Medicines & Healthcare products
Regulatory Agency

10 South Colonnade
Canary Wharf
London
E14 4PU
United Kingdom
gov.uk/mhra

End Point Assessment Portfolio

Level 4 Data Analyst pathway

Shaheer Hussain

Apprentice Details

Name	Shaheer Hussain
ULN	8935304564
Training Provider	Ada, the National College for Digital Skills
Employer	Medicines and Healthcare Regulatory Agency

Table of contents

Table of Figures

ID	Assessment Criteria	Included (Page / Line)
P1	Explains how current, relevant legislation impacts on the safe use of data and how their role contributes to a productive, safe, and secure working environment (K1 B1)	[Evidence statement 1] [Evidence statement 4]
P2	Explains the relevant data policies and procedures for the organisation, and identifies the data standards to be reached (K2)	[Evidence statement 1] [Evidence statement 2]
P3	Describes the fundamentals of data structures and database system design and explains how they are implemented and maintained. (K6)	[Evidence statement 2] [Evidence statement 3]
P4	Explains approaches to combining data from different sources to improve accuracy and / or efficiency and / or maximise benefits to the organisation and / or customer. (K10)	[Evidence statement 1] [Evidence statement 2] [Evidence statement 3]
P5	Describes impact on user experience and domain context on data analysis (S5)	[Evidence statement 2] [Evidence statement 4]
P6	Explains the differences between Structured and Unstructured data (K5)	[Evidence statement 1] [Evidence statement 2] [Evidence statement 3]
P7	Explains the ethical aspects associated with the collation and use of data and justifies why this is important. (K15)	[Evidence statement 2] [Evidence statement 5]
P8	Describes the relevant tools or techniques used for working with the data systems architecture in their organisation. (S9)	[Evidence statement 1] [Evidence statement 2] [Evidence statement 3] [Evidence statement 4] [Evidence statement 5]

ID	Assessment Criteria	Included (Page / Line)
P9	Explains and applies the principles of statistics for analysing datasets, (K13 S10)	[Evidence statement 4]
P10	Identifies and explains challenges in their work and how they overcame them, providing an outline of lessons learned (B6)	[Evidence statement 1] [Evidence statement 2]
P11	Explains how they have applied analytical techniques for data mining and time series forecasting and other modelling techniques (S13)	[Evidence statement 2] [Evidence statement 3]
P12	Identifies areas of work where they adapted to changing contexts within the scope of a project, direction of the organisation or Data Analyst role (B7).	[Evidence statement 3] [Evidence statement 5]
P13	Explains the principles of descriptive, predictive and prescriptive analytics and demonstrates how they have been applied within their own data analysis practice. (K14 S11)	[Evidence statement 3] [Evidence statement 4] [Evidence statement 5]
P14	Demonstrates data analysis activities involving the collation and interpretation of qualitative and quantitative data and displays results using visual representations. (S14)	[Evidence statement 1] [Evidence statement 5]
P15	Explains the principles of user experience and domain context for data analytics. (K7)	[Evidence statement 2] [Evidence statement 4]
P16	Describes how they have appropriately adapted their activities to meet minor, unexpected changes at work. (B2)	[Evidence statement 1] [Evidence statement 4]
P17	Describes how they have ensured the true root cause of any problem is found and a solution is identified which prevents recurrence. (B5)	[Evidence statement 3] [Evidence statement 5]
D1	Critically evaluates the risks and benefits of predictive analytics (K14 S11) **[Distinction]**	[Evidence statement 3] [Evidence statement 5]
D2	Compares and contrasts visual data representation approaches and how they aid understanding by stakeholders (S14) **[Distinction]**	[Evidence statement 1] [Evidence statement 5]
D3	Evaluates the benefits and risks inherent in combining data from different sources (K10). **[Distinction]**	[Evidence statement 1] [Evidence statement 2] [Evidence statement 5]

EVIDENCE STATEMENT [1]

NAME	Shaheer Hussain
TITLE	EVIDENCE STATEMENT **[1]** - **[Training app dashboard]**
PATHWAY	DATA ANALYST

Scenario

Introduction & Business problem

The MHRA does not have an efficient method of booking training sessions for new employees and the current method requires lots of admin work, a new booking app was built in-house, but the problem is monitoring and tracking information about these training sessions to gain more understanding and insight. My role as a Data analyst is to create a Powerbi report to provide insights into the data generated from this booking application.

Aim of the Project

The aim of this project is to provide the client with useful insights about the sessions and courses they provide so that they can make more informed and data driven decisions. This can be done by creating very detailed and specific visualisations that are easy to understand.

Stakeholders of Project

There are many stakeholders within this project, these include the End users who will use this report to gain insight into the training sessions they host, the Previous and current developers who work on the back-end of the application that our report sources data from, the developers of the report which includes me and my colleague, finally other people who have an interest in the project like my manager.

Factors to take into consideration:

There were many factors to take into consideration like the following:

- Project management approach - During this project I chose an agile methodology as the end product will be higher quality and customer satisfaction will be higher due to feedback and transparency with the client.
- Comprehensive planning - I needed to create a detailed plan to understand how we can tackle future hurdles and design the report to avoid planning on the spot.
- Timescale - This was important as it was preferable to the customer if the report would be ready before the Booking app's next deployment and knowing the timescale would help us plan our sprints.
- Budget
- Roles and responsibilities I had to be clear on our roles to avoid confusion which can lead to duplicative work / effort and that becomes a waste of time and a loss of efficiency.
- Value - We had to consider whether features or implementations we planned would add value to the client because if it does not then it can be a waste of time.

- Communication - I considered communication because it would allow us to update stakeholders on our ends of the project but also to discuss issues that occur within the project.
- Miscellaneous - There were many miscellaneous considerations like live connections from Power BI to MS Teams which were important as they were tied to important customer requirements.

Objectives

My Objectives are to complete the following:

- Learn how to extract data from different sources.
- Transform and clean data in Power Query.
- Create relationships between tables / data.
- Use DAX to create useful measures.
- Create visualisations to present data in a useful manner.

Tasks

My role as a Data analyst means I am responsible for the sourcing of appropriate data, cleaning and transforming this data to be only what is needed, modelling relationships to relate data from different sources and visualising this data to give insights.

In this project we used Power BI, Microsoft Teams and Microsoft SharePoint. The general reason for this is that these tools are commonly used within the MHRA but with SharePoint, data can be easily stored and accessed via Power BI so it simplifies the task of sourcing data.

Activities

ES1 - Explains how current, relevant legislation impacts on the safe use of data and how their role contributes to a productive, safe, and secure working environment (K1 B1)

The Medicines and Healthcare products Regulatory Agency (MHRA) works diligently under regulations to safeguard the utilisation of data. These regulations are influenced by, up to date laws such as the General Data Protection Regulation (GDPR) the Data Protection Act 2018 and specific guidelines tailored to sectors. As a data analyst at MHRA grasping the implications of these laws on data usage is vital, for upholding a safe and protected workspace.

The Health Research Authority (HRA) and the Confidentiality Advisory Group (CAG) play roles in overseeing research involving data.The HRA ensures standards are upheld in research while the CAG offers guidance on utilising confidential patient information ethically. It is important for data analysts working with research data to be familiar with these guidelines.

Guidelines provided by the Information Commissioner's Office (ICO) offer insights into data protection practices. Adhering to and implementing these guidelines regularly does not ensure compliance. Also strengthens data security.

As a data analyst my responsibility lies in creating an secure working environment through data governance and compliance measures;

- Policy Implementation : Develop and enforce data governance policies that comply with legal standards, which involve establishing protocols for handling data, managing consent and controlling access to information.
- Audit and Monitoring : Regularly conduct audits to verify adherence to data protection regulations and pinpoint potential risks.
- Data Security : Implementing technical safeguards like encryption, anonymization and secure storage solutions is crucial for safeguarding sensitive information.
- Access Controls : Restrict access, to sensitive data solely to authorised personnel by using access controls based on roles and monitoring access logs diligently.

Maintaining Data Quality and Integrity :

Ensuring Accuracy and Reliability : It is crucial to verify that the data utilised in analyses is precise, current and trustworthy. This necessitates validation and cleansing procedures.
Fostering Transparency and Accountability : Upholding transparency in data processing activities is essential along with ensuring accountability by documenting data workflows and decision making processes.

Promoting Training and Awareness :

Employee Education : Regular training sessions should be provided to staff members on data protection regulations methods for handling data and the significance of maintaining data security.
Raising Awareness Through Campaigns : Organising awareness campaigns can help prioritise data protection within the organisation.

Upholding Ethical Data Practices :

Respecting Informed Consent : Data collection and usage should always respect individuals informed consent honouring their autonomy and privacy rights.
Supporting Ethical Reviews : Participating in or backing ethical review procedures, for research projects involving data helps uphold standards.

ES1 - organisational data and information security standards, policies and procedures relevant to data management activities (K2)

There are many policies and procedures that would be relevant to both this project and my involvement including policies like the Data protection policy (Mentions the processing of data in regards to many topics like consent and mentions data protection principles). I have successfully met data standards by uploading the report within the Power BI service Data & analytics workspace where the majority of other relevant reports are stored and even with the format of the report, I have used the standard MHRA layout to comply with the data standards. The data that has been sourced to the report are all from trusted web connection or SharePoint connections (SharePoint is a commonly used tool in the MHRA).

ES1 - Explains the differences between Structured and Unstructured data (K5)

Structured Data
Characteristics:

Organised and Easily Searchable: Structured data is highly organised and formatted in a way that is easily searchable using basic algorithms. It often resides in a fixed field within a record or file.
Predefined Data Models: It conforms to a predefined data model and often resides in relational databases (e.g., SQL databases) or spreadsheets.

Examples: Structured data includes numbers, dates, and strings, which are stored in databases or Excel files.
Example in Power BI: Relational Data of Training Courses

Source: Relational databases such as SQL Server, MySQL, or Excel files.
Data Schema: The data for training courses might include tables such as Courses, Instructors, Students, and Enrollments, with predefined relationships between them.
Usage in Power BI:
Importing Data: Connect Power BI to the relational database or Excel file to import the structured data.
Data Modeling: Define relationships between tables within Power BI to create a data model that reflects the relational schema.

Visualisations: Use Power BI's visualisation tools to create charts, graphs, and reports that summarise the structured data. For example, you can create a report showing the number of students enrolled in each course, average course ratings, or instructor performance.

