CITY UNIVERSITY LONDON

Analysing Data Protection Complaints:

Data Analytics, Visualisations, Risk Assessment and Mitigation Strategies for ICO's Data Protection Complaints

Master's in Computer Science and Cyber Security

Aadil Saiyad

Department of Computer Science
City University of London
London, United Kingdom, 2024

Analysing Data Protection Complaints:

Data Analytics, Visualisations, Risk Assessment and Mitigation Strategies for ICO's Data Protection Complaints

AADIL SAIYAD

aadil.saiyad@city.ac.uk

aadil-saiyad1@hotmail.co.uk

Supervisor: Nikos Kominos, Department of Computer Science.

Master's Dissertation Project 2024

Department of Computer Science

City University of London

Northampton Square

London EC1V 0HB

Telephone +44 (0)207 040 5060

Word Count: 14864 (Did not Include Figure Descriptions)

Table of Contents

Chapter 1

Introduction

Problem to be solved

The Data Protection Complaints Dashboard is designed to address a problem that is becoming increasingly significant but often overlooked in the realm of cyber security and data regulation, especially with regard to data protection complaints. While there is a plethora of visualisations available for cyber-attacks faced by organizations and the data that has been breached, there is a noticeable absence of visualizations dedicated solely to data breaches by organizations against their users. Additionally, despite the availability of open-source data on data protection complaints, there are no publicly accessible visualizations for such vital information.

The proposed dashboard aims to address this gap by providing a comprehensive and easily accessible and understandable overview of data protection complaints for stakeholders. By making it easier for organizations to identify and mitigate risks, the dashboard can support increased GDPR compliance and ultimately improve data protection practices. Furthermore, the dashboard can help organizations identify areas where they need to be improved to support regulators monitoring organizations' compliance with data security regulations. In conclusion, the Data Protection Complaints Dashboard is very useful for ensuring compliance and further enhancing the protection of personal data.

Objectives

The main objective of this research project is to develop a comprehensive web-based dashboard tool for ICO Data sets of complaints they have handled from members of the public about data protection concerns, offering a user-friendly interface to visualise, analyse, and understand the compiled data.

Sub-Objectives

- Conduct research on data visualizations, exploring various chart types and methodologies, particularly focusing on their application in the context of cybersecurity.
- Conduct research on data protection compliance and GDPR law, including an in-depth understanding of the data protection complaint submission process.
- Perform data analysis, aggregation, cleansing, and preparation of the data protection complaints dataset to ensure accuracy and reliability.
- Implement data visualization techniques to represent data protection complaints effectively, including the creation of interactive charts and visualizations.
- Research and implement risk analysis methods tailored to data protection complaints, aiming to identify and assess potential risks associated with data breaches.
- Develop and implement a GDPR compliance advisor tool, providing guidance and recommendations based on GDPR requirements and data protection best practices.
- Host the website on the Taipy Cloud platform to ensure accessibility and reliability for users.
- Conduct comprehensive testing of the website to ensure optimal performance and usability.

Project Beneficiaries

This tool benefits the general public by providing easy access to transparent overviews of data protection complaints, aiding in informed decision-making regarding personal information sharing online. Understanding potential risks and challenges reported by others enhances awareness of data protection rights, fostering trust between users, organizations, and regulatory bodies like the ICO. Businesses and organizations can utilize the dashboard to tailor practices, ensuring compliance with data protection regulations and building trust through proactive data management. Data Protection Officers within organizations can stay informed about external data protection concerns, aligning internal practices with regulatory standards. Researchers and academia benefit from the dashboard's rich dataset, offering evidence of real-world data protection issues for studies and policy recommendations. Regulators and policymakers use insights from the dashboard to inform policymaking, allocate resources, and benchmark data protection frameworks globally. Legal professionals leverage the dashboard to stay updated on data protection complaints and regulatory actions, advising clients on compliance strategies and conducting thorough risk assessments. For Additional Information on beneficiaries, refer to (Appendix A).

Overall Implementation

For me to achieve the targets of this project, I carried out in-depth research on GDPR compliance and data visualization methods. This involved reading Andy Kirk's book on Data Visualization (Kirk, A. 2019) and researching academic papers, especially on visualisation within Cyber Security. Additionally, I also considered GDPR law, studied articles relating to it and read scholarly papers that show its relationship with personal information as well as how best I can promote compliance, and provide recommendations and remediations to data breaches.

I also undertook a rigorous process of data cleansing using the Pandas Python library. My approach to this involved the identification and rectification of inconsistencies, errors, and missing values, through the implementation of data validation techniques. I also standardised data formats and units, where necessary, to ensure uniformity across the dataset, which would facilitate seamless integration into the visualization within the dashboard. I took great care to ensure that the dataset was properly curated, to maximise its utility for future research endeavours.

After preparing and cleaning the dataset thoroughly, I moved on to the next step by integrating it into the dashboard. Using the Taipy Python library, I created different types of graphs and charts to show the data visually. These included different charts that were carefully designed to show different, but specific aspects of data protection complaints. The goal was to make it easy for users to understand the data and use it to make decisions. By presenting the information visually, I aimed to help users gain insights and act based on what they saw in the graphs.

Additionally, I expanded the project scope that was not initially considered within the Project Development Documentation by incorporating risk analysis methodologies into the dashboard, necessitating research into various risk assessment techniques, their formulas, calculations, and their application to data protection cases. This led to the additional use in the research and addition of normalisation techniques, specifically MinMaxScaler which helped mould the risk data to be used more effectively, providing an increased ability to differentiate between them by expanding its current range from 20-80% to 1-100%.

Furthermore, I introduced another feature, the GDPR compliance advisor, leveraging the research conducted on common GDPR complaints to provide actionable recommendations for organizations. This was possible by taking the most common GDPR complaints made against sectors and subsectors of industry and formulating remediation actions to decrease the likelihood of them occurring if organisations implement or add to those recommendations within their organisations.

Finally, I hosted the project on the Taipy cloud platform and conducted thorough testing to ensure its functionality and reliability. This comprehensive approach enabled me to effectively meet the project objectives, delivering a robust and user-friendly dashboard that empowers stakeholders to navigate complex data protection challenges with confidence.

Assumptions and Limitations

One assumption made in the project is that the data obtained from the ICO is accurate and representative of the actual data protection complaints submitted. I have also assumed that users have basic familiarity with data visualizations and can interpret graphs and charts effectively with internet access and modern web browsers. The dashboard's insights are based solely on the data provided by the ICO and may not capture the entirety of data protection issues. Additionally, the dashboard does not reflect the most current data protection complaint trends in real-time and cannot be updated until the ICO release more data. Furthermore, the risk assessment is based on available data used within the dataset and methodologies implemented, which may not capture all potential risk factors and are based on predefined criteria, which may not fully capture the complexities of each data protection issue. Regarding the Compliance advisor, the recommendations are based on available data and best practices, which may not fully align with an organisation's unique circumstances and does not replace legal consultation or compliance assessments for organisations.

Chapter 2

Output Summary

Project Application	Appendix N/A
Output Type: ZIP File (27.7MB Unzipped)	
Beneficiaries: Examiner and anyone who wishes to further research / develop the project.	

Output Description: Folder containing all files required to run the project. This includes:
• main.py file which contains the main code of the project. (484 lines of code)
• Folder containing 5x Markdown files to display items on the page (265 lines of code):
  ○ root.md - Displaying the navbar over all pages.
  ○ dashboard.md - Visualising the charts on the main dashboard.
  ○ risks.md - Displaying the table and dropdown for organisations.
  ○ advisor.md - Displaying the Sectors and subsectors columns and GDPR.
  ○ instructions.md – Displaying the instructions on how to use the site.
• CSS file required for the colouring of the Risks table (22 lines of code).
• JSON file for the GDPR Compliance advisor page which contains the GDPR articles, description, and recommendations (140 lines of code).
• dataCleansing.py file shows parts of the data preparation stage (22 lines of code).
• CSV dataset contains the Data Protection Complaints dataset used.
• Image of the Sankey Diagram.
• Requirements text file used to host the project on Taipy Cloud and help the system get the necessary modules.	

Purpose: The are the main files required to run the project both locally and on the Taipy Cloud. This includes the data protection complaints dashboard, risk analysis and the GDPR compliance advisor. This can be found within the project confidential submission area.	

Project Setup README File	Appendix N/A
Output Type: txt File (5KB)	
Beneficiaries: Examiner and anyone who wishes to further research / develop the project.	
Output Description: File located within the Project Application folder highlighting the contents of the project application and the requirements needed to run the project.	
Purpose: Helps anyone who wishes to run, research, and further develop the project understand what is needed to run the project and the places to get them.	

Project Requirements	Appendix D
Output Type: Located within the Project Report found in the submission area.	
Beneficiaries: Examiner and anyone who wishes to further research / develop the project.	
Output Description: Project Functional and Non-Functional requirements, highlighting the ideal project results and outputs as well as areas to measure the success of the project on completion by testing.	
Purpose: Defines the scope, objectives, and deliverables of the project. Used as a roadmap, outlining the specific functionalities and features of the project.	

Project Testing Documentation	Appendix K
Output Type: Located within the Project Report found in the submission area.	
Beneficiaries: Examiner and anyone who wishes to further research / develop the project.	
Output Description: Documentation regarding the testing of the system against the final product and the original requirements of the project.	
Purpose: Verifies that the system functions correctly and meets the requirements specified at the start of the project.	

Project Video	Appendix N/A
Output Type: MP4 File located within the Project submission area.	
Beneficiaries: Examiner and anyone who wishes to further research / develop the project.	
Output Description: 15-minute video demonstrating the project deliverable and showcasing the different features it contains. This includes the data protection complaints dashboard, risk analysis and the GDPR compliance advisor and the code used to run certain aspects of it.	
Purpose: Visually demonstrates the project deliverables and highlight its key features and functionalities. This provides users an understanding of how the solution addresses the project requirements.	

Chapter 3

Literature Review

The exponential growth of digital technologies and the scale of data they generate has brought with it a growing number of data protection concerns. EU resident's privacy rights and responsible data handling are the concern of the General Data Protection Regulation (GDPR). Regulatory bodies like the ICO now monitor compliance with GDPR in that they handle public complaints about organizations and enforce the regulations. However, in the absence of tools to visualize these datasets on data protection complaints by the ICO, this literature review outlines the research areas necessary to develop a web-based dashboard to improve GDPR compliance, and bridges this gap by addressing GDPR clauses, the impact of data visualization, and current cybersecurity practices.

ICO's Data Protection Complaints Dataset and GDPR

Developing an online dashboard application for managing complaints submitted to data protection authorities concerning the ICO is a project which aims to implement GDPR clauses and prevent substantial consequences for breaches. (A. Skendžić, B. Kovačić and E. Tijan. 2018). Data protection authorities are required to identify safeguards where personal data is processed, and they can issue significant penalties for non-compliance including fines of up to €20 million or 4% of annual global turnover. The dashboard therefore assists organisations in satisfying an obligation to act when they receive a data protection complaint, following the requirements of the GDPR to handle such complaints 'in a transparent manner' and in accordance with Article 57 which helps them to minimise the potential financial penalties and also the associated reputational damage that it is likely to be caused by coverage of these complaints. (GDPR Info.)

The collection of data protection complaints within the dataset, comes from the ICO submit a formal complaints procedure on their website. (ICO Complaints Tool.) (See Appendix G for full flowchart breakdown) This dataset can be a valuable resource for many stakeholders. It contains detailed information on reported violations, regulatory decisions, and compliance issues, and provides a wealth of applicable analysis. Researchers can use the information to analyse complaint trends over time, identify complaint patterns in specific areas, or assess the effectiveness of the ICO's own regulatory interventions. More broadly, the data could be useful to policymakers wishing to learn more about the real-world impacts of data protection law and the GDPR, or to help inform regulatory decision making. Compliance professionals could use the data to understand what typical, or best practice looks like in any given sector, and how their own organisation's performance compares. The dataset could also increase transparency around the question of how complaints are handled and ultimately resolved by the ICO.

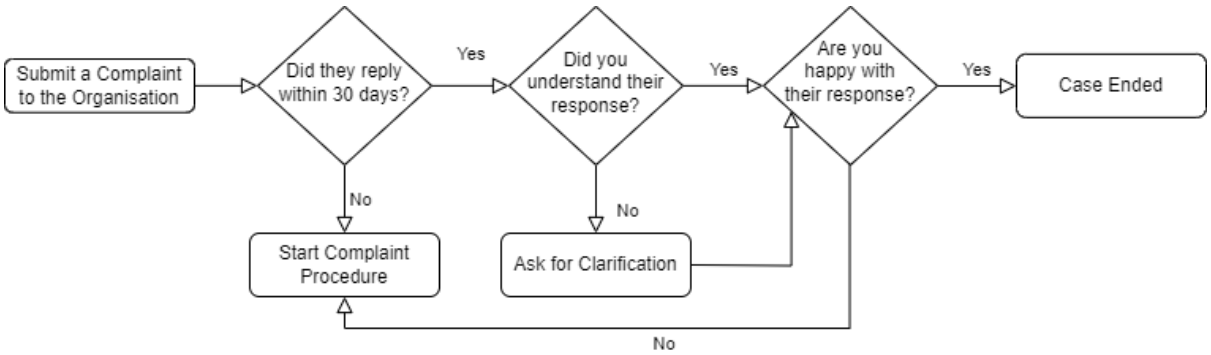


Figure 1: Overall ICO Complaints Submission Flow Chart

ICO's data protection complaints dataset includes multiple columns, each offering insights that can be useful for data visualisation and analysis. The "Sector" and "Sub-sector" columns allow organisations to be categorised according to the industry they operate in. This aids in quickly identifying sectors that have higher complaints volumes or recurring issues. This is essential for prioritising compliance efforts and allocating resources efficiently. (ICO Datasets. 2020)

The "Decision Primary Reason" column provides detail on the specific GDPR Articles breached. This offers an insight into the types of breaches that are occurring the most. Visualising this data will allow compliance teams to quickly see what common compliance pitfalls are and design mitigation strategies. Entries in the "Decision Primary Reason" column, like "Art 15 - Right of access" and "Art 7(3) - Right to withdraw consent," correspond to articles in the GDPR that cover what rights individuals have to their data. Article 15, for example, gives individuals the right to get confirmation that their data is being processed and access to it while Article 7(3) covers under what circumstances individuals may withdraw their consent for processing.

The ICO's Data protection complaints dataset also includes a column called "Submitted about account" containing the full name of the organisations complained about. The column has been published as part of the ICO's commitment to transparency, however there's "potential to damage an organisation's reputation and undermine public trust in that organisation" (Beveridge, C. 2023) as stated by Christopher Beveridge, Head of Privacy & Data Protection at BDO. On the other hand, it can also show organisations that have less complaints and handle their breaches effectively. He also states that "It remains to be seen whether the ICO's approach will actually act as an incentive for stronger compliance" and with the ability to visualise this information may bring a new light to the way we perceive and come to conclusions about these complaints. By categorizing complaints with associated GDPR articles, the "Decision Primary Reason" column gives insight into the specific provisions of the GDPR that are most often being cited in reported breaches, which helps organizations to identify common compliance challenges and prioritize corrective actions to address them. This, in combination with visualizations adds to the benefits to view these easier as well.

Enhancing Data Quality Through Cleansing

Data cleansing is integral for ensuring accurate and reliable data, particularly concerning data protection complaints. (Fakhitah R, et al. 2019) Incomplete information can lead to uncertainties during data analysis, underscoring the importance of meticulous data cleansing. Errors or missing values within datasets can significantly impact business decisions, potentially resulting in financial losses and operational challenges. (Liebchen, G.A. 2010) Gartner underscores the substantial financial implications of poor data quality, estimating an average

annual cost of $12.9 million across industries. (Sullivan, J. 2022) To mitigate these risks, the data cleansing process encompasses several essential stages. This involves investigation, standardization, deduplication/matching, and survivorship stages. Firstly, investigation analyses the dataset for errors and patterns. Then, standardization ensures uniform data formatting. Furthermore, deduplication/matching identifies and removes duplicates. Lastly, survivorship decides data retention post-matching, incorporating stakeholder input for data credibility. These stages ensure accurate, reliable data for informed decision-making. (K. H. Prasad, et al. 2011)

The Use of Data Visualisation

The definition of data visualisation according to Andy Kirk is "The representation and presentation of data to facilitate understanding" (Kirk, A. 2019). This definition is a big significance to the data protection complaints project as we seek to address and mitigate the challenges posed by the vast scale and complexity of the dataset we have. Today, the world's top data practitioners make sense of such enormous datasets through data visualisations. There are "three stages of visualisations. Perceiving, Interpreting, and comprehending". The first, perceiving, requires that you visually decode a representation – noticing patterns, distinguishing shapes, and colours. This is especially helpful when trying to communicate complex data to a wide variety of audiences in an aesthetically pleasing manner, combining different elements to see similarities, differences, and anomalies. Using this to see data protection complaints, we will be able to see how different compliances compare with others. This translates into interpreting a representation, where you understand the information that a visualisation encodes in a dataset. This is where the user can start to base a decision on what they are seeing and whether to take it into a positive or negative manner and form meaning behind it. In the case of our complaint's dashboard, users can assess the severity of complaints, identify trends or patterns, and evaluate the overall compliance landscape. Lastly, that comprehension turns into gaining insight or deriving meaning, where the implications of what you've seen are realised, and that information serves as guide for decisions. This is where we can gain new information to use to make a decision on what we know or what we have found out. This is especially important as "A good design is honest." We must convey our data in a truthful manner and not misinform the audience both accidentally and intentionally which misguides their perception of the data. This matters because each stage informs your next steps with extracting knowledge and understanding from a representation of visual data.

So how is data visualisation used in general complaint management? In a blog, Ernst & Young Financial Services managers stressed the importance of leveraging visualisation of data to improve how firms manage customer complaints. (TIBCO Blog. Spotfire Blogging Team, 2013) They explained how data visualisation allows complaints to be tracked through to resolution, how it helps maintain balanced workloads and identify opportunities to improve their services. They also discussed challenges, such as maintaining the confidence of customers, being able to adhere to regulatory compliance and produce reports for regulators that demonstrate how complaints are being handled and allow for efficient complaint reporting. Visual analytics is a complete game changer for banking institutions, reducing the time to gather, compile and report on data related to customer complaints, while providing deeper insights into key relationships with customers one manager said.

How is Data Visualisation Used in Cyber Security?

In the project of developing a web-based dashboard for ICO datasets of data protection complaints, data visualization plays a crucial role in understanding and addressing cybersecurity challenges. As security analysts and decision-makers seek to manage risks, identify vulnerabilities, and protect information, effective visualization of complaints becomes essential. Given the sophistication and volume of cyber threats, the modern era demands that analysts and decision makers can effectively visualise and understand vast amounts of data, as they seek to manage risk, identify vulnerabilities and threats, and protect sensitive information. “People are habituated to neglecting threats and cyber-crimes that are happening around us thus, leading to more threat-prone days”. B. V. Vikas, N. S. Karthikeya. et al, 2023). Hence, the ability to convey a message that can inform experts in the field, but also everyday members of the public who are also victims of cyber-crime is paramount.

This is helpful for organisations too as discussions of keeping information safe are becoming increasingly commonplace in boardrooms, and it's vital to make business leaders understand that it's not just another conversation. This becomes increasingly significant because reported breach and complaint data now identifies companies by name, removing the anonymity they once had. Traditional metrics can be overwhelming with numerous numbers and tables, potentially causing important details to be missed. Data visualization effectively highlights key points, emphasizing cybersecurity importance. (Balakrishnan, B. 2015). This can then be analysed and evaluated by questioning ourselves with statements like “What are the causes of the incident”, “is there any pattern to notice”, and “how it can be prevented”. This in turn can help us make informed decisions that can help lead to increased security and new regulations. These same questions will be important when assessing the ICO complaint data and what information we can take from it when visualised.

There is a specialised field within cyber security especially for visualisations called VizSec that focuses on the development of visualisations to address complex data analysis problems in cybersecurity. (Staheli, Diane. et al, 2014). The goal of VisSec is to improve a security analyst's ability to recognize threats, make decisions, or take actions by providing informative and actionable visual representations of security data. However, they have said “novel cyber visualizations we have observed to date are either too complex or too basic for the intended users, or too rigid to adapt to different workflows and missions”. Practitioners in the field of VizSec have emphasized the importance of visualizations that strike the right balance between complexity and simplicity which highlights the importance of designing the dashboard with a user-centric approach, ensuring that the visualizations are intuitive, insightful, and adaptable.

The Implementation of Visualisations for Data Compliance

Visualisations are also starting to be used in conjunction with data compliance policies in order to help users understand them better since they are so long winded and complex. Normally, a regular person cannot understand them which results in an ineffective consent procedure which organisations can take advantage of. To counter this, policy makers are attempting to make use of images and graphs to assist the reader. For example, the use of icons as an aid for consent forms and policies. (A. Rossi, M. Palmirani, 2017). This makes use of the emerging study of legal design which focuses on making legal information more understandable and accessible to end-users. This can be directly linked to the proposed task of making a web page displaying ICO data protection complaints as the legal data will be presented in an intuitive and aesthetically pleasing manner that can be understood by those who do not possess legal skills alongside as well as those that do. In doing so the visualisations help make “abstract concepts easier to grasp”, (Rossi, A. 2019) since people have become much more used to visual information than they did only a few decades ago. (Boehme-Neßler, Volker. 2011)

Risk Assessment

Risk assessment plays a pivotal role across various domains that help organizations identify, evaluate, and mitigate risks that would impact their operations, assets or stakeholders, (Irwin, L. 2020), where “The result of the assessment of the total risk gives grounds for deciding on the acceptability of their level”. (V. Mokhor. et al, 2020). This analysis helps in making decisions and distributing resources to manage risks effectively. There are two main ways to do this: qualitative and quantitative. Qualitative methods rely on expert opinions and rankings, while quantitative methods use numerical data and statistics for a more precise understanding of risks. (Tan, D. 2002). Each method has its own advantages and disadvantages. (Kust, M. 2023). In cybersecurity and GDPR compliance, it's important for organizations to consider all risk factors, not just cyber threats. The GDPR emphasizes that personal data can be at risk from accidents or unlawful actions, highlighting the need for a comprehensive approach to risk assessment. The implementation of risk assessment methodology involves several steps, including identifying system performance measures, linking these measures to quantities, collecting relevant data, assessing uncertainties using probabilities, and ultimately calculating and predicting suitable risk levels. This structured approach ensures that organizations can effectively manage risks and enhance their overall resilience in the face of potential threats. (Aven, T. 2012)

Chapter 4

Methodology

Development Lifecycle

When initially considering the development lifecycle for the web-based dashboard project focused on ICO datasets of data protection complaints, I opted for the Incremental Prototyping Model. The reason behind it was that this would allow me to manage the project and divided into small, manageable increments or iterations. Each iteration would consist of the completion of a subset of features or functionalities, allowing for continuous refinement and improvement over time. This would also allow flexibility such that I have the ability to go back to parts of my work and improve the prototypes made. This is tricker to be done with other models such as the Waterfall Model.

However, as development commenced and the project began to take shape, it became apparent that the Agile Scrum method would be better suited for several reasons. Firstly, while Incremental Prototyping focuses on incremental development and refinement, this ended up being not as ideal as I had initially thought since I did not consider implementing changes to the project and additional elements that would be added. For example, when receiving feedback from the weekly meetings with the consultant, the implementation of new features to the project, such as risk analysis, would clash with the use of the Incremental Prototyping Model. Hence, given the dynamic nature of the project and the need for frequent feedback and adjustments, the Agile Scrum methodology offered a more flexible and responsive framework for development.

Secondly, the project was broken down into smaller sprints for continuous feedback, adaptability to changes, and early delivery of requirements. This approach helped to mitigate risks and avoid setbacks, such as learning to use toolkits or errors and bugs during development. Each sprint involved analysing tasks, designing, developing, testing, and critiquing for improvements or changes, resulting in a fully-fledged system for the end product.

Development

The development throughout the project followed the Agile Scrum methodology which made use of sprints. Each sprint consisted of a different part of the project that was to be researched, developed, and tested at the end. The consultant would also ensure that that the sprint is going according to plan and provide feedback.

Sprint 1 – Literature Review

In Sprint 1 of our Agile Scrum methodology, my focus was on research and literature review to lay the groundwork for our web-based dashboard project. This phase was vital for me to grasp the project's scope and requirements fully. I conducted extensive research covering topics like data protection, data visualisation techniques, how visualisations are using in the cyber industry and the utilisation of tools that I may require such as the Avaiga Taipy Python library. By, undergoing this literature review, I gained valuable insights that guided the subsequent stages of development.

Learning Python

I initially planned to use D3.js and JavaScript for my project but realized that building the entire website from scratch using these technologies would be time-consuming. So, I opted for a library that provided pre-built components and hosting solutions. However, this required me to learn Python, which I saw as a worthwhile investment due to its efficiency in web development and visualization tasks. I used resources like the official Python documentation, and platforms such as Codeacademy and W3Schools to learn Python for this project.

Avaiga Taipy and Additional Libraries

This resulted in me taking advantage of Avaiga Taipy, a powerful Python library. It was a fundamental instrument in allowing me to develop a web-based dashboard that was custom-made for the ICO data protection complaint datasets since it consisted of the necessary visualisations that I required as per my prior research of visualisations needed and mock-ups. Taipy also provided necessary pre-built interactive features that my specific data visualisations required which reduced the workload in learning and implementing interactive elements. This, in turn, allowed me to analyse and interpret the data effectively. The library also comes with hosting, allowing me to deploy the dashboard without setting up a separate hosting environment. This saved time and resources, enabling me to focus on improving the dashboard's functionality and appearance. (See Appendix I for Tools used)

Backup and Version Control

Initially, I stored the project code and documentation on OneDrive Cloud. However, upon being reminded of the benefits of using GitHub, I transitioned to saving the code on GitHub regularly. This switch provided a more organised and collaborative environment for the project and ensured version control and easier access.

Requirements & Use-Case Specification

In order to guarantee that our project includes all the necessary features and can potentially outperform other applications, it is necessary to conduct a detailed analysis of the product requirements and use-cases. In my role of the main developer, I have made a well-organised list of components for the app’s dashboard and estimated new functions in terms of their priority. In addition, I have created a table with the project’s features and rated them according to the priority. This approach helped me to understand the needs and preferences of the end user and ensure that I will invest my time in developing the optimal set of features. (See Appendix D for Requirements & Appendix E for Use-cases)

Design

Creating a mock-up of the dashboard and visualisations is an essential part of the development process for multiple reasons. It creates a visual prototype that demonstrates the intended design and function of the tool and allows stakeholders to provide better feedback on the project's direction. By creating this mock-up, it helped me visualise how I wanted to lay out the page as well as what visualisations to display. Additionally, these mock-ups help discover early usability and design issues that would require fixing before the development process becomes too costly in terms of time. This, in turn, helps stick to the project plan and requirements. (See Appendix H for Mock-ups)

Sprint 2 – Analysis of Data Preparation and Cleansing Procedures

In Sprint 2, making a dashboard for data protection complaints requires careful cleaning and preparing data for accuracy. By implementing the methodology researched in the literature review shows how steps like investigation, standardization, and de-duplication stages help improve data quality and avoid mistakes. (K. H. Prasad, et al. 2011). Using tools like Pandas in Python made this work easier.

During the investigation stage of the data protection complaints dataset, thorough auditing was be conducted to make a more detailed analysis of the dataset. This analysis will be conducted to identify different types of errors and patterns present in the data required for assessing the data's quality. The implementation of the investigation stage is linked to the need for a detailed examination of the dataset to discover inconsistencies, inaccuracies, or gaps in the dataset. Various activities, such as data profiling, will help identify the dataset's aspects and characteristics. The results of the investigation stage will help us prepare for the next standardization stage where in this process is to implement the required measures for standardizing the data and improving quality.

After completing the investigation stage, I then focused on the Survivorship stage, which is crucial in identifying the data that needs to be retained from different datasets. Although it is usually the final stage, I have prioritized it as the second stage. This is because each dataset has notable differences, including variations in column order and names. Therefore, it is essential to address these inconsistencies before proceeding with data formatting. This pre-emptive approach ensures optimal efficiency and accuracy in our process, formatting only the data that will be used for the data protection complaints dashboard. Ensuring columns that are not fully populated as well as aligning columns that are not in the same order between the datasets ensures uniformity for easy standardisation in the next stage.

	A	B	C	D	E	F	G	H	I	J	K
1	Status	Legislation	Case Reference	Date Received	Received FY	Received Qtr	Received FY Month	Completed Date	Completed FY	Completed Qtr	Completed Month
2	Completed	Hybrid	IC-72725-F9V0	12/11/2020	2020/21	Q3	08-Nov	01/01/2021	2020/21	Q4	10-Jan
3	Completed	Hybrid	IC-71646-P6T8	16/06/2020	2020/21	Q1	03-Jun	01/01/2021	2020/21	Q4	10-Jan
4	Completed	Hybrid	IC-51808-P4Y0	29/06/2020	2020/21	Q1	03-Jun	01/01/2021	2020/21	Q4	10-Jan
5	Completed	Hybrid	IC-72196-R7Y9	13/05/2020	2020/21	Q1	02-May	01/01/2021	2020/21	Q4	10-Jan

Figure 2: Dataset Columns A-K

L	M	N	O	P	Q	R	S
Sector	Sub Sector	Decision Articles	Decision Primary Reason	Submitted About Account	Decision	Decision Detail1	Decision Detail2
Retail and manufacture	Supplier of services	Art 32	Art 32 - Security of proce	Arnold Clark Automobiles	No Further Act	No action	Insufficient informati
General business	Motor Trade	Art 15	Art 15 - Right of access	Avis Budget EMEA Ltd	No Further Act	No action	Insufficient informati
Land or property services	Building and property	Art 15	Art 15 - Right of access	BAM Construct UK Limitec	No Further Act	No action	Insufficient informati
Online Technology and Te	Service providers	Art 15	Art 15(3)(1) - Provide a cc	British Telecommunicatio	No Further Act	No action	Insufficient informati

Figure 3: Dataset Columns L-M

During the Standardization Stage, the main focus is to transform the available data into a uniform and consistent format to be processed into visualisation for the data protection complaints. The separate datasets would be merged, however, it becomes apparent that certain columns, such as "Decisions" and "Submitted about account," have variations in capitalization for identical data entries, such as "No further action" and "No Further Action." To address this issue, Python case functions, such as str.title(), are utilised to standardize the text format and ensure consistency across the dataset. Similarly, the "Date Submitted" and "Completed Date" columns, originally in string format, need standardization to a date format using Pandas. This conversion provides the ability to make use temporal visualizations and analysis.

