the defined colour range.

A second issue in the application is with the GeoFacet visualisation, where data for four counties did not display. The dataset contains no missing values and functions properly on another system, this is likely due to package versions, data encoding, or file handling.

6.4 Evaluation of Final System

The final system was evaluated against the initial objectives and success criteria to determine which requirements were achieved.

6.4.1 Specific Objectives

The first objective was to identify, collect, and prepare relevant datasets relating to smoking, health, population, and geographic boundaries, was fully achieved. Census datasets and shapefiles from CSO and MapGenie were successfully gathered, cleaned, and processed for use within the dashboard.

The second objective was to develop interactive Leaflet maps to display smoking rates across Small Areas and Local Electoral Areas was also achieved. Two interactive map layers were implemented using colour-coding and legends with a dropdown menu to swap between the data displayed.

The third objective was to create additional visualisations such as GeoFacet charts and data tables. This objective was also successfully achieved. A GeoFacet grid was developed to display health data for each county, along with an interactive data table with filtering and exportation functionalities.

The fourth objective involve regression modelling, correlation analysis, and supporting charts was also achieved. A liner regression model was implemented to display the relationship between smoking and other selected variables. Correlation analysis and supporting visualisations, including pairwise plots and a graph to display predicted values vs actual values, were also successfully implemented.

The final objective was to integrate all visualisations into a user friendly R Shiny dashboard, which was successfully achieved. All maps, charts, and tables were combined into a single interactive application with tabs for navigation.

6.4.2 Success Criteria

The success criteria were also met by the final application. A complete and functional R Shiny dashboard was successfully developed, with all maps, charts, and tables rendered correctly.

Users were able to navigate between different tabs, accessing maps, the GeoFacet visualisation, data table, and regression output without issue.

Datasets were correctly joined with shapefiles, which allowed for proper visualisation of data for maps.

Leaflet maps were successfully updated with different layers when selected, however, small visual issues occurred in come Local Electoral Area polygons where colour did not render correctly.

Visualisations effectively used colour scales, legends and labels to communicate health and smoking data and was easily interpreted by users.

Regression model, correlation plot, and other model charts were successfully implemented, which allowed for further analysis of smoking and health relationships.

The application ran successfully within the local Rstudio environment, while minor performance issues occurred when loading bigger shapefiles, this did not impact usability or functionality.

7 Project Management

7.1 Introduction and methodology

The Project Management section discusses how the project was managed throughout each section. This project implemented a Sprint methodology which helped divide the work into manageable development cycles. How this project was conducted are outlined in the following sections.

7.2 Sprint 1 – Researching different mapping techniques

7.2.1 Goal

The primary goal of Sprint 1 was to investigate and evaluate different mapping techniques that could be used to visualise geographical data for this project. This project required spatial representation from census information, it was important to find a way to map this data that was technically suitable and user-friendly.

7.2.2 Learning about different mapping techniques

During this sprint, research was conducted into many different mapping techniques and technologies available within an R environment. This included static mapping libraries such as ggplot2, as well as interactive libraries like Leaflet and Mapview. Compatibility with shapefiles, GeoJSON data, and Shiny dashboards was also ensured.

Geofacet was also explored as another method of visualisation to show regional statistics in addition to the more traditional approaches. Hafen (2018) describes Geofacet as a technique used to take data representing different geographic entities and apply a visualisation method to each entity, with the resulting set of visualisations arranged in a grid that mimics the original geographic topology as closely as possible. This was considered useful for displaying county level health indicators in a clear and structured format.

The advantages and limitations of each technique was assessed based on interactivity, customisation, and ease of implementation. Interactive mapping was identified as the most suitable option for spatial exploration, while Geofacet was selected was chosen to compare county level health trends.

By the end of Sprint 1, Leaflet was chosen as the main mapping framework for this project, with Geofacet chosen as an additional analytical and visualisation tool.

7.3 Sprint 2 – Gathering and Processing Data

7.3.1 Goal

The main goal of this sprint was to gather and prepare the datasets required to support the visualisations and analysis within the application. This included sourcing data related to health, population, smoking prevalence and the geographical boundary files necessary for spatial mapping. Ensuring that accurate and reliable datasets were available at an early stage was essential.

7.3.2 Data Sources

The data implemented in this project was obtained from publicly available datasets. Statistical information relating to health, smoking, population and other data was sourced from the Central Statistics Office (CSO) databases. These datasets are compiled from a range of official government sources, which include national surveys and census data from the most recent census release.

In addition to statistical datasets, geographic boundary data was required to support the spatial mapping component of this project. Shapefiles representing administrative boundaries were downloaded in GeoJSON format from Tailte Eireann MapGenie, an official Irish geospatial data platform. These boundary files included small areas, Local electoral areas, and Counties, which were later integrated with the census data to enable the creation of interactive choropleth maps.