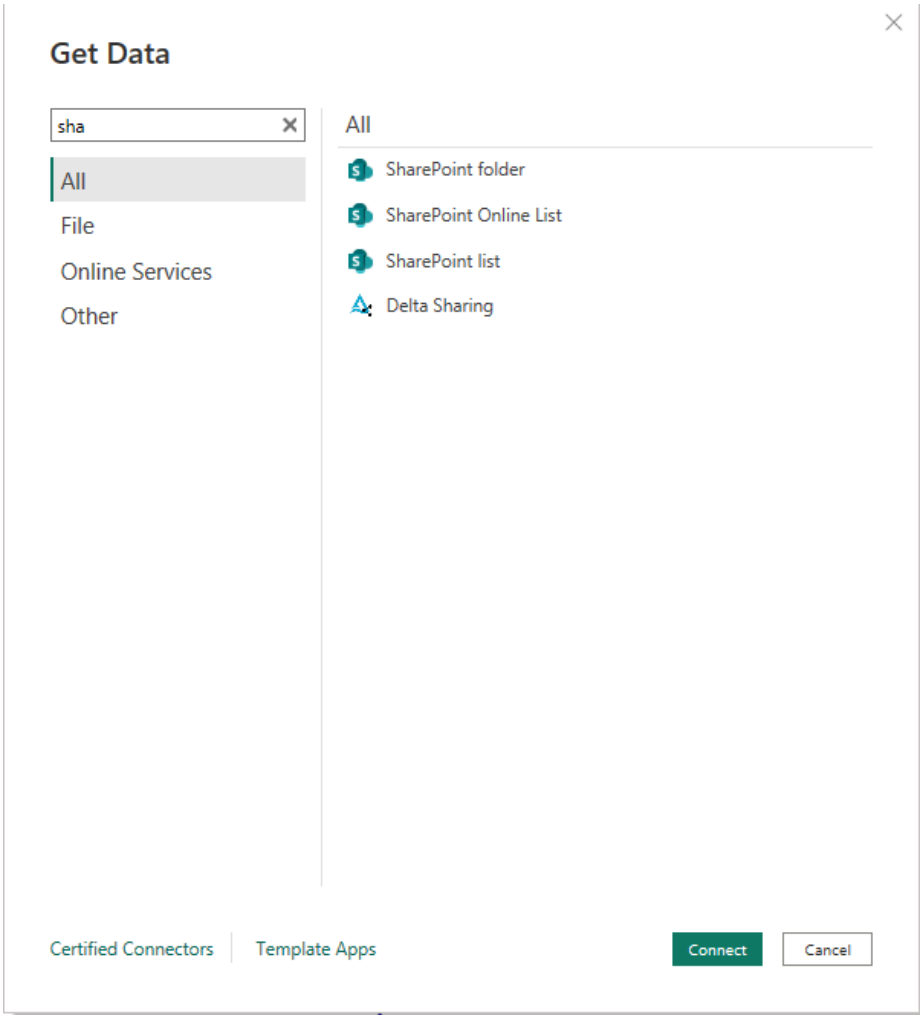
Unstructured Data
Characteristics:

Lack of Predefined Structure: Unstructured data does not follow a specific format or structure, making it more complex to store and analyse. It includes text, images, audio, and video files.
Diverse and Varied: The data can come in a variety of formats and does not fit neatly into tables or rows.
Examples: Unstructured data includes emails, social media posts, documents, multimedia files, and text from feedback forms.
Example in Power BI: Text from Google Feedback Forms

Source: Text data collected from feedback forms, which may include open-ended responses.
Data Schema: There is no predefined schema : each response can vary in length and content.
Usage in Power BI:
Data Import: Use Power BI to connect to data sources that store unstructured data, such as Google Forms responses stored in Google Sheets or a text file.
Data Processing:
Text Analytics: Utilise Power BI's integration with Azure Cognitive Services or other text analytics tools to process and analyse the unstructured text data. This can include sentiment analysis, keyword extraction, and language detection.
Data Transformation: Use Power Query to clean and transform the text data, extracting relevant information for analysis.

Visualisations: Create visualisations to summarise the insights gained from the text data. For example, you can generate word clouds to visualise common feedback themes, sentiment analysis charts to show the overall sentiment of responses, or bar charts to display the frequency of certain keywords or topics.

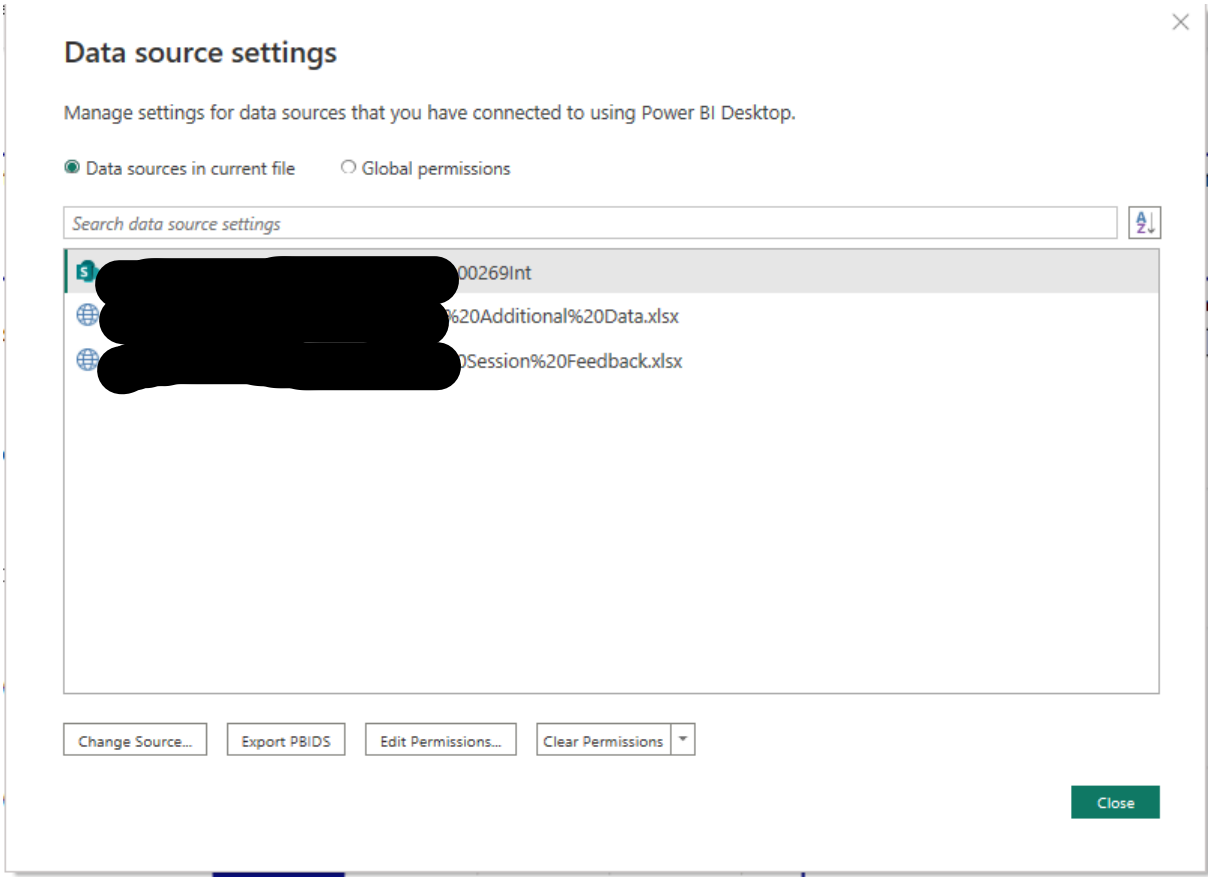
ES1 - approaches to combining data from different sources. (K10)



ES1 Figure 1: Using Get data function in Power BI to source data.

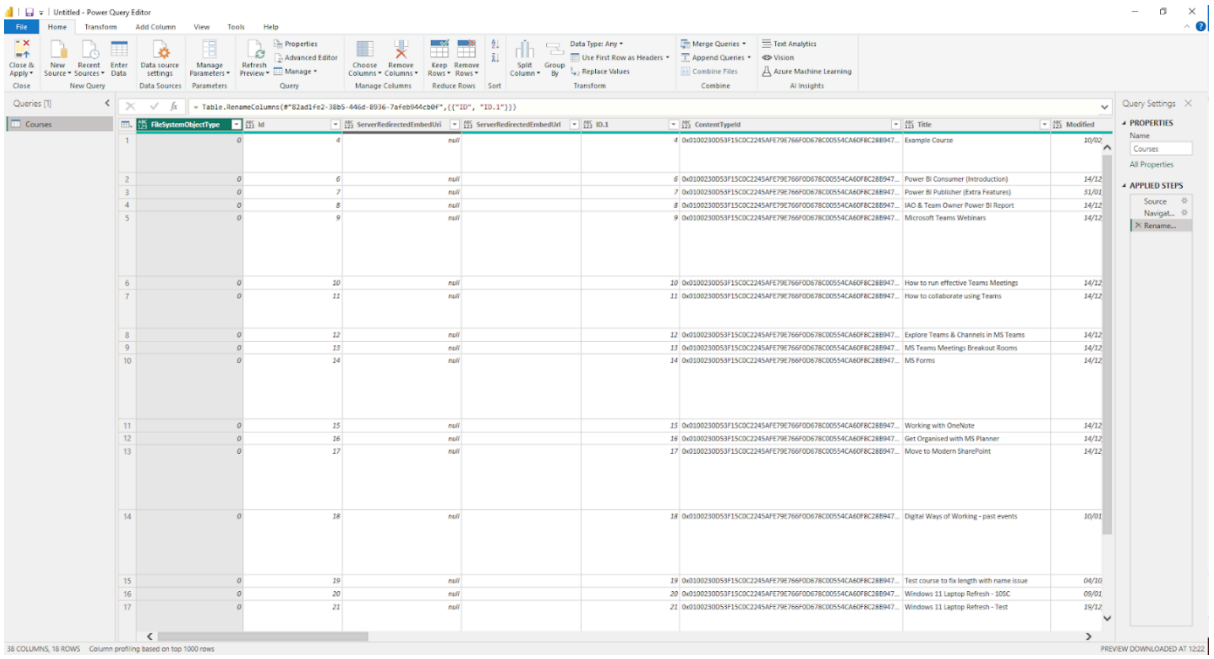
To meet the aim of providing insight and the objective of extracting data I used the “Get data” function in Power BI (As shown in Figure 1) to search up the appropriate connections which where SharePoint Lists and Web, this allowed me to extract data from different sources which would aid me in combining this data. Combining data from different sources will improve efficiency as you can provide more detailed visuals so the user does not have to look through multiple pages or visual to answer their query instead, they can look at one visual, I combined the feedback form data and the Courses data, I also combined the Courses and sessions data, I did this to use the data from the different sources together in visuals that will provide a more useful insight, for example I combined the average ratings of the presenters and sessions with the courses data to

show these ratings amongst the different courses as shown in Figure 11.There are also other approaches to combining data like using SQL to combine two table which will then combine the data inside them.

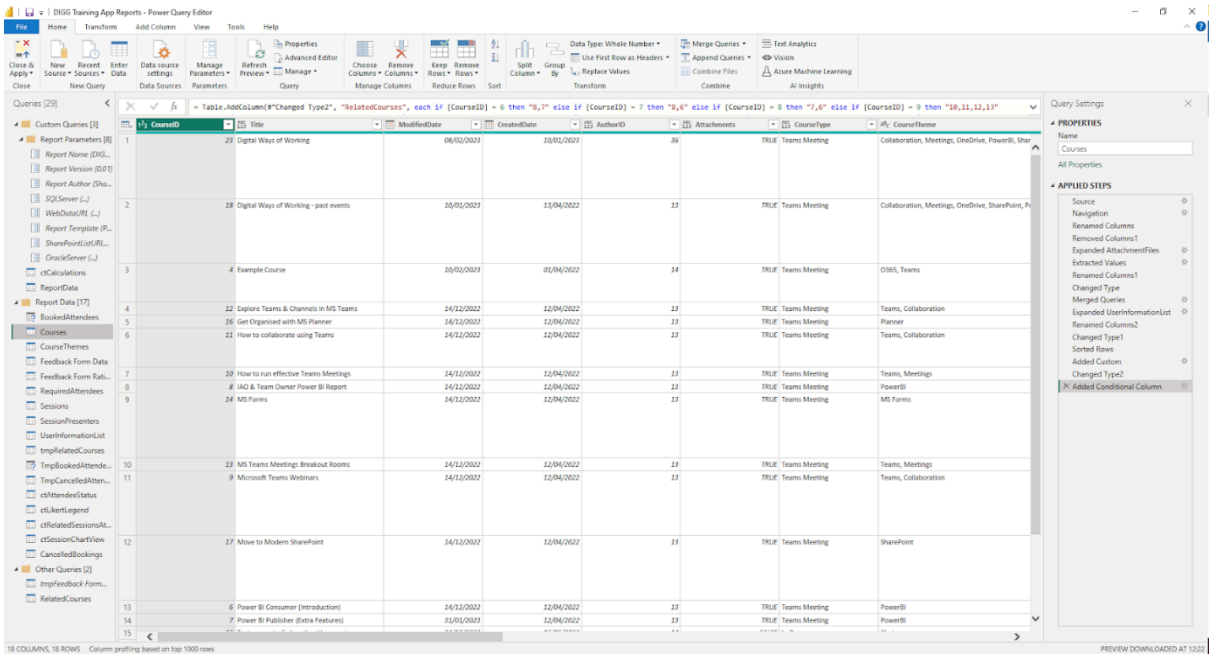


ES1 Figure 2: List of Data sources for the report I built.