During the final stage of the data cleaning process, known as the De-duplication Stage, the standardized data from the previous stages is used to identify similar or duplicate records that may exist within or across datasets. However, in our particular dataset, no duplicate entries were found since each case was distinct in its separate dataset. If a case ID appears more than once, it indicates that the case was reopened for further investigation, rather than being a duplicate entry.

Sprint 3 – Implementation of Visualization Techniques

The dashboard for data protection complaints was developed using dynamic visual elements created with Taipy Python library. A list of potential visualisation analysis documentation and mock-ups were used to display these visualisations, outlining the chart types, values in the dataset, and intended insights. The backend comprises Python methods for processing data and making visualisations, while the frontend is published as pages using Markdown code. The production cycle followed principles outlined in Andy Kirk's Data Visualization book, encompassing Data Representation, Interactivity, Annotation, Colour, and Composition. (Kirk, A. 2019). (See Appendix F for Potential Visualisations Documentation)

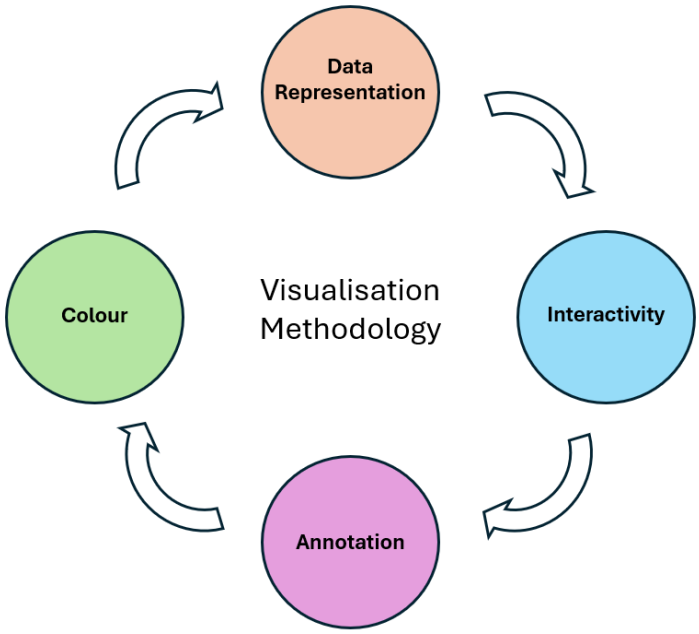


Figure 4: Visualisation Methodology Implemented

When it comes to the data representation level, I used visual encoding to construct the charts and design the dashboard. Marks and attributes were used to decode the data for users to visualize. I used the “CHRTS” methodology (Kirk, A. 2019) to select the right type of chart to represent the data and the insights based on the nature of the dataset and the target audience to establish clarity and understandability. This will aid in the sense that the audience needs to see both overview information presented at once and examine a certain point thoroughly at another time. The dashboard will contain a collection of chart types to choose from the available options and the use of editorial thinking will be used to showcase relevant composition choices to present information in a unified page layout, ensuring it delivers valuable insights in a user-friendly manner.

Moving to Interactivity, this is important for engaging user experiences. Leveraging the possibility to make use of dynamic charts, the implementation of Taipy's selection and filtering functionalities within the graphs allow users to interact with the data protection complaints, allowing users to tailor their analysis based on specific criteria, enhancing their ability to derive useful insights.

The next stage, Annotation, involves enriching graphs with additional context or explanations to aid users in interpreting the visualizations effectively. This stage is pivotal for the dashboard in bridging the gap between the data protection complaints representation and user understanding. The implementation of clear and informative annotations to the charts, ensures that users could gain the insights conveyed by the visualizations effortlessly. Whether through titling the graphs or providing detailed explanations for key elements, effective annotation enhances the usability and accessibility of the dashboard.

In the colour stage, careful consideration of colour theory is essential to ensure effective interpretation of visualizations as if used incorrectly, can impact how insights are perceived. When incorporating colour into the dashboard, I prioritized data legibility, selecting colours that facilitate differentiation of quantitative values. Opting for the RGB colour model over CMYK was deemed appropriate for digital work. Additionally, I utilized the HSL (Hue, Saturation, Lightness) model to differentiate colours, providing flexibility and enabling a wide range of colour variations. This approach offers versatility in selecting desired colours, enhancing visual appeal and clarity.

Lastly, Composition addresses the spatial arrangement and hierarchy of visual elements within the dashboard. This stage is instrumental in guiding user attention and facilitating seamless navigation. By scaling down smaller graphs like pie charts and strategically placing them alongside other visuals, I optimized space utilization while maintaining coherence and readability. Incorporating design principles such as partial visual cues further enhance user engagement and encourages exploration of the dashboard's content.

Sprint 4 – Implementing Risk Assessment

To evaluate the risks surrounding data protection complaints against various organisations, careful consideration was given to various methodologies, with qualitative and quantitative risk assessments being the primary contenders. Within risk assessment, the qualitative can be used as an effective tool for subjectively assessing and giving value to the different variables, based on expertise, experience, and individual perception. (Tan, D. 2002) . However, because of the high volume of complaints within the data protection dataset, this assessment appeared to be unreasonable in this case. In particular, the subjective assessment may contain certain biases and inaccuracies, which can distort the risk data by undermining its reliability and validity, especially when it comes to a large amount of information. (Kust, M. 2023). Consequently, the focus shifted towards quantitative risk assessment, which offers a more objective and systematic approach to analysing risks by utilizing numerical data and mathematical calculations.

Using the risk assessment methodology highlighted in the literature review (Aven, T. 2012), each key factor, such as sector, sub-sector, decision article, and submitted account, was identified as a system performance measure. These measures were put into equations in order to derive a deterministic model that relates complaints quantities. To obtain individual risks associated with each key factor, details about quantities were collected and then merged to find total risk factor. Uncertainty wasn't factored in because the model relied on precise data and deterministic relationships, making explicit consideration of uncertainty unnecessary, which resulted in my methodology skipping the last part of the risk analysis methodology. This is a quantitative way of looking at how many risks are connected with data protection complaints that organizations would use to rank the risks, accordingly, allocate resources appropriately and come up with corrective actions based on available information. (Kust, M. 2023).

This can be done using the formula:

$$Risk\ Probability = \frac{Decision\ (Action\ Taken)}{Total\ Decisions}$$

Where Risk Probability is the probability of a Risk occurring, Decision (Action Taken) is the number of times the event has occurred in the past, and Total Decisions is the total number of complaints, to calculate the probability of a risk based on past experience.

Subsequently, it was crucial to determine the areas for risk analysis within the dataset. The broader the scope of these areas deemed suitable for analysis, the more granular and detailed the resulting assessment would be.

The sector column provides information on the industry or sector to which the organization belongs, allowing for sector-specific risk analysis. This enables organizations to identify high-risk sectors and tailor their compliance efforts accordingly. Similarly, the sub-sector column delves deeper into specific segments within each sector, uncovering nuanced risk profiles and patterns that may require targeted interventions.

The decision article column highlights the specific GDPR articles breached in the complaints, offering valuable insights into the legal and regulatory implications of non-compliance. By understanding which articles are most frequently violated, organizations can prioritize remediation efforts and implement controls to mitigate future breaches.

Finally, the submitted account column provides data on the frequency and volume of complaints lodged against organizations, indicating the level of scrutiny and potential regulatory action they may face. By quantifying the volume of complaints, organizations can gauge the extent of their exposure to regulatory risk and take proactive measures to address underlying compliance issues.

The overall risk factor was calculated by combining the individual risks from each factor and dividing by the number of risk areas used, then multiplying by 100 to obtain a percentage. The formula for calculating the overall risk factor can be represented as follows:

$$Total\ Risk\ Factor = \frac{Sector\ Risk + SubSector\ Risk + Article\ Risk + Account\ Risk}{4} \times 100$$

This percentage provides insight into the proportion of cases where action was deemed necessary, indicating the severity of issues within that category. Higher percentages suggest a higher severity level, as a larger proportion of cases required action.

However, these risk factors are initially calculated based on how and often cases occur for each category. Yet, the value range for those risk factors may differ significantly due to the nature of data. Thus, I had applied an algorithm called MinMaxScaler, which ensures they take the same scale. MinMaxScaler is a frequently used normalisation algorithm in machine learning and data preprocessing which normalises features or variables to a given range, usually from 0 to 1, while preserving their relationships to one another based on how data is distributed. In our case, MinMaxScaler was used to normalise the calculated risk factors in the data protection complaints for the sector, decision articles, sub-sector, and submitted accounts. MinMaxScaler uses this formula:

$$Risk\ Scaled = \frac{Total\ Risk\ Factor - \min\ (Total\ Risk\ Factor)}{\max(Total\ Risk\ Factor) - \min\ (Total\ Risk\ Factor)}$$

This led me to transform the risk factors within the range of 0 and 100, Making the risk of all the cases easy to interpret and comparable against one another. Seralouk (2022). This relatively scaled representation of risk factors will help the users understand the severe cases within a category of implementation and therefore, make decisions based on the expected level of risk factors.

Sprint 5 – Implementing the GDPR Compliance Advisor

Within the GDPR compliance advisor page, it allows users to select their organisations sector and sub sector and provide recommendations on how to ensure they are compliant with GDPR, based on data protection complaints on organisations within those sectors.

In order to provide appropriate and correct advice to organisations, a study of the articles of the GDPR was the first step towards understanding the general principles and rules of data protection. (ICO. Personal Data). The GDPR consists of numerous articles to protect individuals' privacy rights and regulate organisations' processing of personal data. (Ekdahl , A. and Nyman, L. 2018). Therefore, familiarity with these articles was a prerequisite for developing proposals for

the most common compliance issues that organisations in various sectors might be more focused on resolving. Combining the insights of all the articles with knowledge and understanding in the field of cybersecurity, the proposals cover the most vulnerable areas. (Y. -S. Martin and A. Kung. 2018)

Data aggregation began by grouping the dataset by sector and subsector using the pandas library in Python. The most common GDPR breaches were identified within each subsector. The processed data was structured into a nested dictionary named decision_dict for efficient data retrieval and GUI integration. Initial dropdown values were pre-set for user convenience.

This then involved identifying the key components necessary for effective compliance guidance, including GDPR articles, their descriptions, and associated recommendations. This information was organised into a structured JSON file, to facilitate easy retrieval and processing. Once the data structure was established, the implementation of the advisor began with the creation of a function responsible for retrieving GDPR articles based on user inputs. This function utilised a search mechanism to match the user-selected sector and subsector with corresponding GDPR articles. If an exact match is found, the relevant article details, including the identifier, description, and recommendations, are extracted, and stored in the advisor's state.

In cases where an exact match is not found, the advisor has an implemented fallback mechanism using regular expressions to extract article numbers from the provided decision primary reasons. This approach ensured that even if the exact reason was not available, the advisor could still provide relevant guidance based on the extracted article numbers.

Finally, error handling mechanisms were implemented to address scenarios where neither the exact reason nor the extracted article number matched any entries in the GDPR database. In such cases, appropriate error messages were generated to inform the user that the requested information could not be found.

Unit testing

Thorough testing was conducted on each individual sprint throughout the development of the project. Since each individual part of the system was created separately, each unit was uploaded to the Taipy cloud to ensure that it was working as intended. Before uploading, each unit was tested locally to ensure the outputs and processes are correct and accurate.

Integration testing

Due to the incremental nature of the project development, after each unit was completed, other completed units were tied together into one coherent system and tested to ensure they are working in conjunction correctly. This was mainly ensuring that each page including the dashboard, organisational risk and GDPR compliance advisor were integrated together effectively.

System testing

After finalising the overall system, comprehensive testing was conducted to verify both functional and non-functional requirements, ensuring high standards of performance and functionality across the entire project. Details of this testing can be found in the testing documentation. (See Appendix K for Testing)

Chapter 5

Results

Dashboard Design

The design of the dashboard was prepared with detail to ensure a high level of user experience while navigating it. This was achieved through the use of the preliminary mock-ups that served as a blueprint into how each section of the site was organized. These models enabled me to prepare the draft text and also assisted in chart positioning. The “Getting Started” page was also drawn in detail to describe the features of the application and give a brief view of each available design. This ensures that users using the site with zero background knowledge are aided in using the dashboard. The navbar with the list of pages ensured that users could navigate around the site quickly. The main page on the dashboard was designed to convey a story through the graphs. I carefully selected each chart to provide users with a gradual understanding of the data related to protection complaints. To achieve this, I utilised two different methodologies: the general data visualization methodology by Andy Kirk, as described in his Data Visualization book, (Kirk, A. 2019) and the Security Data Visualization methodology paper by GIAC as seen in the Literature Review. (Balakrishnan, B. 2015) Using these two techniques, I obtained a comprehensive and specific view of how to visualise the data in my project.

Data Preparation and Cleansing

Using the cleansing methodology researched, (K. H. Prasad, et al. 2011), I started with the investigation phase for developing a dashboard to address data protection complaints, I conducted a thorough audit of the datasets containing these complaints. This analysis revealed that the datasets were separated, covering a period from Q4 2020/2021 to Q3 2023/2024, and consisted of 12 different datasets. This separation posed a significant challenge in consolidating the data for visualization purposes and when merged would result in 101,896 rows. I also noted inconsistencies in the column order across the various datasets, which required careful merging to ensure that the data is aligned in the master dataset. While all values were filled within the datasets, I observed that some entries were labelled as "unassigned," which represented information that was not known by ICO regulators, such as sector type. I opted to keep this data as it was a valid data value. Moreover, variations in column structures, including the presence of new columns in some datasets, necessitated the dropping of irrelevant columns to maintain data consistency across the merged dataset.

Figure 5: Data Preparation and Cleansing Methodology

The next step focused on the implementation of the Survivorship stage. This involved identifying and eliminating excess columns in the original datasets that were unhelpful for any visualisation of data protection complaint cases. As a result, it made the original dataset more usable and helped to create a more focused dataset structure. For instance, the column 'Received FY Month', was irrelevant due to its inconsistent data and lack of relevance for visualization. In this case, the column was dropped from the final merged dataset. Moreover, another obstacle faced at this stage was the inconsistency of the order of columns in the datasets. In other words, to create a consistent and appropriate merged dataset, it was essential to carefully align columns and check their order and placement. Since the columns were ordered differently in each dataset, this process had to be performed manually, which was effectively supported by Excel. While being time-consuming, the manual alignment was critical to ensure that each column is placed properly and ready to maintain integrity within the entire merged dataset.

The standardization stage as discussed in the methodology was effectively implemented using the Pandas library, (Pandas tutorial), significantly enhancing the integrity and usability of the data protection complaints dataset. By applying "data["Decision"]= data["Decision"].str.title()", instances of duplicated values due to capitalization inconsistencies, such as "No further action" and "No Further Action," were rectified, which ensured uniformity and accuracy. This approach not only merged identical values but also addressed potential inconsistencies that may have gone unnoticed within the same column. Similarly, the "Submitted About Account" column underwent standardization using a similar function to capitalize the first letter of each word, reducing the occurrence of duplicate entries of organization names. This standardization process was especially beneficial for the Risks page, when sorting cases by organizations and ensuring that all cases against a specific organization are aggregated under a single name. Furthermore, the conversion of the "Date Submitted" and "Completed Date" columns from string to date format using pandas provided the dataset with standardized temporal data. This conversion enabled the use of temporal visualizations and analysis, enabling further insights to be gained from the data.

Dashboard Visualisations

This section showcases the dashboard's visuals, revealing details from the ICO complaints data. Before, this info was just raw data and not presented graphically. Now, interactive charts unveil key patterns and trends, aiding data protection decisions. These visuals give stakeholders evidence and a clear view to enhance GDPR compliance and manage risks.

Statistical Analysis of Data Protection Complaints

Users can obtain crucial insights into data protection complaints through the brief statistics on the dashboard while making decisions and setting priorities. Take for instance, total cases including reopening of 726 out of 101,895 complaints. This gives an insight into the magnitude of the problem and recurrent non-resolutions. There is also the statistic on the average time to process a complaint. Knowing that it takes on average 106 days to process a complaint can assist users, organizations, and regulators in finding systemic issues and optimizing resource allocation. Additionally, analysis of reopened cases enables firms to fine-tune their data protection policies and responses. This also helps authorities develop better enforcement mechanisms and identify typical patterns that may necessitate further investigation. Individuals benefit from this information by recognizing the importance of active communication with organizations to ensure satisfactory resolution of data protection issues and the protection of their rights and interests.

Figure 6: Statistical Visualisations Within the Dashboard

Complaints per Sector

I used Andy Kirk's advice (Kirk, A. 2019) to create a bar chart of the number of complaints per sector. The chart helps identify sectors with high complaint volumes, and sorting was implemented to aid readability. Although axis labeling was limited due to long sector names, Taipy's interaction allows users to see specific values for each sector. The bar chart was created by grouping data protection complaints by sector using Python's Pandas library. The groupby() function aggregated the data based on unique values in the 'Sector' column and counted complaints using the .size() function. The resulting series was converted into a dataframe with 'Sector' and 'Count' columns using the .reset_index(name='Count') method. To improve clarity, the dataframe was sorted in descending order by complaint count using the .sort_values() function.

```
dataset_2 = dataset.groupby(['Sector']).size().reset_index(name='Count').sort_values(by=['Count'], ascending=False)
```

Figure 7: Code for Setting Up the Complaints per Sector Chart

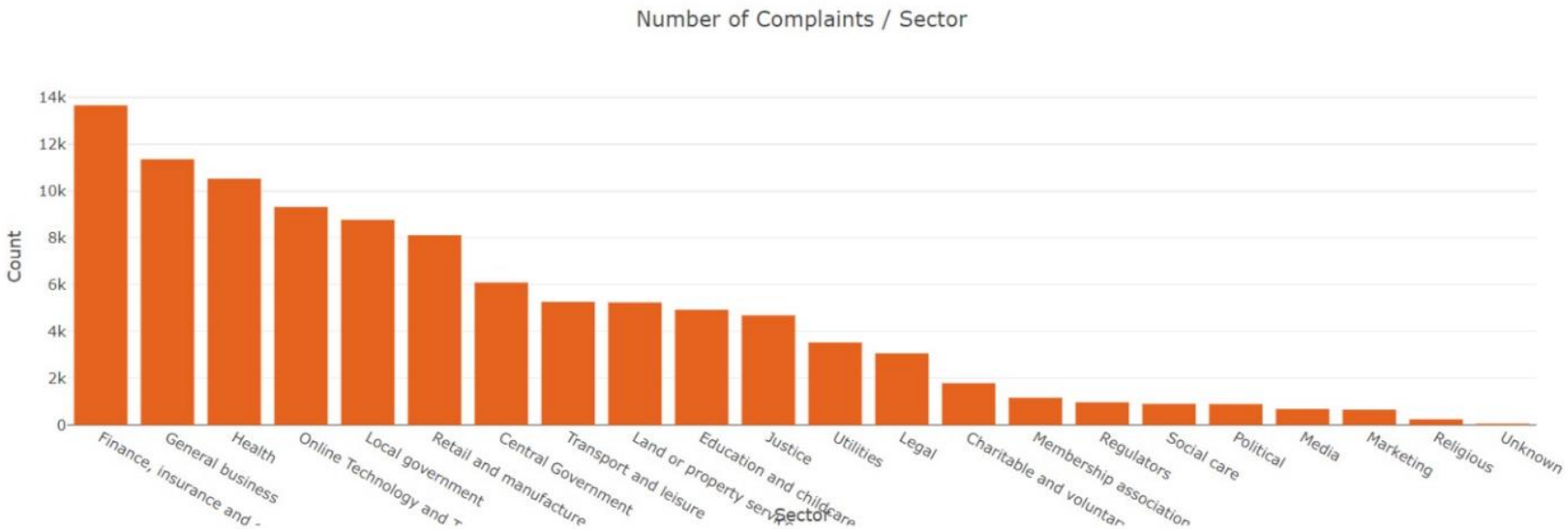


Figure 8: Complaints per Sector Chart

Using this chart, we can gather a wide range of useful information. Firstly, the bar chart provides a straightforward and intuitive representation of the distribution of complaints across different sectors. By visually comparing the heights of the bars, users can easily identify which sectors have the highest and lowest numbers of complaints.

The prominence of the finance and insurance, general business, and healthcare sectors indicates several noteworthy insights. Firstly, the finance and insurance sector as the highest position suggests that it is particularly susceptible to data protection complaints, possibly due to the sensitive nature of financial information handled by organisations within this sector. This could be due to several reasons such as the storing and handling of personal banking details as well as insurance claims and financial transactions. However, this could also be due to the wide range of use and necessity of this type of data by many companies that may use a financial organisation as a middleman to process transactions.

Similarly, the high number of complaints within the general business sector may stem from its broad scope, encompassing a wide range of businesses and industries, each with its own set of data protection challenges. The generalised nature of the sector could lead to a wide array of complaints related to data handling practices.

Lastly, the healthcare sector is a significant area which underscores the importance of data protection in the handling of sensitive medical and personal information. Healthcare organisations are entrusted with highly confidential data, such as medical records. This may result in users seeking information from organisations on the data they have about them (Article 15 – Right of access) or the removal of their data (Article 17 – Right of erasure). Sectors that are not as high in terms of complaints may be due to the more nuanced nature of these areas such as religious or social care, however, should not be discarded as the percentage of complaints may result in more data protection breaches.

While the bar chart effectively visualizes sectors with the highest frequency of data protection complaints, it's essential to recognize that sectors with fewer complaints aren't necessarily less significant. It's plausible that these sectors are more niche, focusing on specific activities that may generate fewer complaints. To delve deeper into understanding the severity of complaints across sectors, I explored creating another graph to illustrate the distribution of action taken versus no action taken within each sector.

Overview of Complaint Resolution

To provide a comprehensive overview of complaint outcomes, I incorporated a pie chart illustrating the overall decisions made on complaints. This visual representation highlights the distribution of decisions, offering insight into the proportion of cases resulting in action taken versus no action taken. The implementation of the pie chart follows a similar process to the bar chart, as both involve grouping and counting data to visualise distributions. However, unlike the bar chart, the pie chart does not require sorting of the data since it represents categorical proportions rather than ordinal relationships, therefore, grouping the dataset by the Decision column.

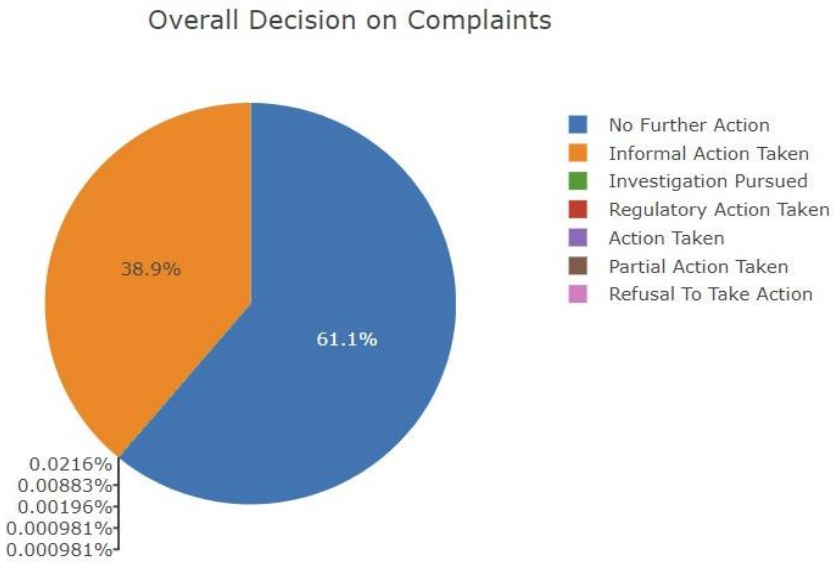


Figure 9: Overall Decision on Complaints Pie Chart

I have refined the design of the pie chart by following advice from Andy Kirk's book on Data Visualization. To avoid overcrowding I avoided filling small, angled sections with other information, in order to make charts easier to read. Taipy also automatically places the first slice at 12 o'clock and a counterclockwise arrangement of sectors helps us interpret it better. Despite the fact that only two values take up most of the graph, it is important to keep lesser ones in order not to delete data as this could result in incomplete findings and partial analysis. This is addressed through adding a line displaying more categories in a smaller area with their corresponding percentages even though they are less than zero. Furthermore, Taipy allows for filtering options within the pie chart, enabling users to hide larger values temporarily to focus on differentiating between the smaller values more effectively.

The analysis of the pie chart, particularly the prevalence of "No Further Action" (61.1%) and "Informal Action Taken" (38.9%), provides valuable insights into GDPR compliance practices. Firstly, the majority share allocated to "No Further Action" could indicate the complex nature of data protection complaints adjudication. In other words, one may conclude that there are a lot of complaints that cannot be verified due to a lack of evidence or does not actually fall under a specific GDPR article. While it may reflect a rigorous screening process aimed at conserving regulatory resources and prioritizing high-impact cases, it also raises concerns about accessibility and transparency in the complaint resolution process. Thus, complainants may feel that their concerns are not taken seriously or, at least, are not clearly scoped, which damages institutional and regulatory trust.

On the other hand, the allocation of 38.9% of “Informal Action Taken” can indicate the proactive behaviour towards data protection issues. This category includes measures that vary from simply a warning, to making several recommendations about an organization’s improvement. When opting for the informal measures, the regulatory body (ICO) signifies its intention to enable organizations to secure compliance, remedy the infractions, and engage in a constructive manner with the data controller. This category does not only help to eliminate the immediate risks but will also help to educate and help organizations to be proactive regarding data protection in the future.

Both complaints submitters and organizations can benefit from the information provided by pie charts. For users, understanding the prevalence of "No Further Action" underscores the importance of providing substantial evidence and clear documentation when lodging complaints. Regulatory bodies can provide guidance on what constitutes sufficient evidence and how to articulate concerns effectively which can help users to navigate the complaint process more confidently. Likewise, this data can be useful for organizations to improve their data protection practices in order to reduce regulatory risks. Knowledge of the typical violations resulting into non-compliant actions can enable organizations implement focused procedures as it will also lead to compliance improvement process by promoting a proactive culture of safeguarding personal information.

Distribution of Decisions per Sector

The stacked bar chart I implemented attempts to offer a clear view of how data protection complaints are handled in different sectors, combining insights gained from the bar chart on complaints per Sector and Pie chart on Decision outcomes and adding a new level of analysis, it shows the proportion of actions taken versus no actions taken, helping stakeholders understand how complaints are addressed within them. By allowing users to sort and filter data by year, this chart provides valuable insights into trends over time, aiding in future decision-making and planning.

In implementing the stacked bar chart to illustrate the distribution of 'Action Taken' versus 'No Action Taken' per sector, firstly, I utilized two functions, taken() and not_taken(), to categorize each entry in the dataset based on whether an action was taken or not. These functions assessed the values in the 'Decision' column and returned 1 if the decision corresponded to the respective category and 0 otherwise. Next, I aggregated the dataset by the 'Sector' column which grouped the data by sector and summed up the counts of 'Action Taken' and 'No Action Taken' decisions for each. By consolidating the data in this manner, I obtained an overview of the distribution of decisions across different sectors. Subsequently, I computed the total number of decisions for each sector and calculated the percentages of 'Action Taken' and 'No Action Taken' decisions. This step provided insights into the proportion of decisions taken within each sector, enabling comparisons across different sectors.