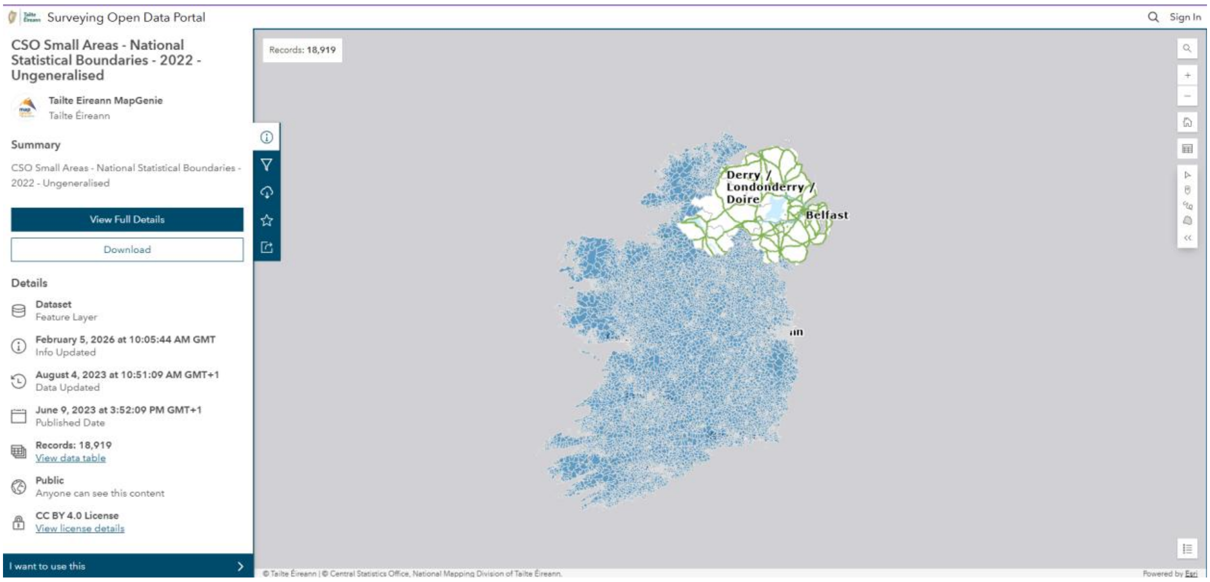

Figure 6 Tailte Eireann MapGenie

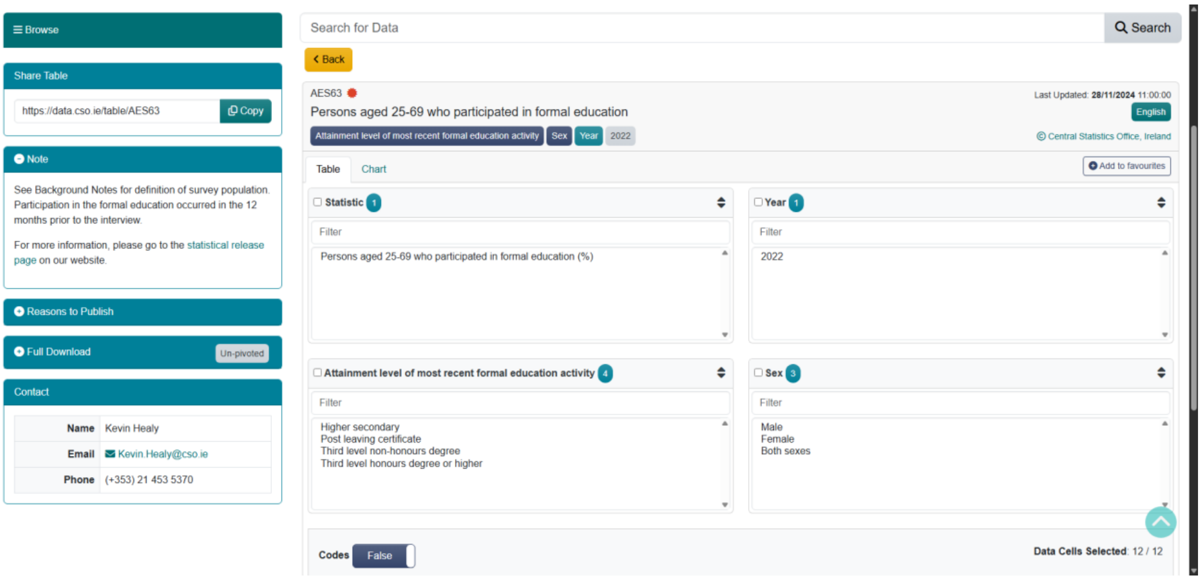

Figure 7 CSO database example

After gathering the data, the datasets were prepared to ensure consistency in structure, variable names, and geographical identifiers. This processing stage was necessary to allow accurate merging of spatial and statistical datasets in the following sprints.

7.4 Sprint 3 – implementing data in tableau

7.4.1 Goal

The main goal of this sprint was to integrate the prepared datasets into Tableau and develop interactive visualisations to allow users to explore health and smoking data across geographical boundaries, this involved combining spatial and statistical datasets, creating calculated fields, and creating filters to support data exploration. Hospital locations were also implemented to enhance visualisation and provide more context for public health analysis. This sprint also helped serve as basic feasibility testing, helping evaluate whether the datasets, spatial joins, and visualisation approach were suitable and effective for supporting this project.

7.4.2 Processing data for visualising and plotting

This stage of the sprint focused on preparing and processing data required for effective visualisation within Tableau. The small areas statistical dataset was merged with the corresponding small areas shapefiles, enabling each geographic boundary to be directly associated with its relevant health and demographic indicators. This spatial integration is essential for choropleth visualisations.

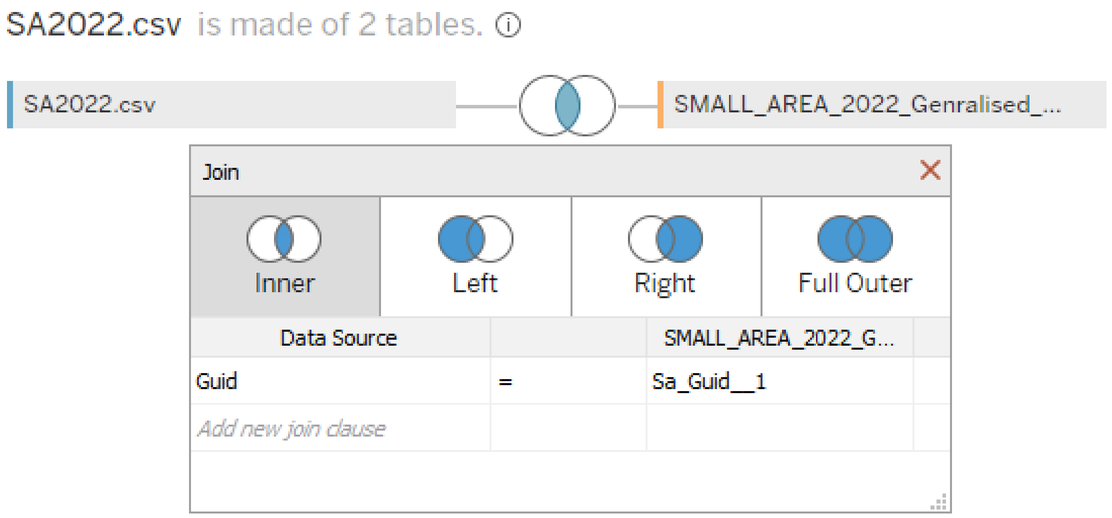

Figure 8 Small Area shapefile joined to data file

Calculated fields were then created to represent key health indicators, including the total number of the population within each category (very bad, bad, fair, good, very good), as well as smoking prevalence. These calculated fields allowed raw census data to be transformed for analysis and comparison.