In Figure 2 we can see a list of the data connections that my report had with two being a web connection and one being a direct SharePoint connection. I made these connections using the “Get data” function in Power BI extract data from different sources which will enable me to visualise the data, this means that establishing data connections would mean that I meet the objective of extracting data from different sources but it would also help me meet my other objectives by acting as a stepping stone, also it will help me meet the overall aim of the project of providing useful insights.



ES1 Figure 3: Courses data from booking app before transformation.



ES1 Figure 4: Courses data from booking app after transformation.

Figure 3 and Figure 4 portrays the before and after transformation of the courses data, Figure 4 shows us combining data from the UserInformationList and Attachmentfiles tables, it also shows us the transformation of data. I transformed the data to get rid of any unnecessary data that will consume memory and promote inefficiency. I also renamed columns to make them more readable and easier to work with. This will also improve efficiency as the user data will be in the Courses table so there is no need to relate the UserInformationList and Attachmentfiles tables, combining data will also improve accuracy as you can

streamline the information in a table to answer questions more accurately and this will maximise the usage or benefit of this report to the client and agency.

ES1 - Evaluates the benefits and risks inherent in combining data from different sources (K10). **[Distinction]**

Benefits of Combining Data from Different Sources

Enhanced insight and Decision Making :

Comprehensive Perspective : Bringing together data offers a view of an issue aiding in thorough analysis and informed decision making.

Relationships and Causes : Enables the discovery of connections and potential cause and effect links that may not be obvious when looking at individual data sets.

Improved Data Integrity :

Data Verification : Checking data from sources helps spot discrepancies and enhance data accuracy.

Eliminating Duplication : Merging data can get rid of repetition ensuring that the information remains current and precise.

Operational Streamlining :

Efficient Workflows : Unified data systems can simplify business operations cutting down on time and resources required for collecting and analysing data.

Cost Efficiency : Reduces the necessity for data storage systems and the associated upkeep expenses.

Encouraging Innovation and Progress :

Fresh Possibilities : Access to data sets can inspire creativity encouraging the creation of products, services or business strategies.

Advanced Analysis : Integrating sources of information enables sophisticated analytics and machine learning applications enhancing predictive capabilities.

Risks of Combining Data from Different Sources

Challenges with Data Consistency and Quality :

Variety in Data Formats : Different sources using formats, structures and standards can make it difficult to integrate the data seamlessly.

Data Accuracy Concerns : Inaccurate or inconsistent data from a source can impact the integrity of the data.

Privacy and Security Considerations :

Risk of Exposing Sensitive Information : Consolidating data from sources may expose information raising privacy issues.

Threats of Data Breaches : The abundance and diversity of data can attract cyberattacks posing risks to security.

Legal and Regulatory Matters :

Adherence to Regulations : Varied data protection laws across jurisdictions may lead to non compliance when merging data sets.

Protection of Intellectual Property Rights : Unauthorised integration of data could lead to legal conflicts.

Technical Hurdles :

Scalability Challenges : Managing amounts of data from sources may present scalability obstacles.

Integration Complexity : Ensuring compatibility and integration, among data systems can be both intricate and resource demanding.

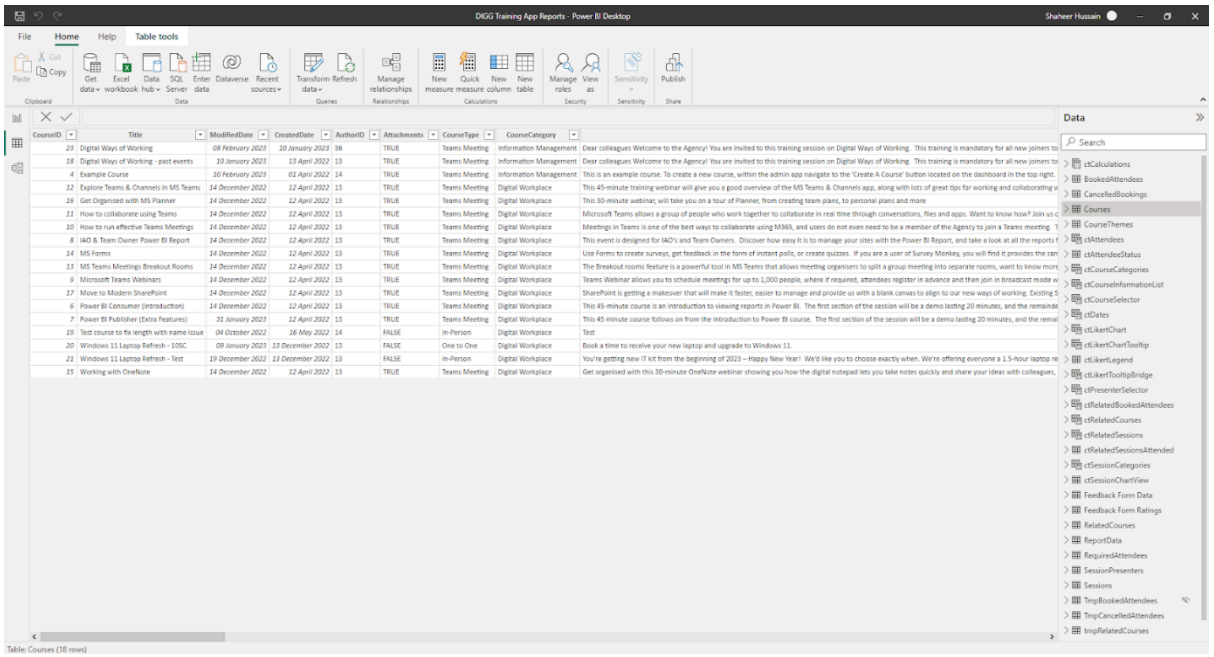
Risk Assessment and Mitigation Strategies

- Creating Data Standards : Establishing formats and protocols for data collection and storage.
- Ensuring Data Integrity : Implementing thorough data cleansing procedures to detect and rectify inconsistencies.
- Enhancing Security Measures :Utilising encryption methods to safeguard data during transmission and storage.
- Enforcing Access Controls :Implementing stringent access controls and authentication protocols to prevent entry.
- Protecting Privacy : Anonymizing data whenever feasible to ensure privacy is maintained.
- Ensuring Compliance : Regularly conducting audits to verify adherence to data protection regulations.
- Legal Protection Measures : Drafting agreements defining data usage rights and obligations.
- Seeking Legal Counsel : Consulting professionals for guidance on navigating data integration complexities and ensuring compliance.
- Utilising Solutions : Leveraging integration platforms that are scalable and compatible with diverse data sources.
- Monitoring System Performance : Regularly overseeing the integrated system, for efficiency issues and updating it as needed to address challenges.
- Managing Risks Effectively : Performing risk assessments before integration to identify pitfalls.
- Developing Risk Mitigation Strategies : Creating customised plans to mitigate identified risks effectively.

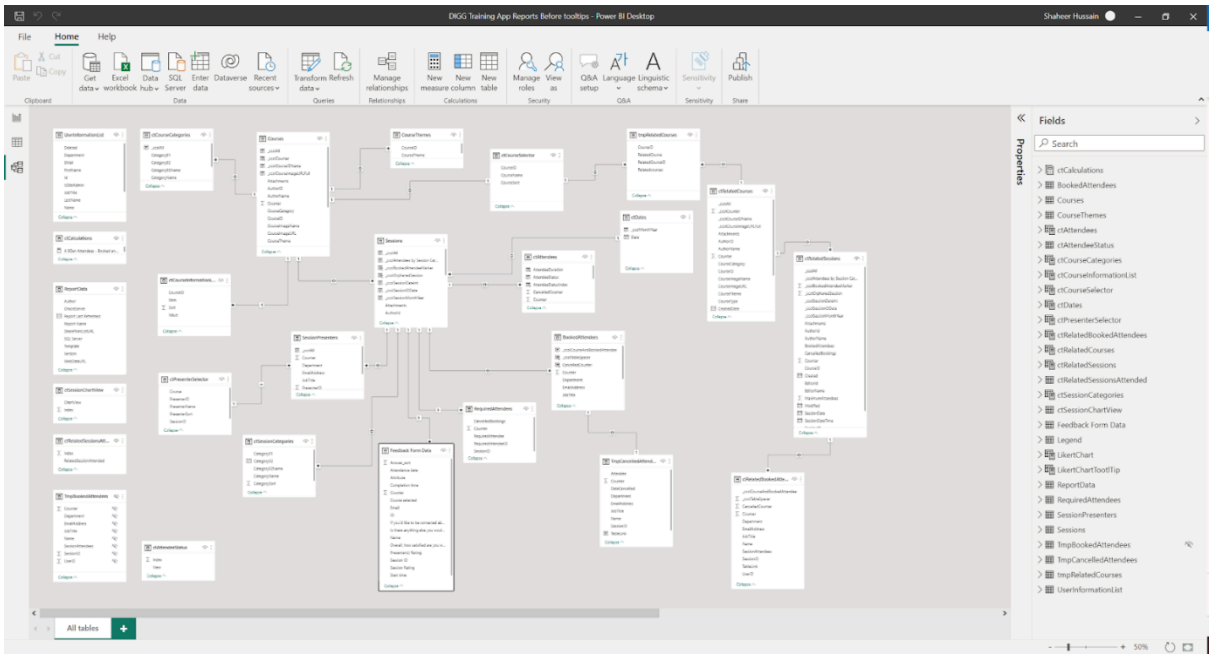
ES1 - apply organisational architecture requirements to data analysis activities (S9)

The tools relevant to the MHRA and this project include Power BI Desktop, SharePoint and MS Teams, however as this project continues to develop, other tools like a PowerBI to MS Teams live connector may be used or potentially MS Graph. Power BI Desktop is where the main tasks will be completed to build the report (Tasks like data sourcing, data transformation, data modelling and data visualisation), this report will mainly utilise SharePoint and Web connections to source data and MS Teams is just used for communication between stakeholders. From Power BI I can directly publish the report to the agency’s Data & analytics workspace on power BI service (Power BI Service is where a lot of reports made for the agency are stored and can be managed).