Moreover, I incorporated functionality to enable filtering the data by year. This feature allowed users to select a specific year, updating the dataset accordingly. By implementing buttons for selecting different years, users could explore how the distribution of decisions varied over time, providing additional context to the analysis.

```
##### Stacked barchart for decision taken per sector over time #####
dataset_18 = dataset.groupby(['Sector']).agg({'Decision (Action Taken)':'sum', 'Decision (No Action Taken)':'sum'}).reset_index()
dataset_18["Total Decisions"] = dataset_18['Decision (Action Taken)'] + dataset_18['Decision (No Action Taken)']
dataset_18['Decision (Action Taken)'] = round((dataset_18['Decision (Action Taken)'] / dataset_18['Total Decisions'])*100)
dataset_18['Decision (No Action Taken)'] = round((dataset_18['Decision (No Action Taken)'] / dataset_18['Total Decisions'])*100)
```

Figure 10: Stacked Bar Chart Code

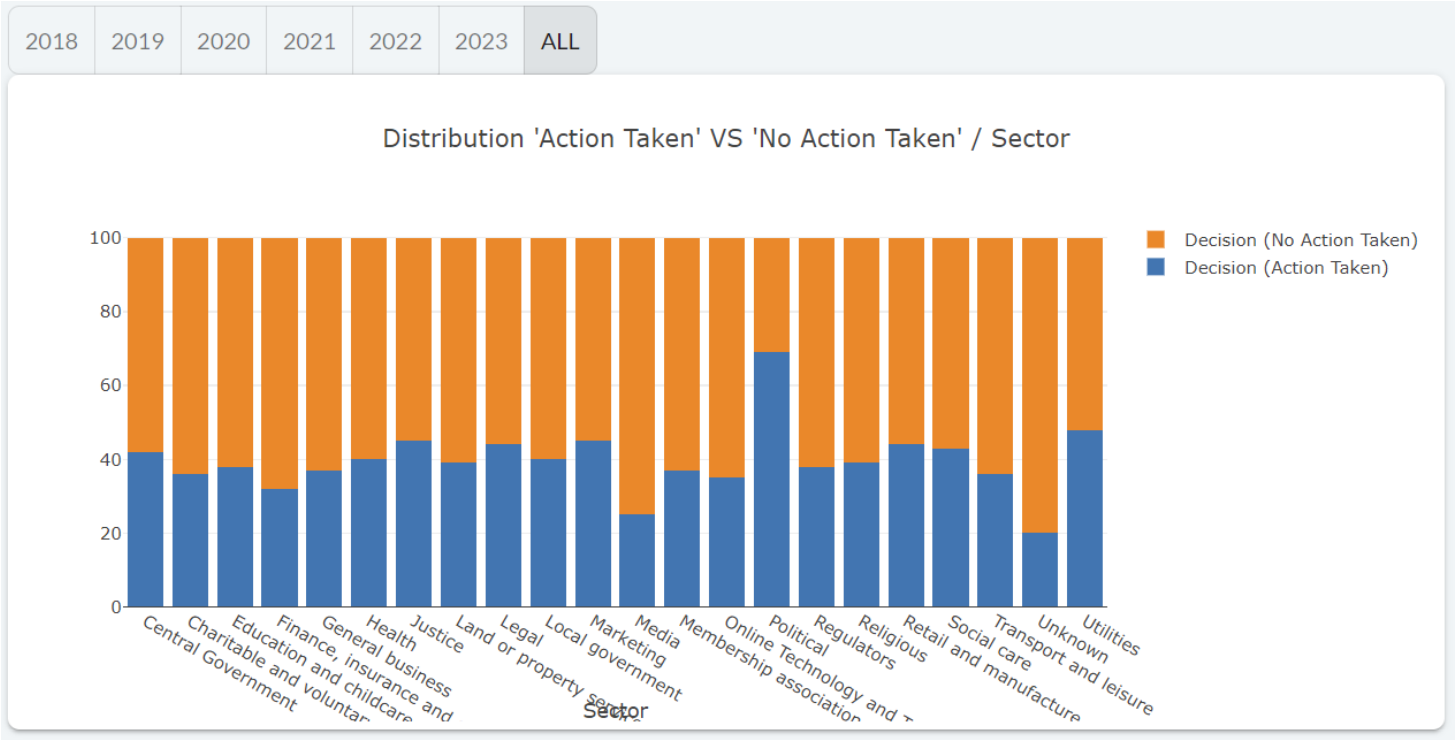


Figure 11: Stacked Bar Chart

I used Andy Kirk's suggestions for stacked bar charts to improve the distribution of 'Action Taken' versus 'No Action Taken' per sector in my implementation. This included addressing the annotation aspect. I used concise value axis labels instead of direct value labelling to avoid cluttering the chart, particularly when comparing multiple major categories, ensuring clarity and avoiding overwhelming the viewer. Taipy uses gridlines to mark key units like 20%, 40%, 60%, 80%, and 100%, which is helpful for a 100% stacked bar chart and adds context to the distribution of actions taken. The default colour scheme in Taipy clearly shows the different colours that were assigned to each stacked part for easy observation of their bar lengths. This improved the representation of categorical nominal data.

When analyzing the aggregated data from all years, it becomes clear that most sectors have over 50% complaints with no action taken. This suggests potential insights into the handling of data protection complaints. The high proportion of complaints resulting in no action taken may indicate a systemic issue with regulatory enforcement or compliance, suggesting an overwhelmed or under-resourced regulatory authorities. This may be due to limited resources, budgets, or conflicting priorities within regulatory bodies, resulting in fewer actions taken for less severe complaints. Additionally, some complaints may lack enough evidence, especially if they involve minor incidents that do not pose risks to data protection, leading to no action taken. In these cases, regulators may prioritize serious breaches or systemic issues affecting data protection.

The differentiation between years did show an interesting pattern in the utilities sector, with the number of regulatory actions increasing yearly from 2020 to 2023. There may be several reasons behind growing regulatory action against utility companies. For instance, utility infrastructure has become increasingly digitized in recent years, which also increases the number of cyber risks. Regulatory bodies may need more points of control to supervise critical infrastructure and protect consumer data. Moreover, utilities store vast amounts of massive data on consumers, with frequent occurrences of sensitive information being stolen from company databases. “The number of cyberattacks on utilities rose steadily between 2020-2022, with the number of weekly cases more than doubling in three years’ time to 1,101”. (King, C. 2023). Heightened regulatory scrutiny reflects the need to address these risks and comply with data protection regulations like GDPR.

Secondly, escalating regulatory actions indicate a growing recognition of the sector's importance in safeguarding sensitive data and critical infrastructure. As utilities handle vast amounts of personal and operational data, regulators prioritize enforcement to mitigate risks to data security and privacy, ensuring resilience in essential services. Companies in the utilities sector can learn from the increasing regulatory actions and improve their data protection capabilities. They can prioritize the most common areas of non-compliance and invest in cybersecurity technologies to mitigate this. They could also conduct comprehensive risk assessments and enforce effective data protection policies, training their employees and seek guidance from other industry players and regulatory bodies.

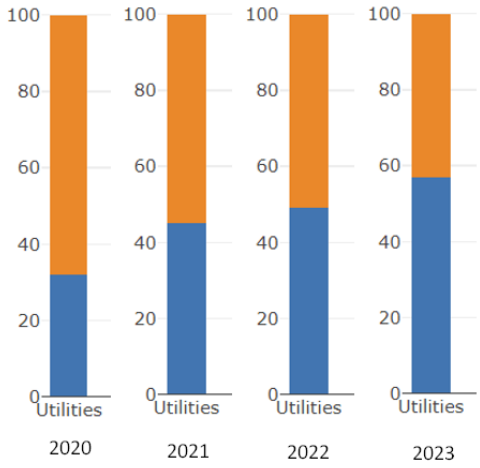


Figure 12: Changes In Utilities Sector Chart

Another insight we can gain from the stacked bar chart is the decline in the number of no regulatory actions taken against the central government sector each year. This unveils a notable shift in regulatory enforcement dynamics. However, it's crucial to consider the context provided by other graphs, such as the line chart tracking complaints over time. In particular, the limited number of complaints submitted in 2018 and 2019 may influence the accuracy of the results for these years. Nevertheless, the trend of decreasing no regulatory actions persists from 2020 to 2023, indicating a consistent pattern.

We're seeing fewer compliance breaches, possibly due to stricter regulations and increased focus on data protection. Governments are likely taking proactive steps to address compliance gaps and uphold public trust. Additionally, better regulatory oversight and data analysis could be helping to identify and investigate breaches more effectively.

The trend towards stronger data protection governance and regulatory oversight is beneficial for all concerned parties. For example, users sharing their data with the sector can feel more secure about the protection of their personal data, while the central government can demonstrate their commitment to data protection and regulatory compliance. Regulatory bodies also benefit from increased compliance, reinforcing their role in upholding data protection standards and fostering a culture of accountability and transparency. The decreasing trend in no regulatory actions against the central government sector reflects positive strides towards stronger data protection governance and regulatory oversight.

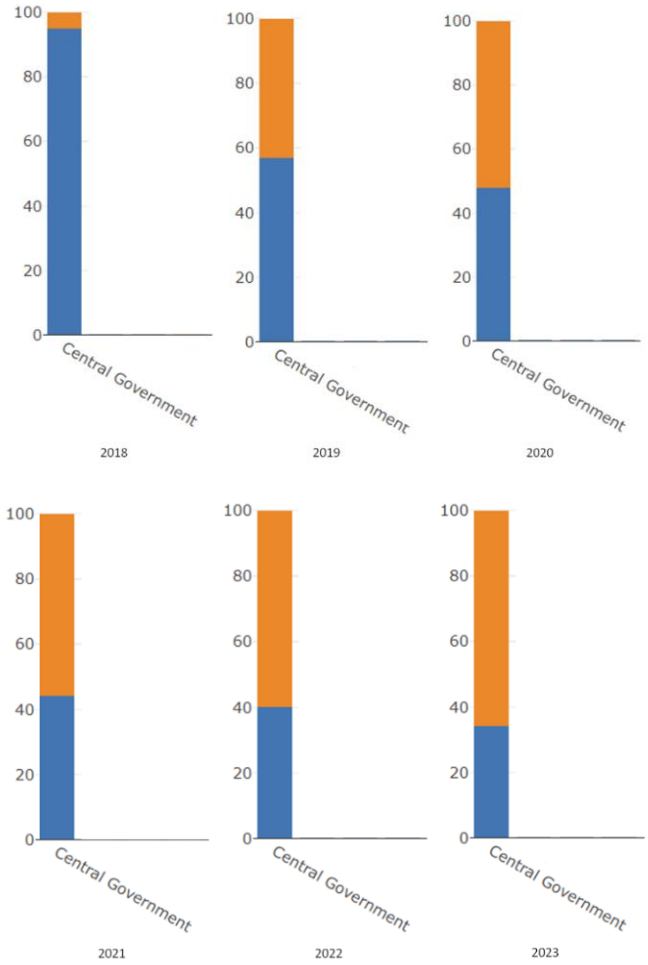


Figure 13: Changes in Central Government Sector in Stacked Bar Chart

Examining the Distribution of Breached GDPR Articles in Complaints

The creation of a pie graph displaying the Distribution of GDPR Complaints arises from a crucial need to comprehend the landscape of data protection breaches. By visualizing the most common GDPR articles breached, the graph aims to be a strategic tool for identifying common vulnerabilities and areas of non-compliance.

The implementation of visualising the pie chart and displaying the Distribution of GDPR Complaints involves several important steps in data preprocessing and visualization. Initially, regular expressions were utilized to parse the "Decision Primary Reason" column, extracting relevant GDPR article names and numbers from each entry. This parsing ensures that only relevant data, such as converting entries like "Art 15 - Right of access" to "Art 15," is retained for analysis. Additionally, the data is grouped by the "Decision Articles" column, enabling the calculation of the frequency of each breached GDPR article. Unassigned values are excluded from the dataset, providing a more focused analysis of relevant data points. To enhance the clarity of the pie chart, less common articles are aggregated into a single category labelled "Other." This collection simplifies the visualization by reducing clutter and highlighting the most common breaches. Moreover, filtering techniques are applied to prioritize significant breaches by excluding articles with low frequencies. By setting a threshold (e.g., 500 occurrences), the visualization focuses on articles with substantial impact, providing a clear representation of the Distribution of GDPR Complaints.

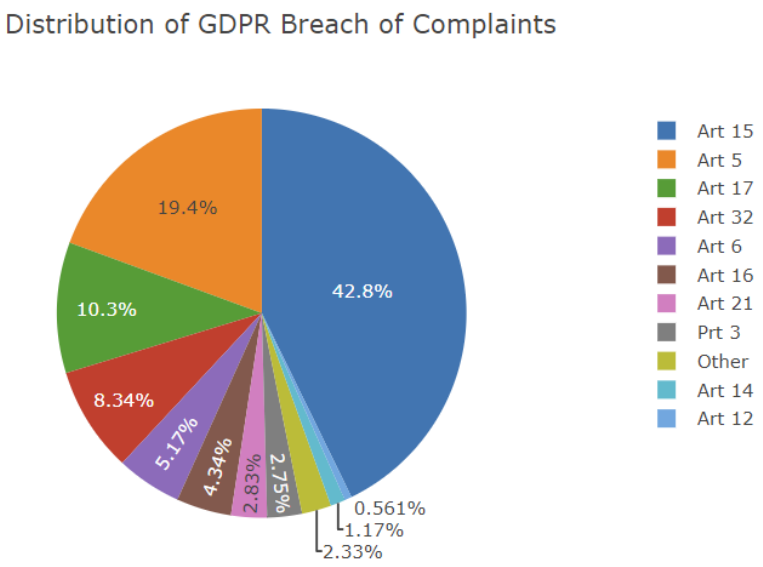


Figure 14: Distribution of GDPR Articles for Complaints Pie Chart

This pie chart visualizing the proportion of GDPR articles breached becomes a useful graph for understanding the areas in which most organizations struggle to maintain compliance with the relevant data protection standards. The chart reveals that Article 15 is almost at the heart of all breaches, comprising 42.8% of all breaches reported. This insight suggests that this article's detailed requirements require further clarification, as the legislative acts covered by these requirements protect an individual's right of access to data for data controllers. The fact that it accounts for such a large fraction of complaints suggests that many people struggle to obtain their personal information from data controllers which indicates a lack of transparency and ineffective data access procedures.

Article 5 follows closely with 19.4% of breaches, a high proportion that indicates a major deficit in compliance with the general data protection regulations. Article 5 lays down the principles of lawfulness, fairness, and transparency in data processing, as well as the accuracy, storage limitation, integrity, and confidentiality of personal data. The significant frequency of breaches indicates the need for organizations to develop comprehensive data processing regulations and policies to ensure they abide by these data protection principles. These breaches can not only expose considerable legal and regulatory risks but also identify the extent to which some of the data processing practices may be unethical.

Furthermore, Article 17 breaches, making up 10.3% of the total, sheds light on challenges related to the right to erasure. This article grants individuals the right to request the deletion or removal of their personal data when there is no compelling reason for its continued processing. The high incidence of breaches under Article 17 suggests that organizations may struggle with fulfilling requests for data erasure in a timely manner or may lack adequate mechanisms to effectively manage and respond to these requests.

Investigating the Temporal Distribution of Submitted Cases

Implementing a line graph to track the timeline of data protection cases submitted offers a concise yet powerful visual representation of the evolving landscape. This visualization aims to illuminate trends over time, providing users of the dashboard with valuable insights for informed decision-making and proactive intervention in data protection challenges. This was done by using the 'Date Received' column and plotting its entries.

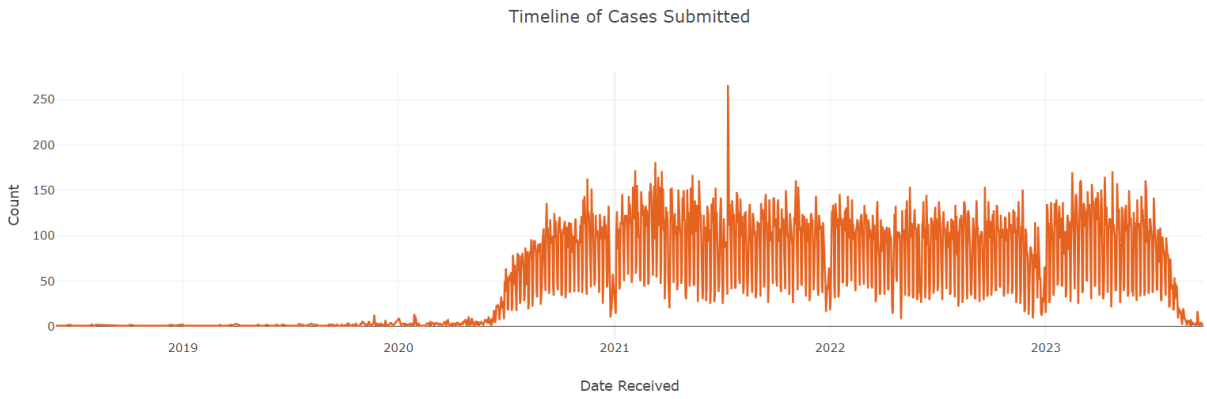


Figure 15: Timeline of Cases Submitted Line Graph

Notably, the graph reveals a significant reduction of cases during 2018, 2019, and the beginning of 2020, followed by a notable uptick in submissions post-Q4 of 2020. This can be attributed to the timing of the ICO's release of the data protection complaints dataset. Prior to this release, comprehensive data on complaints may not have been readily accessible or available for analysis, resulting in the absence of cases during the earlier periods.

The graph shows a regular decrease in submissions at the end of each year, likely due to holidays and reduced business activities. This trend may reflect the shifting organizational priorities and fewer resources allocated to data protection during this time. Additionally, weekdays see over 100 complaints, indicating high operational activity, while weekends have fewer complaints, likely due to reduced staffing and operational capacities.

Deciphering Decision Pathways in Complaint Outcomes

The Sankey diagram in the dashboard provides users with a valuable visualization tool for understanding the complex relationships between decision categories and their underlying reasons. Refer to appendix for additional information on the Sankey implementation. (See Appendix J for more information)

Leveraging the capabilities of Flourish, I imported my dataset and specified the "Decision" and "Decision Detail 2" columns as the source and target nodes for the Sankey diagram. This allowed me to generate the Sankey diagram I envisioned, albeit outside of the Python environment. After generating the Sankey diagram in Flourish, I made adjustments to ensure that all nodes were labelled, sorted, and coloured appropriately, adhering to the established visualization methodology used throughout the project. While this approach involved a departure from the Python-based workflow, it enabled me to overcome the limitations encountered with Plotly and successfully integrate the Sankey diagram into the dashboard. While this approach sacrifices the interactive functionality typically associated with Sankey diagrams, it still enables users to gain insights and make informed decisions based on the visualized data.

The Sankey diagram reveals that a significant proportion of unaddressed complaints were due to the regulatory body's lack of adequate information, corroborating prior research that highlights the challenges posed by strict regulatory conditions and limited information availability, which impedes the progress of complaints. This underscores the importance of providing comprehensive and well-supported information to regulatory bodies by individuals when submitting complaints and requires the establishment of clear guidelines by regulatory bodies to articulate the information needed to review a complaint. Organizations, too, must recognize the damaging consequences of inadequate information, which could lead to noncompliance. Accordingly, organisations should manage data effectively and communicate transparently with regulatory authorities to minimize compliance risks. By leveraging this insight, regulatory authorities can extend their information-gathering mechanisms, offer clearer guidance to complainants on the information necessary for complaint submission, and, in turn, streamline complaint-handling processes while enhancing regulatory effectiveness.

The Sankey diagram is a valuable tool in understanding the reasons behind regulatory action taken against organizations for data protection issues. One key finding is that many organizations face regulatory action due to their inadequate or non-existent responses to complaints. This underscores the importance of prompt and effective responses from organizations when addressing data protection concerns. For users, this means it is important to prioritize engaging with organizations that have a track record of less timely and unsatisfactory responses to data protection issues. This can help foster accountability and build trust in data handling practices, ultimately ensuring that their privacy is protected. For organizations, the diagram serves as a reminder to strengthen their response mechanisms, highlighting the importance of swift acknowledgement and resolution of complaints. This can help mitigate regulatory repercussions and safeguard their reputation. It also promotes a culture of accountability and responsiveness in data protection governance, which can foster greater trust with customers. Regulatory bodies can use this information to identify areas of non-compliance and enforce measures that ensure timely and adequate responses from organizations. This can help create a regulatory environment that encourages organizations to prioritize data protection and respond effectively to complaints.

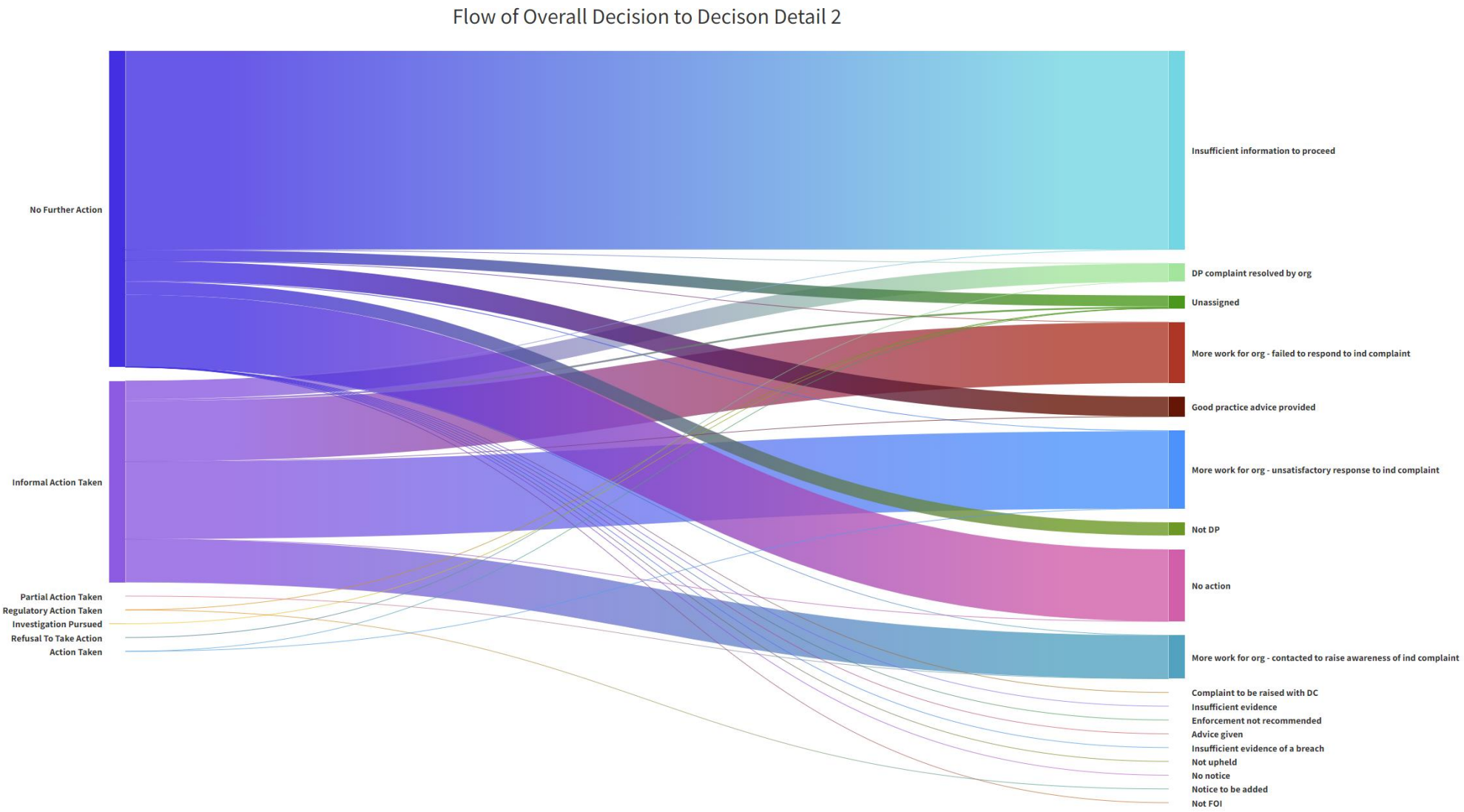


Figure 16: Sankey Diagram of the Flow of Decisions and Their Reason

Sectoral Insights into Monthly Complaint Trends

Incorporating a heatmap into the dashboard offers a dynamic way to visualize the distribution of data protection complaints across different sectors submitted throughout the year. This visualization provides insights into the patterns and trends of complaint submissions, allowing users to discern any fluctuations or concentrations in activity across various months.

To create the heatmap, I first mapped month names to numbers using calendar.month_name. This mapping generated a new 'MonthNum' column in the dataset. Then, I grouped the dataset by sector, month name, and month number, calculating complaint counts for each combination. Next, I sorted the dataset by the numerical day and eliminated the 'unknown' column, which contained only partial data. I also changed the color scheme of the heatmap from a continuous Red Blue scale to the Portland color scale for improved clarity. The Portland color scale offers a smoother transition between colors, making it easier to interpret the heatmap, especially when values are closely distributed.

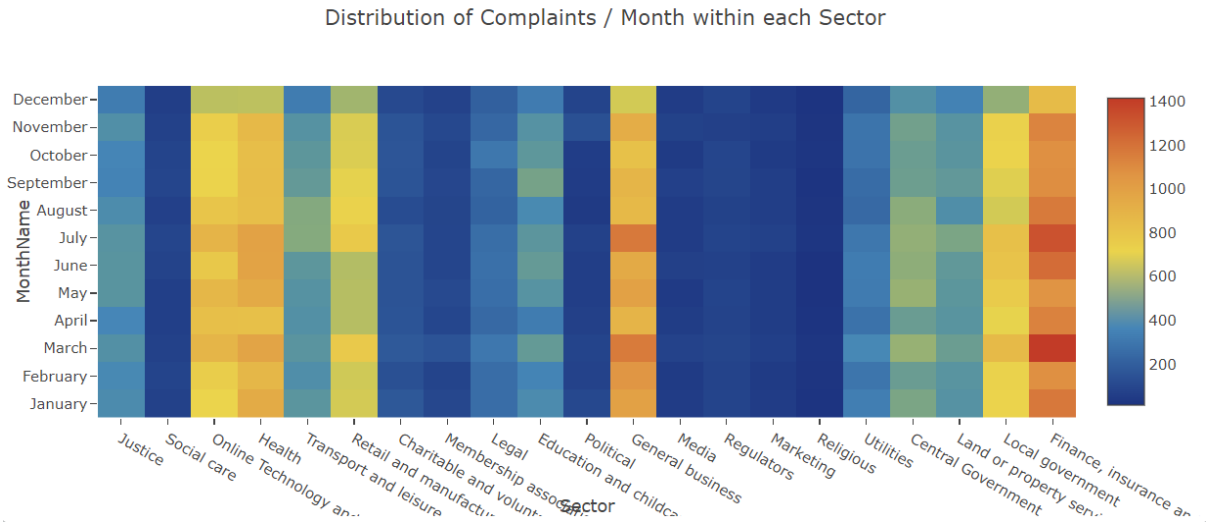


Figure 17: Distribution of Complaints per Month Within Each Sector Heat Map

This yielded valuable insights, notably revealing an increase in data protection complaints across nearly all sectors during the months of March, July, and November. This trend is more evident in the General Business and Finance, Insurance, and Credit sectors.

The rise in data protection complaints during months like March, July, and November may be due to seasonal trends, increased data processing activities during financial reporting periods, or regulatory compliance reviews. For instance, March marks the end of the fiscal year for many businesses, leading to heightened data processing, while July could see spikes due to mid-year audits or reviews. November's increase may stem from stakeholders addressing concerns before the holiday season. These insights into seasonal trends can help users, organizations, and regulators enhance data protection practices and compliance. Organizations can allocate resources effectively and conduct internal audits. Regulatory bodies can prioritize enforcement efforts and interventions based on sector or organizational risk levels during specific months.

Other Implemented and Disregarded Visualisations

Due to word limitations, not all implemented and disregarded visualisations could be analyzed in detail. However, the sub-sector bar chart revealed 'supplier of goods' as the highest complaint sector, offering valuable insights into sector-specific distribution. Additionally, the detailed pie chart showcased prominent breaches like Article 15(3)(1) - 'Provide a copy of personal data', assisting organizations in addressing compliance gaps proactively. Some visualisations were excluded due to reasons such as ineffective insights, unsuitable graphs, or better alternatives being utilized. These analyses are available in the Appendix. (See Appendix J for Additional and Un-used Visualisations)

Enhancing Complaints Dashboard with Risk Analysis

The process of calculating and implementing risk factors for each data protection complaint as discussed in the methods section involved leveraging Python code to manipulate and analyse the dataset. Initially, the dataset was segmented into different categories, including sectors, decision articles, subsectors, and the frequency of complaints submitted against specific organisations in alignment with the quantitative risk assessment methodology. This segmentation was achieved using the **groupby** function, which grouped the dataset based on specified columns.

To calculate risk factors for sectors, the groupby function was applied to the 'Sector' column to aggregate data. Using the agg function, the 'Decision (Action Taken)' and 'Decision (No Action Taken)' columns were summed to determine the total actions and no actions taken decisions within each sector, aligning with the outlined formula. Subsequent calculations derived the risk factor for each category. For sectors, the ratio of actions taken to total decisions was computed to determine the sector risk factor.

```
##### Risk Factor data #####

# Group the dataset by sector and calculate the sum of action taken and no action
taken decisions
temp_data_1 = dataset.groupby(['Sector']).agg({'Decision (Action Taken)':'sum',
'Decision (No Action Taken)':'sum'}).reset_index()
temp_data_1["Total Decisions"] = temp_data_1['Decision (Action Taken)'] + temp_data_1
['Decision (No Action Taken)']
temp_data_1["Sector Risk Factor"] = (temp_data_1['Decision (Action Taken)'] /
(temp_data_1['Decision (Action Taken)'] + temp_data_1['Decision (No Action Taken)']))
```

Figure 18: Grouping Sector Column for the Risk Factor data

Risk factors for decision articles, subsectors, and complaints against specific organizations were determined through grouping, aggregating, and calculating. After calculating risk factors for each category, relevant information was structured alongside corresponding categories by selecting and formatting specific columns from the aggregated datasets. These individual datasets were merged into a single dataframe using the pd.merge function in Python, integrating risk factor information from each category into the final dataset. Once merged, the total risk factor for each entry in the dataframe was computed by averaging risk factors from all categories and then multiplying by 100 to obtain a percentage. However, this introduced an unforeseen problem: the resulting percentages ranged from 20 to 80 percent, creating a compressed scale that failed to adequately reflect variation in risk across complaint data. This constrained range made it difficult to differentiate subtle differences in risk levels, potentially obscuring critical insights.

```
# Merging the groups
data_merged = pd.merge(company_data,temp_data_1,left_on='Sector',right_on='Sector')
data_merged = pd.merge(data_merged,temp_data_2,left_on='Decision Articles',
right_on='Decision Articles')
data_merged = pd.merge(data_merged,temp_data_3,left_on='Sub Sector',right_on='Sub
Sector')
data_merged = pd.merge(data_merged,temp_data_4,left_on='Submitted About Account',
right_on='Submitted About Account')
```

Figure 19: Merging the Columns for Overall Risk Factor

This led me to the integration of the MinMaxScaler function for the risk factor calculation process which involved several key steps outlined in the methods section. (scikit). This function which was crucial for normalising the risk factor values to a predefined range, helped the comparability across the different datasets which implements the Risk Scaled formula as researched in the methodology. Firstly, the MinMaxScaler object is initialised with a specified feature range, and was chosen to be between 1 and 100, which allows for a more intuitive interpretation of the scaled risk factors. Next, the **fit_transform** method is applied to transform the original risk factor values into the 1 to 100 specified range. This method calculates the minimum and maximum values of the input data and then scales each feature to the specified range. This ensured that all the risk factor values fall within the same numerical range, irrespective of their original scale.

```
##### MinMaxScaler Use #####

# Setting the range of normalisation
scaler = MinMaxScaler(feature_range=(1, 100))
data_merged['Scaled Risk Factor'] = scaler.fit_transform(data_merged['Total Risk
Factor'].values.reshape(-1,1))
data_merged[["Scaled Risk Factor"]] = data_merged[["Scaled Risk Factor"]].round(2)
data_merged_filtered = data_merged[["Case Reference", "Submitted About Account" ,
"Date Received", "Completed Date", "Sector", "Sub Sector", "Decision Primary
Reason", "Decision", "Decision Detail2", "Scaled Risk Factor"]]
```

Figure 20: Implementation of MinMaxScaler

To ensure clarity and consistency in the dataset, the scaled risk factor values were rounded to two decimal places, simplifying interpretation and analysis. Sorting the table by the 'Submitted About Account' column allowed for filtering based on specific organizations. After selecting the first company name, relevant data was filtered, and unnecessary columns were removed for clarity. Standardizing, formatting, and reducing duplicates were achieved by capitalizing the first letter of each word in company names using the process_string function applied via the Pandas apply function.