Figure 9 Calculated field to show % of smokers

In addition, filters were implemented using calculated fields to allow users to select and view specific percentage ranges. This further enables data exploration, allowing patterns to be examined across different health or smoking thresholds.

7.4.3 Creating maps using joined data

Once the datasets were integrated and processed, they were used to construct interactive maps within Tableau. The joined spatial and statistical data allowed health and smoking indicators to be displayed across geographic boundaries using different colour-coding.

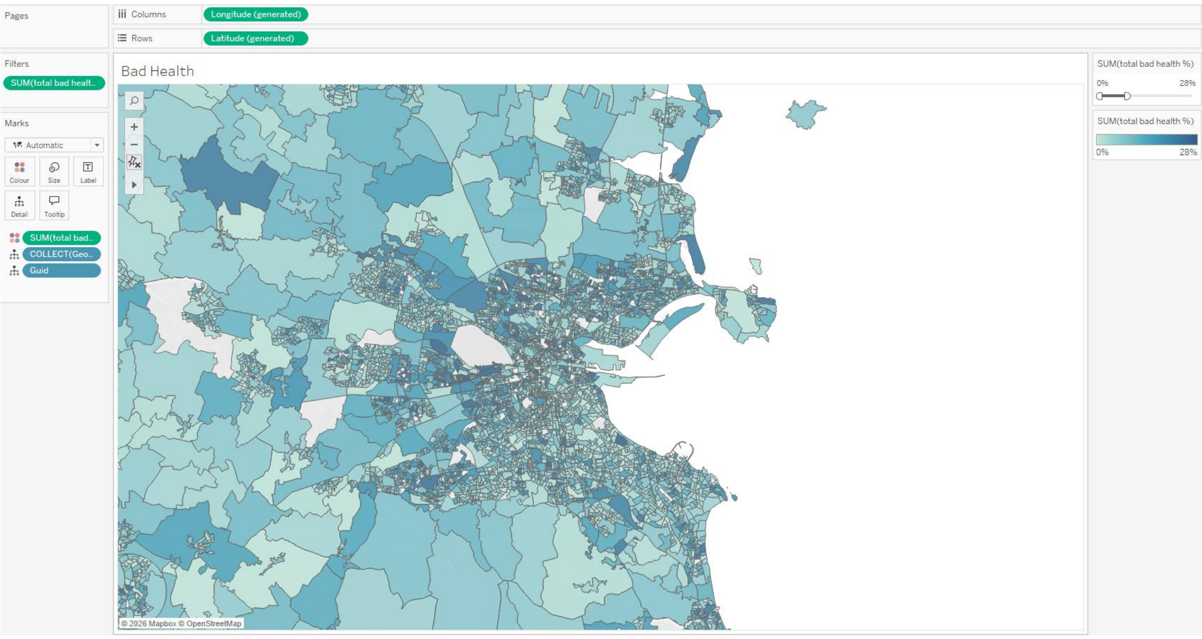

Figure 10 Tableau map

An additional dataset containing hospital locations was also merged. This dataset was layered onto existing health maps as point markers, to visually compare hospital distribution with regional health outcomes. This provided an additional analysis to the dashboard by linking healthcare infrastructure with population health indicators.

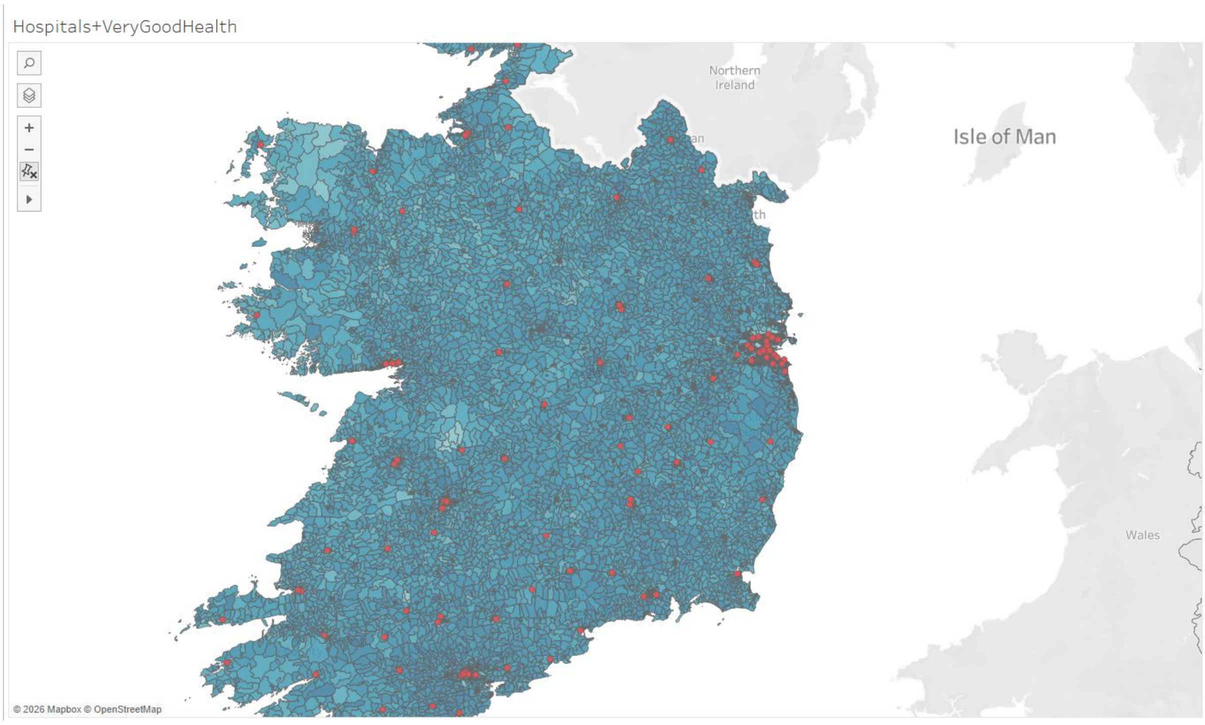

Figure 11 Map comparing Good Health and Hospital locations

Overall, this sprint resulted in a fully interactive dashboard within Tableau, which combined boundary based health data with point based hospital information to further support geographical analysis.