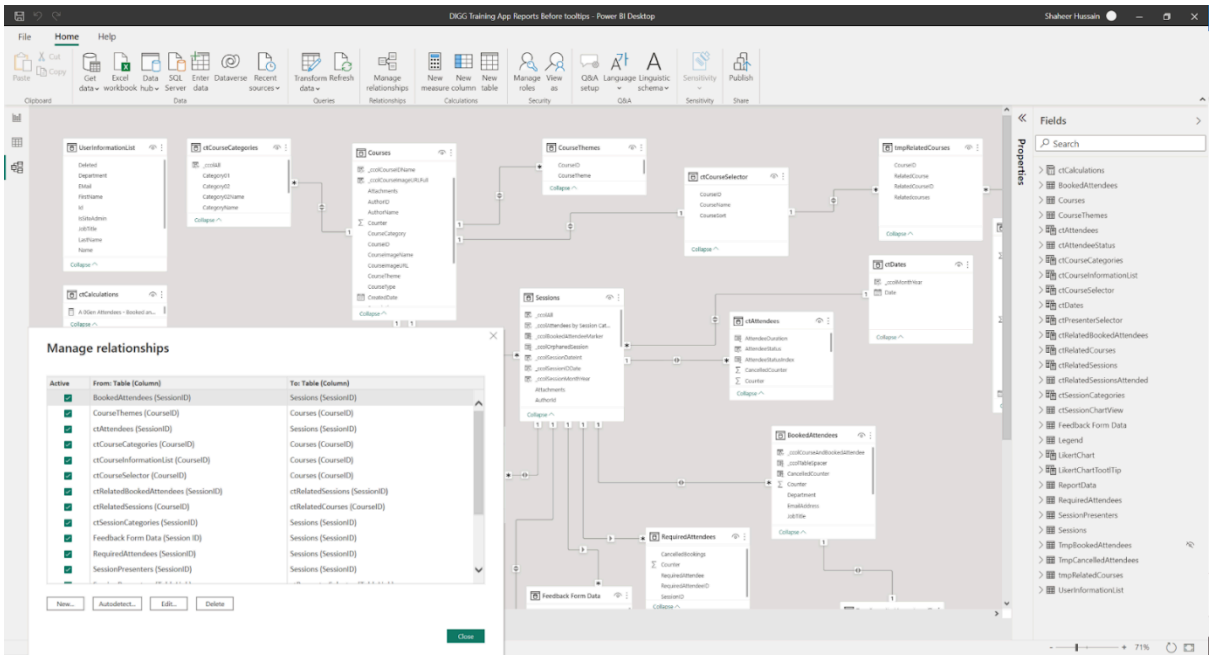
ES1 - collate and interpret qualitative and quantitative data and convert into infographics, reports, tables, dashboards and graphs (S14)



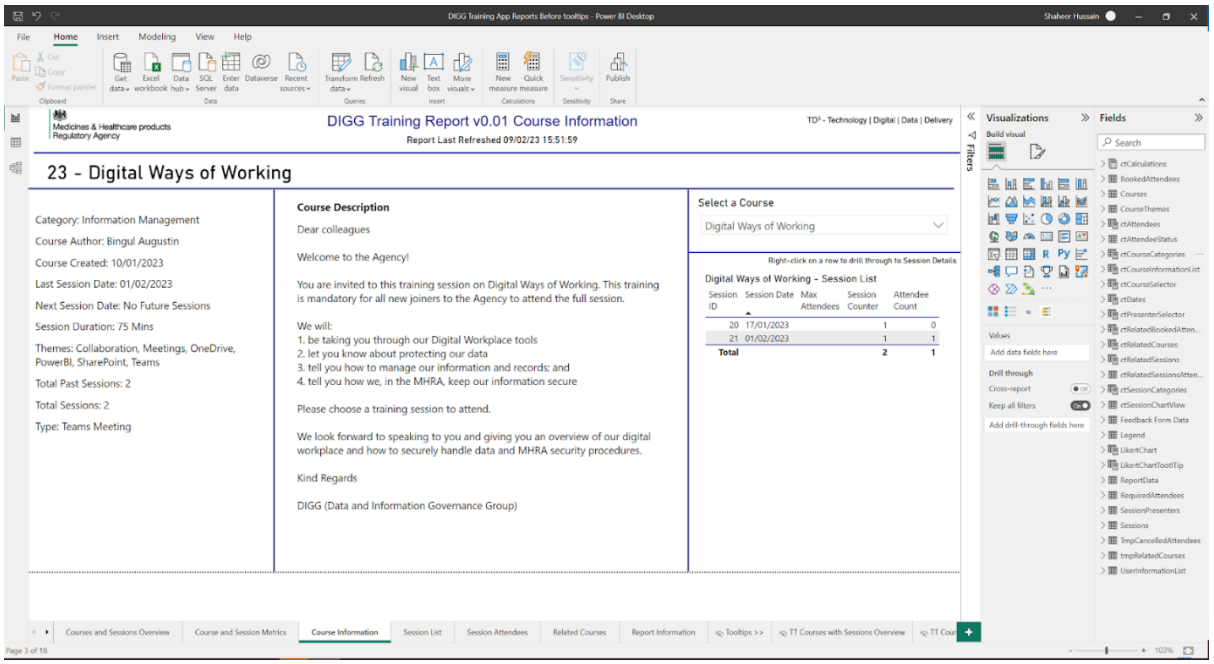
ES1 Figure 5: Courses data in table view (Power BI).



ES1 Figure 6: Modelling view of report showing the relationship between the different dimensions and fact tables.



ES1 Figure 7: List of relationships between tables.

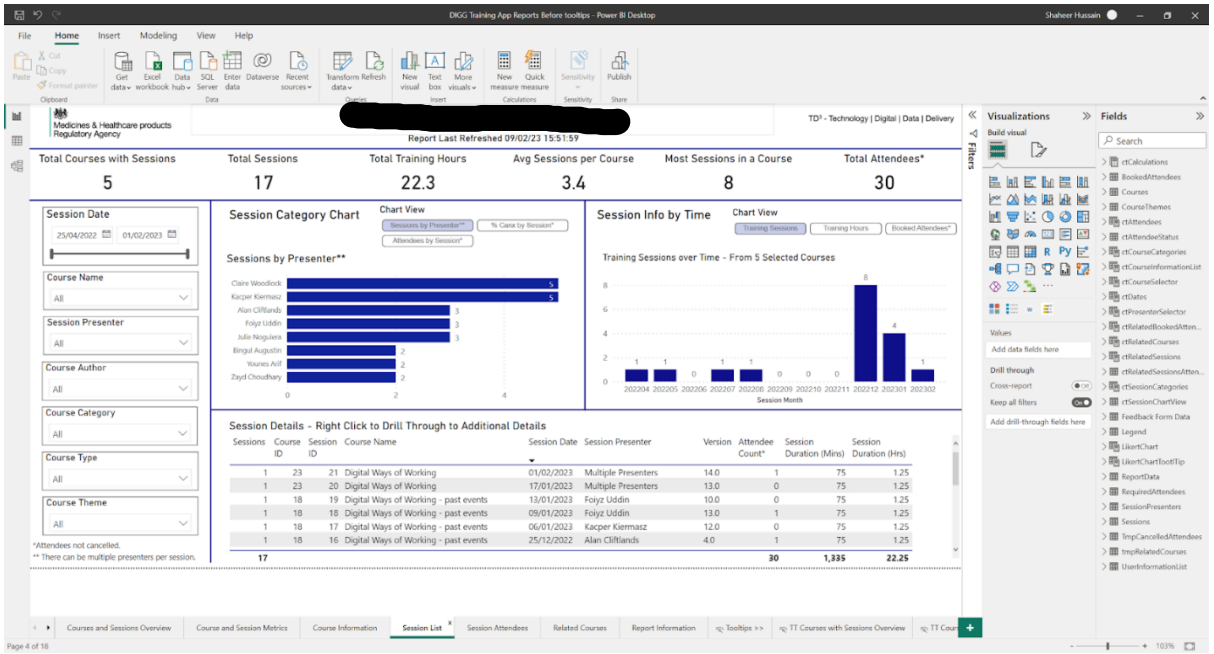


ES1 Figure 8: Course Information page (User can select a course and various details about the course will display).

S14: collate and interpret qualitative and quantitative data and convert into infographics, reports, tables, dashboards and graphs.

Figure 8 shows the visualisation of qualitative and quantitative data that have been collated and converted into a information page for the courses, this will allow me to meet the objective of creating useful visualisations and meet the aim of providing insight, this was done by using tables and slicers where data from the courses, sessions and some of the calculated tables were dropped into these tables. We can clearly see

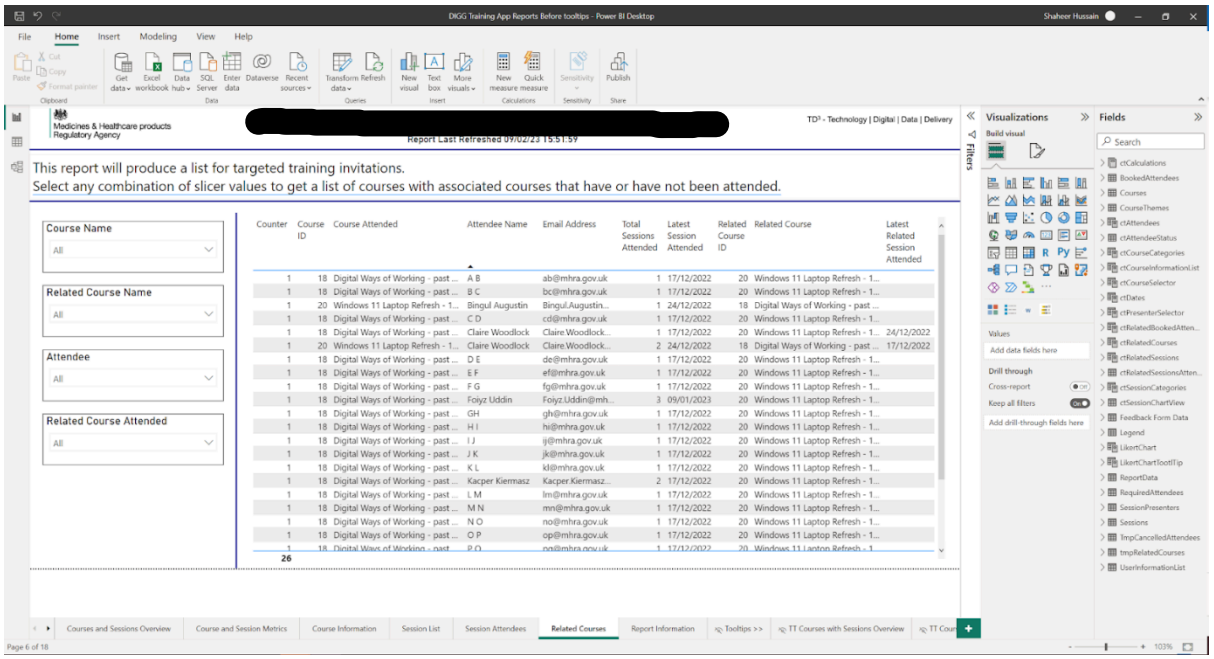
qualitative data where the section for course details is and there is quantitative data in the left side of the page (e.g. “Total sessions: 2”) as well as the session list table to the right. I created this page so that the end user can easily look through the courses that they are providing and gather information quickly.