The dataset initially lacked visual indicators for risks, making it difficult to detect critical cases and trends. To resolve this, I created the risk_style function, which assigns colour codes using predefined ranges. This helps visually interpret and analyze risk factors. I chose this incremental colour scheme because of the colour connotation. It creates a sense of urgency for complaints per organization which helps viewers understand the data without context. However, as seen in the literature review, a fundamental principle of effective design is honesty. Using red for critical risks could create a negative image for organizations. Offering clear context and explanations with colour-coded risk indicators can help address this issue. To improve the understanding of risk indicators, it is vital to include a clear legend with the colour-coded risk levels assigned by the risk_style function as users may struggle to understand complaint severity without knowing the meaning of each colour.

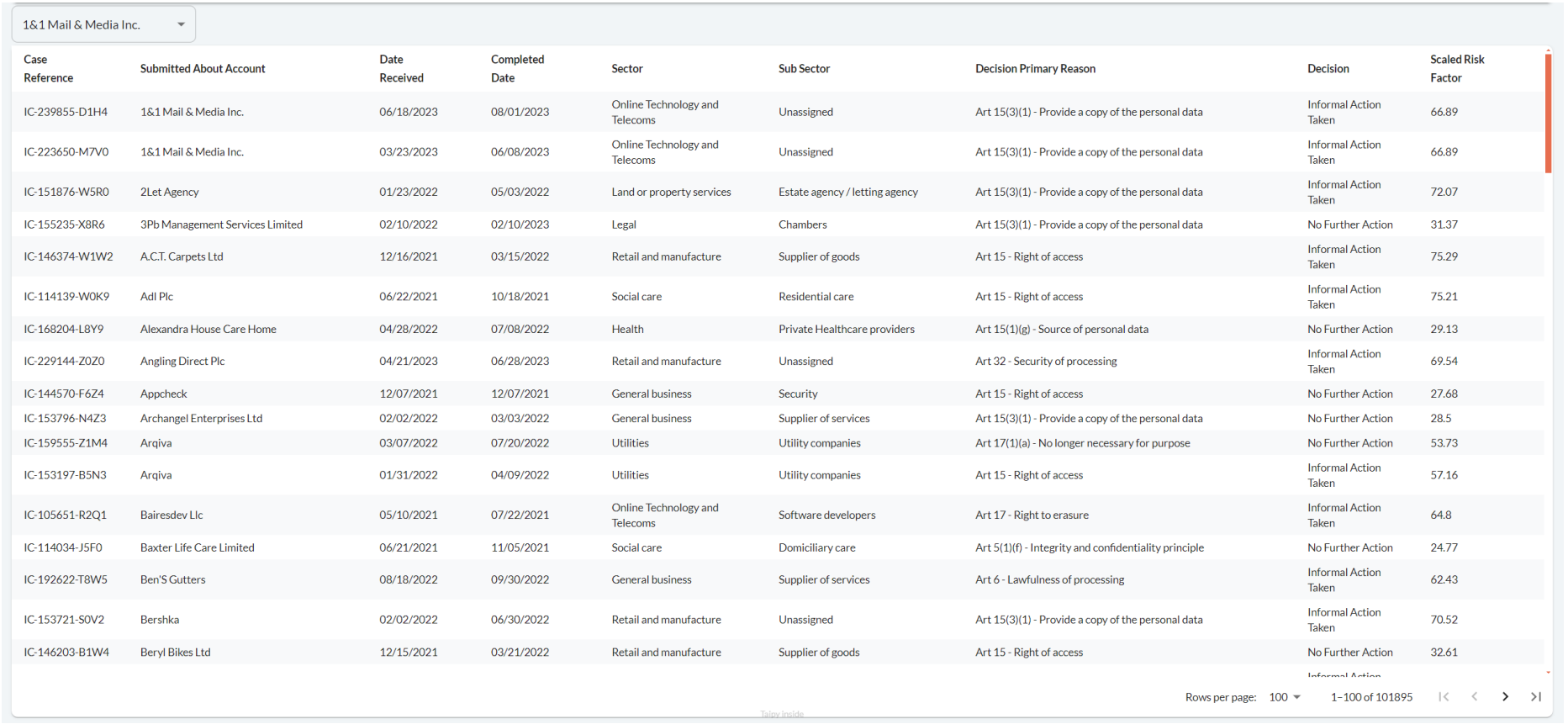
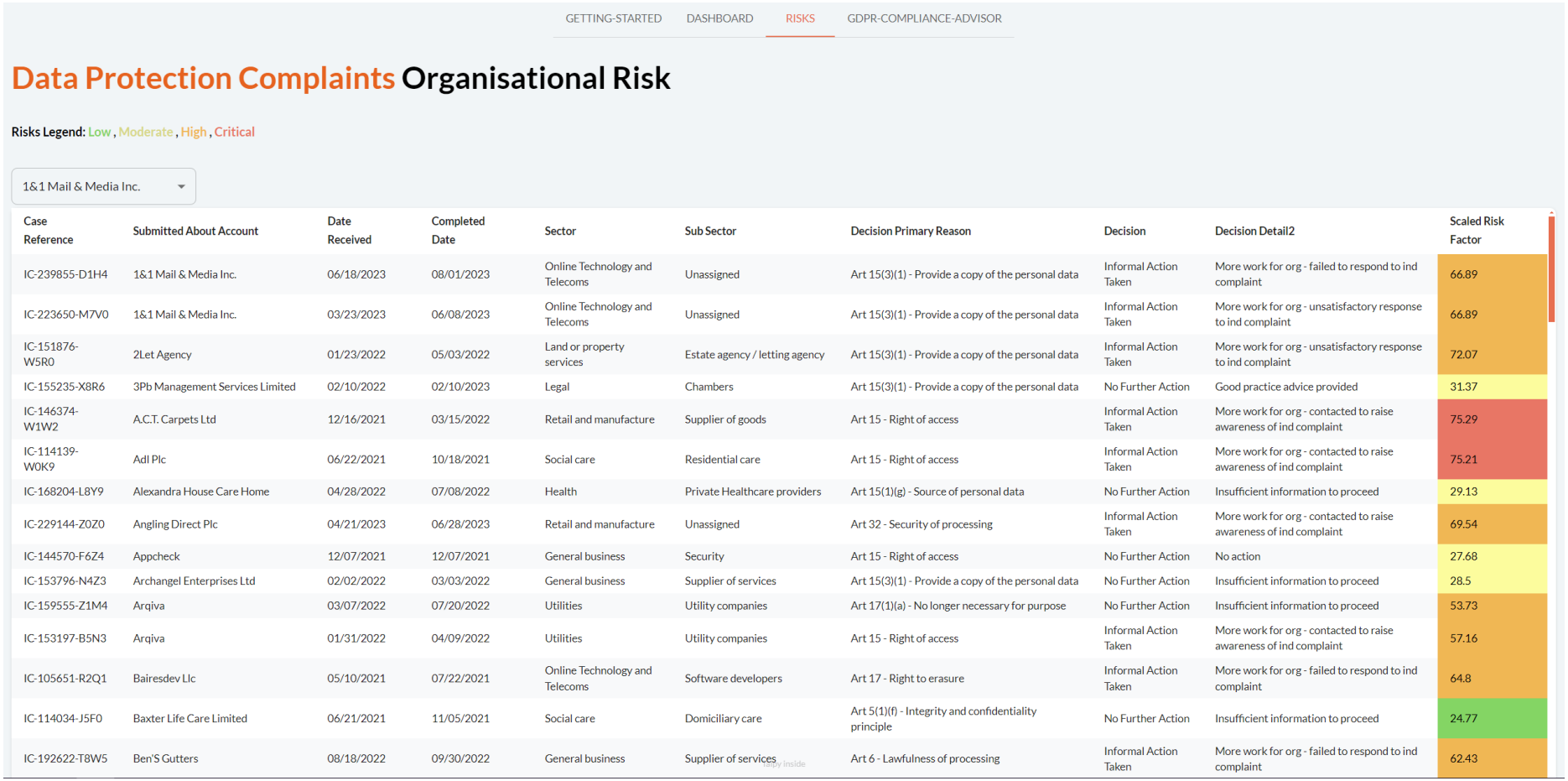


Figure 21: Un-coloured Dashboard



Case Reference	Submitted About Account	Date Received	Completed Date	Sector	Sub Sector	Decision Primary Reason	Decision	Decision Detail2	Scaled Risk Factor
IC-239855-D1H4	1&1 Mail & Media Inc.	06/18/2023	08/01/2023	Online Technology and Telecoms	Unassigned	Art 15(3)(1) - Provide a copy of the personal data	Informal Action Taken	More work for org - failed to respond to ind complaint	66.89
IC-223650-M7V0	1&1 Mail & Media Inc.	03/23/2023	06/08/2023	Online Technology and Telecoms	Unassigned	Art 15(3)(1) - Provide a copy of the personal data	Informal Action Taken	More work for org - unsatisfactory response to ind complaint	66.89
IC-151876-W5R0	2Let Agency	01/23/2022	05/03/2022	Land or property services	Estate agency / letting agency	Art 15(3)(1) - Provide a copy of the personal data	Informal Action Taken	More work for org - unsatisfactory response to ind complaint	72.07
IC-155235-X8R6	3Pb Management Services Limited	02/10/2022	02/10/2023	Legal	Chambers	Art 15(3)(1) - Provide a copy of the personal data	No Further Action	Good practice advice provided	31.37
IC-146374-W1W2	A.C.T. Carpets Ltd	12/16/2021	03/15/2022	Retail and manufacture	Supplier of goods	Art 15 - Right of access	Informal Action Taken	More work for org - contacted to raise awareness of ind complaint	75.29
IC-114139-W0K9	Adl Plc	06/22/2021	10/18/2021	Social care	Residential care	Art 15 - Right of access	Informal Action Taken	More work for org - contacted to raise awareness of ind complaint	75.21
IC-168204-L8Y9	Alexandra House Care Home	04/28/2022	07/08/2022	Health	Private Healthcare providers	Art 15(1)(g) - Source of personal data	No Further Action	Insufficient information to proceed	29.13
IC-229144-Z0Z0	Angling Direct Plc	04/21/2023	06/28/2023	Retail and manufacture	Unassigned	Art 32 - Security of processing	Informal Action Taken	More work for org - contacted to raise awareness of ind complaint	69.54
IC-144570-F6Z4	Appcheck	12/07/2021	12/07/2021	General business	Security	Art 15 - Right of access	No Further Action	No action	27.68
IC-153796-N4Z3	Archangel Enterprises Ltd	02/02/2022	03/03/2022	General business	Supplier of services	Art 15(3)(1) - Provide a copy of the personal data	No Further Action	Insufficient information to proceed	28.5
IC-159555-Z1M4	Arqiva	03/07/2022	07/20/2022	Utilities	Utility companies	Art 17(1)(a) - No longer necessary for purpose	No Further Action	Insufficient information to proceed	53.73
IC-153197-B5N3	Arqiva	01/31/2022	04/09/2022	Utilities	Utility companies	Art 15 - Right of access	Informal Action Taken	More work for org - contacted to raise awareness of ind complaint	57.16
IC-105651-R2Q1	Bairesdev Llc	05/10/2021	07/22/2021	Online Technology and Telecoms	Software developers	Art 17 - Right to erasure	Informal Action Taken	More work for org - failed to respond to ind complaint	64.8
IC-114034-J5F0	Baxter Life Care Limited	06/21/2021	11/05/2021	Social care	Domiciliary care	Art 5(1)(f) - Integrity and confidentiality principle	No Further Action	Insufficient information to proceed	24.77
IC-192622-T8W5	Ben'S Gutters	08/18/2022	09/30/2022	General business	Supplier of services	Art 6 - Lawfulness of processing	Informal Action Taken	More work for org - failed to respond to ind complaint	62.43

Figure 22: Coloured and Contextualized Dashboard

The Data Protection Organisational Risks page not only serves as a crucial tool for organizations navigating GDPR compliance but also has the potential to be a valuable resource for various stakeholders, as highlighted in the literature review. The page's primary function is the representation of a table which provides information related to the complaints concerning GDPR. Each record is assessed in terms of the risk level and is defined as a percentage within the range from low to critical. This method of organising the complaints on data protection risks allows organisations to have an overview of the threats and make holistic, risk-based decisions about their mitigation. The table's filtering options allow users to analyse data with additional layers of insights. Sorting risks by severity enables organizations to focus on critical issues first and prioritize remediation efforts. This helps to differentiate risks across different organizations, sectors, and based on the breach recorded.

Case Reference	Submitted About Account	Date Received	Completed Date	Sector	Sub Sector	Decision Primary Reason	Decision	Scaled Risk Factor ↓
IC-145546-G8V5	Portsmouth Liberal Democrats	12/13/2021	06/21/2022	Political	Political Party	Art 21 - Right to object	Informal Action Taken	97.11
IC-94469-Z1X3	Croydon Conservative Federation	03/12/2021	09/10/2021	Political	Political Party	Art 21(2) - Objection to direct marketing	Informal Action Taken	97.11
IC-227338-V9V5	Brighton Pavilion Constituency Labour Party	04/12/2023	05/05/2023	Political	Political Party	Art 21(2) - Objection to direct marketing	Informal Action Taken	97.11
IC-194923-B2D1	Fareham Conservative Association	10/01/2022	11/03/2022	Political	Political Party	Art 14 - Right to be informed	Informal Action Taken	96.3
IC-210008-Z3L6	Westminster North Conservative Association	01/05/2023	03/24/2023	Political	Political Party	Art 14 - Right to be informed	Informal Action Taken	96.3
IC-181513-G9N1	Stoke Central Constituency Labour Party	07/15/2022	11/15/2022	Political	Political Party	Art 14 - Right to be informed	Informal Action Taken	96.3
IC-52994-B5Q7	Ashfield Independent Party	08/12/2020	11/23/2022	Political	Political Party	Art 13 - Information to be provided when collected from DS	Informal Action Taken	95.78
IC-145477-D1L8	Labour Party	12/13/2021	04/06/2022	Political	Political Party	Art 12(2)(2) - Unable to identify DS	No Further Action	95.36
IC-47578-T7R9	Labour Party	07/22/2020	03/02/2021	Political	Political Party	Art 12(3)(1) - Within 1 month	Informal Action Taken	95.36
IC-139483-F8Z6	Momentum Campaign (Services) Ltd	11/09/2021	12/06/2021	Political	Political Party	Art 17 - Right to erasure	Informal Action Taken	94.92
IC-189838-J1Y8	Momentum Campaign (Services) Ltd	09/01/2022	10/27/2022	Political	Political Party	Art 17(1)(b) - Consent withdrawn	Informal Action Taken	94.92
IC-119126-Z7T1	Sandwell Conservatives	07/18/2021	08/10/2021	Political	Political Party	Art 17(1) - Right to erasure	Informal Action Taken	94.92
IC-58370-V7F8	Momentum Information Ltd	08/07/2020	01/12/2021	Political	Political Party	Art 17(1) - Right to erasure	Informal Action Taken	94.92
IC-59180-S1P6	Spelthorne Conservative Association	08/27/2020	03/24/2021	Political	Political Party	Art 17 - Right to erasure	Informal Action Taken	94.92
IC-111098-S6P3	Redditch County Conservative Association	06/07/2021	10/21/2021	Political	Political Party	Art 5(1)(b) - Purpose limitation principle	Informal Action Taken	94.5
IC-98490-J1H3	Monmouth Conservative Association	04/03/2021	08/20/2021	Political	Political Party	Art 5(1)(c) - Data minimisation principle	Informal Action Taken	94.5

Figure 23: Sorted Table by Scaled Risk Factor, Showing Prominence in Political Sector

For example, when sorting the table by risk factor, all of the top-risk complaints are within the political sector, with action taken. This analysis highlights the heightened vulnerability of the political sector to data protection risks. Given the sensitive nature of political activities and the wide use of personal data involved in political campaigns, these organisations operating within this sector are inherently exposed to various data privacy challenges. This is shown by the large number of regulatory actions taken within these complaints which suggests systemic issues or compliance gaps within the political sector, needing urgent reforms to mitigate risks and ensure compliance. This could be due to insufficient compliant protocols implemented by these organisations. This aligns with GDPR analysis findings within the literature, emphasizing the critical role of risk assessment and mitigation strategies in reducing breaches and avoiding significant fines and penalties for non-compliance.

The primary utility added to this page is the option to filter the table by organization. This helps users identify patterns and recurring complaints for specific organizations, aligning with findings from the literature review regarding the benefits and drawbacks of the "Submitted about account" column. While enhancing transparency, there's a risk of damaging an organization's reputation and undermining public trust, as noted by Christopher Beveridge. However, the positives outweigh the negatives as this could help facilitate a more cost-effective targeted approach to investigations and other actions based on the severity and frequency of complaints made.

One significant insight we can gain from filtering by organisation is the number of complaints against NHS England that do not proceed because of inadequate information provided by the complainant. While it is correct that this indicates a large number of complaints, it is also an apparent concern that vital complaints remain unaddressed or ignored due to poor information submitted. This underscores the role of transparency and completeness in complaint-making. This makes it known to users that if they are submitting a complaint against the NHS, they are to provide as much information as possible. Furthermore, this allows regulatory bodies and the NHS themselves, ensuring users are adequately informed of what is expected of them when reporting a complaint and the information to be provided.

Submitted About Account	Decision	Decision Detail2
Nhs England	No Further Action	Insufficient information to proceed
Nhs England	No Further Action	Insufficient information to proceed
Nhs England	No Further Action	Insufficient information to proceed
Nhs England	No Further Action	Insufficient information to proceed
Nhs England	No Further Action	No action
Nhs England	No Further Action	No action
Nhs England	No Further Action	No action
Nhs England	No Further Action	No action
Nhs England	No Further Action	Insufficient information to proceed
Nhs England	No Further Action	Insufficient information to proceed
Nhs England	No Further Action	Insufficient information to proceed
Nhs England	Informal Action Taken	DP complaint resolved by org
Nhs England	No Further Action	No action
Nhs England	No Further Action	Insufficient information to proceed
Nhs England	No Further Action	Insufficient information to proceed

Figure 24: Sorted Table by the NHS Organisation for Insights

Moreover, the complaints table is a useful tool for organisations and startups which might want to learn from already identified and constructed complaints filed against other bodies from the same industry. For instance, a bank or financial company can leverage this freely accessible information to identify common pitfalls. This can be seen by the prevalence of data protection complaints against Barclays Bank resulting in regulatory action, where 14.4% of the cases, Barclays failed to respond to complaints and 8.5% gave an unsatisfactory response. By analysing these patterns, organisations can proactively implement protocols to avoid similar breaches and enhance their compliance efforts.

Submitted About Account	Decision	Decision Detail2 ↑
Barclays Bank Plc	Informal Action Taken	More work for org - contacted to raise awareness of ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint
Barclays Bank Plc	Informal Action Taken	More work for org - failed to respond to ind complaint

Figure 25: Sorted Table by Barclays Bank Plc Organisation for Insights

Furthermore, organisations can utilise the table to review their own complaint history, facilitating analysis of past shortcomings and opportunities for improvement. For instance, many of the Labour Party complaints where the Party is contacted to raise awareness. This is where the regulatory body are making the party aware of an individual's complaint but do not have evidence of an infringement or Cases where the organisation has advised that they have done something, but the complainant says they have not. This analysis enables organisations to identify potential areas of concern and refine their response strategies accordingly. By learning from past mistakes and addressing deficiencies in their complaint-handling processes, organisations can improve their compliance efforts and mitigate future risks.

Submitted About Account	Sector	Sub Sector	Decision Primary Reason	Decision	Decision Detail2	Scaled Risk Factor
Labour Party	Political	Political Party	Art 32 - Security of processing	Informal Action Taken	More work for org - contacted to raise awareness of ind complaint	89.52
Labour Party	Political	Political Party	Art 32 - Security of processing	Informal Action Taken	More work for org - failed to respond to ind complaint	89.52
Labour Party	Political	Political Party	Art 15 - Right of access	Informal Action Taken	More work for org - unsatisfactory response to ind complaint	90.5
Labour Party	Political	Political Party	Art 15 - Right of access	Informal Action Taken	More work for org - unsatisfactory response to ind complaint	90.5
Labour Party	Political	Political Party	Art 15 - Right of access	Informal Action Taken	More work for org - failed to respond to ind complaint	90.5
Labour Party	Political	Political Party	Art 15 - Right of access	Informal Action Taken	More work for org - unsatisfactory response to ind complaint	90.5
Labour Party	Political	Political Party	Art 12(3)(1) - Within 1 month	Informal Action Taken	More work for org - unsatisfactory response to ind complaint	95.36
Labour Party	Political	Political Party	Art 14 - Right to be informed	Informal Action Taken	More work for org - unsatisfactory response to ind complaint	88.44

Figure 26: Sorted Table by Labour Party Organisation for Insights

GDPR Compliance Advisor

The development of the GDPR Compliance Advisor page involved utilizing Python libraries and data processing techniques. Initially, compliance recommendations were sourced from a JSON file named 'recommendations.json', organized by article names, descriptions, and recommendations, formatted into a list named ‘recommendationsList’. The dataset was then grouped by sector and subsector using the 'groupby' function, and a lambda function computed the most common primary reason within each group.

The processed data was converted into a nested dictionary named ‘decision_dict’ for user interaction in the GUI interface which organises primary reasons based on sector and subsector. Both Sector and Subsector classes were defined and initialized to retrieve individual sectors and subsectors and link them to primary reasons and recommendations. Recommendations were then loaded based on the decision primary reason of the first subsector.

The function ‘startSubSectorSelection’ controlled subsector changes where it generated a new list of subsectors based on the selected sector. It assigned the first subsector in the list as the selected variable. The ‘getArticleFromJson’ function fetched article information based on the selected subsector's decision primary reason, facilitating users' understanding of GDPR compliance requirements and recommendations.

This then resulted in ‘getArticleFromJson’ looping through the ‘recommendationsList’, comparing the ‘articleIdentifier’ to the decision primary reason of the selected subsector. If a match was found, the function populated the article info attributes. If no exact match was found, it applied regular expressions to extract the article number. Default values were assigned if no match was found in either case, ensuring users had access to relevant article information for compliance recommendations.

The GDPR Compliance Advisor page aims to be an integral source of help for companies concerned with finding their way through data protection laws. It helps fill out the most significant gaps in compliance with sectoral advice and insights and encourages active decision-making through a user-friendly interface.

This makes it possible for everyday users to know their right and the process of making informed decisions about personal information disclosure and helps hold organisations accountable and encourages transparency in their data handling processes. The GDPR Compliance Advisor page allows ordinary users for example, patients or research subjects to understand health sector-specific data protection concerns around health research. In case someone wants to know how their health research organizations handle their personal data, they can then visit this page for the top three breaches of GDPR in this subsector. This way they can more effectively advocate for their privacy rights relating to personal information while engaged in a medical study. Specifically, if one is taking part in a health research study, they may ask for their personal data as explained by Art 15 present in the page. If they have security concerns about their data, they can ask about measures to comply with Art 32. Furthermore, lacking the required copy of their personal data under Art 15(3)(1), individuals can leverage this knowledge to urge the organization to meet this obligation.

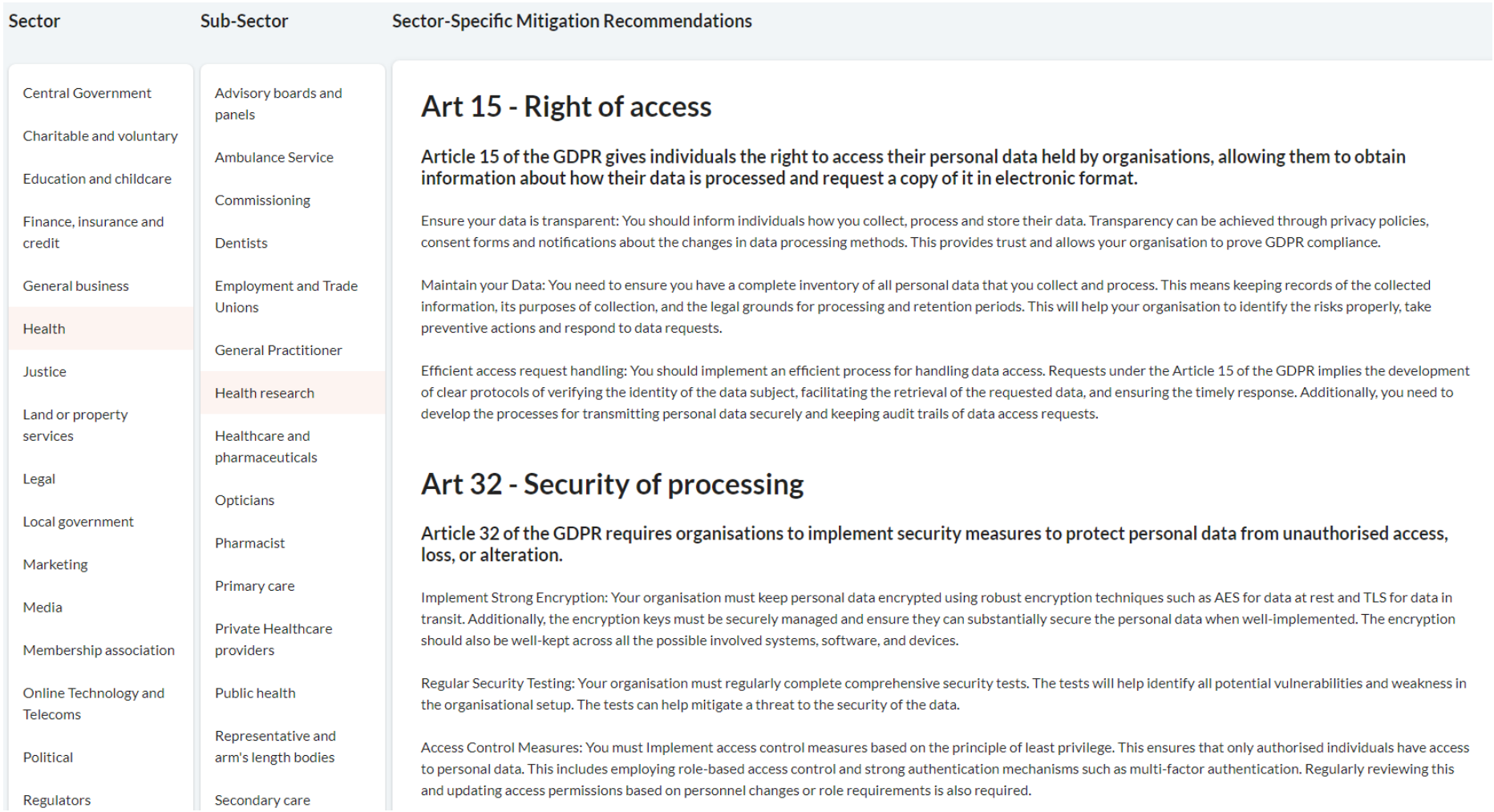


Figure 27: Health Research Sub Sector Within Advisor Page

Another example is in the Justice Sector, specifically the Government department subsector, the most common breach identified is Art 16 - Right to rectification. This insight holds significance for professionals operating within this sector, as it sheds light on the prevalent data protection challenges, they face which they may have previously not been aware of. Using this knowledge, individuals working in the Justice Sector can take proactive measures to address the root causes of breaches and strengthen their compliance efforts by leveraging the recommendations provided by the GDPR Compliance Advisor page. The stakeholders can implement the targeted procedures and protocols aimed at upholding GDPR compliance and safeguarding the rights of data subjects.

Figure 28: Government Department Sub Sector Within Advisor Page

Utilising Taipy Cloud

The dashboard has been successfully uploaded to Taipy Cloud and is fully operational. Although it runs a bit slower than when executed locally, this performance difference was anticipated. To enable its functionality on Taipy Cloud, I configured the machine and created a requirements text file. This file ensures that Taipy recognizes and installs the necessary modules needed to run the application smoothly. Overall, the upload process went smoothly, and the dashboard is now accessible on Taipy Cloud.