7.5 Sprint 4 – Loading Data into Rstudio

7.5.1 Goal

The main goal of this sprint was the initialise the first development environment in R Studio, load the required datasets, and begin prototyping project components outside the Shiny Dashboard. This allows each individual feature to be developed and tested independently before being integrated into the final application.

7.5.2 Downloading libraries and packages

During this sprint, the required R libraries and packages were installed to support data processing, mapping, statistical analysis, and dashboard development. Once this environment has been prepared, the relevant datasets were imported into R Studio, including both statistical datasets and spatial boundary files.

```
library(shiny)
library(readr)
library(sf)
library(tidyverse)
library(leaflet)
library(dplyr)
library(ggmap)
library(htmltools)
library(mapview)
library(htmlwidgets)
library(readxl)
library(csodata)
library(leaflet.extras)
library(leaflet.extras2)
library(leafgl)
library(leafem)
library(dbscan)
library(geofacet)
library(DT)
library(ggpubr)
library(ggcorrplot)
library(GGally)
```

Figure 12 Libraries used during development

Each file was validated to ensure they loaded correctly, variable names were recognised, and the data structures were suitable for later use within the dashboard.

Data	
Census_Very_Bad_Health	62 obs. of 20 variables
corranalysis	num [1:5, 1:5] 1 -0.707 -0.226 0.532 0.182 ...
correlationanalysis	31 obs. of 5 variables
Geofacet_all_health	248 obs. of 55 variables
healthmodel	List of 12
LEAshp	31 obs. of 3 variables
r	num [1:11, 1:11] 1 0.424 -0.226 0.532 NA ...
regressionhealth	31 obs. of 11 variables
regressionhealth_Copy	31 obs. of 6 variables
Values	
predicted_values	Named num [1:31] 14.5 13.3 14.2 13 13.1 ...

Figure 13 Loaded data in RStudio

As well as data preparation, R Studio was used as a separate development environment to start building early visualisation components, such as the first Leaflet map. Developing these features separately allowed functionality to be tested and refined before being transferred into the R Shiny dashboard.

By the end of this sprint, datasets were successfully loaded into Rstudio.

7.6 Sprint 5 – Creating Leaflet map to display smoking rates

7.6.1 Goal

The main goal of this sprint was to develop an interactive Leaflet map capable of displaying smoking rates across different geographic regions. The object was to transform the prepared census and spatial datasets into accessible visual format that would allow users to identify regional variants in smoking prevalence. A further aim was to provide multiple geographic levels by enabling users to switch between different administrative boundary levels.

7.6.2 Implementing first map – Small Areas

The first stage of development focused on creating a map based on Small Areas, which represents the most detailed geographic units available to download. Small areas are based on factors such as containing between 50 and 200 dwellings and generally comprised either of complete townlands or neighbourhoods (Central Statistics Office, n.d.). These Small Area shapefiles were joined with the matching Small Area census smoking data which allowed each boundary to display the percentage of individuals who smoke within that area.

A choropleth map was implemented using the addPolygons() function within Leaflet. Each polygon represented a small area boundary and was colour-coded according to smoking prevalence using a custom colour palette generated through colorBin(). This allows areas with different smoking percentages to be visually contrasted through graduated shading. Interactive labels were also added so that users could hover over each area and see the corresponding smoking percentage.