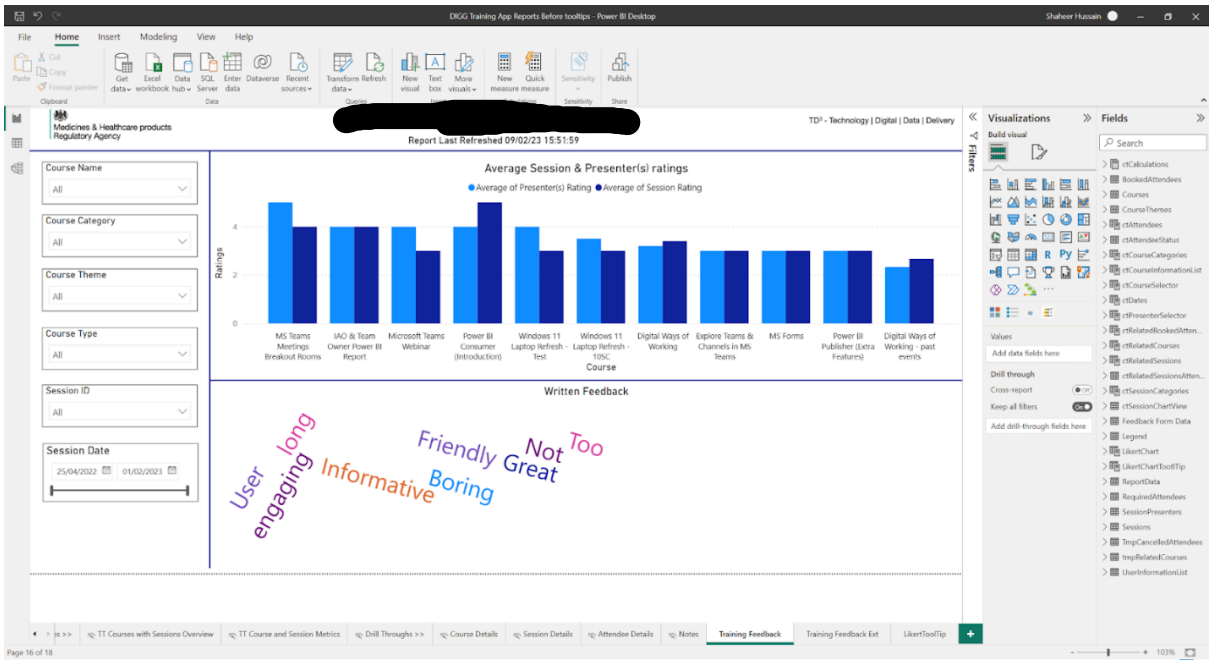
ES1 Figure 9: Session Information page (User can view the different sessions taking place as well as some stats over time).

S14: collate and interpret qualitative and quantitative data and convert into infographics, reports, tables, dashboards and graphs.

Figure 9 is another example of qualitative and quantitative data being visualised as we can see Quantitative data in the two graphs as well as in the top bar of the page where data like “Total attendees” is shown. This was done by using a series of card visualisations as well as clustered column charts and stacked bar chart visualisations to show the number of sessions hosted by each presenter and the number of sessions hosted over time. I also used a custom visual called a chiclet slicer which can be seen where the title “Chart view” is present, This will allow you to change the view of the visualisations to other stats like attendees by session or training hours over time which will help me meet the aim of providing useful insights to the client by giving the client more information to explore. I created this page to allow the user to explore more stats about the training sessions which may be of interest for the end user.



ES1 Figure 10: Related courses page (Requested by client).

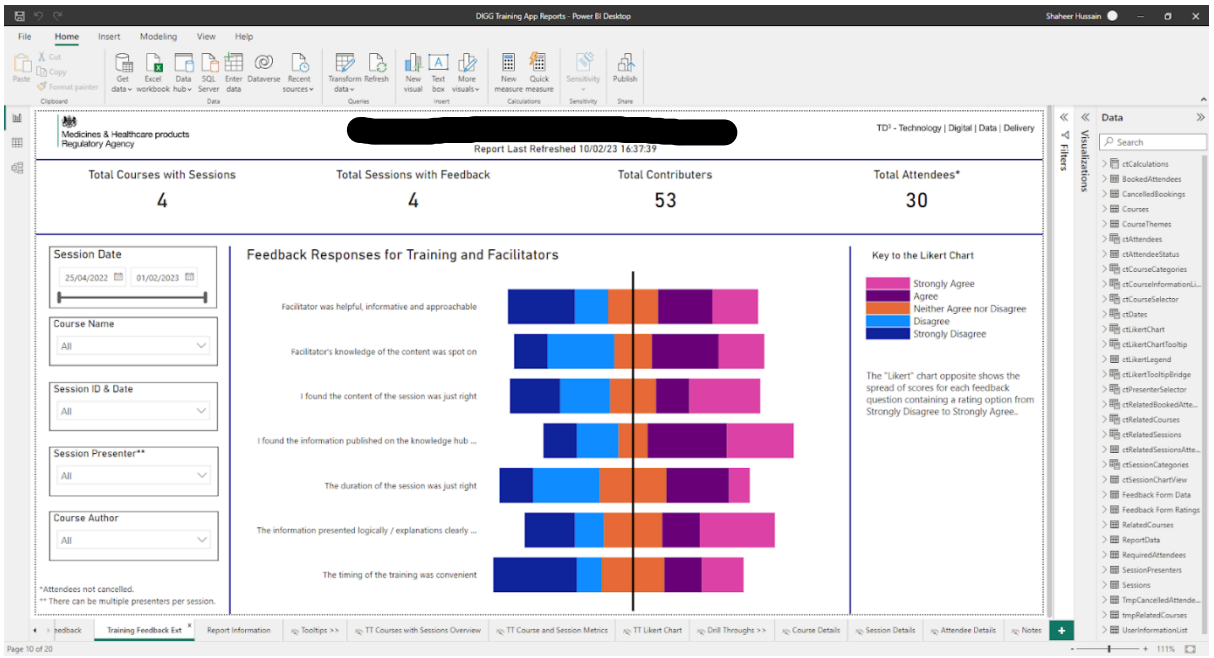


ES1 Figure 11: Training feedback analysis (Requested by client).

Figure 11 shows the average presenter and session rating (Quantitative data) by Course and it shows the most frequent written feedback from the feedback form (Qualitative data).I created this page by copying the basic format of another page whilst altering it to suit needs, I was able to copy over the slicers from another page but I had to establish a relationship between the feedback form data and the Courses/Sessions data to allow the slicer to actually work as you can see in Figure 6 and Figure 7, For the top section I simply used a clustered column chart and dropped the presenter and session ratings into the visualisation, I also included courses from the Courses table to separate the ratings by course, For the bottom section I was able to use a word cloud visualisation which is a custom visual that allows you to display the most frequent words that populate a column. I created this page because it provides a useful overview of the Training feedback since the end user

can see the most frequent words associated with the sessions, they can also see the direct ratings of presenters and sessions, this page enables me to meet the aim of providing useful insight as the end user will understand what the trainees think of their sessions, in addition to this the user has many filter options on the left side allowing for specific searches.

There are a few different stakeholders for this project and data visualisation should be specialised so that different stakeholders can understand the information. For example, I do not know if the client / end user has a deep statistical background so I may use clustered column charts like in Figure 11 to aid their understanding of the data whereas a stakeholder like my manager will have a more technical or statistical back which will enable me to use visualisations like the Likert chart in Figure 12 or other visualisations like a distribution graph or box plot diagram.



ES1 Figure 12: Training feedback analysis (Likert chart).

Figure 12 shows a custom Likert chart made from a bar chart visualisation using DAX (Data analysis expression) which means this helps me meet the objectives of using DAX to create useful measures and creating visualisations to present information in a useful manner, this will also enable me to meet my aim of providing useful insight as it will show the end user the distribution of trainees who Strongly agree, agree, neither agree nor disagree, disagree or strongly disagree with the statements on the left side of the Likert chart. This Likert chart presents qualitative data in the form of the statements and colour code but also quantitative data in the form of the bars that populate the chart. I created this Likert chart with the DAX shown in Figure 15. As you can see I use the UNION function because I will be taking the feedback forms ratings table twice to return a table which I can edit. I used a counter to assign values to the different values like agree and disagree, this will be explained later. In the first part of the UNION I select the columns; ID, Contributor, CourseID, Attribute, Response, Sort and Likertbarvalue from Feedback forms rating only if response is Disagree or strongly disagree and if so, the counter will be multiplied by -1 so that these responses appear on the left side of the Likert chart where it belongs. The second part of the UNION is very similar as it selects the same columns, but it is filtered to the response being equal to “Neither Agree nor Disagree'' and the counter will be divided by 2, this is so that these responses will appear closer to the middle of the Likert chart.

Reflective Report

SD

Scott, Donald

To Hussain, Shaheer

09:13

Shaheer Hussain - Reflective Report.docx
19 KB

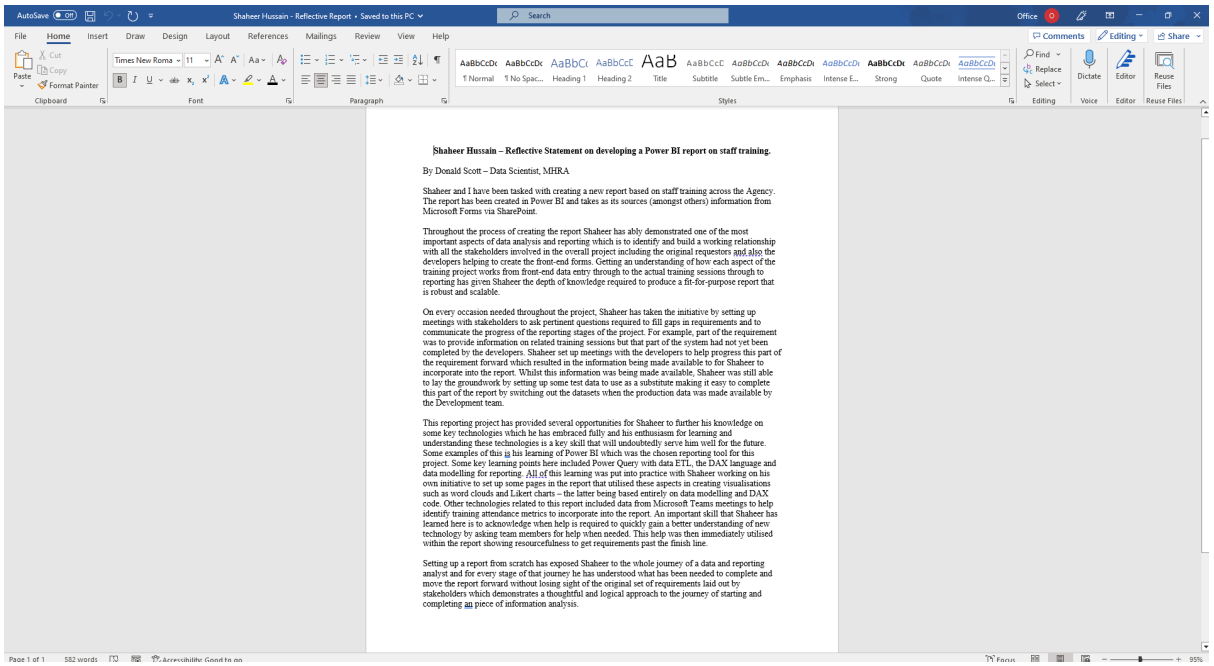
Morning Shaheer – hope you had a good weekend!

I've attached a reflective statement – let me know if it's in the right format etc as I can easily amend it as needed.

Cheers!

Don.

ES1 Figure 13: Reflective statement from colleague.



Shaheer Hussain – Reflective Statement on developing a Power BI report on staff training.

By Donald Scott – Data Scientist, MHRA

Shaheer and I have been tasked with creating a new report based on staff training across the Agency. The report has been created in Power BI and takes as its sources (amongst others) information from Microsoft Forms via SharePoint.