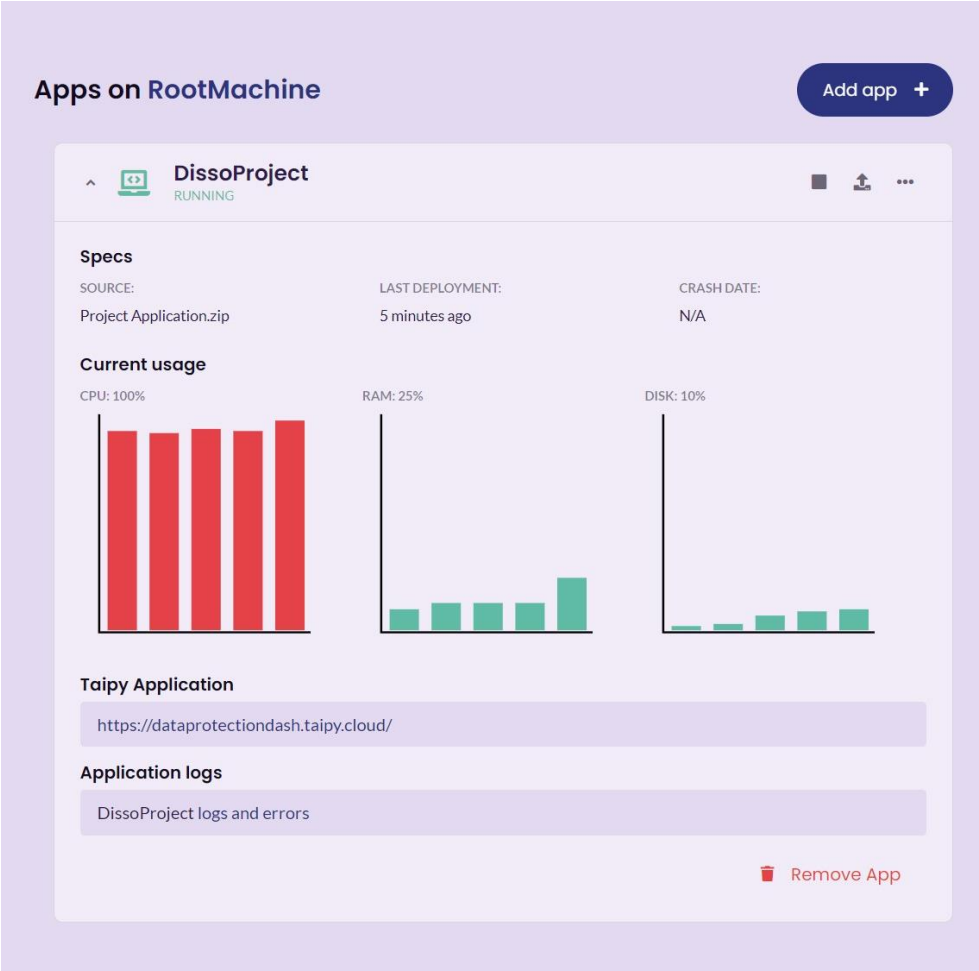


Figure 29: Project Running on Taipy Cloud Successfully

Testing

Systematic testing was done to ensure reliability and functionality of the Data Protection Complaints Dashboard. The unit testing thoroughly examined individual components and verified performance both locally and on the Taipy cloud. Integration testing brought together completed units such as the dashboard, risk advisor, and GDPR compliance advisor, ensuring combined functionality. System testing verified both functional and non-functional requirements, ensuring high performance and functionality throughout the project. Test cases were executed for all graphs such as displaying decisions, highest complained sectors, and risk assessment. The testing documentation including test cases and results, can be found in the appendix, providing an overview of the testing for the project's development lifecycle. (See Appendix K for Testing Documentation)

Chapter 6

Conclusion & Discussions

The Conclusions and Discussion chapter summarizes the research and development done in this project on data protection complaints by evaluating the achievement of project objectives, requirements, and research questions. The main goal of this project was to create a user-friendly web dashboard for ICO complaint data. To achieve this, various sub-goals were set and completed.

Firstly, the project aimed to ensure the accuracy and reliability of the data protection complaints dataset by performing data analysis, aggregation, cleansing, and preparation. To achieve this, the dataset underwent thorough cleansing and preparation using methodologies explored within the literature such as investigation, survivorship, and standardization. Despite challenges arising from inconsistencies in the ICO's recording methods, the results largely met expectations. Although complete cleansing within the given timeframe was not achieved, manual efforts improved the dataset's usability. These outcomes directly link to the project's objectives, as they laid the foundation for effective data visualization and risk analysis. Ultimately, this contributed to the development of a comprehensive dashboard tool for GDPR compliance.

The project also strived to provide visualizations of data protection complaints. As a result, the dashboard serves to educate users, organizations, and regulators on trends, patterns and common compliance pitfalls that will improve their communication for prioritizing corrective actions. Through the dashboard’s visuals and analysis, not only does it identify areas where complaints are not processed due to lacking evidence but also sheds light on cases where action was taken against organizations for inadequate responses. This emphasises the need for improved communication channels between organizations and users, spelling out the importance of transparency.

This made use of multiple different visualisation methodologies highlighted within the literature review, aiding the insights gained from the graphs. Many of these procedures leveraged Andy Kirk's book on Data Visualization, facilitating effective graph creation using techniques such as Data Representation, Interactivity, Annotation, Colour, and Composition. This helped create the most effective graphs to aid the user in gaining the most effective insights. This methodology was also adapted to suit the cyber security context, using insights from VisSec and GIAC procedures, ensuring effective security anomaly detection and visualization techniques to help portray complex theories in a simple manner, keeping them “informative” and “summative”.

The implementation of risk assessment methodologies within the dashboard has yielded valuable benefits and insights. The aim of the risk assessment implementation was to identify and assess potential risks associated with data breaches in order to inform decision-making and mitigate potential threats. By providing specific details on organizations and their responses to data protection complaints, the risk analysis component successfully offers a deeper understanding of the potential risks associated with breaches. This benefits everyday users, organizations, and regulators alike by providing a comprehensive understanding of potential risks associated with data protection complaints. For users, it offers insights into the severity of complaints and empowers them to make informed decisions regarding their data privacy. Organizations can utilize this information to prioritize compliance efforts and allocate resources effectively, ultimately enhancing their data protection practices. Additionally, regulators can leverage the risk analysis to identify trends, assess compliance levels, and inform regulatory interventions, thus promoting transparency and accountability in the data protection landscape.

Extensive research into the decision to use quantitative risk assessment which was guided by its need and usefulness in data protection. Quantifying risks associated with complaints enables organizations to prioritize compliance efforts, allocate resources effectively, and provide regulators with valuable insights into compliance trends for targeted interventions. Thus, robust risk assessment methodologies improve data protection practices and promote transparency and accountability in data handling.

In developing the GDPR compliance advisor, the aim was to provide actionable guidance based on GDPR requirements and best practices. The results were successful, effectively calculating and recommending actions for common GDPR breaches. This implementation also contributes to the emerging field of legal design by presenting GDPR compliance policies in an intuitive and aesthetically pleasing manner. This approach aligns with the literature's emphasis on making legal information more understandable and accessible to end-users, bridging the gap between legal requirements and practical implementation. This success aligns with the project's objectives and requirements, demonstrating the tool's utility for users, organizations, and regulators in navigating GDPR compliance challenges. By fulfilling this objective, the project contributes to enhancing data protection practices and fostering transparency in data handling, in line with the overarching goal of promoting GDPR compliance and accountability.

Lessons Learnt

The project progressed smoothly, adhering closely to the agile-scrum project development cycle and schedule. I successfully delivered the project requirements within the specified timeframe which demonstrates effective time management and organization. However, certain restrictions and limitations presented challenges that affected the completeness of the requirements. Despite these obstacles, I remained proactive in problem-solving and sought ways to mitigate their impact. This ensured that the project moved forward as planned.

Overall, the project’s outcomes provided valuable lessons, such as the use of Python. Learning a new language, which was expected to increase productivity, turned out to elongate the project lifecycle due to the learning curve. If more time were allocated to the project, the investment in learning Python would eventually pay off, rather than compromising for the deadlines. Secondly, the Taipy library provided limited opportunities for visualization. The fixed number of charts available in the library meant that all additional functionalities would need to be done via other libraries, which often incorporates compatibility issues with other Python libraries. For instance,

attempts to implement the Sankey diagram were hindered by performance issues. Given more time and resources, alternative solutions such as using D3.js or established applications like Power BI could address these limitations effectively by offering a more established visualization platform for future projects.

The ICO data protection complaints dataset revealed significant limitations due to inconsistent values and column structures, requiring extensive cleansing efforts. Addressing alternate spellings and varying column structures across quarterly datasets took considerable time for data accuracy. Challenges persisted with organizations listed under different names due to spelling discrepancies or additional material like "LTD" or "Limited." A potential solution involves using official company numbers for standardized identification, retrieved through an API from the government website. This would streamline the process, ensuring data integrity even with organization updates, offering a more efficient means of logging complaints.

While the current risk assessment implementation provides valuable insights, there is room for improvement to achieve greater accuracy. The assessment relies on existing values within the dataset, which may not capture all relevant factors influencing the risk of a data protection complaints. To enhance accuracy, incorporating additional information from external sources, such as GDPR fines within each sector, could provide a more comprehensive understanding of risk factors. Additionally, collecting supplementary data from the ICO and other sources could better the risk assessment process, leading to a more accurate representation of the risks associated with data breaches. Additionally, the ability to implement a search feature for organisations rather than a drop down would greatly improve usability and streamline the filtering process, making it easier for users to locate and select organizations of interest.

If more time were available, the next step for the dashboard would be to transform it into a primary tool for users to report data protection complaints directly to the ICO. This enhancement would enable users to have direct communication with the ICO and track the progress of their cases in real-time. Implementing live updates on the graphs would allow all stakeholders to monitor trends and patterns as they emerge, allowing timely actions to mitigate data protection complaints. Additionally, integrating features such as automated notifications and personalized recommendations based on case specifics could further improve and streamline the complaint handling process.

In conclusion, this project has effectively addressed the gap in attention given to data protection complaints, shedding light on breaches from the client side and providing valuable insights into instances where individuals' data rights have been compromised. By developing a comprehensive web-based dashboard and implementing tools for data analysis, visualization, risk assessment, and GDPR compliance guidance, the project has contributed significantly to understanding and addressing data protection concerns.

References

Kirk, A. (2019) | “*Data Visualisation: A Handbook for Data Driven Design*.”

B. V. Vikas, N. S. Karthikeya, G. S. Chandu, G. Raja and N. R. Sai. (2023) | *"Study of Enhancing Usage of Data Visualization in Cyber Security- Quick, Efficient, and Complete"* | Available at: https://ieeexplore.ieee.org/document/10073891

Balakrishnan, B. (2015) | *“Security Data Visualization.”* | Available at: https://www.giac.org/paper/gslc/7633/security-data-visualization/122613

A. Rossi and M. Palmirani. (2017) | "*A Visualization Approach for Adaptive Consent in the European Data Protection Framework"* | Available at: https://ieeexplore.ieee.org/document/8046282

Rossi, A. (2019) | *“Legal design for the General Data Protection Regulation. A methodology for the visualization and communication of legal concepts”* | Available at: https://amsdottorato.unibo.it/9060/1/rossi_arianna_tesi.pdf

TIBCO Blog. Spotfire Blogging Team. (2013) | “How Data Visualization Helps Customer Complaint Management” | Available at: https://www.tibco.com/blog/2013/02/12/how-data-visualization-helps-customer-complaint-management/

A. Skendžić, B. Kovačić and E. Tijan. (2018) | *"General data protection regulation - Protection of personal data in an organisation."* | Available at: https://ieeexplore.ieee.org/document/8400247

Staheli, Diane; Yu, Tamara; Crouser, R. Jordan; Damodaran, Suresh; Nam, Kevin; O'Gwynn, David; Harrison, Lane; and McKenna, Sean. (2014) | *"Visualization Evaluation for Cyber Security: Trends and Future Directions"* | Available at: https://scholarworks.smith.edu/csc_facpubs/94

Beveridge, C. (2023) | *"The ICO publishes complaints and concerns data sets"* | Available at: https://www.bdo.co.uk/en-gb/insights/advisory/risk-and-advisory-services/the-ico-publishes-complaints-and-concerns-data-sets

ICO. Personal Data | *"What is personal data?"* | Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-is-personal-data/

GDPR Info. | *"General Data Protection Regulation (GDPR)"* | Available at: https://gdpr-info.eu/

Boehme-Neßler, Volker. (2011) | *"Pictorial Law. Modern Law and the Power of Pictures"* | Available at: https://www.researchgate.net/publication/233961397_Pictorial_Law_Modern_Law_and_the_Power_of_Pictures

Kust, M. (2023) | *"How can you calculate the probability of a risk in project management?"* | Available at: https://www.linkedin.com/advice/3/how-can-you-calculate-probability-risk-project

Tan, D. (2002) | *"Quantitative Risk Analysis Step-By-Step"* | Available at: http://130.18.86.27/faculty/warkentin/SecurityPapers/Robert/Others/Tan2003_SANS_QuantitativeRiskAnalysis.pdf

Seralouk (2022) | *"Can someone explain to me how MinMaxScaler() works*?" | Available at: https://stackoverflow.com/questions/62178888/can-someone-explain-to-me-how-minmaxscaler-works

King, C. (2023) | *"Utilities industry defending against increased cyber-attacks"* | Available at: https://cybermagazine.com/articles/utilities-industry-defending-against-increased-cyber-attacks

Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon. (2019) | *"A Review on Data Cleansing Methods for Big Data"* | Available at: https://www.sciencedirect.com/science/article/pii/S1877050919318885

Liebchen, G.A. (2010) | *"Data cleaning techniques for software engineering data sets"* | Available at: https://bura.brunel.ac.uk/handle/2438/5951

Ekdahl , A. and Nyman, L. (2018) | *“A Methodology to Validate Compliance to the GDPR”* | Available at: https://gupea.ub.gu.se/bitstream/handle/2077/62557/gupea_2077_62557_1.pdf?sequence=1&isAllowed=y

Pandas tutorial | Available at: https://www.w3schools.com/python/pandas/default.asp

K. H. Prasad, T. A. Faruquie, S. Joshi, S. Chaturvedi, L. V. Subramaniam and M. Mohania. (2011) | *"Data Cleansing Techniques for Large Enterprise Datasets”* | Available at: https://ieeexplore.ieee.org/document/5958082

Irwin, L. (2020) | *“Why risk assessments are essential for GDPR”* | Available at: https://www.itgovernance.co.uk/blog/why-risk-assessments-are-essential-for-gdpr-compliance

Sullivan, J. (2022) | *“The cost of poor data quality, British Retail Consortium”* | Available at: https://brc.org.uk/news/customer/the-cost-of-poor-data-quality/

ICO Datasets. (2020) | *“Data Protection Complaints - data sets”* | Available at: https://ico.org.uk/about-the-ico/our-information/complaints-and-concerns-data-sets/data-protection-complaints/

ICO Complaints Tool. | *“Data Protection and Personal Information Complaints Tool”* | Available at: https://ico.org.uk/make-a-complaint/data-protection-complaints/data-protection-complaints/

Y. -S. Martin and A. Kung. (2018) | *"Methods and Tools for GDPR Compliance Through Privacy and Data Protection Engineering"* | Available at: https://ieeexplore.ieee.org/document/8406568

Aven, T. (2012) | *“Foundations of Risk Analysis”* | Chichester: John Wiley.

V. Mokhor, S. Honchar and A. Onyskova, (2020) | *"Cybersecurity Risk Assessment of Information Systems of Critical Infrastructure Objects"* | Available at: https://ieeexplore.ieee.org/document/9467957

scikit. | *“Sklearn.preprocessing.MinMaxScaler”* | Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Appendices

Appendix A: Project Definition Document

Appendix B: Reuse Summary

Appendix C: Meeting Log

Appendix D: Requirements

Appendix E: Use-Case Specification

Appendix F: Potential Visualisations

Appendix G: Complaints Procedure Flow Charts

Appendix H: Mock-ups

Appendix I: Tools Used

Appendix J: Additional and Un-used Visualisations

Appendix K: Testing

Web-Based GDPR Data Protection Complaints Dashboard

by
Aadil Saiyad
Aadil.Saiyad@city.ac.uk

Individual Project
Consultant: Andrey Povyakalo

City, University of London
MSci Computer Science with Cyber Security
2024

Word Count: 2200

Table of Contents

Abstract

Data protection is essential to safeguard individual privacy in today's digital age. In an era of widespread data processing, regulations ensure that organisations handle personal data transparently, with individuals having control over their information, encouraging transparency and responsible data management. With strict penalties for non-compliance, these regulations incentivise organisations to prioritise data protection. This also means that people who believe that their privacy has been breached, can confidently report this to regulators who handle their query with the necessary actions. The ability to view the trends in these complaints help the public and organisations make informed decisions on how they can improve their own privacy and the privacy of the general public and consumers.

Problems to be Solved

The main problem my project will tackle to solve is to provide a visual insight into the vast cases of data protection complaints handled by the information commission office from members of the public about data protection concerns. The lack of a user-friendly dashboard makes it difficult for the general public, businesses, and other stakeholders to access and understand the raw data on data protection complaints handled by the Information Commissioner's Office. A dashboard provides a visual and comprehensible interface, promoting transparency and making the information more accessible to a broader audience.

This will be a web-based dashboard on ICO datasets of complaints which will serve as an interactive platform, making information easily accessible. By providing a visual representation of data protection complaints, the dashboard offers a clear and concise overview of the types of issues individuals are raising with the ICO. Visualizations such as charts, graphs, and trends enable users to grasp complex information quickly.

Due this data being accessible in a raw format, this dashboard will solve the problem of not being able to explore and manipulate the data according to the user's specific needs. This allows users to control the data, apply filters, and generate personalised visualisations, tailoring the information to their unique requirements which is currently not possible. Enabling the users to sort and filter data based on relevant parameters is beneficial for consumers looking to understand industry-specific complaints, researchers focusing on particular trends, or legal professionals seeking insights into specific types of cases.

Creating this dashboard, we can help support GDPR compliance efforts. By promoting transparency with the public, facilitating risk identification and mitigation methods to organisations, and supporting regulatory oversight to authorities, the dashboard contributes to a data protection framework that can help uphold the principles and requirements of the GDPR compliances. This is because the dashboard will have the ability to display GRPR clauses relating to the data protection complaint, helping see the pattern between them and how these can be mitigated.

The dataset I will be using is the Data protection complaints data sets highlighting complaints the ICO have handled from members of the public about data protection concerns. Each of these datasets are based on a quarter of a year starting from 2020 to 2024. Within these datasets there includes the status of complaints made by people against entities that handle their data. The reason for the complaint is not specific, but highlighted within a legislation reason which gives us a sense of the general reason why the complaint had arisen. Furthermore, the sector of the entity is specified as well as its subsector, allowing us to see the trends of complaints within these areas of work. The decision made in brief is also documented within the dataset by the ICO which also gives us an idea of whether an infringement occurred and how often they do over the recorded time.

The datasets ICO provide are free to be re-used without charge. However, I must comply with data protection regulations to ensure the use of this information is not misused. This means compliance with GDPR law, transparency of the data used, attribution to the ICO since they are providing this data and using it in an ethical manner. **(ICO, 2024)**.

Project Objectives

The main objective of this project is to develop a comprehensive web-based dashboard tool for ICO Data sets of complaints they have handled from members of the public about data protection concerns, offering a user-friendly interface to visualise, analyse, and understand the compiled data.

Sub-objectives table

Sub-objectives	Testing
Analyse, aggregate and clean all datasets into one singular dataset.	Run data integrity checks, cross-referencing with ICO official records and verify that all relevant data fields are included and accurately represented.
Create the web-based user interface that is appealing and easy to navigate.	Implement user testing sessions where I can receive feedback on how the UI performs.
Data visualisation development which can display graphs and trends related to the compiled dataset.	Ensure implemented visualisations are providing the correct information that correlates to the data in the dataset.
Implement data manipulation and customisation features including filters.	Ensure relevant customisation options are available and operable as well as correct results when filtering.
Ease of use for various devices and browsers.	Test the dashboard on different devices (desktops, laptops, and smartphones) and browsers to confirm compatibility and responsiveness.
Website accessibility compliance to ensure usability for users of all abilities and needs.	Use tools such as WAVE and Lighthouse to ensure the site is compliant.
Ensuring the security of the site is up to date	Test the system against security vulnerabilities and ensure data compliance is adhered to.

Project Beneficiaries

One beneficiary of this tool will be the general public. This is because it will help individuals seeking information on data protection complaints easy access to a transparent overview of the types of concerns raised by others. This allows them to make informed decisions about how they share their personal information online, understanding the potential risks and challenges reported by others. This also allows them to understand their rights in terms of data protection by seeing examples of complaints and resolutions. This can help them become more aware of the protections afforded to them under data protection regulations and increase their awareness of current and past privacy issues. This in-turn can help increase the trust between users and organisations and also with regulatory bodies such as the ICO as they can see how complaints are handled and resolved. Furthermore, if the results of these cases prove to be inadequate, the public can seek clarification from organisations, and advocate for stronger data protection measures.

Furthermore, this means that the businesses and organisations handling the data can use the dashboard to assess the common data protection challenges faced by individuals. This insight aids in tailoring organisational practices to address specific concerns, ensuring better compliance with data protection regulations especially within the scope of their sector. This helps build trust with their consumers and stakeholders through proactive and responsive data management. This tool can mainly be made use of by Data Protection Officers within an organisation by staying informed about the external landscape of data protection concerns which in turn, allows them to provide relevant advice to their organizations, ensuring that internal practices align with external expectations and regulatory standards.

Another beneficiary this tool can be useful for is researchers and academia. Researchers can find value in the dashboard as it will provide a rich dataset for studying trends and patterns in data protection complaints. This can help offer evidence of real-world data protection issues reported by individuals where researchers can use this evidence to support their studies. They can also use this evidence to inform researchers about practical challenges faced by individuals in data protection. Researchers can use this information to formulate policy recommendations, contributing to discussions on enhancing privacy legislation and regulatory enforcement. This also means that the dashboard can be used for creating educational material for students and professionals studying data protection as it provides real-world examples within a rich dataset.

Regulators and policymakers are also a beneficiary of this tool. These regulatory bodies can use the dashboard to identify trends and patterns in data protection complaints. These insights can help inform and contribute to policymaking, enabling the development of targeted regulations that address specific challenges faced by the public. It may also benefit resource allocation and regulatory interventions which will help these regulators be more strategic with this data. This also means that international regulatory bodies from different parts of the world to benchmark their own data protection frameworks against the challenges faced by individuals. This also facilitates collaboration and information exchange among regulatory bodies globally where they can share practices, discussing effective enforcement strategies, and collectively address common issues faced by individuals amongst regulatory compliance. Additionally, another beneficiary can be legal professionals where they can leverage the dashboard to stay updated on the evolving landscape of data protection complaints and regulatory actions. This information is crucial for advising clients on compliance strategies, ensuring legal practices align with the changing expectations of data protection authorities and the public. This also means that these professionals can analyse the types of complaints and can conduct a more thorough risk assessment for their clients. This insight helps in identifying potential areas of non-compliance and promoting a more proactive rather than a reactive approach to implementing measures to mitigate legal risks.

Work Plan

The following table will show the schedule of what needs to be completed in my project. I will use this table in order to keep track of the tasks that needs to be done. The date when I've started working on a specific task and how many days it took to complete it.

Work Plan Table

Task	Status	Start Date	End Date	Days	Priority 1-4
Changes to PDD Document	Not Started	05/02/24	11/02/24	7	4
Conduct research on tools	Not Started	05/02/24	07/02/24	2	2
Conduct research on visualisations	Not Started	05/02/24	07/02/24	2	2
Analyse and aggregate dataset	Not Started	05/02/24	11/02/24	7	4
Create web-based UI	Not Started	TBA	TBA	TBA	TBA
Test web-based UI	Not Started	TBA	TBA	TBA	TBA
Implement Data Visualisation capabilities	Not Started	TBA	TBA	TBA	TBA
Test Data Visualisation capabilities	Not Started	TBA	TBA	TBA	TBA
Implement compatibility multi-device for dashboard	Not Started	TBA	TBA	TBA	TBA
Ensure dashboard is accessible	Not Started	TBA	TBA	TBA	TBA
Ensure dashboard is secure	Not Started	TBA	TBA	TBA	TBA
System testing and final additions	Not Started	TBA	TBA	TBA	TBA
Project Submission	Not Started	31/04/24	31/04/24	1	4

Research on tools to be used to produce this project are to be conducted. This includes the programming language and libraries that will be used to help create the dashboard and visualisations. The current plan is to use an open-source Python library named "Taipy" **(Avaiga, 2021)** which aids with web development and data visualisations. However, the use of a JavaScript library named "D3" **(Mike Bostock and Observable, Inc., 2024)** is also being considered. This is a library specifically for visualising data which is interactable and flexible. However, using D3 will mean I will need to use a library to manipulate data into a webpage such as "node.js" **(node.js, 2019)** or "flask" **(pallets, 2010)**. In doing so may make my project complicated and unfeasible to complete within the given time frame so additional research is to be conducted to ensure that this is not the case.

Project Risks Table

Risks	Likeliness (1 Low – 4 High)	Severity (1 Low – 4 High)	Mitigation
Inaccurate or incomplete data from the datasets.	1	4	Ensure the combined datasets align with the official ICO's records and cross reference to see if any data is missing or incomplete.
Scope creep where I spend more time on a certain part of the project than intended.	2	3	Stick to project plan and ensure that that the parameters of the project are adhered to.
Technology Risk (Unable to implement a tool or system I have planned to use).	2	4	Familiarise myself with the tools, languages, and systems I plan to use on the system.
No participants to take part in user testing.	2	2	Ensure I give apt time and notice to participants before taking part and ensure they are ready for when the system is up to test.
Time Crunch where project overruns the deadline.	2	4	Implement buffer zones in the project plan. Perform research on the tasks I am planning for to ensure that it I'm giving myself a realistic amount of time to perform the task.
Web-based dashboard is not running smoothly / efficiently.	1	3	Optimise the code of the dashboard and ensure that the graphics and graphs being used are generated efficiently.
Inaccurate and unreliable data representation in the dashboard graphs.	1	4	Ensure that the graphical representation of the data provides a clear correlation of data by testing beforehand to see if they align with the data being used.
Usability of the dashboard is not up to standard and provides an inadequate user experience.	1	3	User testing provides feedback on improvements that will help provide the best dashboard implementation.
Security vulnerabilities breach confidentiality, integrity and availability of the dashboard and data.	1	4	Ensure the dashboard is tested with security protocols against common attacks.
Hardware failure leading to the loss of work and code.	1	4	The use of the cloud to upload documents such as drive and the use of GitHub to ensure code is not lost.

There is no main potential risk to the users of the tool. This is because it does not collect and personal information or store personal data. This negates the need for the protection of confidentiality of data. Necessary security measures will be taken to protect the integrity of the data set and ensure that attackers cannot alter the dataset to skew the visualisations to their benefit. Furthermore, the dashboard will be protected against attacks that target the availability of the data where the site becomes slow or inaccessible.

Ethics Checklist

Checklist Part A

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure, try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners, and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**

2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos, or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological, or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**

A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**YES**

Checklist Part B

B.3 Attachments **ALL of the following documents MUST be provided to supervisors if applicable.** **All must be considered prior to final approval by supervisors.** **A written record of final approval must be provided and retained.**	***YES***	***NO***	***Not Applicable***
Details on how safety will be assured in any non-University location, including risk assessment if required (see B2)			**X**
Details of arrangements to ensure that material and/or private information obtained from or about the participating individuals will remain confidential (see B1.5)	**X**		
Full protocol for any workshops or interviews**			**X**
Participant information sheet(s)**	**X**		
Consent form(s)**	**X**		
Questionnaire(s)** *sharing a Qualtrics survey with your supervisor is recommended.*	**X**		
Topic guide(s) for interviews and focus groups**			**X**
Permission from external organisations or Head of Department** *e.g. for recruitment of participants*	**X**		

References

Data Protection Complaints - data sets, ICO. Available at: https://ico.org.uk/about-the-ico/our-information/complaints-and-concerns-data-sets/data-protection-complaints/ (Accessed: 12 February 2024).

Avaiga Taipy: Turns data and AI algorithms into production-ready web applications. *GitHub*. Available at: https://github.com/Avaiga/taipy (Accessed: January 2024).

D3 by Observable | The JavaScript library for bespoke data visualization. Available at: https://d3js.org/ (Accessed: January 2024).

Node.js. Available at: https://nodejs.org/en (Accessed: January 2024).

Welcome to Flask. Available at: https://flask.palletsprojects.com/en/3.0.x/ (Accessed: January 2024).

Appendix B: Reuse Summary

I didn't reuse any code except for the standard modules and their components, like the charts in Taipy or the MinMaxScaler in scikit-learn. Everything that was used from these modules was properly referenced and documented within the code.

To make my report better, I used Grammarly. It helped fix spelling mistakes and gave me tips on how to write more clearly and directly. It suggested different words I could use and helped rephrase sentences for better flow. By using Grammarly regularly, I improved how I expressed my ideas and made sure my report was easy to understand and professional.

For example:

Initial Paragraph written manually:

To provide a good overview of complaints, I used a pie chart showing the overall decisions made on complaints. This visual representation highlights the distribution of decisions, providing insight into the proportion of cases resulting in action taken versus no action taken. The coding of the pie chart followed a similar process to the bar chart, as both used grouping and counting data to visualise distributions. However, unlike the bar chart, the pie chart does not require sorting of the data since it represents sections rather than ordinal relationships, therefore, grouping the dataset by the Decision column.

Paragraph when corrected and improved:

To provide a comprehensive overview of complaint outcomes, I incorporated a pie chart illustrating the overall decisions made on complaints. This visual representation highlights the distribution of decisions, offering insight into the proportion of cases resulting in action taken versus no action taken. The implementation of the pie chart follows a similar process to the bar chart, as both involve grouping and counting data to visualise distributions. However, unlike the bar chart, the pie chart does not require sorting of the data since it represents categorical proportions rather than ordinal relationships, therefore, grouping the dataset by the Decision column.

This was done for the majority of the project to ensure it is easy to read and is informative.