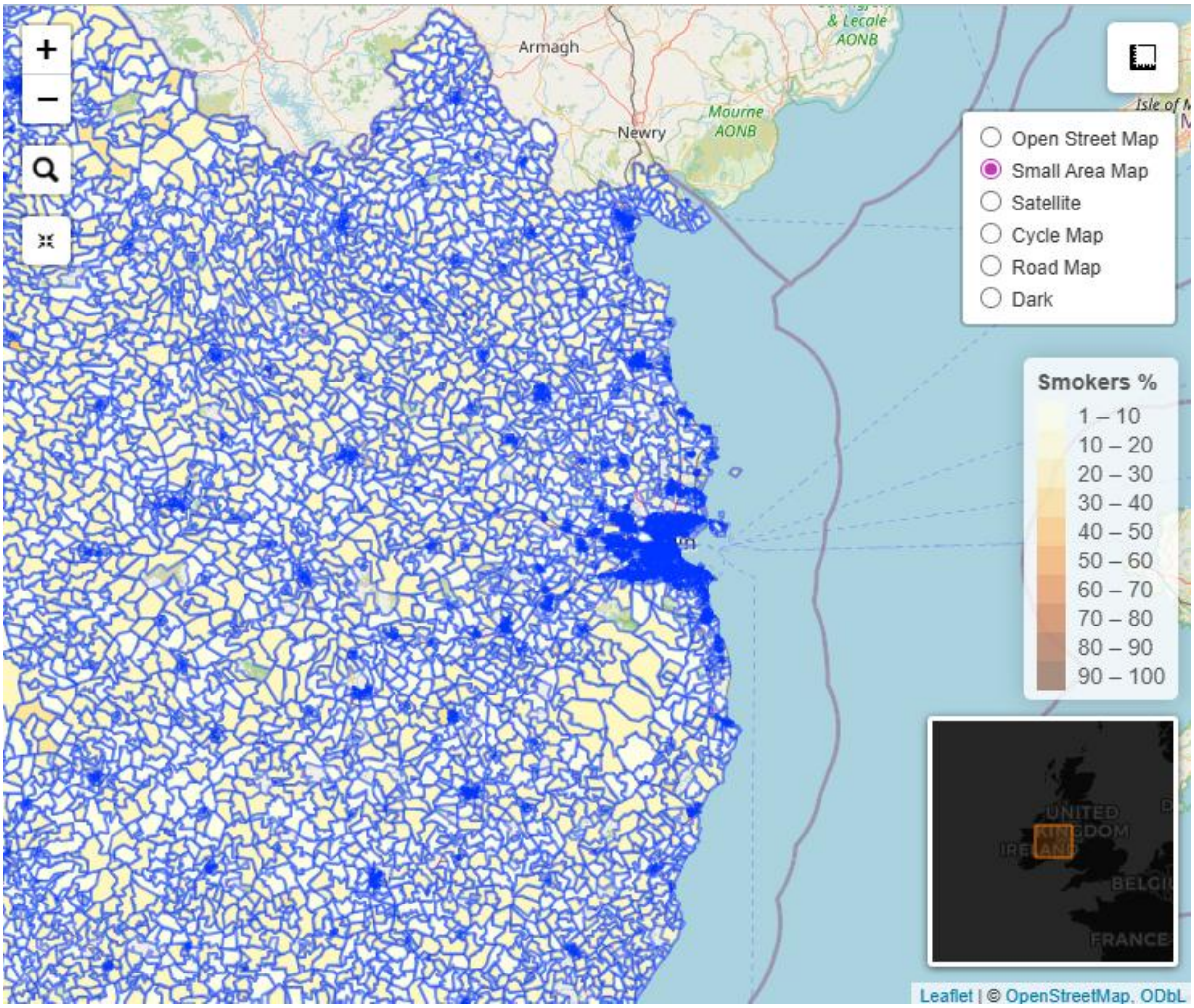

Figure 14 Leaflet Map Example

```
m <- leaflet(JoinSAshpCensus.new) %>%
  addTiles(group = "Open Street") %>%
  addMeasure() %>%
  addSearchOSM() %>%
  addResetMapButton() %>%
  addMiniMap(tiles = "CartoDB.DarkMatter") %>%
  addProviderTiles("Esri.WorldImagery", group = "Satellite") %>%
  addProviderTiles("MtbMap",group = "Cycle Map") %>%
  addProviderTiles("OPNVKarte", group = "Road Map") %>%
  addProviderTiles("CartoDB.DarkMatter",group = "Dark") %>%
  setView(lng= -6.2587,lat= 53.3482,zoom = 14) %>%

  addPolygons(data = JoinSAshpCensus.new,
              weight =2,
              fillColor = pal(JoinSAshpCensus.new$percentx),
              fillOpacity = 1,
              highlight = highlightOptions(
                weight = 5,
                # color = ~ qpal(percentx),
                color = "#666666",
                fillOpacity = .1,
                bringToFront = TRUE),
              label = lapply(labels,HTML),
              group = "Small Area Map")  %>%

  addLegend("bottomright",pal = pal,values = ~JoinSAshpCensus.new$percentx, title = "Smokers %") %>%

  addLayersControl(
    baseGroups = c("Open Street Map","Small Area Map","Satellite","Cycle Map","Road Map","Dark"),
    options = layersControlOptions(collapsed = FALSE))

m
```

Figure 15 Code Snippet

7.6.3 Implementing second map – Local Electoral Areas

Following the successful implementation of the Small Area map, a second version was developed using Local Electoral Areas. Local Electoral Areas are formed by aggregating Electoral Divisions, and in some cases, Electoral Divisions are split between Local Electoral Areas in order to render them suitable for production of statistics (Central Statistics Office, n.d.). These larger administrative units provided a broader regional overview and allowed smoking rates to be examined at a higher geographic level.

As with the first map, LEA boundary data was merged with the corresponding smoking dataset and displayed using addPolygons(). A separate colour scale was implemented to better reflect the data range at this level, ensuring that the shades remained visually clear. This second map gave a different perspective by highlighting broader regional trends that may not be immediately visible as Small Area level.

7.6.4 Implementing interactive and visual features

To improve user interaction and navigation, a dropdown selection menu was created using selectInput() within the Shiny user interface. This control allowed users to switch dynamically between the Small Area map, Local Area map, and a standard OpenStreetMap base layer.

The dropdown was linked to leafletProxy(), allowing map layers to be shown or hidden without reloading the full map. This improved application performance and created a smoother user experience. In addition, a dynamic legend was implemented using addLegend(), which updated depending on the map layer selected. This legend explained the colour-coding system used to represent smoking percentages, making the visual output easier to interpret.

The implementation of selectable layers and legends enhanced the comparative analysis and overall visual clarity between the different geographic levels.

By the end of this sprint, a fully interactive mapping component has been developed with selectable geographic layers, colour-coded polygons, and dynamic legends, providing one of the core visual features of the final dashboard.

7.7 Sprint 6 – Creating Geofacet map

7.7.1 Goal

The main goal of this sprint was to develop a GeoFacet visualisation to display and compare health outcomes across the different counties in a structured grid format. This required preparing a dataset in a compatible format with the GeoFacet framework, as well as the transformation of raw census data into comparable percentage indicators across regions and demographics.

7.7.2 Creating Excel file for GeoFacet data

A significant part of this sprint was spent constructing a structured Excel dataset to support the GeoFacet visualisation. The dataset was built using county level codes which served as identifiers for each geographic unit.

The raw census data contained a breakdown of health status across both genders, ranging from "very bad" to "very good" health categories. For each category, values were separated into male and female totals. These values were then normalised by dividing each group by the total corresponding population of the area, producing percentage-based measures. This allowed for comparison across counties of different population sizes.

An additional preprocessing step was required to address overlapping administrative regions. For example, Dublin was split into multiple local authority areas such as Dublin City, Dún Laoghaire-Rathdown, Fingal, and Soth Dublin. These were aggregated to produce a single combined county level value to ensure compatibility with the GeoFacet grid structure, which operated at county level only.

Once aggregation and normalisation were completed, the dataset was reshaped into a long format suitable for GeoFacet visualisation. Each county code was repeated across multiple rows, representing each health category and gender combination. The final dataset included a variable for county code, health by gender, and the corresponding percentage values for each health category. This was the final input data for the GeoFacet map.