Throughout the process of creating the report Shaheer has ably demonstrated one of the most important aspects of data analysis and reporting which is to identify and build a working relationship with all the stakeholders involved in the overall project including the original requestors and also the developers helping to create the front-end forms. Getting an understanding of how each aspect of the training project works from front-end data entry through to the actual training sessions through to reporting has given Shaheer the depth of knowledge required to produce a fit-for-purpose report that is robust and scalable.

On every occasion needed throughout the project, Shaheer has taken the initiative by setting up meetings with stakeholders to ask pertinent questions required to fill gaps in requirements and to communicate the progress of the reporting stages of the project. For example, part of the requirement was to provide information on related training sessions but that part of the system had not yet been completed by the developers. Shaheer set up meetings with the developers to help progress this part of the requirement forward which resulted in the information being made available to for Shaheer to incorporate into the report. Whilst this information was being made available, Shaheer was still able to lay the groundwork by setting up some test data to use as a substitute making it easy to complete this part of the report by switching out the datasets when the production data was made available by the Development team.

This reporting project has provided several opportunities for Shaheer to further his knowledge on some key technologies which he has embraced fully and his enthusiasm for learning and understanding these technologies is a key skill that will undoubtedly serve him well for the future. Some examples of this is his learning of Power BI which was the chosen reporting tool for this project. Some key learning points here included Power Query with data ETL, the DAX language and data modelling for reporting. All of this learning was put into practice with Shaheer working on his own initiative to set up some pages in the report that utilised these aspects in creating visualisations such as word clouds and Likert charts – the latter being based entirely on data modelling and DAX code. Other technologies related to this report included data from Microsoft Teams meetings to help identify training attendance metrics to incorporate into the report. An important skill that Shaheer has learned here is to acknowledge when help is required to quickly gain a better understanding of new technology by asking team members for help when needed. This help was then immediately utilised within the report showing resourcefulness to get requirements past the finish line.

Setting up a report from scratch has exposed Shaheer to the whole journey of a data and reporting analyst and for every stage of that journey he has understood what has been needed to complete and move the report forward without losing sight of the original set of requirements laid out by stakeholders which demonstrates a thoughtful and logical approach to the journey of starting and completing an piece of information analysis.

ES1 Figure 14: Reflective statement from colleague.

ES1 - show initiative, being resourceful when faced with a problem and taking responsibility for solving problems within their own remit. (B2)

During this project I had encountered a series of issues aligning with what the client wanted from the report, the main issue was being able to show MS Teams call attendance data in Power BI; this would be needed to track stats like how long someone is present in a call. I took it upon myself to do the research on this topic as this would help me meet the aim of the project which is providing insights useful to the end user. I came to the conclusion that a lot of the solutions I could provide would require manual steps but a solution could possibly be engineered through Microsoft Graph, however MS Graph is out of my skill range and capability, it also did not look very user friendly when I looked into learning how to use the tool so I had to adapt my research towards a live connector between the two programs, this is when I stumbled across connectors from companies like CDATA and Sharegate which could solve our issue. Some of the other issues include not having a way to link a course to its related courses and establishing a page for the Knowledge hub but as my colleague has stated in Figure 13 and Figure 14, I set up quite a lot of meeting between stakeholders from the end users to the developer behind the app, allowing me to explain what is needed or to gain clarity in what is required of the report.

ES1 - Identifies and explains challenges in their work and how they overcame them, providing an outline of lessons learned (B6)

I had encountered a number of challenges during this project as it its my first big project where I am directly involved in producing a deliverable, One challenge I had was getting used to PowerBI which was a completely new program to me so there was quite a steep learning curve, I had overcome this issue by investing the hours into learning the basics of PowerBI using the MHRA's Pluralsight subscription (Pluralsight is a website that hosts a number of learning courses of differing difficulties on programs like PowerBI) as well as other online courses and resources. The main lesson I had learned here was to continue investing the time into the skill Im trying to develop because although the learning curve may be steep, there are a lot of resources for me to improve on my skills and make it easier for myself to work on deliverables.I also had a couple of technical challenges however I had overcome these challenges in a similar manner, looking at online resources, conducting research and gathering advice from my manager so although these challenges were different, they reinforced the lesson of making the most of my resources to learn more and improve as much as i can to tackle the issues I face at work.

Review

Output, Outcomes, and their evaluations

The output of this project is a report that will allow the end user to see useful information like the feedback on their sessions and gain a easily understandable insight into the training sessions / Courses that they host, The report has not been deployed yet but the predicted outcomes of this report include the ability to track who has or not attended their training session and / or a related training session in a much more efficient manner, The ability to understand what the trainees think about the training sessions and adapt / improve the courses, view statistics like the number of sessions over time and even viewing much simpler information like Course details or who should be attending a session. All of these outcomes are intended to massively improve the efficiency of the workflow for the end user who currently has to rely on an admin-heavy process to carry out these functions whereas this report will host these functions all in one place.

Lessons learned | WWW & EBI

The main lessons I have learned from this project are more technical, this includes how to get data into Power BI, cleansing and transforming the data to get rid of unnecessary data to align the data to meet requirements, model relationships between tables to make the data relational (Allowing me to pull data from different tables into one visualisation without the visualisation breaking, this means that I can create more specific visualisations that hold more value to the end user), Using Dax to create some simple and complex functions like a custom Likert chart and a custom tooltip that appears when hovering over the Likert chat, finally I learned how to make visualisations that add value and would meet the clients requirements e.g. the clustered column chart to show average presenter and session rating shown in Evidence 11.

What went well?

The report has been quite successful in function as it has working live connections that successfully display accurate information about the sessions / courses in an easy to understand manner with a lot of functionality e.g. the filters allowing the user to streamline their search, making the value of this report very high due to its capabilities. Also, I was able to collaborate quite a lot with my colleague even with obstacles such as Power BI versions causing incompatibility issues as well as communicating openly with stakeholders like the end user for feedback and the developer for implementation suggestions.

Even better if …

We were able to find a way to create a live connection between Power BI and MS Teams without paying for a third-party connector or using very complex and not user-friendly tools like MS Graph, The implementation of certain functions like the “Related courses” feature being present earlier would allow me to have more time to test it in the report, Finally having more clarity about the requirements of the report from the beginning of the project would give me a better understanding of what would’ve needed to be done to create a robust solution.

Changes I would make.

If I were to redo this project, I would have dedicated a little more time into researching 3rd party connectors for Power BI and Microsoft Teams so I could learn more about the capabilities of these connectors and

understand whether they could be used elsewhere within the agency to improve the efficiency of the workflow.

Appendix

```
ctLikertChart =
    UNION(
      SELECTCOLUMNS(
        'Feedback Form Ratings'
        ,"ID", 'Feedback Form Ratings'[ID]
        ,"Contributer", 'Feedback Form Ratings'[Contributer]
        ,"CourseID", 'Feedback Form Ratings'[CourseID]
        ,"Attribute",'Feedback Form Ratings'[Attribute]
        ,"Response",'Feedback Form Ratings'[Response]
        ,"Sort",'Feedback Form Ratings'[AnswerSort]
        ,"LikertBarValue",IF('Feedback Form Ratings'[Response] IN {"Disagree","Strongly Disagree"}
                  ,'Feedback Form Ratings'[Counter] * -1
                  ,IF('Feedback Form Ratings'[Response] = "Neither Agree nor Disagree"
                    ,('Feedback Form Ratings'[Counter] /2)*-1
                    ,'Feedback Form Ratings'[Counter]))
        ,"SessionID", 'Feedback Form Ratings'[SessionID]
        ,"Counter", IF('Feedback Form Ratings'[Response] = "Neither Agree nor Disagree"
                  ,.5
                  ,1))
      ,CALCULATETABLE(
        SELECTCOLUMNS(
          'Feedback Form Ratings'
          ,"ID", 'Feedback Form Ratings'[ID]
          ,"Contributer", 'Feedback Form Ratings'[Contributer]
          ,"CourseID", 'Feedback Form Ratings'[CourseID]
          ,"Attribute",'Feedback Form Ratings'[Attribute]
          ,"Response",'Feedback Form Ratings'[Response]
          ,"Sort",'Feedback Form Ratings'[AnswerSort]
          ,"LikertBarValue",'Feedback Form Ratings'[Counter] / 2
          ,"SessionID", 'Feedback Form Ratings'[SessionID]
          ,"Counter", .5)
        ,FILTER(
          'Feedback Form Ratings'
          ,'Feedback Form Ratings'[Response]="Neither Agree nor Disagree")))
```

ES1 Figure 15 – DAX for the custom Likert chart visualisation

EVIDENCE STATEMENT [2]

NAME	Shaheer Hussain
TITLE	**[Communications Dashboard]**
PATHWAY	**[DATA ANALYST]**

Scenario

Introduction & Business problem

A request was sent to the data team (My team) around the 13th of June about a report that included information about project and non-project work. I took this request as an opportunity to gather some experience on problem solving and work on something new. The stakeholder of this request was the Lead Application Operations Engineer and he wanted to modify a few pages on this report to combine the project and non-project data into visualisations. This request had a timeline of between 1 - 2 weeks.

The issue was that the stakeholder tried to model the tables in Power BI however, the stakeholder did not do a correct job which explains why they were not able to combine this data. The structure of the data did not allow for a cardinality to be established between the projects and non-projects tables so the dilemma was figuring out a way to combine data from these two unrelated tables.

Tasks

My role as a Data analyst means I am responsible for the sourcing of appropriate data, cleaning and transforming this data to be only what is needed, modelling relationships to relate data from different sources and visualising this data to give insights. However, focussing on this project my role was to find a modelling solution to combine the two tables and create effective visualisations that will provide the insight that the stakeholder needs. For this project, Power BI was the only tool used as the issues were all contained within

Power BI, meaning that there were no other issues like data source issues. I chose to take a agile methodology as i will be able to continuously improve the report and gather feedback from the stakeholder to make sure the insights i am providing align with that is required.This will also help me deliver a higher quality report and avoid any risks of developing a report that is not what the stakeholder wants.

Activities

ES2 - Explains the relevant data policies and procedures for the organisation, and identifies the data standards to be reached (K2)

The Medicines and Healthcare products Regulatory Agency (MHRA) in the UK is responsible for ensuring that medicines and medical devices work and are acceptably safe. The MHRA follows stringent data policies and procedures to uphold these responsibilities, ensuring the safety, efficacy, and quality of medicines and medical devices.

Relevant Data Policies and Procedures, for MHRA
Data Protection and Privacy :

Following the guidelines set by GDPR the MHRA abides by the General Data Protection Regulation (GDPR) and the Data Protection Act 2018 to ensure that personal data is processed lawfully, fairly and transparently.

Protection of Data Subject Rights : Individuals are empowered to access their data, rectify any inaccuracies and request the removal of their data. The MHRA guarantees that these rights are respected.

Safeguarding Data Security : The MHRA implements both organisational measures to safeguard data from unauthorised access, accidental loss or destruction.

Clinical. Research Data :

Upholding Good Clinical Practice (GCP) : The MHRA ensures adherence to guidelines to oversee conduct in clinical trials and maintain accurate and reliable data.
Compliance with Clinical Trials Regulation (CTR) : The MHRA follows the EU Clinical Trials Regulation to regulate the approval and execution of trials ensuring standards for data quality and integrity.

Pharmacovigilance :

Monitoring Adverse Drug Reactions : The MHRA has established protocols for gathering, analysing and reporting drug reactions. This process includes utilising the Yellow Card Scheme, for healthcare professionals and the public to report any side effects they encounter.Monitoring Signals and Managing Risks : Consistently keeping an eye on data to spot any safety concerns related to medications and taking regulatory steps.