Appendix C: Meeting Log

ID	Date	Duration (Minutes)	Meeting Consultant	Notes of Meeting	Actions Agreed
1	23/11/23	30	Nikos	Brief discussion on what kind project to pick and how to pick a viable project.	Look into research that has already been conducted such as papers on IEEE Explore
2	13/12/23	30	Nikos	Discussed current list of around 10 - 15 ideas noted down. Assessed the pros and cons of each and generally what they all require.	Ensure these projects are solving a problem for a certain audience and not to be so vague.
3	29/01/24	15	Nikos	Discussed a reduction of 2 potential projects that I would like to pick. Analysed each one and what each one may need to satisfy the PDT requirements.	Ensure that the GDPR Compliance tool is unique and research current tools and what they have or may not have. This can then be used to improve my project and generate more ideas.
4	12/02/24	30	Nikos	Discussed PDD submission and what is good and what changes need to be made.	Agreed to go more in depth into GDPR for the PDD and add the questionnaire template into the document.
5	19/02/24	30	Nikos	Went over the literature review and discussed the improvements needed to be made.	Need to make sure the literature is not general and relates back to the topic at hand.
6	08/03/24	30	Nikos	Discussed additional ideas to be added to the project and reviewed current work done.	Suggested a few additions to the project such as adding a severity or risk value to the cases.
7	22/03/24	30	Nikos	Showed current work done and improvements to be made to the visualisations such as colour for heat map and any additional functions to be added to the site.	Change colour of heatmap to potentially improve the visualisation as well as start of implementing additional features.

8	05/04/24	30	Nikos	Discussed the implementation of the risks page and its features. Also looked over the methods section for this implementation.	Be more specific and ensure I am referring to papers and materials used to justify the decisions and implementations made. Also add a scale to the risks.
9	15/04/24	30	Nikos	Showed the code and project functionality. Discussed the results section and the areas written down so far.	Ensure that the results refer back to the literature review and methods section. Also ensure code is referenced.
10	22/04/24	30	Nikos	Project draft discussion. Spoke about heading and the need to ensure all areas are added to the document such as appendix and references.	Ensure graphs are annotated as well as ensure images are correctly sized. Add necessary heading to paragraphs and make them meaningful as mentioned in the project documentation.

Appendix D: Requirements

Functional Requirements

Requirement ID: Dataset Cleansing 1
Description: The dataset of data protection complaints shall be aggregated from multiple sources into a unified dataset.
Rationale: A merged dataset is necessary to provide a comprehensive overview of data protection complaints.
Originator: Project Creator
Fit Criterion: The dataset contains information from all sources and does not miss any data points.
Priority: High
Requirement ID: Dataset Cleansing 2
Description: The dataset shall be cleansed, removing duplicates, inconsistencies, and irrelevant data entries.
Rationale: Clean and accurate data is essential for accurate analysis and visualization of data protection complaints.
Originator: Project Creator
Fit Criterion: The dataset undergoes a cleansing process resulting in a reduction of duplicate entries and the removal of uncompleted columns.
Priority: High
Requirement ID: Data Visualisation
Description: The system shall generate charts to visualise data based on the settings given within the code.
Rationale: The charts are used to gain insights from the dataset.
Originator: Project Creator
Fit Criterion: The charts present the data accurately and can be used to gain insights on data protection complaints.
Priority: High
Requirement Chart Interactivity
Description: The system shall allow users to make use of chart interactions such as Download plot as PNG, Zoom, Pan, Box select, Lasso select, zoom in, zoom out, auto scale, and reset axes.
Rationale: The ability to interact with the chart
Originator: Project Creator
Fit Criterion: Enhances user engagement and provides a more interactive experience, allowing users to explore data dynamically and gain deeper insights.
Priority: Medium
Requirement ID: Risk Assessment Calculation
Description: The system shall calculate risk values for each data protection complaint based on predefined criteria.
Rationale: Risk values provide insights into the severity of data protection concerns and prioritize actions accordingly.
Originator: Project Creator
Fit Criterion: Risk values are accurately calculated and displayed for all of the complaints.
Priority: High

Requirement ID: Risk Assessment Implementation
Description: The system shall display risk values on the risks page for each complaint, alongside relevant details within the table.
Rationale: Visible risk values enable stakeholders to identify high-priority complaints quickly and take appropriate actions.
Originator: Project Creator
Fit Criterion: Risk values are displayed adjacent to each complaint with clear labels e.g. colouring and formatting for easy interpretation.
Priority: High
Requirement Compliance Advisor 1
Description: The system shall provide recommendations for GDPR compliance based on common breaches identified for each sector/subsector.
Rationale: Recommendations assist organisations in addressing common GDPR compliance challenges and improving data protection practices.
Originator: Project Creator
Fit Criterion: At least the top 10 most common GDRP articles breached should have their article description written out in a simple manner as well as 3 recommendations for each.
Priority: High
Requirement ID: Compliance Advisor 2
Description: The system shall display GDPR compliance recommendations on the compliance advisor page, categorised by sector and subsector.
Rationale: Visible recommendations enable organizations to access relevant guidance easily and implement necessary actions.
Originator: Project Creator
Fit Criterion: Recommendations are displayed in a structured format with clear headings and concise descriptions for each sector and subsector.
Priority: High
Requirement ID: Hosting on Taipy Cloud
Description: The system shall deploy and host the dashboard on the Taipy Cloud platform for accessibility and scalability.
Rationale: Cloud hosting ensures reliable access to the dashboard and accommodates increased user demand over time.
Originator: Project Creator
Fit Criterion: The dashboard is accessible via a secure URL hosted on the Taipy Cloud platform.
Priority: Medium

Non-Functional Requirements

Requirement ID: Performance 1
Description: The system shall respond to user interactions within two seconds under normal operating conditions.
Rationale: Users expect the dashboard to be responsive and provide timely feedback during interactions.
Originator: Project Creator
Fit Criterion: The user interactions result in a response time of less than two seconds.
Priority: High
Requirement ID: Performance 2
Description: The system loads all pages in less than 2 seconds under normal operating conditions.
Rationale: Users expect the dashboard to be responsive and have pages that are accessible within a timely manner.
Originator: Project Creator
Fit Criterion: The pages load within two seconds of transitioning to that page.
Priority: High
Requirement ID: Browser Compatibility
Description: The system shall be compatible with modern web browsers, including Chrome, Firefox, Safari, and Edge, on desktop.
Rationale: Users expect the dashboard to function consistently across different browsers and devices for seamless access and usability.
Originator: Project Creator
Fit Criterion: The dashboard displays correctly and functions without errors on at least 70% of modern, popular web browsers and devices.
Priority: High

Appendix E: Use-Case Specification

Use-case Table:

Use Case Name: Dashboard Page
Use Case ID: 1
Description: This use case involves users accessing a web-based dashboard where they are presented with various data protection visualisations, such as charts and graphs. The users can manipulate and customise these visualisations using interactive elements provided by the dashboard interface, allowing them to filter and adjust the data to their preferences.
Primary Actors: End User
Secondary Actors: None
Pre-Conditions: The Dashboard is up and running on the site. The user has internet access and a compatible web browser.
Main Flow: 1. The user navigates to the website hosting the data protection complaints web-based dashboard. 2. Upon arrival, the user is directed to the dashboard page, which displays various data protection complaint visualisations, such as charts, graphs, and tables. 3. The user reviews the default visualisations presented on the dashboard. 4. The user may utilise interactive elements (e.g., dropdown menus, check boxes, buttons) provided by the dashboard interface to filter, manipulate, or customise the visualisations according to their preferences. 5. As the user applies filters or manipulates the visualisations, the dashboard dynamically updates to reflect the changes made. 6. The user continues to explore and interact with the visualisations, adjusting filters and manipulations as desired. 7. Once satisfied with their exploration, the user may choose to save or export the customised view of the dashboard for future reference.
Alternative Flow: 1. If the user experiences technical difficulties or encounters errors while interacting with the dashboard, they may need to refresh the page or try accessing the dashboard at a later time.
Post-Conditions: The user has successfully interacted with the complaint's dashboard visualisations, filtering and manipulating them to their liking. Any customisations made by the user are saved as an image on their device.

Use Case Name: Risks Page
Use Case ID: 2
Description: This use case enables users to navigate to the Risks page within the data protection complaints website and filter complaints by company name. Users can then review a table of complaints, each assigned a calculated risk factor, and apply a filter to view complaints associated with a specific company. This functionality provides the users with insights into data protection risks associated with individual companies, which may aid in risk assessment and mitigation efforts.
Primary Actors: End User
Secondary Actors: None
Pre-Conditions: The Dashboard is up and running on the site. The user has internet access and a compatible web browser.
Main Flow: 1. The user navigates to the Risks page within the data protection complaints website. 2. Upon arrival, the user is presented with a table containing all registered complaints, along with associated details such as complaint ID, date filed, company name, and nature of complaint. 3. A calculated risk factor is automatically applied to each complaint based on predefined criteria, indicating the severity or likelihood of the complaint resulting in a data protection breach. 4. The user utilises the filtering functionality provided by the dashboard interface to filter complaints by company name. 5. The user enters the desired company name into the filter input field and submits the query. 6. The dashboard dynamically updates the table to display only the complaints associated with the specified company, along with their respective details. 7. The user reviews the filtered complaints to gain insights into the data protection risks associated with the selected company.
Alternative Flow: 1. If the user experiences technical difficulties or encounters errors while interacting with the dashboard, they may need to refresh the page or try accessing the dashboard at a later time.
Post-Conditions: The user has successfully navigated to the "Risks" page and filtered complaints by company name, gaining insights into data protection risks associated with the selected company.

Use Case Name: Recommendations Page
Use Case ID: 3
Description: This use case enables users to navigate to the "Recommendations" page within a data protection complaints website and select their organisation's Sector and Subsector. By choosing their industry Sector and Subsector, users receive tailored information on the most common data protection breaches within their specific industry. The website then provides personalised mitigation recommendations, offering actionable steps and best practices to address these breaches effectively. This functionality assists organisations in improving their data protection practices and minimising potential risks associated with data breaches.
Primary Actors: End User
Secondary Actors: None
Pre-Conditions: The Dashboard is up and running on the site. The user has internet access and a compatible web browser. The user has a listed Sector and Subsector to pick
Main Flow: 1. The user navigates to the Recommendations page within the data protection complaints website. 2. Upon arrival, the user is presented with a list of sectors relevant to their organisation, such as healthcare, finance, education, etc. 3. The user selects the sector that best represents their organisation's industry from the provided list. 4. Upon selecting a sector, the dashboard dynamically populates a dropdown menu with subsector options specific to the chosen sector. 5. The user selects the subsector that aligns with their organisation's activities from the dropdown menu. 6. After selecting the subsector, the dashboard automatically calculates and generates a list of the most common data protection breaches within that subsector. 7. The dashboard provides catered information and recommendations on how to mitigate these breaches, offering actionable steps and best practices tailored to the user's subsector.
Alternative Flow: The user does not find a Sector or Subsector that represents their organisation correctly The user does find their Sector and Subsector but the output for the recommendations gives an error due to not enough information about the breach and provides a link for more information.
Post-Conditions: The user has successfully navigated to the "Recommendations" page, selected their organisation's subsector, and viewed catered information and mitigation recommendations tailored to their industry.

Use-case Diagram:

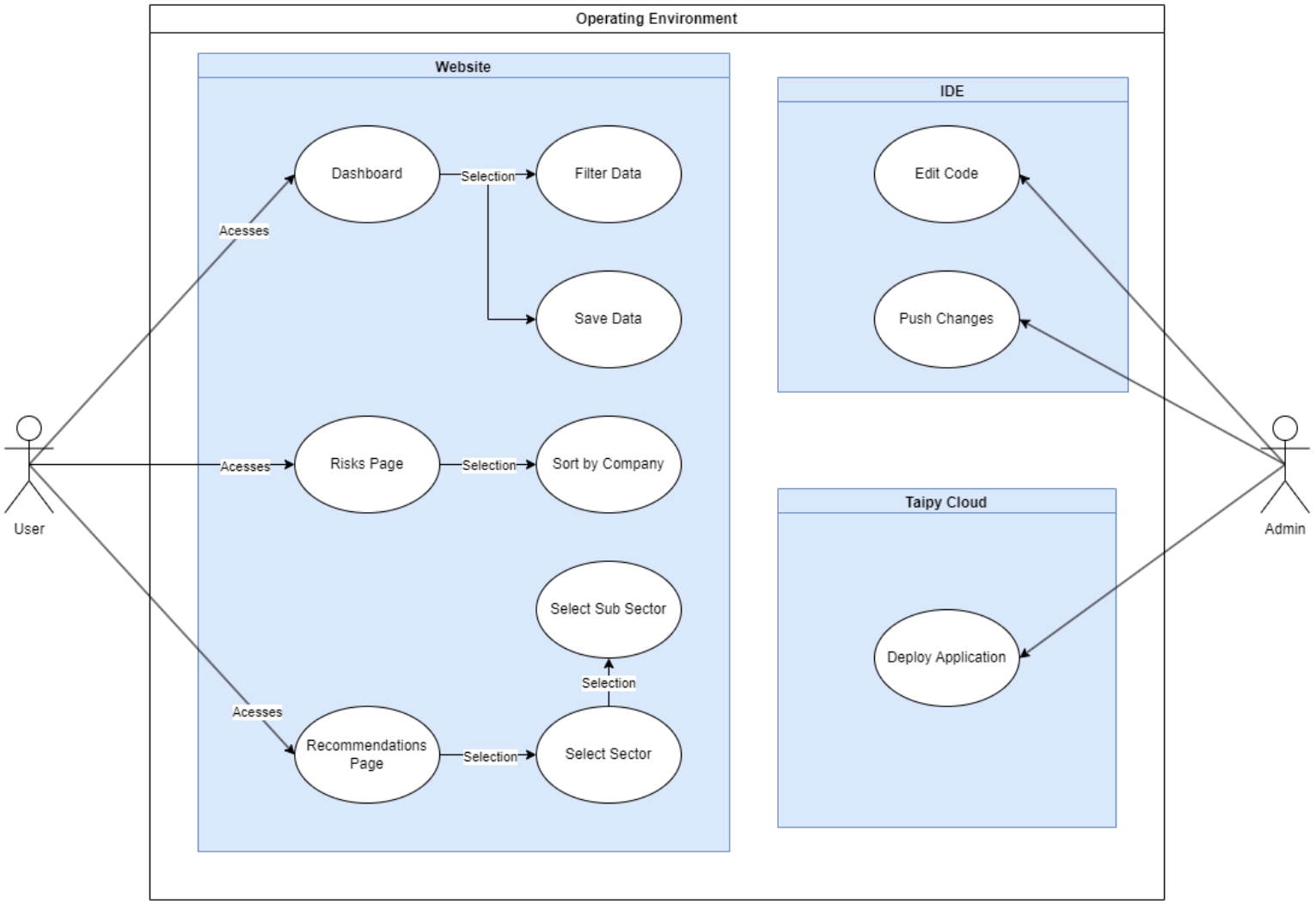


Figure 30: Use Case Diagram

Appendix F: Potential visualisations

Potential Visualizations Table:

ID	Visualisation	Type	Data Used	Description	Used
1	Sector Count / Case	Bar Chart or Pie Chart	Sector, Count	Visualising the total number of cases per sector to see if any sectors stand out compared to others.	**X**
2	Overall Decisions	Bar Chart or Pie Chart	Decision, Count	Visualising the overall decisions made to see the outcome of cases.	**X**
3	Total Cases	Number Representation	Case Reference, Count	Displaying the overall cases recorded to show how many have been processed.	**X**
4	Re-opened Cases	Number Representation	Case Reference, Count	Shows the cases that have been re-opened for further investigation which may show how re-investigations occur.	
5	Article Breached Count	Bar Chart, Pie Chart	Decision Article, Count	Shows the main GDPR articles breached for the submitted complaints.	**X**
6	Sector Count / Case with subsectors	Tree map	Sector, Count, Sub Sector	Can see the size difference between Sectors as well as sub sectors within one another.	
7	Complaints Submitted / Time	Line Graph	Date Received	Can see the cases submitted over time showing peaks and troughs.	**X**
8	Complaints Received / Quarter	Bar Chart or Line Graph	Received Qtr, Count	Can see the cases received per quarter of the year potentially showing trends in areas of the year	
9	Average Time to process a complaint	Bar Chart or Number Representation	Date Received, Date Completed, Average	Can show the time taken for a case to be processed which may show them to be too long or quick.	
10	Average Time to process a complaint / Sector	Bar Chart	Date Received, Date Completed, Average, Sector	Shows which sector takes the longest for a case to be processed.	**X**

11	Number of Complaints / Sub Sector	Bar Chart	Sub Sector, Count	Visualising the total number of cases per sub sector to see if any sub sectors stand out compared to others.	
12	Comparison of Decisions made and no Decisions made / Sector	Stacked Bar Chart	Sector, Decision	Showing the difference between the decisions made per sector and how they vary.	**X**
13	Cases / Day or Month	Heatmap	Sector, Day, Month	Shows the cases submitted on each day or month to see how they vary.	**X**
14	Plot date received vs date completed on the same graph	Line Graph	Date Received, Date Completed	Can potentially show a trend between	
15	How each case flows into sectors and subsectors	Sankey Diagram	Sector, Sub Sector, Count	Shows how cases are distributed within sectors and subsectors showing the flow of information.	**X**

Appendix G: Complaints flow charts

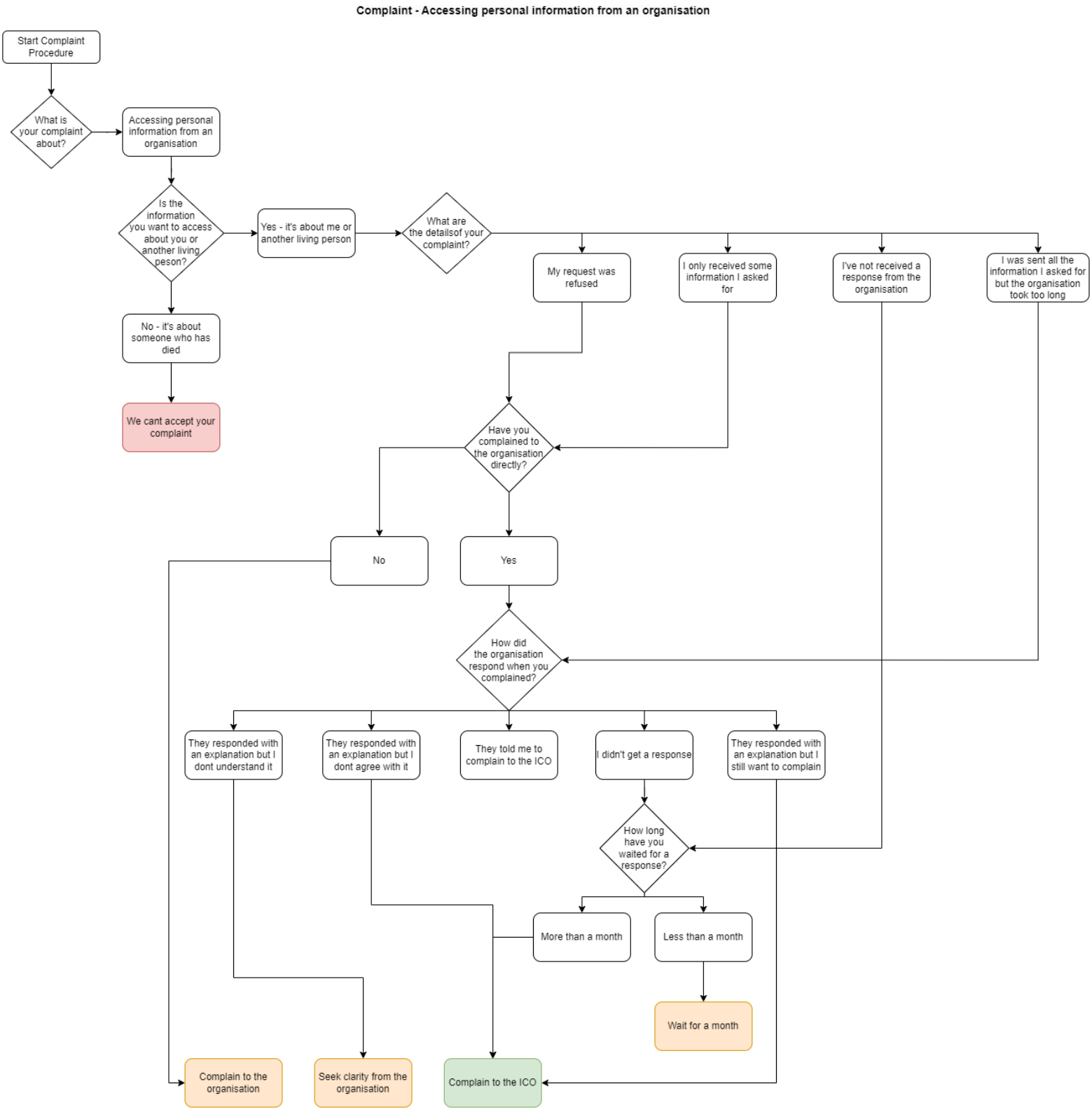


Figure 31: Accessing Personal Information from an Organisation Complaint Flow

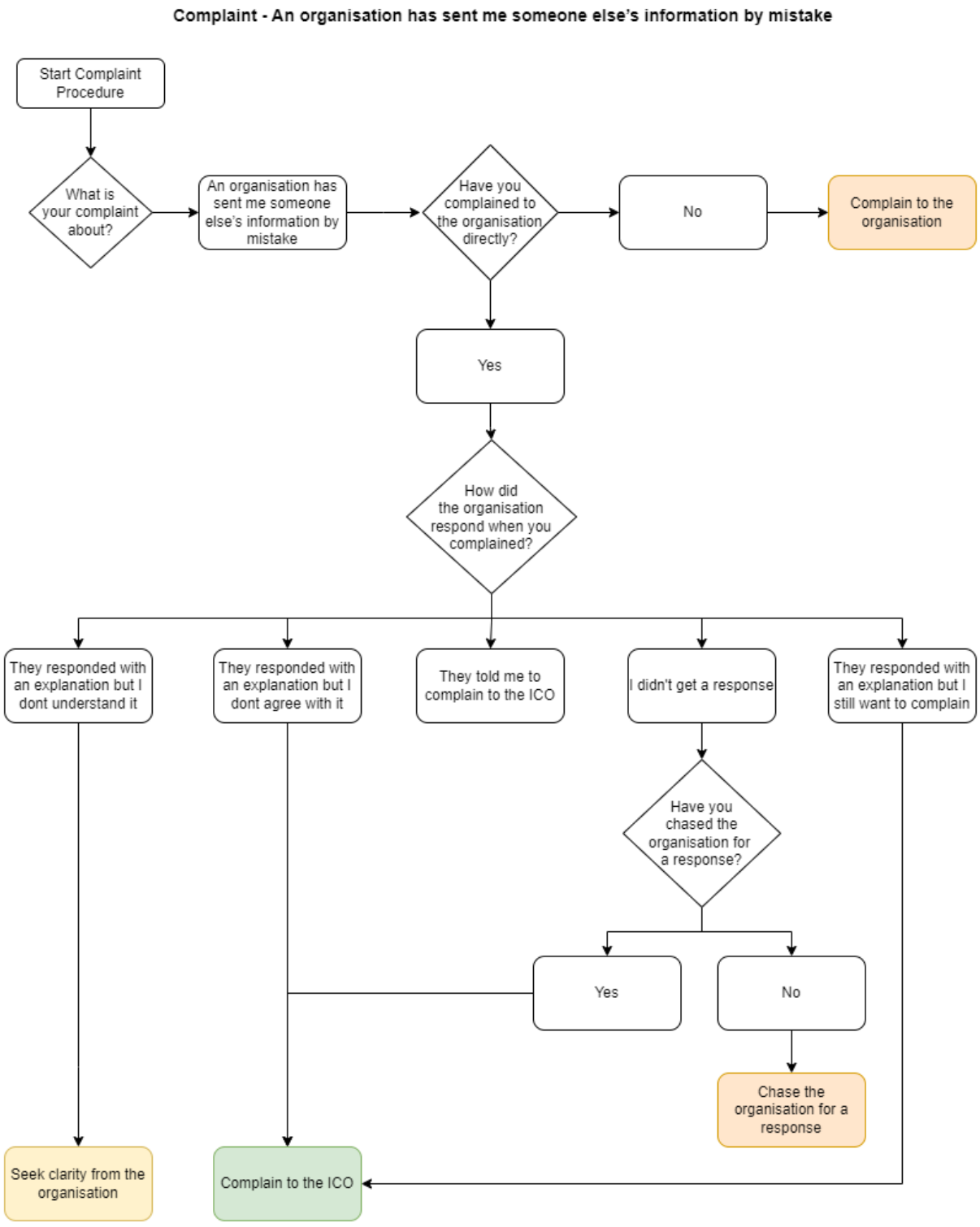


Figure 32: An Organisation Has Sent Me Someone Else's Information By Mistake Complaint Flow