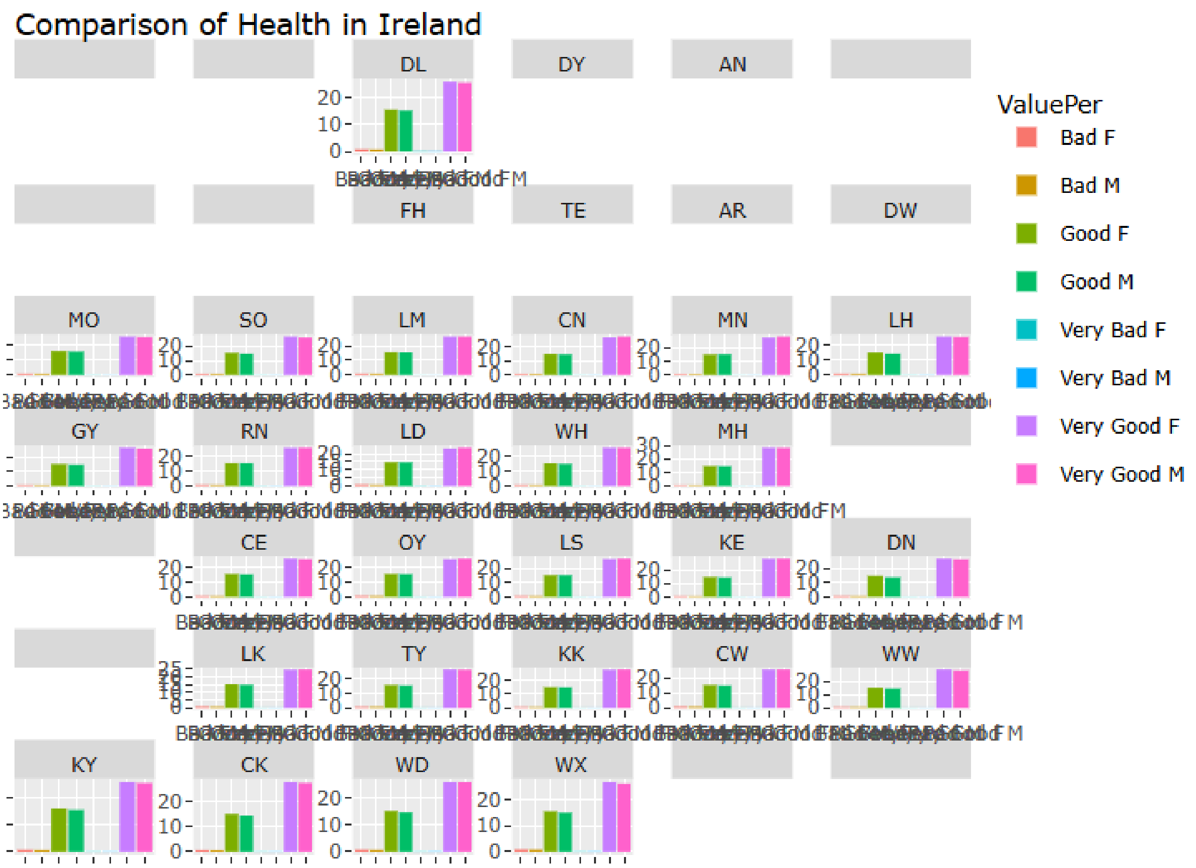

Figure 16 Original Geofacet Visual

```
interactive <- ggplot(Geofacet_all_health, aes(Percent,ValuePer, fill = ValuePer)) +
  geom_col(position = position_dodge()) +
  coord_flip() +
  #theme_bw() +
  facet_geo(~ CODELPer,grid = "ie_counties_grid3",scales = "free")+
  labs(title = "Comparison of Health in Ireland",
       subtitle = "Counties, 2015-16 to 2016-17",
       caption = "Source: R", x = "", y = "")

ggplotly(interactive)
```

Figure 17 Geofacet Code Snippet

7.7.3 Implementing data into a GeoFacet map

The prepared dataset was then imported into R Studio and used to create a GeoFacet visualisation using grid layout 3, which approximates the geographic positioning of Irish counties.

Each facet in the grid represents a county, with individual plots displaying health distribution percentages across the categories. The long format structure of the dataset allowed for seamless mapping of health indicators to visual elements with each grid panel. This enables comparison of health outcomes across counties while keeping a spatially meaningful layout.

By the end of this sprint, a fully structured GeoFacet dataset was developed and successfully implemented into a grid based visualisation, providing a different perspective to Leaflet base mapping.

7.8 Sprint 7 – Creating a table for GeoFacet data

7.8.1 Goal

The main goal of this sprint was to create a structured and interactive data table to present the processed GeoFacet dataset. This was implemented to compliment to GeoFacet visualisation by allowing users to view, filter and explore the numerical values in a clear way.

7.8.2 Creating table to display GeoFacet data.

During this sprint, the final GeoFacet dataset was used to construct an interactive table within R Studio using a dynamic data table framework. This table included all variables.

The table was designed to support user interaction, allowing for sorting and searching to enable more detailed exploration of specific counties or health categories. This ensured users could analyse the dataset independently from the GeoFacet visual.

In addition, a function to allow users to download the dataset as in CSV or Excel formats was implemented, which allows users to access the raw structured output used in the visualisation.

The table was aligned with the GeoFacet dataset structure, ensuring consistency between graphical and tabular representations of this data.

By the end of this sprint, a fully interactive and exportable data table had been implemented, providing both analytical flexibility and improved accessibility for users.

7.9 Sprint 8 – Creating models, correlation plots, and charts

7.9.1 Goal

The main goal of this sprint was to develop statistical models and supporting visualisations to analyse the relationship between smoking prevalence and a range of socioeconomic and health-related variables. This included building a linear regression model, generating correlation analysis, and producing multiple plots to evaluate the model performance and variable relationships.

7.9.2 Model Data

During this sprint, a linear regression model was constructed to examine the relationship between smoking prevalence and several exploratory variables, including very good health, long term unemployment, employment in construction related sectors, highest education level completed by people over the age of 15 (primary education), and unskilled social class.

The dataset used for modelling prepared using census data and processed to ensure consistency. This allowed the model to focus on key socioeconomic indicators that are commonly associated with health outcomes.