Maintaining Data Integrity :

Following ALCOA+ Principles : Making sure that data is identifiable, readable, recorded in time, precise, complete, uniform, long lasting and accessible.

Tracking Changes with Audit Trails : Keeping records of changes made to ensure the integrity of data throughout its lifespan.

Submitting Regulations :

Using eCTD (Electronic Common Technical Document) : The MHRA mandates the use of eCTD, for submitting regulations to ensure organised data for review processes.

Adhering to ISO IDMP Standards : Following ISO IDMP guidelines for identifying and describing products to enhance data exchange and compatibility.

Promoting Transparency and Public Health :

Sharing Data Publicly : The MHRA is committed to transparency by sharing trial information safety reports and regulatory decisions to build trust and aid healthcare choices.

Establishing Data Sharing Agreements : Forming agreements, for sharing data with bodies, healthcare providers and researchers while ensuring ethical use that benefits public health.Information security management standards are in place to protect the integrity, confidentiality and availability of data.

HL7 Standards :

HL7 FHIR (Fast Healthcare Interoperability Resources) facilitates the exchange of healthcare information to improve data interoperability.

HL7 CDA (Clinical Document Architecture) is used to structure documents for interoperability and data sharing.

ICH Guidelines :

ICH E6 (R2) GCP is a guideline from the International Council for Harmonisation that ensures quality and reliable data in trials through Good Clinical Practice.

ICH E2E Pharmacovigilance Planning provides guidelines for planning and managing pharmacovigilance activities to ensure safety monitoring.

CDISC Standards :

CDASH (Clinical Data Acquisition Standards Harmonization) standardises the format of data collected in trials to ensure consistency and comparability.

SDTM (Study Data Tabulation Model). Formats data for submission to authorities ensuring data quality and compliance.

Data Interoperability Standards :

SNOMED CT (Systematized Nomenclature of Medicine. Clinical Terms) establishes terminology for better communication, in healthcare settings.

LOINC (Logical Observation Identifiers Names and Codes) standardises laboratory and clinical observation data for interoperability.

ES2 - K5:The differences between structured and unstructured data

Structured data is clearly defined, can be organised and can be mapped into predefined fields. This often includes quantitative data which usually consists of hard numbers or items that can be aggregated. We often store structured data in data warehouses which is the endpoint of the data's journey in its ETL (Extract, transform and load) pipeline or it could be stored in relational databases. Finally, another difference is the format of structured data which consists of text and numbers.

Unstructured data is more scattered, variable and undefined. This often includes qualitative data which can't be analysed or processed using conventional methods and tools. Unlike structured data, unstructured data is usually stored in data lakes which is like an unlimited repository where data is stored in its original format or after undergoing a “cleansing” process. Unstructured data can come in a number of different forms including audio, imagery and video to email and sensor data. There is no sort of data model for unstructured data, instead, the data is stored in a data lake which doesn't require any transformations.

Examples of structured data in the MHRA could include data that is stored within the Oracle SQL database. This is structured as the data will be in the form of columns and rows, the data is usually broken up into tables with Primary and Secondary keys to establish relationships between them.

Examples of Unstructured data in the MHRA could include text descriptions like when a medical device breaks, the manufacturer of the medical device will have to give a description of what is wrong with the device, another example is when we have an adverse drug reaction, a patient's medical history will have to be searched up and a person's medical history will also be unstructured data. There are many instances where we store data in JSON files which is short for JavaScript Object Notation.

This is relevant to the work i did on this report as i had to create some bridge tables to make the connection between the Non-project hours and Project/Non-project data, this means that i had to include the Non-project data into the relational structure which connected the entire dataset together, this relational structure is an example of structured data.

ES2 - Describes the fundamentals of data structures and database system design and explains how they are implemented and maintained. (K6)

Principles of Data Structures, in the MHRA

1. Overview of Data Structures

Data structures are methods used to store and manage data efficiently for access and modification. Within the MHRA, a variety of data structures are utilised to handle amounts of information, clinical trial data, drug safety records and other healthcare related data.

2. Commonly Used Data Structures

- Arrays and Lists : These are employed for storage of data where elements are assigned indexes.
- Hash Tables : They are utilised for retrieval of data using key value pairs proving beneficial for indexing databases.
- Trees : Specifically binary trees and B trees are used for managing data structures and facilitating search, insertions and deletions.
- Graphs : These structures come into play when dealing with relationships like tracking interactions between drugs or healthcare entities.
- Stacks : These are applied in processing tasks by managing them in, out (FIFO) or last in first out (LIFO) manner.

Database System Design within the MHRA

1. Principles of Database Design

Data Integrity : Ensuring the precision and consistency of data throughout its lifespan.

Scalability : The capability to manage increasing volumes of data and user traffic effectively.

Security : Safeguarding healthcare information, from access.

Normalisation : Structuring data to minimise redundancy while enhancing data integrity.Denormalization is implemented when the performance gains outweigh the advantages of normalisation in databases that're read intensive.

There are two types of databases :

Relational Databases : These utilise Structured Query Language (SQL) and follow a table based structure. Examples include MySQL, PostgreSQL and Oracle Database.

NoSQL Databases : These are used for data and require scalability. Examples include MongoDB, Cassandra and Elasticsearch.

In the context of the MHRA, a combination of NoSQL databases is likely used to handle its datasets. Database Management Systems (DBMS) assist in defining, creating, managing and manipulating databases.

When it comes to schema design :

Entity Relationship Diagrams (ERD) visually represent the database schema by mapping out entities like drugs, clinical trials and healthcare providers along with their relationships.

Normalisation is employed to remove data redundancy and ensure integrity by organising data into tables and establishing relationships through keys.

Data indexing involves creating indexes on columns to enhance the speed of data retrieval operations. Hash indexes are utilised for exact match lookups while B tree indexes are commonly used for range queries.

Data Segmentation

Horizontal Segmentation (Sharding) : Spreads out the rows of a table across database nodes.

Vertical Segmentation : Divides the columns of a table into tables typically done to segregate less frequently accessed data.

Maintenance of Data Structures and Database Systems

1. Routine Updates and Patching

Databases receive updates, with the security patches and software upgrades to defend against vulnerabilities.

2. Backup and Restoration

Implementing solutions to prevent data loss from hardware malfunctions, software glitches or other emergencies.

It is crucial to have routines and test recovery procedures.

3.. Fine tuning

Continuous monitoring of database performance measures (such as query efficiency, disk usage).

Optimization involves indexing tuning queries. Adjusting database settings based on observed performance.

4. Data Reviewing and Security Compliance

Conducting data reviews to ensure adherence to healthcare regulations and standards.

Enforcing access controls and encryption for safeguarding information.

Practical Illustration : MHRA Adverse Event Reporting System

Consider an event reporting system :

Data Structures : Utilises a hash table for retrieval of reports by patient ID, a tree structure for managing hierarchical categorization of adverse events and lists, for maintaining report sequences.

Database Design :

Utilising a database to store organised information, on incidents with standardised tables for patients, reports, medications and symptoms. In cases of text data from reports a NoSQL database could be employed.

For upkeep : Routine backups of the database updating security patches optimising indexing, for efficiency and conducting data audits to adhere to standards.

ES2 - Explains the principles of user experience and domain context for data analytics. (K7)

User experience (UX) and domain context play a role in data analytics in a regulatory and health setting like the Medicines and Healthcare products Regulatory Agency (MHRA). When creating data visualisations for use

such as bar charts in Power BI to showcase work hour distributions, understanding these principles is key to making the data accessible and actionable. This tailored explanation is crafted specifically for the MHRA and my project.

Key User Experience (UX) Principles in Data Analytics :

Clear Communication and Simplicity :
In the context of MHRAs focus on regulation and health products, clarity is essential to prevent misinterpretations that could lead to compliance issues.
For my project I used labels and simple bar charts to help the stakeholder easily grasp the distribution of work hours. Keep things neat by concentrating on metrics.

Consistency Matters :
Within the MHRA environment, maintaining design consistency aids users in getting familiar with data presentation, which is crucial for analysis and decision making. For my project application, I ensured colours for project hours versus project hours across all bar charts. Stick to the scale and formatting standards throughout.

Relevance Is Key :
I made sure that I presented data that directly aligns with users requirements,assisting them in making choices regarding compliance and resource allocation.
Tailoring the bar charts to display breakdowns like employees and division is essential to address different inquiries from various user groups.

Interaction :
Interactive elements boost the user engagement and facilitate in depth analysis without complicating the view in the MHRA setting.
For application, I incorporated features in Power BI such as filters and slicers. This allows users to explore data based on time frames, projects or departments.

Domain Context in Data Analytics :

Regulatory Environment :
Given the environment within MHRA operations ensuring data accuracy and compliance is paramount. For project implementation purposes I validated that data sources for work hours are precise and current. Implement validation checks within Power BI to maintain data integrity.

Operational Efficiency :
In an MHRA context comprehending resource allocation across projects is critical, for efficiency and meeting regulatory deadlines.
I Designed bar graphs that showcase inefficiencies, like project hours in specific departments to assist managers in effectively reallocating resources.

Regarding Data Security and Privacy :
In the MHRA context, handling details about employees and projects demands compliance with data security procedures. To ensure adherence to MHRAs data protection policies I made sure that any employee specific information is either anonymized or access controlled in Power BI.

ES2 - K10: approach to combining data from different sources

There are many approaches to Combining data from different sources. if we look on the Power BI side, we can create relationships between tables by establishing a cardinality between them (e.g. One to Many), this will allow me to combine data from two different tables which can be from two different data sources (e.g. CSV file on SharePoint and Oracle database), In Power BI we can even combine data from two unrelated tables by utilising the UNION function in DAX (Evidence 1), this will take two tables that have the same number of columns and combine them

However, there are also other ways of combining data from different sources, for example using SQL we can make use of Inner joins (Combines rows that have matching values in two or more tables), Left joins (Includes unmatched rows from the table that is specified here the join clause) and Right joins (Includes all records from right table and only matching rows from left table in a result table) to combine tables and therefore the data within them. Combining data maximises efficiency as you can have two or more variables in one visualisation (Evidence 2) which can make it quick for a viewer to gain insight and it can save space on a page/dashboard for more visualisations. Making the insight more efficient can also reduce human error, improving accuracy, this is because the user doesn't have to keep switching between visualisations to grasp an insight.

ES2 - Evaluates the benefits and risks inherent in combining data from different sources (K10). [Distinction]

Combining data from different sources will provide many advantages; however, it can also come with its own drawbacks. I will provide an extensive evaluation of the benefits and drawbacks of combining data from different sources:

Some of the benefits include:

- Insights will be greatly enhanced when combining data because data analysts will be able to create a more holistic and comprehensive view of the subject that they are trying to convey information about, the integration of data leads to deeper insights which then leads to a more complete understanding of the subject at hand and this is always good for the stakeholder who will use the report.
- There will be more context to the data as integrating diverse data from different data sources will create a broader and deeper understanding of the subject matter and what the data is actually about. Data integration opens up opportunities for exploring the correlations, patterns and relationships between data from different sources which won't be present when analysing isolated datasets.
- Combining data from different sources can vastly improve data accuracy as it will help point out errors or anomalies which are present within the individual datasets and through validation processes, the validity and accuracy of the final dataset can be enhanced greatly. This can also go on to improve data quality as data integration can help fill in information that

is missing, therefore reducing information gaps and resulting in a more comprehensive dataset.
- Combining data from different sources will expand the scope of analysis and a data analyst will be able to widen their scope of analysis beyond an individual dataset, this will further help to identify outliers, trends and new opportunities which would have otherwise remained undiscovered.