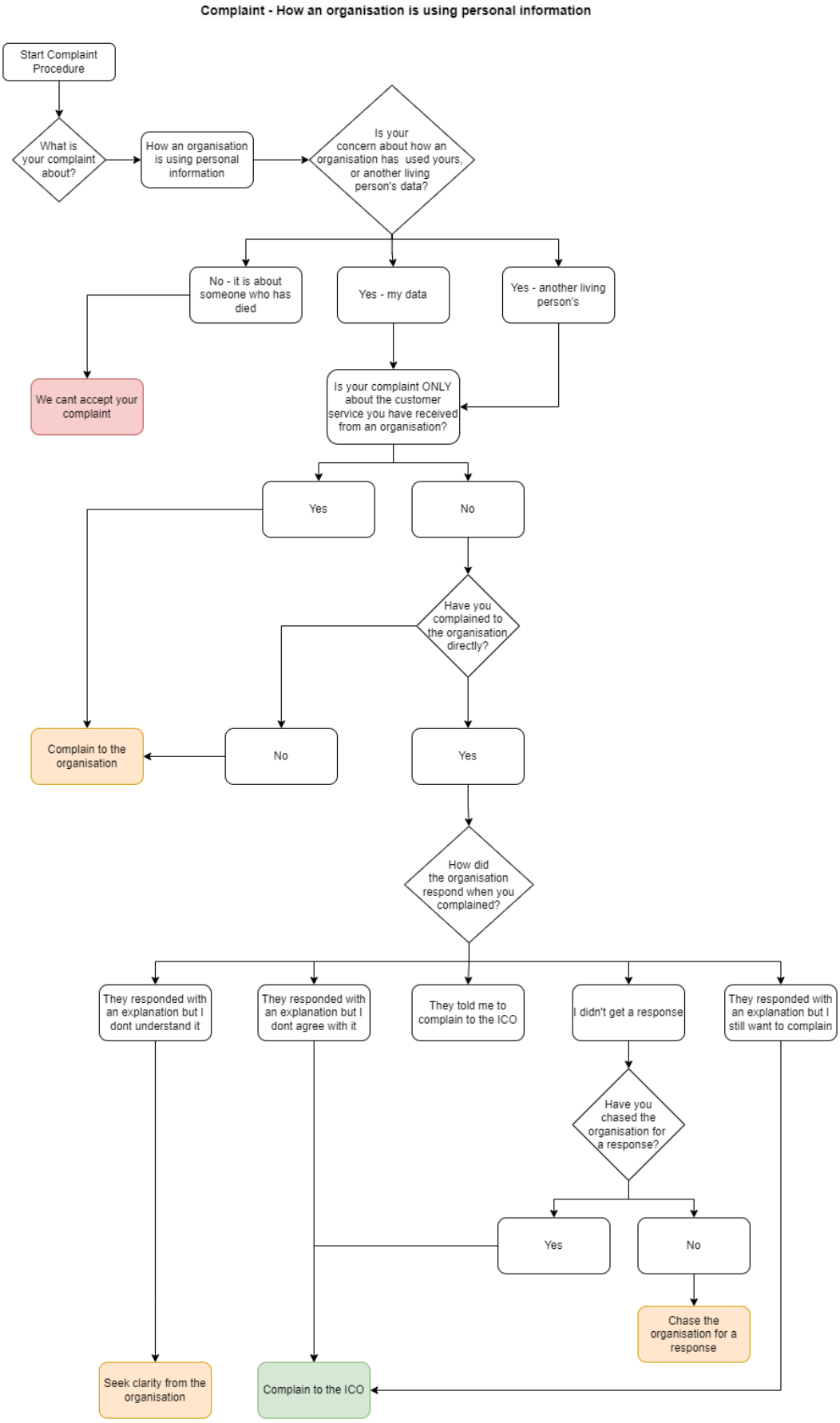


Figure 33: How an Organisation is Using Personal Information Complaint Flow

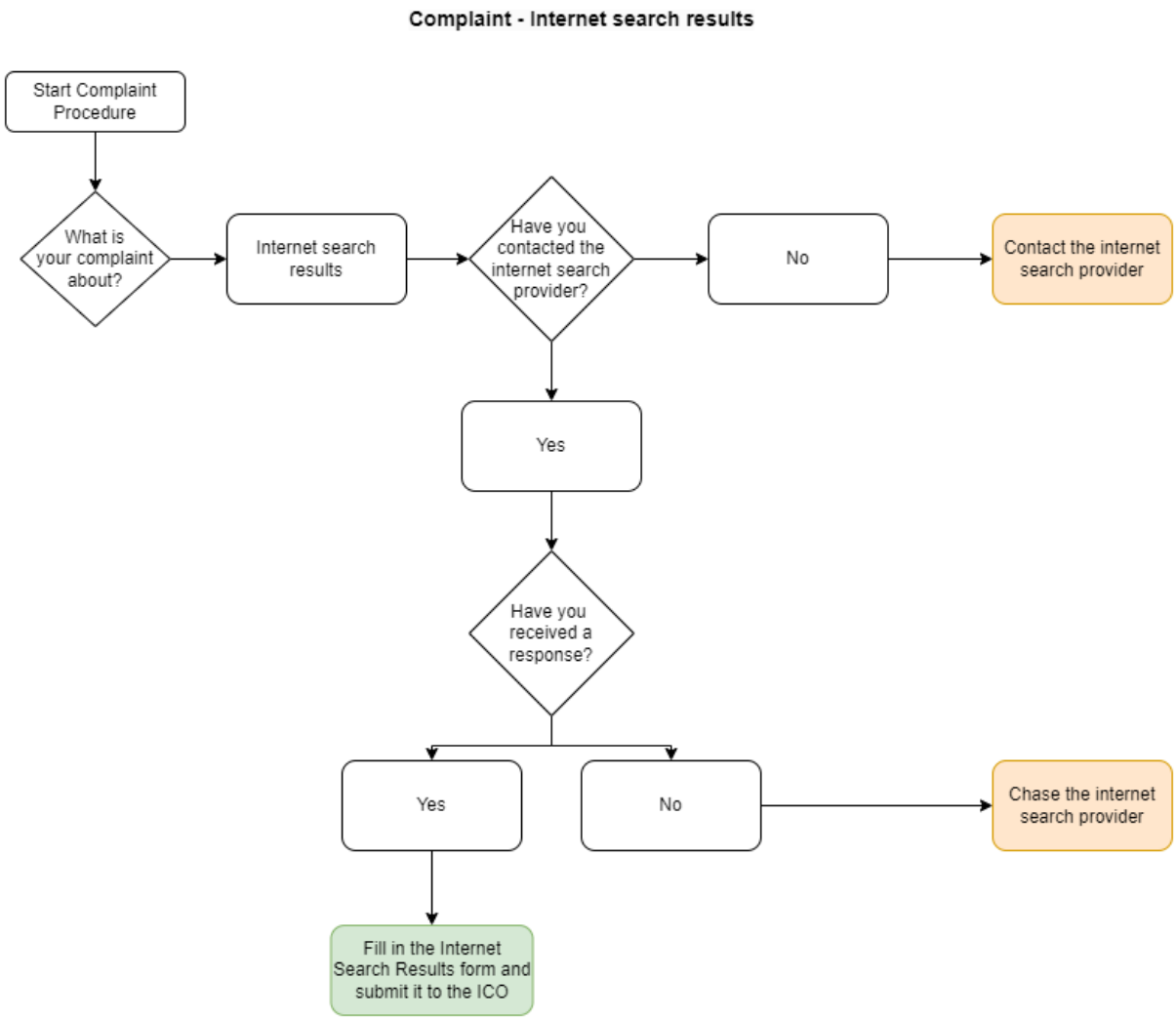


Figure 34: Internet Search Results Complaint Flow

Appendix H: Mock-ups

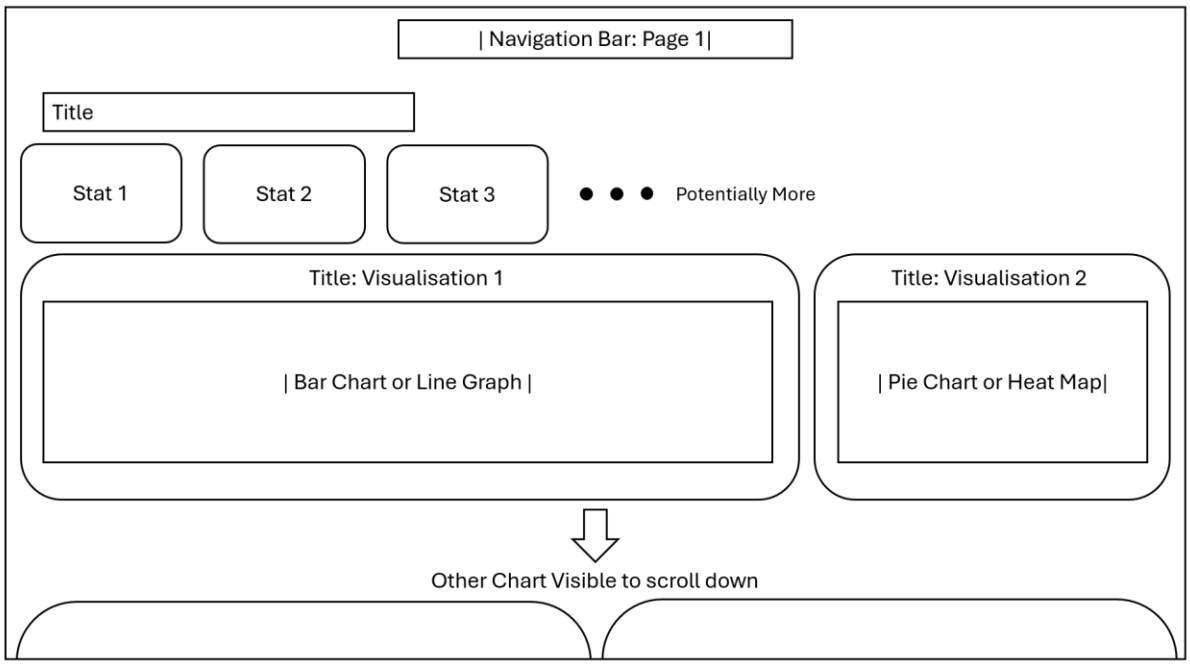


Figure 35: Main Page Mock-up

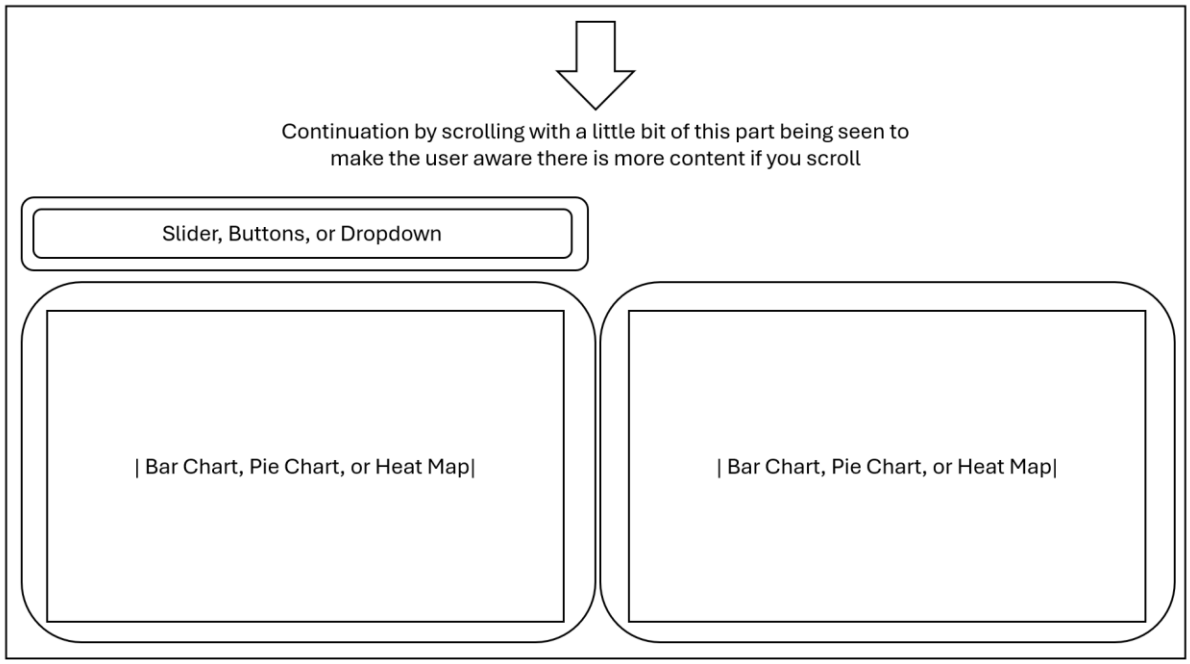


Figure 36: Main Page Scrolled Mock-up

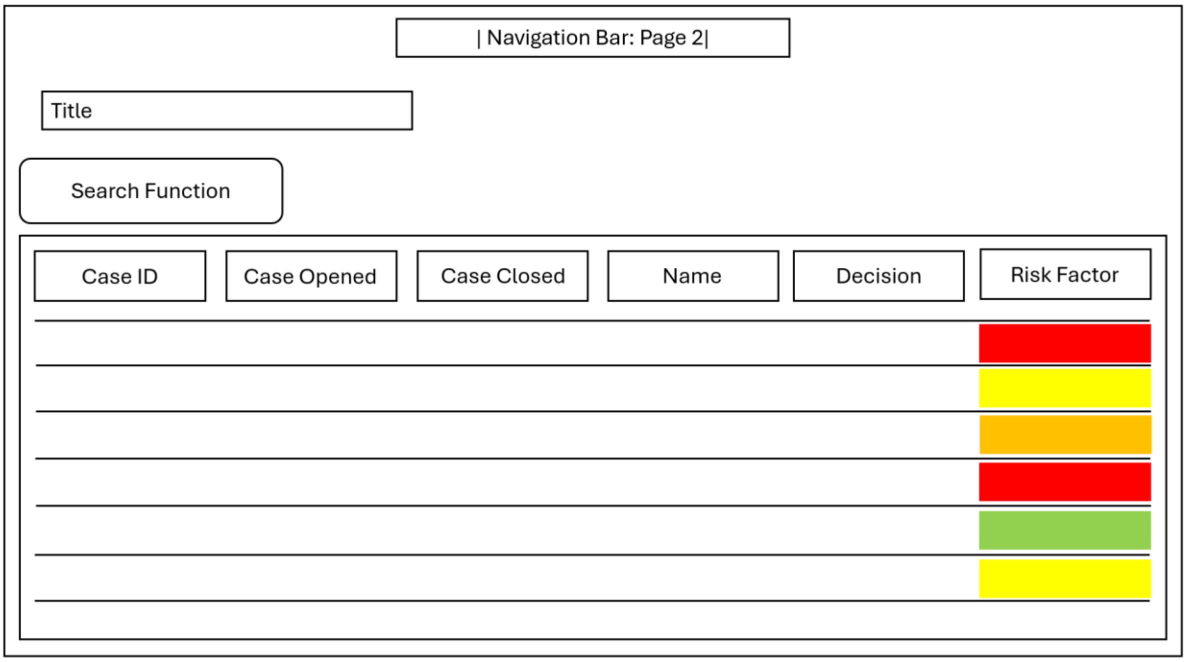


Figure 37: Risk Assessment Page Mock-up

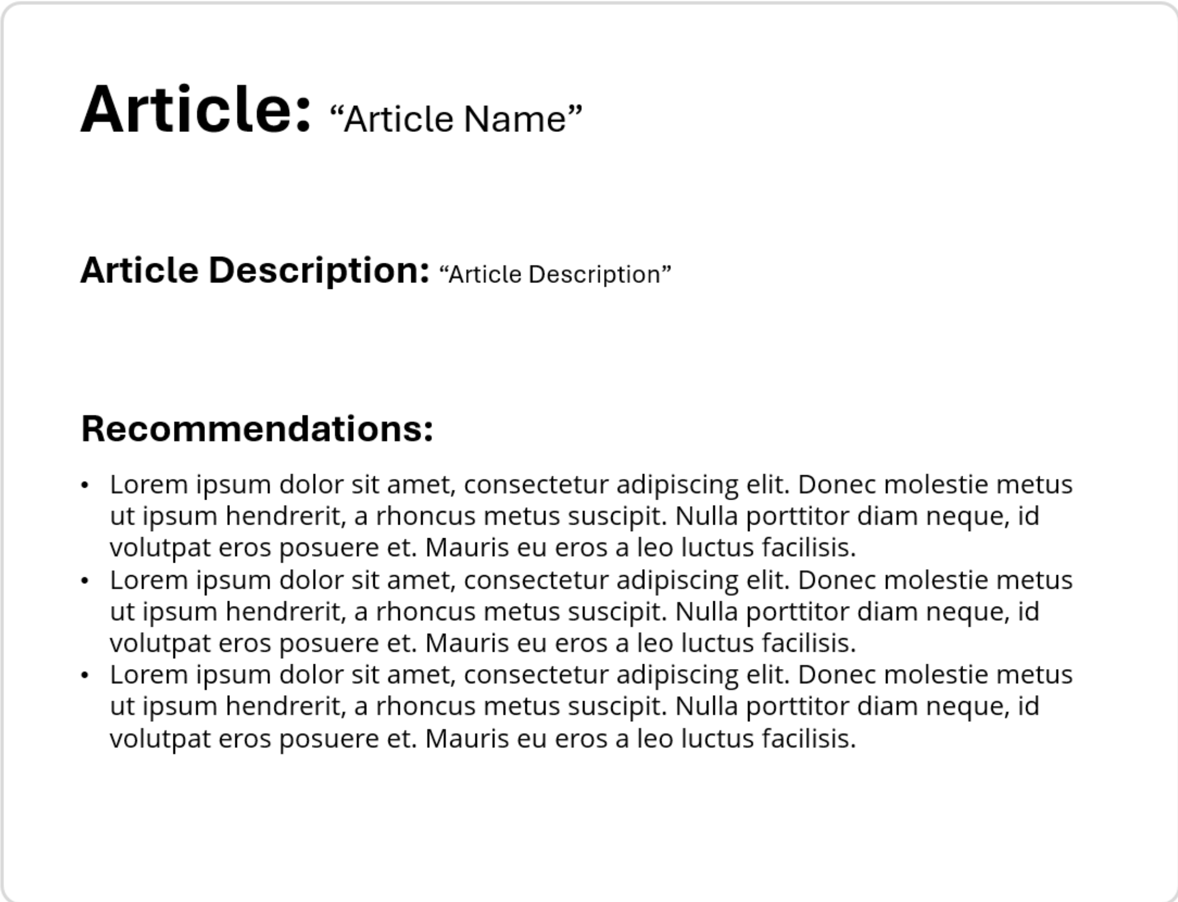


Figure 38: Advisor Article Mock-up

Appendix I: Tools Used and Project Structure

Project Structure:

The code:

- One main.py file which contains the main code of the project.
- 5x Markdown files containing the markdown code needed to display items on the page.
  - root.md - Contains the markdown for displaying the navbar over all pages.
  - dashboard.md - Contains markdown for visualising the charts on the main dashboard.
  - risks.md - Contains the markdown for displaying the table and dropdown for organisations.
  - advisor.md - Contains the markdown for displaying the Sectors and subsectors columns and GDPR recommendations.
  - instructions.md - Contains the markdown for displaying the instructions on how to use the site.
- One CSS file required for the colouring of the Risks table.
- One JSON file required for the GDPR Compliance advisor page which contains the GDPR articles, description and recommendations.
- One dataCleansing.py file not needed to run but showcases parts of the data preparation stage.

The dataset:

- One CSV dataset containing the Data Protection Complaints data needed to visualise the data.

Images:

- One Image of the Sankey Diagram.

Cloud File:

- One requirements text file used to host the project on Taipy Cloud and help the system get the necessary modules.

Requirements and Tools of the Project:

- IDE: The IDE used during this project was Visual Studio Code Windows 64bit so the same would be ideal.

https://code.visualstudio.com/

- Python: The Python version used in development was 3.12.2 64bit for Windows so the same would be ideal.

https://www.python.org/downloads/

- Pip: Pip is included by default if you use Python 3.4 or later. Otherwise, you can follow the official installation page of pip to install it.

https://pip.pypa.io/en/stable/installation/

- Taipy: The preferred method to install Taipy is by using pip. "pip install taipy". If this does not work, I implemented it using "py -m pip install taipy". However, there are multiple ways to install using this tutorial depending on setup.

https://docs.taipy.io/en/release-3.1/installation/

- Scikit learn: Scikit learn is required to implement the MinMaxScaler normalisation for the risks table. This is done using the command "pip install -U scikit-learn". If this does not work, I implemented it using "py -m pip install scikit-learn".

Further information can be found on their webpage.

https://scikit-learn.org/stable/install.html

- Pandas: Data maipulation and analysis tool within Python used for data cleansing, preperation, and visualisation. This is installed using the command "pip install pandas".

Further information can be found on their webpage.

https://pandas.pydata.org/docs/index.html

- Taipy Cloud Account: In order to host the site on the Taipy cloud, a free account must be created which gives you access to host an application on a machine for 6 hours a day.

- Creating a Machine: Once the account is created, you can set up your machine with the details you desire, however, ensuring the Python version is set to 3.11.

- Adding the Application: Once the machine is ready, you can add the application to it, ensuring that the project is within a ZIP file and the entry point of the file is set to "main.py". Furthermore, the requirements file should be set to "requirements.txt".

Running this application should then successfully host the Data Protection Complaints Dashboard on the Taipy Cloud URL you set for anyone who has the link to use.

Complaints per Sub Sector

The inclusion of subsectors as an additional bar chart offers deeper insights into the distribution of data protection complaints, enabling targeted interventions in areas that require heightened attention and resources. I followed a similar methodology to create the subsector bar chart, focusing on the top 10 subsectors due to the large number of sub sectors listed.

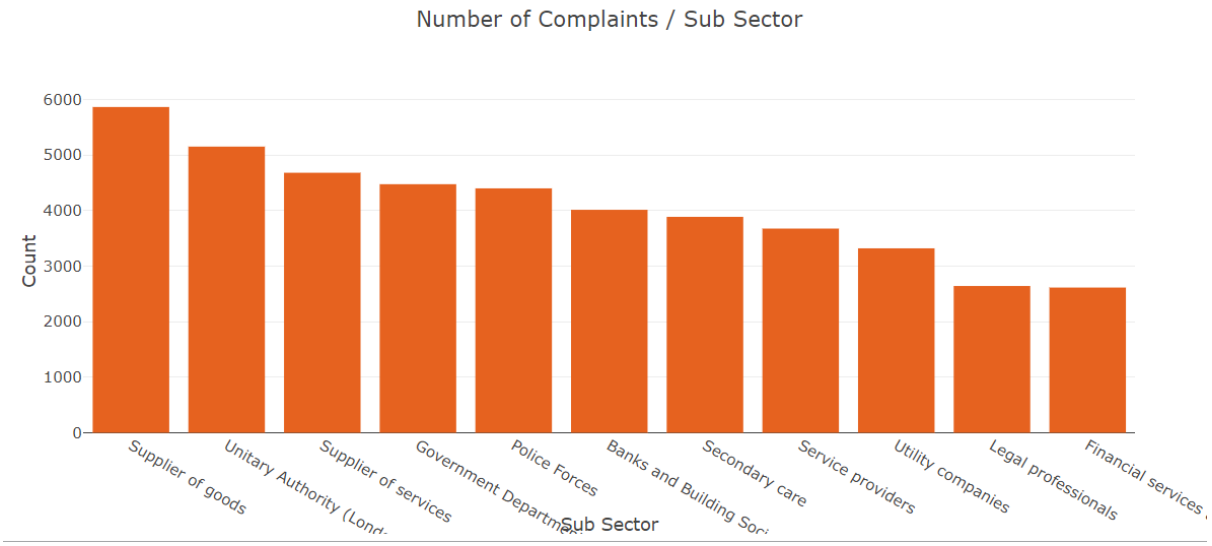


Figure 39: Complaints per Subsector Bar Chart

The analysis reveals that the "Supplier of Goods" subsector received the highest count of complaints with 5866, which indicates potential issues with product quality, customer service, or data handling practices. The "Unitary Authority" subsector follows with 5153 complaints, highlighting challenges in managing and protecting sensitive information in urban areas. The "Supplier of Services" subsector had 4682 complaints, highlighting the critical role of these entities in managing customer data securely, potentially due to the broad nature of the subsector. This indicates potential shortcomings in data protection policies or incident response procedures. These findings suggest the need for closer examination and remedial actions to address the concerns and enhance compliance.

Insights from these specific sub sectors can help improve practices, guiding users, organizations, and regulatory bodies. Knowing which subsectors have the most complaints informs users and mitigates data risks. Organizations can prioritize resources and address compliance gaps within subsectors using this data as well as enhance data protection policies and training programs. This approach reduces breaches and penalties, while building trust among customers. Regulatory bodies, and more particularly those functioning at the subsector level, are responsible for compliance and standards enforcement. More specifically, it is the responsibility of these bodies to monitor complaints and their type within the given subsectors and look for trends or systemic problems. This data is used to prioritize resource allocation and enforcement with the focus on subsectors and the problems most important for user's protection.

Detailed Pie Chart for Distribution of GDPR Articles

In implementing another pie chart focusing on the "Decision Primary reason," I aimed to provide users with more detailed insights into the specific sections of the GDPR articles being breached. This additional granularity not only enhances user understanding but also empowers organizations to identify the precise areas where compliance issues are most prevalent. For instance, with 20.8% attributed to Art 15(3)(1), users can see that the right of access under Article 15, Section 3, Subsection 1 is frequently violated, prompting a closer examination of data access procedures and protocols. By breaking down the reasons behind decisions, the pie chart serves as a valuable tool for organizations to prioritize corrective actions and bolster compliance efforts. Moreover, for users less familiar with the GDPR articles, this detailed breakdown offers educational value, providing a better understanding of the regulation's nuances and implications. This aligns with practices in data transparency and accountability, supplying stakeholders with insights derived from the data analysis. As organizations attempt to navigate the complex landscape of data protection regulations, the implementation of detailed visualizations like the Decision Primary reason pie chart becomes increasingly indispensable, providing informed decision-making and proactive compliance strategies.

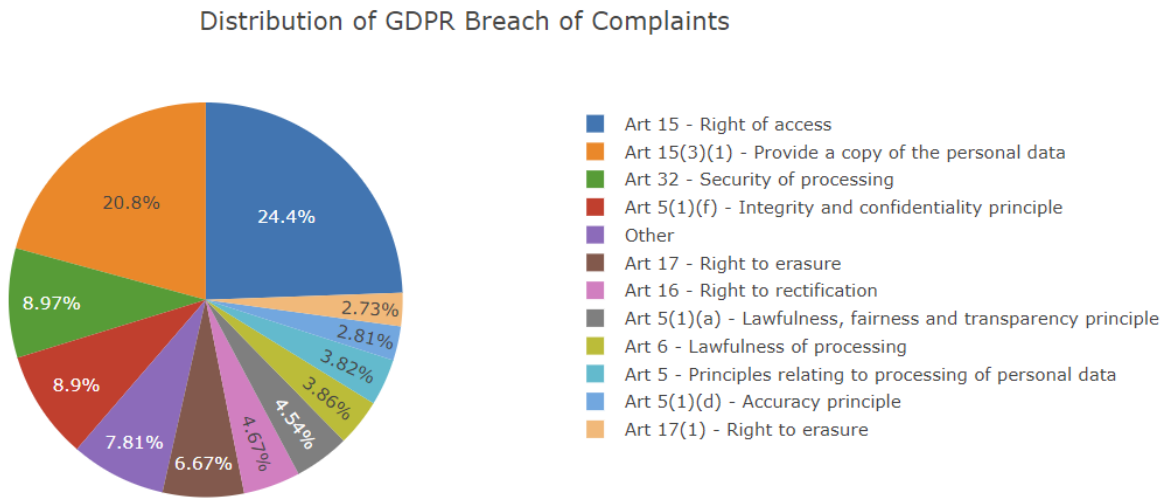


Figure 40: Detailed Pie Chart Pie Chart for Distribution of GDPR Articles

Distribution of Complaints per Day Heatmap

Incorporating a heatmap into the dashboard offers a dynamic way to visualize the distribution of data protection complaints across different sectors submitted throughout the week. This visualization provides insights into the patterns and trends of complaint submissions, allowing users to discern any fluctuations or concentrations in activity across various days of the week. To create the heatmap, I first mapped day names to numbers starting from 1 using calendar.day_name. This mapping was applied to the dataset, generating a new 'DayNum' column. Then, I grouped the dataset by sector, day name, and day number, calculating complaint counts for each combination.

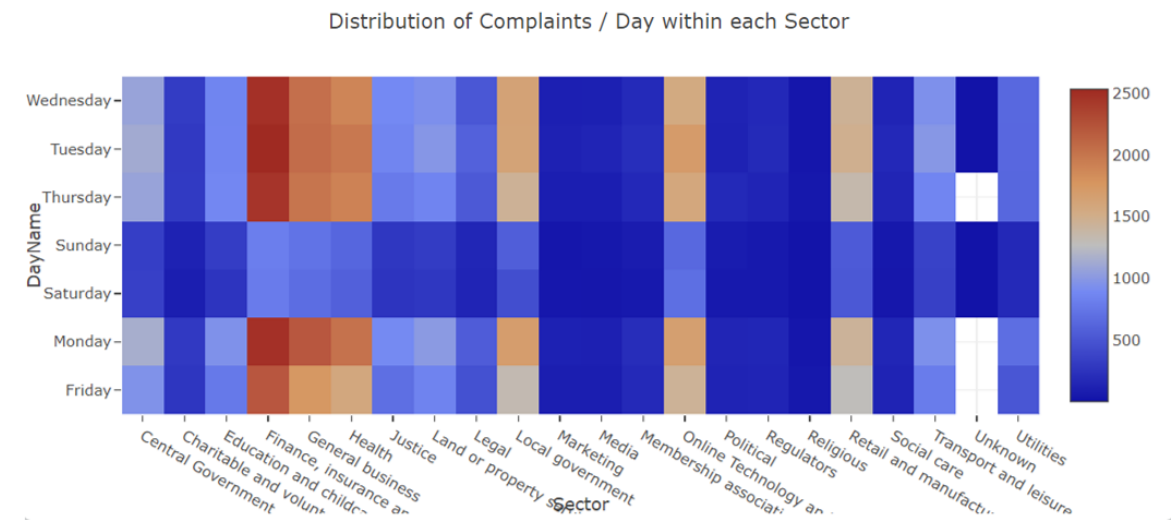


Figure 41: Distribution of Complaints per Day Heatmap

I then sorted the dataset by the numerical day since the days were listed in an unordered manner. Additionally, I eliminated the 'unknown' column, as it contained only partial data. Moreover, I opted to change the color scheme of the heatmap. The continuous Red Blue color scale initially used made it challenging to distinguish values closely positioned to one another, especially since the majority of values were closely distributed. Consequently, I selected the Portland color scale for improved clarity. The Portland color scale offers a smoother transition between colors and provides better differentiation between adjacent values, facilitating easier interpretation of the heatmap.

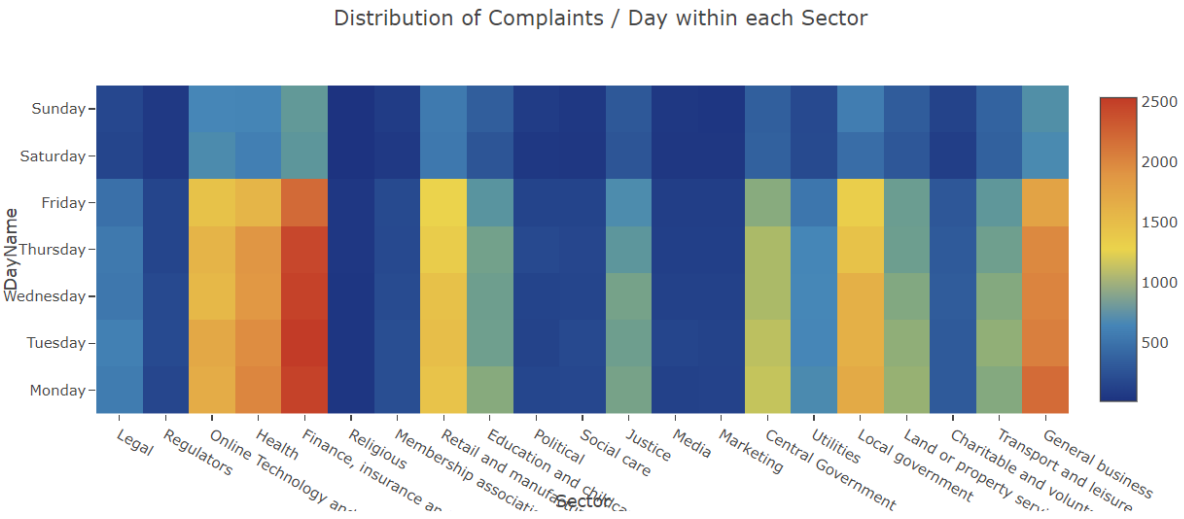


Figure 42: Distribution of Complaints per Day Heatmap Re-Coloured

The daily distribution of complaints across sectors did not yield significant insights, as complaints remained relatively constant throughout the week, except for weekends, likely due to holidays. This could have been used under better circumstances if there were anomalies present within the heatmap that could help us lead to a conclusion we may not have been aware of, however, the data visualized does not show any data worth visualizing since its generic and could be assumed without the need of a graph.

Sankey Diagram Issues

The implementation of a Sankey diagram into my dashboard was initially met with challenges due to the limitations of the available libraries. While Taipy, the primary library used for data visualization in the dashboard, did not offer a Sankey chart option, I sought out alternative solutions to incorporate this valuable visualization tool. After exploring various options, I decided to utilize Plotly, a Python library known for its extensive capabilities in creating interactive visualizations.

However, integrating the dataset with the Sankey diagram posed several difficulties. The dataset contained a large number of values, totalling over 101,000 entries. When attempting to render the Sankey diagram with this dataset, Plotly attempted to draw each node individually, resulting in significant performance issues and high resource utilization. The sheer volume of data overwhelmed the rendering process, causing the site to slow down considerably. In an effort to address this challenge, I experimented with grouping sections of the Sankey diagram to reduce the number of individual nodes rendered. However, I encountered limitations in this approach, as many existing examples of grouped Sankey diagrams relied on manually assigning weights to categories, a process that was impractical given the size of the dataset. Furthermore, both source and target nodes were mixed which meant the data presented was inaccurate.