7.9.3 Correlation Matrix

A correlation matrix was generated to examine the relationships between all selected variables prior to and alongside model development. This helped identify the strength and direction of associations between smoking rates and socioeconomic.

The correlation matrix provided an initial overview of multicollinearity and variable relationships, supporting the interpretation of the regression model and ensuring strong predictors were identified early.

Figure 18 Early Model Heatmap

7.9.4 Model Plot

A regression based comparison plot was created to evaluate the performance of the model by comparing predicted smoking values against actual observed values. This visualisation allowed for clear assessment of how well the model compared variation in smoking prevalence across different regions.

The plot highlights the alignment between predicted and actual values, with deviations indicating areas where the model may under or overestimate smoking rates.

7.9.5 ggPairs Plot

To further explore the relationship between variables, a pairs plot was generated using ggPairs. This provided a visual overview of pairwise relationships between all selected variables in the regression model.

This plot allows for inspection of correlations, distributions, and interaction patters at the same time, enabling deeper insight to the structure of the dataset and supporting model interpretation.

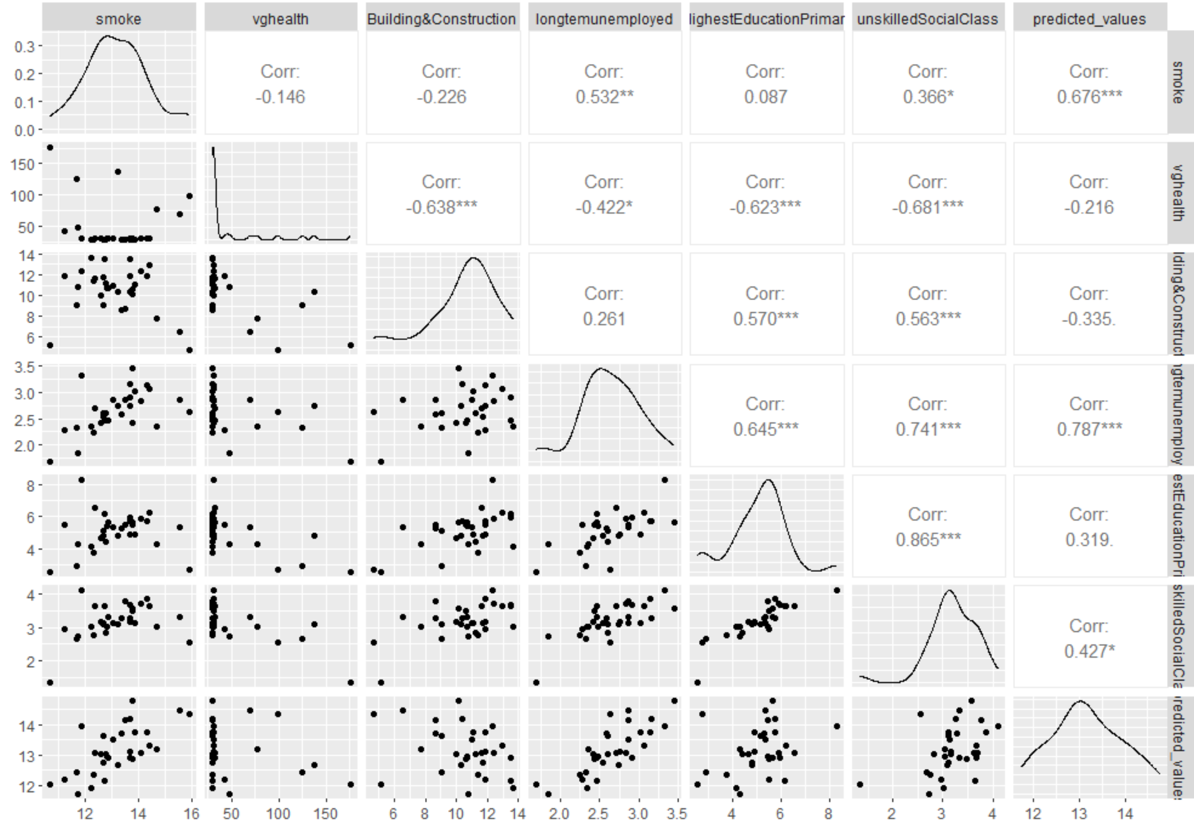

Figure 19 ggPairs Plot

7.9.6 Model Summary

A full statistical summary of the regression model was produced to evaluate overall model performance. This included coefficients, significance levels, and measures of model fit.

The summary output provided an interpretation of how strongly each independent variable affected smoking prevalence, and whether these relationships were statistically significant. This step was essential in validation the effectiveness of the model and understanding the contribution of each variable.

7.10 Sprint 9 – Creating Rshiny dashboard to display all maps and charts.

7.10.1 Goal

The main goal of this sprint was to integrate all previously developed prototypes into a single interactive R Shiny dashboard. This involved combining the Leaflet maps, GeoFacet visualisation, data table, and statistical models to a unified application. A further objective was to ensure all components were functional with a reactive environment and users could seamlessly interact with different forms of data analysis from one central platform.

7.10.2 Implementing existing maps and charts into the dashboard

During this sprint, the previously developed visualisations were integrated into the Shiny framework. This includes the Leaflet based smoking rate maps, the GeoFacet visualisation, the interactive data table, and the regression analysis outputs, including correlation plots, model plots, ggPairs plot, and model summary.

Each component was adapted from their standalone R scripts into the reactive outputs compatible with Shiny's server structure. This required reconfiguring plotting functions and ensuring all the datasets were loaded and accessible within the server environment. The integration process also involved resolving layout and rendering issues to ensure that all visualisations displayed correctly within the dashboard interface.

7.10.3 Creating tabs to swap between the data being shown

To improve usability and navigation, tabs were implemented using Shiny's tabsetPanel() structure. This allows users to switch between different sections of the dashboard, including maps, GeoFacet analysis, data tables, and regression outputs, without needing to reload or navigate away from the application.

Each tab was dedicated to a specific analytical component, ensuring the dashboard remained organised and easy to navigate. The structure also supported a logical workflow for users, moving from Leaflet spatial analysis to statistical modelling.