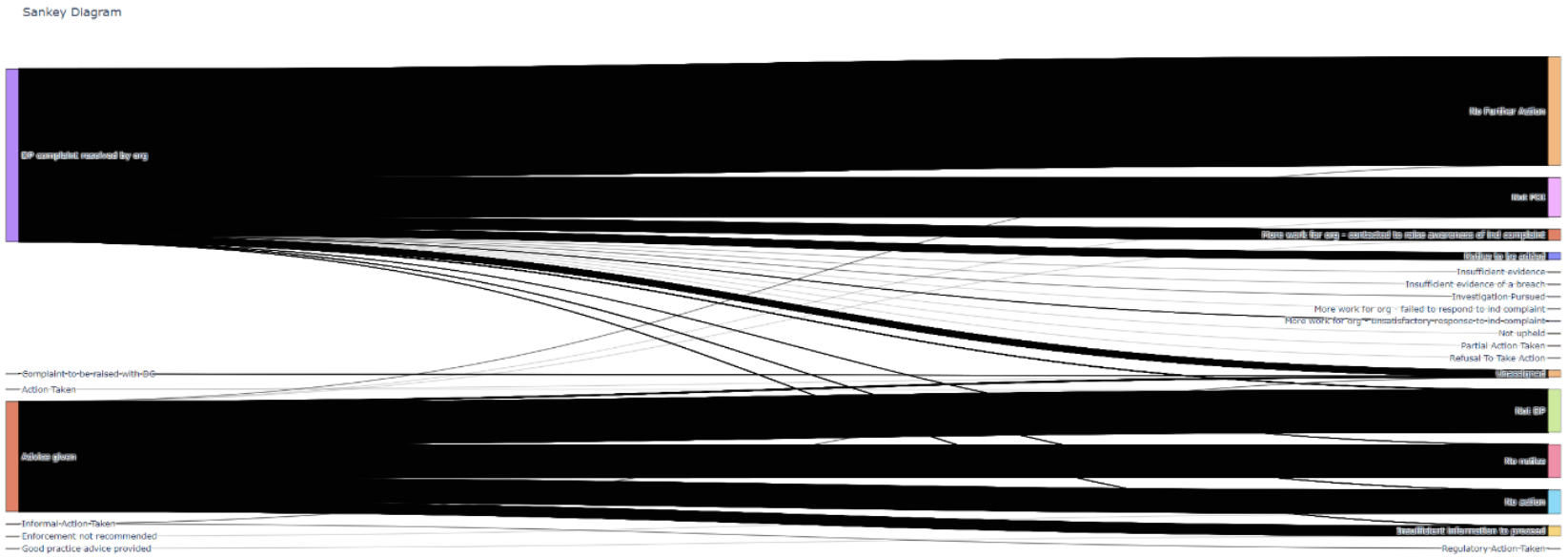


Figure 43: Unused Sankey Diagram Due to Performance Issues

Un-used Visualisations

Many visualisation prototypes were experimented with while developing the web-based dashboard for data protection complaints. However, several visualisation approaches were developed but were not included in the final product for various reasons. These prototypes varied in design and purpose, aiming to represent different aspects of the data. However, upon evaluation, it became apparent that some visualisations did not effectively convey the intended insights or were not aligned with the project's objectives. Additionally, as the project progressed, alternative visualisation techniques were discovered that better suited the dataset or provided clearer interpretations of the data. As a result, these unused prototypes were set aside in favor of more suitable and impactful visualizations for enhancing data understanding and decision-making.

Cases Submitted per Quarter

One of the visualizations I experimented with but ultimately decided not to include in the dashboard was the display of the count of data protection complaints per quarter of the year. Initially, this visualization seemed promising as it had the potential to reveal trends or significant variations in the number of complaints submitted throughout the quarters. To implement this, I calculated the count of complaints and organized them by quarter, representing the numerical value of each quarter as a bar chart across all years in the dataset. However, upon analysis, it became evident that the results were not informative. The counts for each quarter were strikingly similar, differing by only around 1000 complaints across all quarters. Consequently, the resulting bar chart displayed similar heights for each quarter, failing to provide valuable insights or discernible patterns. Had the data exhibited more substantial variations within each quarter, this visualization could have potentially offered insights into periods of heightened or reduced complaint submissions throughout the year.

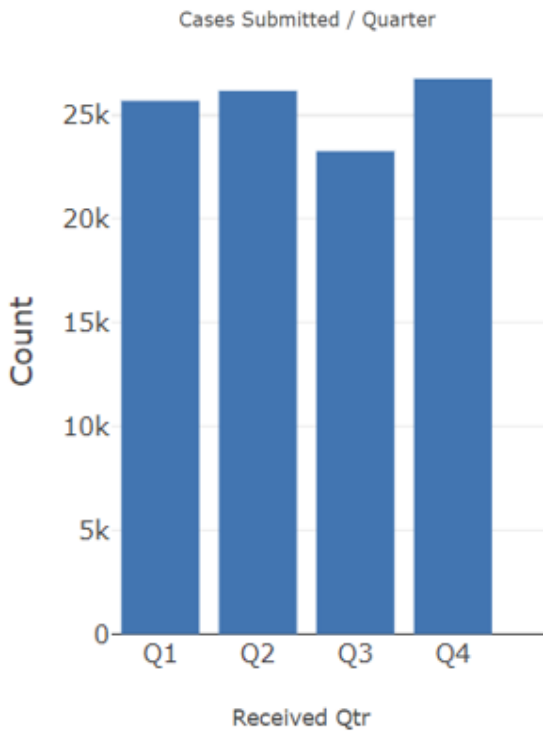


Figure 44: Cases Submitted per Quarter Chart

Average Time to Process a Complaint

Another visualisation I developed but chose not to include was the representation of the average time to process a complaint. Initially, I thought it could be helpful to see how long it takes for a complaint to be dealt with, maybe to spot if things were taking too long or being resolved too quickly. To achieve this, the time delta was computed by determining the difference between the submission date and completion date of each complaint. The resulting time deltas were then sorted in ascending order and visualised as a histogram, projecting the distribution of different time ranges and their frequency. However, upon examination, I observed that the distribution across time intervals was fairly uniform, making the visualisation less informative. Nonetheless, I recognised the significance of this data, and opted to present the overall average processing time for complaints. This simplified the metric and provided a straightforward means of understanding the efficiency of the how complaints are processed, potentially helping to identify areas of improvement.

Figure 45: Average Time to Process a Complaint Chart

Average Time to Process a Complaint per Sector

Furthermore, I decided not to represent the average time to process a complaint per sector. This visualisation would have been insightful as it would indicate how much the processing of a complaint might vary depending on its sector. Similar to the previous visualisation, to build it, I calculated the time delta between the complaint submission and resolution dates, and for each sector, I also calculated the average delta. Therefore, the data I used to plot this visualisation was shown as a bar graph, which depicted the delta ranges and how many times a complaint required this much time in this sector of complaints. Once I analyzed this visualisation, I noticed that many sectors had similar counts under the bars for the same ranges of deltas. If the counts varied significantly per delta range, the visualisation would be much more useful as it would enable identifying more efficient sectors in complaint resolution.

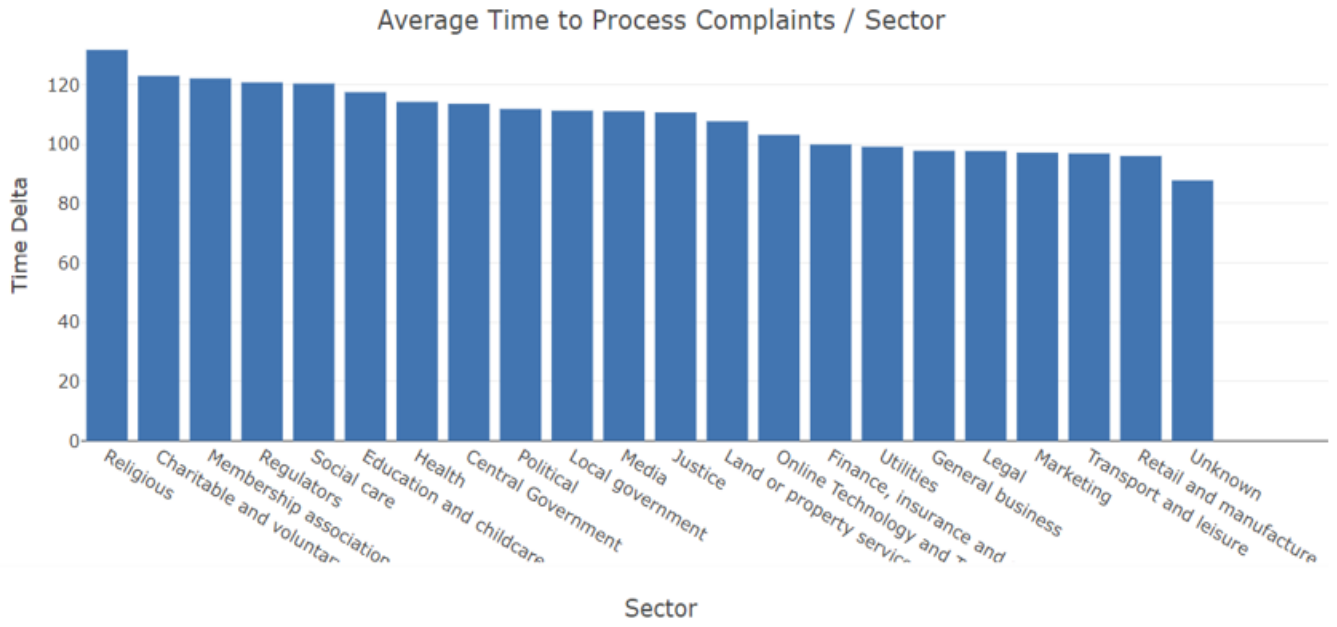


Figure 46: Average Time to Process a Complaint per Sector Chart

Full Graph for Complaints per Subsector

Additionally, another visualisation not included was the depiction of complaint counts per subsector. Initially, this was considered useful to identify subsectors more susceptible to complaints, aiding in tailored mitigation efforts just like I had done with the Sectors. The count of complaints per subsector was calculated and organised, then displayed as a bar graph, sorted in ascending order to illustrate the distribution of complaints across subsectors. However, due to the large number of subsectors, presenting all of them in a single graph was deemed inefficient. Narrowing the focus to the top 5 or 10 subsectors would offer more actionable insights regarding which areas might necessitate additional training and attention.

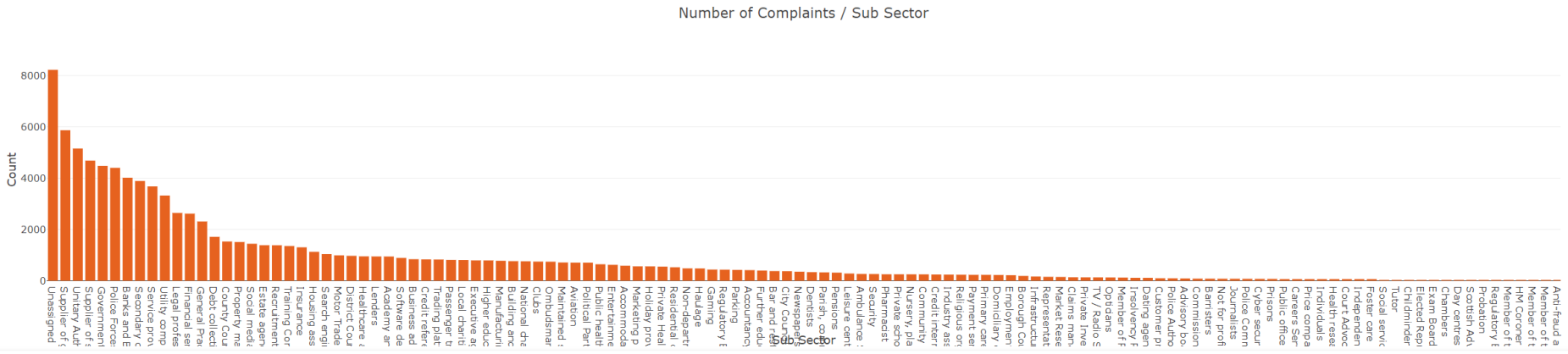


Figure 47: Full Graph for Complaints per Subsector Chart

Sector Tree Map

Another type of chart that was considered in regard to the distribution of sectors which could be implemented in the data protection complaints dashboard was a tree map. A tree map would allow for visually demonstrating different sizes of sectors and comparing them to one another easily. This was implemented in Taipy as a chart type, with the columns to count specified from sectors. As a result, the chart outputted a large map of sectors with rectangles, each of differing size and color. However, the tree map was not used, ultimately, due to the fact that the information can be provided more effectively using a bar chart when displaying sectors on their own. However, under better circumstances, the tree map could have been employed if it was possible to have subsectors under the sectors. This could offer a better understanding of how the sectors and subsectors relate to each other, although it could also increase the complexity and reduce the ease of use of the visualization. However, this could be reduced through how interaction on the chart takes place.

Figure 48: Sector Tree Map

Distribution of Articles per Sector

Another chart that I had attempted to implement was a heat map, intended to show the distribution of articles against the sectors the complaints are submitted against. By putting it in a heatmap, the original intention was to see what articles are prominent against certain sectors, aiming to see trends and anomalies. The graph was implemented using the Taipy heatmap chart, plotting the articles column against the sectors column. However, the results provided were not what was expected. This is due to the fact that there are many articles, so not all of them were plotted and due to the fluctuation of the distribution of articles submitted against organisations, there were many areas blank since they are no submitted complaints of that article against the sector. Furthermore, most of the values that were mapped were of a similar colour, not providing much useful information on how the values of the number of articles submitted against the organisation differ. In better circumstances, mapping articles that are positively noted to have been submitted against all sectors may provide better results as well as potentially limiting the number of articles mapped to the top 10 most common articles by sorting them using the Pandas python library, reducing the blank areas on the chart.

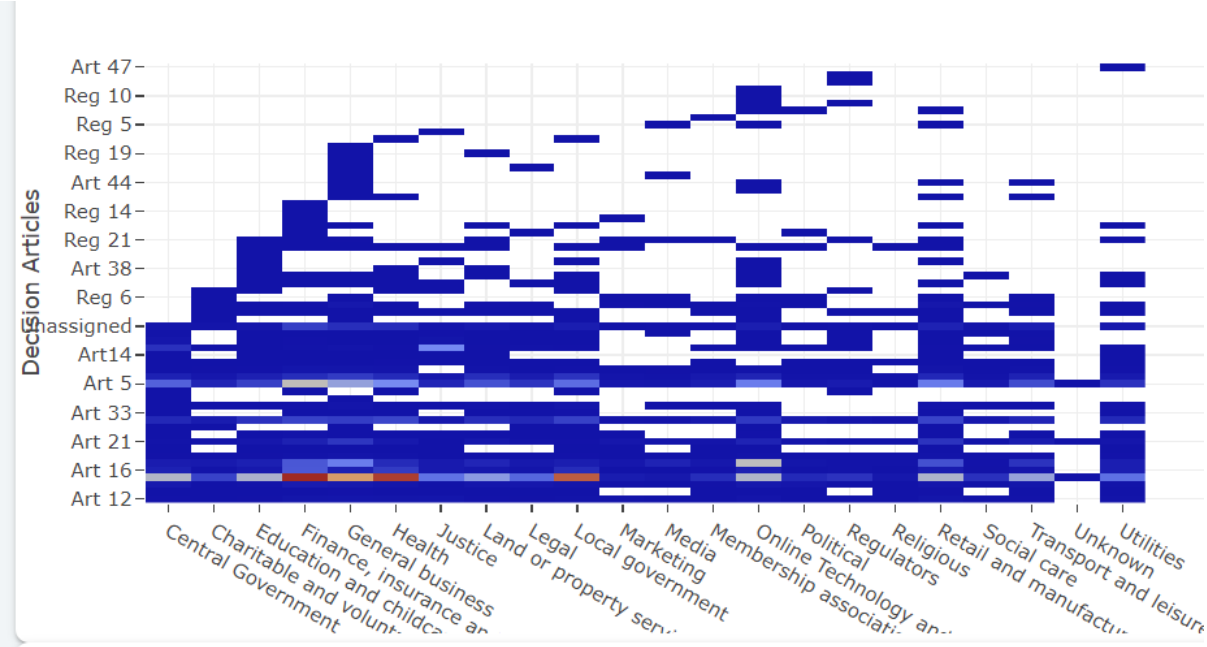


Figure 49: Distribution of Articles per Sector Heat Map

Appendix K: Testing

Unit Testing

Test ID: Displaying Total Number of Cases
Summary: Verify that the dashboard accurately displays the total number of data protection cases as 101,895.
Pre-requisite: The dashboard must be accessible and loaded with the dataset.
Test Actions: Access the dashboard, locate the statistic displaying the total number of cases and verify that the displayed total matches the expected value of 101,895.
Expected Result: The dashboard correctly displays the total number of cases as 101,895.
Actual Result: The dashboard displays the total number of cases as 101,895.
Status: Pass
Test ID: Displaying Average Time to Process a Complaint
Summary: Ensure that the dashboard correctly presents the average time to process a data protection complaint as 106 days.
Pre-requisite: The dashboard must be accessible and loaded with the dataset.
Test Actions: Access the dashboard, locate the statistic displaying the average time to process a complaint and verify that the displayed average matches the expected value of 106 days.
Expected Result: The dashboard accurately displays the average time to process a complaint as 106 days.
Actual Result: The dashboard displays the average time to process a complaint as 106 days.
Status: Pass
Test ID: Displaying Number of Re-opened Cases
Summary: Confirm that the dashboard correctly presents the number of re-opened data protection cases as 726.
Pre-requisite: The dashboard must be accessible and loaded with the dataset.
Test Actions: Access the dashboard, locate the statistic displaying the number of re-opened cases and verify that the displayed count matches the expected value of 726.
Expected Result: The dashboard accurately displays the number of re-opened cases as 726.
Actual Result: The dashboard displays the number of re-opened cases as 726.
Status: Pass
Test ID: Bar chart: Complaints per Sector
Summary: Test bar chart for number of complaints per sector
Pre-requisite: Data import functionality must be functioning correctly.
Test Action: Observe the bar chart displaying the number of complaints per sector.
Expected Result: The bar chart accurately represents the distribution of complaints across different sectors.
Actual Result: The bar chart displays the number of complaints per sector correctly in descending order as specified in the code.
Status: Pass
Test ID: Pie chart: Distribution of Complaint Decisions
Summary: Test pie chart for distribution of complaint decisions
Pre-requisite: Data import functionality must be functioning correctly.
Test Action: View the pie chart illustrating the distribution of complaint decisions.
Expected Result: The pie chart accurately represents the proportion of different decision outcomes for complaints.
Actual Result: The pie chart displays the distribution of complaint decisions correctly.
Status: Pass

Test ID: Line chart: Timeline of Cases Submitted
Summary: Test line chart for timeline of cases submitted
Pre-requisite: Data import functionality must be functioning correctly.
Test Action: Review the line chart depicting the timeline of cases submitted over time.
Expected Result: The line chart accurately illustrates the trend of case submissions over the specified time period.
Actual Result: The line chart displays the timeline of cases submitted correctly.
Status: Pass
Test ID: Stacked Bar chart: Distribution of Action Taken per sector over time
Summary: Test stacked bar chart for distribution of action taken and no action taken per sector which can be filtered per year using buttons.
Pre-requisite: Data import functionality must be functioning correctly.
Test Action: Interact with the stacked bar chart to filter data by different years.
Expected Result: The stacked bar chart allows users to filter data by different years, accurately representing the distribution of action taken and no action taken per sector.
Actual Result: The stacked bar chart filters data by different years as expected, displaying the distribution of actions correctly.
Status: Pass
Test ID: Heatmap: Distribution of Complaints per Month
Summary: Test Heatmap for distribution of complaints per month
Pre-requisite: Data import functionality must be functioning correctly.
Test Action: Analyse the heatmap displaying the distribution of complaints per month for each sector.
Expected Result: The heatmap accurately visualizes the distribution of complaints across months for different sectors.
Actual Result: The heatmap displays the distribution of complaints per month correctly.
Status: Pass
Test ID: Bar chart: Complaints per Subsector
Summary: Test bar chart for number of complaints per subsector
Pre-requisite: Data import functionality must be functioning correctly.
Test Action: Examine the bar chart illustrating the number of complaints per subsector.
Expected Result: The bar chart accurately represents the distribution of complaints across different subsectors and sorts them in descending order.
Actual Result: The bar chart displays the number of complaints per subsector correctly.
Status: Pass
Test ID: Pie Chart: Distribution of Detailed GDPR Breaches
Summary: Test pie chart for distribution of detailed GDPR breaches
Pre-requisite: Data import functionality must be functioning correctly.
Test Action: Review the pie chart showing the distribution of detailed GDPR breaches for complaints.
Expected Result: The pie chart accurately represents the distribution of detailed GDPR breaches across complaints and aggregates smaller values into one part name “Other” to reduce the number of slices on the pie chart.
Actual Result: The pie chart displays the distribution of detailed GDPR breaches correctly.
Status: Pass
Test ID: Sankey: Flow of Decision to Decision Detail 2
Summary: Test Sankey diagram for flow of overall decisions on complaints
Pre-requisite: Data import functionality must be functioning correctly.
Test Action: Analyse the Sankey diagram illustrating the flow of overall decisions on complaints to decision detail 2.

Expected Result: The Sankey diagram accurately visualizes the flow of decisions on complaints, providing insights into decision pathways.
Actual Result: The Sankey diagram displays the flow of decisions on complaints correctly.
Status: Pass
Test ID: Checking Table Columns on Risks Page
Summary: Verify that the table on the risks page contains the specified columns.
Pre-requisite: Access to the risks page on the dashboard and dataset integration.
Test Action: Inspect the table displayed on the page and check if the table contains the following columns: "Case Reference", "Submitted About Account", "Date Received", "Completed Date", "Sector", "Sub Sector", "Decision Primary Reason", "Decision", "Decision Detail2", "Scaled Risk Factor".
Expected Result: The table on the risks page contains all the specified columns.
Actual Result: The table on the risks page contains all the specified columns.
Status: Pass
Test ID: Verifying Calculated Risks in the Table
Summary: Ensure that the calculated risks are correctly displayed in the table.
Pre-requisite: Access to the risks page on the dashboard and dataset integration.
Test Actions: Identify a data protection case in the table with a known risk factor. Cross-reference the calculated risk factor with the value displayed in the terminal for the corresponding case.
Expected Result: The calculated risk factor matches the value displayed in the table for the corresponding case.
Actual Result: The calculated risk factor matches the value displayed in the table for the corresponding case.
Status: Pass
Test ID: Checking Table Colour Coding for Risk Levels
Summary: Verify that the table on the risks page color-codes the risk values correctly.
Pre-requisite: Access to the risks page on the dashboard and dataset integration.
Test Actions: Inspect the table and identify cases with different risk levels. Verify that risk values below 25 are coloured green (Low risk), between 25-50 are coloured yellow (Moderate), between 50-75 are coloured orange (High), and above 75 are coloured red (Critical).
Expected Result: The table correctly color-codes the risk values based on their respective risk levels.
Actual Result: The table correctly color-codes the risk values based on their respective risk levels.
Status: Pass
Test ID: Testing Table Filtering by Organisation
Summary: Ensure that filtering the table by organisation using the drop-down button presents the correct data.
Pre-requisite: Access to the risks page on the dashboard and dataset integration.
Test Actions: Use the drop-down button to select a specific organisation. Verify that the table displays only the data protection cases associated with the selected organisation.
Expected Result: The table correctly filters and displays only the data of the selected organisation.
Actual Result: The table correctly filters and displays only the data of the selected organisation.
Status: Pass

Test ID: Compliance Advisor: Checking Sectors Dropdown
Summary: Verify that all sectors are displayed in the dropdown menu.
Pre-requisite: Access to the GDPR compliance advisor page on the dashboard and dataset integration
Test Actions: Locate the dropdown menu for selecting sectors and check if all sectors are listed in the dropdown menu.
Expected Result: The dropdown menu displays all available sectors.
Actual Result: The dropdown menu displays all available sectors.
Status: Pass
Test ID: Compliance Advisor: Checking Subsectors Display for Selected Sectors
Summary: Ensure that selecting a sector displays all relevant subsectors and updates correctly when the sector is changed.
Pre-requisite: Access to the GDPR compliance advisor page on the dashboard and dataset integration
Test Actions: Select a sector from the dropdown menu then check if all relevant subsectors for the selected sector are displayed. Change the selected sector and verify that the displayed subsectors update accordingly.
Expected Result: Selecting a sector displays all relevant subsectors, and changing the sector updates the displayed subsectors accordingly.
Actual Result: Selecting a sector displays all relevant subsectors, and changing the sector updates the displayed subsectors accordingly.
Status: Pass
Test ID: Compliance Advisor: Verifying Top 3 Articles Display
Summary: Ensure that the top 3 GDPR articles displayed are correct.
Pre-requisite: Access to the GDPR compliance advisor page on the dashboard and dataset integration
Test Actions: Check the output of the top 3 GDPR articles displayed on the page and cross-reference the displayed articles with the expected values printed in the terminal.
Expected Result: The top 3 GDPR articles displayed match the expected values.
Actual Result: The top 3 GDPR articles displayed match the expected values.
Status: Pass

Functional Testing

Test ID: Incorporation of all pages
Summary: Ensure that all pages appear and function correctly when selected from the navbar.
Pre-requisite: Access to the dashboard with the navbar visible.
Test Action: Check if the navbar contains links to all pages: Getting Started, Dashboard, Risks, and GDPR Compliance Advisor. Click on each link in the navbar and verify that the corresponding page loads correctly. Repeat for each page in the navbar.
Expected Result: Clicking on each link in the navbar navigates to the corresponding page, and each page loads correctly without errors.
Actual Result: Clicking on each link in the navbar navigates to the corresponding page, and each page loads correctly without errors.
Status: Pass
Test ID: Dashboard charts
Summary: Test Generation of Different Types of Charts
Pre-requisite: Data import functionality must be functioning correctly.
Test Action: Generate various types of charts (e.g., bar charts, pie charts, line graphs) based on the imported data.
Expected Result: The charts accurately represent the underlying data and display the desired visualization type.
Actual Result: Different types of charts are generated based on the imported data, accurately representing the underlying dataset.
Status: Pass
Test ID: Chart Interactivity
Summary: Test Interactive Features of the Dashboard
Pre-requisite: Data import functionality and chart generation must be functioning correctly.
Test Action: Interact with the dashboard elements, including buttons, filtering, sorting, and drill-down capabilities.
Expected Result: Users can interact with the dashboard elements as intended, applying filters, sorting data, and drilling down into details as well as working buttons.
Actual Result: Interactive features of the dashboard work as expected, allowing users to manipulate data effectively.
Status: Pass
Test ID: Risks Page
Summary: Verify that risk values are assigned to all cases and the filtering button by organisation functions correctly.
Pre-requisite: Access to the risks page of the dashboard and imported dataset with populated and calculated risk values as well as organisation information.
Test Actions: Navigate to the risks page of the dashboard and inspect the table displaying cases and associated risk values. Then verify that each case entry in the table has a corresponding risk value assigned. Click on the drop-down button and select an organisation from the drop-down list. Verify that the table updates to display only cases associated with the selected organisation. Repeat for multiple organisations to ensure consistent functionality.
Expected Result: All cases in the table should have an associated risk value displayed. When selecting an organisation from the drop-down list, the table should filter and display only cases related to that organisation.
Actual Result: All cases are displayed with corresponding risk values. Upon selecting an organisation, the table filters correctly to display only cases associated with that organisation.

Status: Pass
Test ID: Compliance Advisor
Summary: Verify that the GDPR Compliance advisor page functions correctly by displaying all sectors, updating subsectors based on sector selection, and presenting the correct top 3 articles with recommendations.
Pre-requisite: Access to the GDPR Compliance advisor page of the dashboard and a dataset containing GDPR breach information and recommendations for each sector and subsector.
Test Actions: Navigate to the GDPR Compliance advisor page of the dashboard and interact with the Sector column. Verify that all sectors are listed in the drop-down menu. Then select a sector from the menu and check the Subsector column. Verify that the subsectors listed in the drop-down menu update based on the selected sector. Then select a subsector from the drop-down menu. Inspect the displayed top 3 articles with recommendations and cross-reference the displayed articles and recommendations with the expected values. Repeat this for multiple sectors and subsectors to ensure consistent functionality.
Expected Result: All sectors are listed in the column and upon selecting a sector column updates with relevant subsectors. The top 3 articles with recommendations are displayed accurately based on the selected sector and subsector.
Actual Result: All sectors are listed in the column and the subsectors update correctly based on the selected sector. The displayed top 3 articles and recommendations align with the expected values for each sector and subsector.
Status: Pass
Test ID: Accessibility
Summary: Test Accessibility Features of the Dashboard
Pre-requisite: Accessibility features must be implemented in the dashboard.
Test Action: Assess the dashboard's compliance with accessibility standards (e.g., WCAG) using assistive technologies.
Expected Result: The dashboard is accessible to users with disabilities, allowing them to navigate and interact effectively using assistive technologies.
Actual Result: The dashboard meets accessibility standards, providing an accessible user experience for all users.
Status: Pass
Test ID: Hosting Cloud Functionality
Summary: Ensure the site functions correctly when hosted on Taipy Cloud and accessed via a URL.
Pre-requisite: Project uploaded to Taipy Cloud to host and access to the hosted site URL.
Test Action: Access the hosted site URL and then verify that the pages load without errors. Click on each link in the navbar to navigate to different pages (Getting Started, Dashboard, Risks, GDPR Compliance Advisor) and verify that each page loads correctly without errors. Interact with various elements on each page to ensure functionality (e.g., filtering on the dashboard, sorting on the risks page) and repeat multiple times to ensure consistent behaviour.
Expected Result: The site loads correctly, and all page’s function as expected without errors. Navigation between pages is smooth, and all interactive elements work as intended.
Actual Result: The site loads correctly, and all page’s function as expected without errors. Navigation between pages is smooth, and all interactive elements work as intended.
Status: Pass

Non-functional testing

Test ID: Performance
Summary: Test Dashboard Performance
Pre-requisite: Dataset must be imported into the dashboard.
Test Action: Measure the loading time of the dashboard and its responsiveness during interactions.
Expected Result: The dashboard loads within a reasonable time frame and responds promptly to user interactions, even with large datasets.
Actual Result: The dashboard loads quickly and maintains responsiveness during interactions. However, when testing the performance on the Risks page, sorting by organisation takes a considerable amount of time which can be classed as “Too long”.
Status: Failed
Test ID: Browser Compatibility
Summary: Test Dashboard Compatibility Across Web Browsers
Pre-requisite: None
Test Action: Access the dashboard using different web browsers (e.g., Chrome, Firefox, Safari, Edge).
Expected Result: The dashboard functions correctly across all supported web browsers, maintaining consistency in appearance and functionality.
Actual Result: The dashboard is compatible with various web browsers, displaying consistent performance and functionality.
Status: Pass