By the end of this sprint, all previously developed components had been successfully integrated into a fully functional R Shiny dashboard. This marked the completion of the development phase and resulted in a cohesive and interactive application for exploring health and smoking data across Ireland.

7.2 Tools

Project management and development was done through Trello, which was used to organise and monitor the progress of sprints and provided a structured way to manage tasks and track development stages.

8 Conclusion and Future Work

8.1 Summary of Findings and Reflection on Project Aim

8.1.1 Summary of Findings

Using the Small Area Leaflet map, I found that higher rates of smoking tend to cluster around areas close to hospitals, especially in Dublin. This pattern might be influenced by higher population density in urban environment, where hospitals are commonly located. Another factor may be that urban areas usually contain more diverse population groups with different health behaviours compared to rural regions.

Using the Local Electoral Area map, I found that higher smoking percentage areas were usually clustered around larger cities and towns. This suggest a strong divide between urban and rural areas when it comes to smoking prevalence, which could be linked to differences in conditions like socioeconomic status, employment opportunities, education levels, or other lifestyle factors.

From the GeoFacet health by county visualisation and supporting data table, I found differences in self-reported health outcomes between genders and geographic regions. Women generally reported higher levels of "very good" health in major urban counties such as Dublin, Galway, and Cork. In contrast, less populated counties such as Cavan and Longford showed a higher proportion of males reporting "very good" health. While this census data used is reliable, in some cases the census is filled out by a single individual on behalf of the household. This may slightly skew the data, as reported health status for other members in the household may be based on perception rather than self-reported.

From the regression model, I found that the overall relationship between smoking prevalence and the selected variables is moderately strong. It has an adjusted R2 of 0.5507 which indicates that just over half of the variation in smoking rates can be explained and related to the other variables.

I found that long term unemployment had a significantly positive relationship with smoking rates. This could suggest that areas with a higher number of unemployment would tend to have higher smoking rates. Unskilled social class was also positively associated with smoking, which established a link between lower socioeconomic statis and increased smoking patterns.

In contrast to this, higher levels of primary education showed a significant negative relationship with smoking rates. This would suggest that improved education is associated with lower smoking prevalence. The building and construction variable also showed a significant negative relationship.

I found that factors such as employment status, education level, and social class has a stronger influence on smoking prevalence that general health indicators alone.

8.1.2 Reflection on Project Aim

Overall, I found that the project successfully met its original aim of developing an interactive R Shiny dashboard to visualise and analyse smoking and health data across Ireland. The completed application integrates spatial mapping, statistical modelling, and interactive visualisations into a single platform to allow users to explore differences in smoking prevalence and related factors across different regions. the Leaflet map achieved the aim of visualising smoking rates across different geographic levels, with colour-coded polygons and legends to aid interpretation of patterns.

I also found that the supporting visualisations, including the GeoFacet grid, data table, regression model, and correlation analysis extended the dashboard beyond simple mapping.

8.2 Critical Reflection of Development Methodology and Personal Learning

8.2.1 Reflection

This project was developed using a sprint methodology. I found that this was an effective approach, as each component could be developed and tested separately before being integrated into the final Rshiny dashboard. This helped reduce the risk of errors and resolve issues at each stage of development. However, I found that integrating multiple spatial data files other components within Shiny took longer than expected.

8.2.2 Personal Learning

From a personal learning perspective, I developed a stronger understanding of spatial data handling, especially when it came to joining census data with shapefiles and managing datasets. This project also helped improve my ability to debug issues within R, especially in when working with Leaflet maps in Shiny. I also gained experience in statistical analysis, including regression modelling and correlation analysis, and how these can be applied to real world health and other datasets.

Overall, I found this project significantly enhanced my technical skills in data visualisation, spatial analysis, and interactive dashboard development, while also improving my timekeeping skills and my ability to manage structured sprint during a development process.

8.3 Limitations of Current Application and Future improvement

One of the main limitations of the current application is the GeoFacet visualisation. As of now, GeoFacet doesn't support interactive filtering, meaning you can't select which variables to display through checkboxes. All variables are fixed within the output, meaning any changed to the displayed data must be made to the code directly.

Another limitation is the regression model, which is currently based on a small set of socioeconomic and health variables. While this model does help in exploratory analysis, it does not account for a wider range of potential influencing factors. In terms of future development, the regression analysis could be improved by adding a broader range of factors to better explain smoking percentages and overall health outcomes.

Further development could include exploring alternative spatial analysis techniques to provide further insight into geographic patterns and relationships between the data.

8.4 Overall Conclusion

I found that overall, this project achieved the aim of developing an interactive Rshiny dashboard to visualise and analyse smoking and health data across Ireland. The final application includes spatial mapping and statistical analysis in a clear and accessible way.

From a technical perspective, I found that the project showed effective use of shapefiles, census data, and R libraries and packages to create interactive maps and visualisations. While some limitations were identified, the core function of the project was met and performs as intended within a local environment.

Personally, I found that this project improved my skills in data handling, spatial analysis, and Rshiny development. This provided valuable experience in building a full data visualisation from raw datasets to the final interactive dashboard.

References

Central Statistics Office. (n.d.). *Census 2022 small area population statistics*. CSO. https://www.cso.ie/en/census/census2022/census2022smallareapopulationstatistics/

Hafen, R. (2018, March 10). *Introducing geofacet*. Ryan Hafen. https://ryanhafen.com/blog/geofacet/

Ofori, M. A., Lartey, S., Durneva, P., Jha, N., Mittal, N., Roy, S., ... & Joshi, A. (2025). Visual communication of public health data: a scoping review. *Frontiers in Digital Health*, *7*, 1555231.

Tan, X., Suo, X., Li, W., Bi, L., & Yao, F. (2025). Data visualization in healthcare and medicine: a survey. *The Visual Computer*, *41*(5), 3037-3058.

World Health Organization. (2025, October 6). *WHO tobacco trends report: 1 in 5 adults still addicted to tobacco*. https://www.who.int/news/item/06-10-2025-who-tobacco-trends-report-1-in-5-adults-still-addicted-to-tobacco

World Health Organization. (2025, June 25). *Tobacco*. https://www.who.int/news-room/fact-sheets/detail/tobacco