The drawbacks include:

- Risk of raising concerns in regards to data privacy and security. It will be incredibly key to ensure compliance with the relevant rules and regulations to protect sensitive data and prevent any unauthorised misuse or access to the data.
- Combining data sources will potentially result in more data inconsistencies when it comes to definitions, data formats, data quality and units. These inconsistencies can cause errors or even biases within the analysis and this can be detrimental as it can lead to misleading insights which could then result in the wrong actions being taken by stakeholders.
- The complexity of the data when combined will increase as the analyst will be handling much larger volumes of data, maintaining the data consistency/quality and managing any updates to the data. This can become extremely challenging and resource intensive.
- There will be more challenges with data integration as combining data from different sources will require more complex processes, especially when dealing with tough data formats or structures. Challenges in data transformation, data cleansing and data merging may arise, and require a higher level of expertise to deal with.

Two of the Points above that relate back to this project include improved insights as i was not only able to show off the different projects that employees are working on by the resources that are being allocated to each of these projects, this combination will help to provide more understanding to factors like employee performance or even identify any outliers where e.g. a project may not be getting enough resources allocated to it.However, the complexity of the data has also increases as I had do deal with creating a relationship between the Project/Non-Project hours and the actual Project/Non-Project data through DAX.

ES2 - Describes impact on user experience and domain context on data analysis (S5)

User experience focuses on having a great understanding of users and what they value, what they need and their abilities, this can also extend to include the way that users interact with the data analysis results. User experience can take into account aspects like ease of use, intuitiveness, clarity and satisfaction.

Domain context refers to the industry in which the data analysis is being conducted, different industries or fields will have their own challenges, requirements and characteristics.Understanding the domain context is very important as it ensures I am able to create effective analysis which aligns with the domain's specific goals and needs.

The domain context for this report is communications / HR or workforce management, I came to this conclusion as this report will help the user derive insights about employee productivity as it shows how employees are spending their time (whether on projects or non-project work), being able to track the number of project hours also enables the stakeholder to allocate resources more efficiently. There are a number of other reasons why I believe this is the domain context like performance evaluation, time management and project management.

The impact of user experience on data analysis is that attention to user experience will make the data analysis more user friendly, this means that it will be easier to understand, this will prevent the stakeholder from making wrong decisions due to misunderstanding the insights. A positive user experience will help the stakeholder build trust in the data analysis process, this will mean that the stakeholder is likely to rely on the data analysis process more and therefore make more informed data driven decisions.

Understanding the domain context has a huge impact on data analysis as it enforces me to provide analysis or insights which are tailored to the specific needs for communications and HR. This will allow me to have more relevant data selection when creating visualisations to directly address the domains challenges and opportunities. The domain context can have an impact on interpreting insights as certain patterns or trends may seem irrelevant but to certain field or industry, it can be a major indicator of something important.If positive user experience can build trust between the stakeholder and the data analysis process, understanding the domain context will ensure that the stakeholders planning, policy changes,process improvements and resource allocation are more strategic and fit for the field.

ES2 - Describes the relevant tools or techniques used for working with the data systems architecture in their organisation. (S9)

Tools	This tool was used to…	Reason for selection
PowerBI	import the data, transform and cleanse the data, create visualisations and a dashboard to provide insights	This is a widely used tool within the agency and it is very intuitive which makes it easy for a beginner to create useful dashboards

Microsoft teams	Communication and meetings	Another heavily used tool in the MHRA but it's very effective when trying to get a message across very quickly which makes communication around the project very easy.

ES2 -Explains how they have applied analytical techniques for data mining and time series forecasting and other modelling techniques (S13)

Here is analytical technique i used for the descriptive analysis of this project :

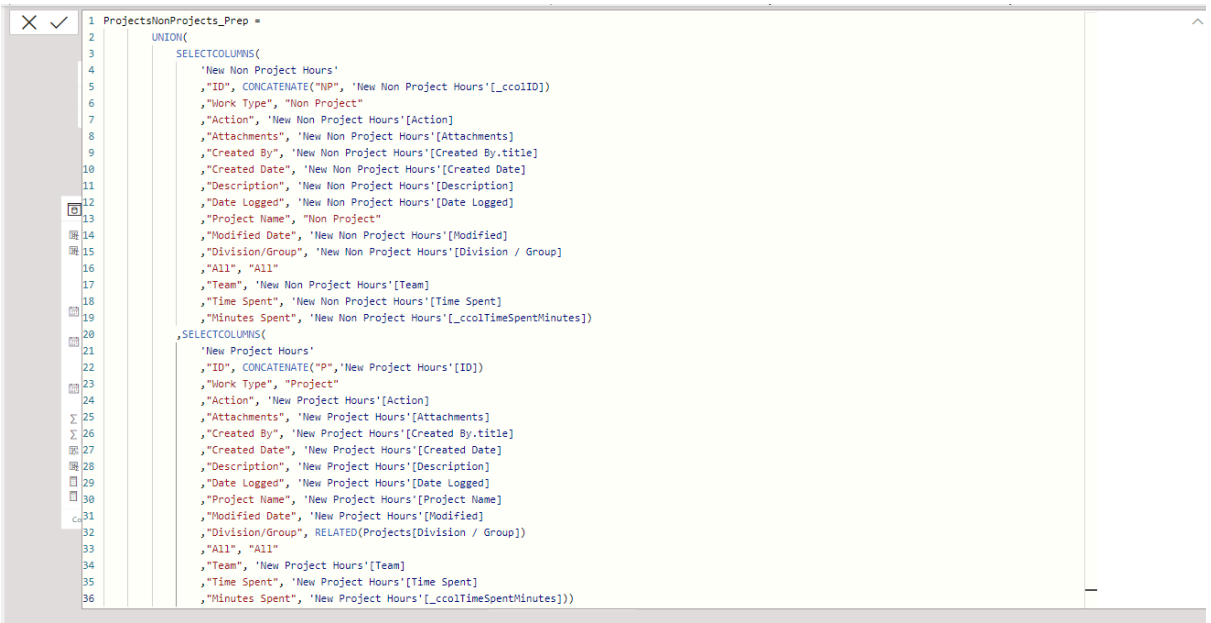
1. Data Collection and Preparation:
    a. Data Collection: I gathered data from the csv files that included project and non-project hours
    b. Data Cleaning: Handle missing values, remove duplicates, and correct errors in the data.
    c. Data Transformation: Aggregate, and encode data to ensure it's in a suitable format for analysis.
2. Exploratory Data Analysis (EDA):
    a. Visualisation: I made use of a lot of bar charts as it's easily understood and perfect for this project and scenario in terms of providing effective insights into the data .
3. Model Building:
    a. I made use of the UNION function in DAX in order to combine data from two unrelated tables and this allowed me to compare the project hours with the non-project hours
4. Model Deployment:
    a. Integration: Implement the model in a production environment where it can interact with real-time data which in this case is SharePoint.
    b. Monitoring: Continuously monitor model performance and update it as necessary to maintain accuracy and relevance.

The steps I took in the above can be seen in Evidence 1, Evidence 2, Evidence 3, Evidence 4, Evidence 5, Evidence 6 and Evidence 7.

ES2 - Explains the ethical aspects associated with the collation and use of data and justifies why this is important (K15)

There are many ethical aspects to consider when dealing with the collation and use of data, I will include some examples below:

- When dealing with personal or sensitive information it is important to obtain informed consent as it will ensure the individuals are aware of how their data is being used or how the MHRA's data is being used. The purpose behind its collection and any potential risks that would be involved in sharing data. This will also empower individuals to make more informed decisions about their data and it promotes data transparency which is best practice.
- Data analysis and any decision-making processes should avoid any discrimination and promote fairness as biases in data analysis will perpetuate any existing unjust treatment or inequalities.
- Another aspect is the security of data as safeguarding data against any breaches, unauthorised access and cyber threats is imperative. Organisations as well as individuals behold the responsibility to carry out appropriate security measures to protect data. Failure to secure data will have severe consequences which can include financial loss, identity theft and even damage to the MHRA's reputation.
- Adherence to the principles of data minimisation which is to only collect the necessary data for a specific purpose. Collecting irrelevant or excessive data is very questionable in ethics because it exposes more data to privacy risks and the potential for unintended use or unauthorised access. Minimising the data to collect only what is needed means there is less data exposed to these risks.
- Finally another very important aspect is respecting the rights and decisions of individuals to withdraw their data or request the deletion of their data when their consent has been provided. It is ethical practice to allow individuals to be in control of their data which allows them to exercise their rights and own choices over their data.



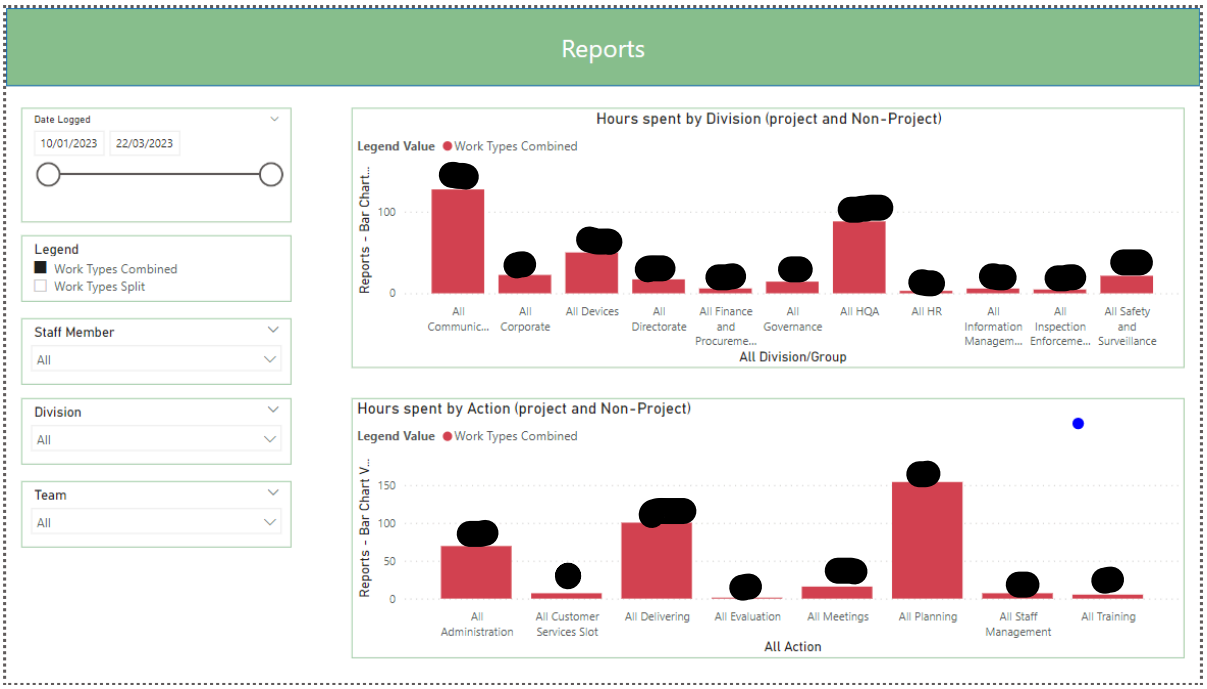
```
ProjectsNonProjects_Prep =
        UNION(
            SELECTCOLUMNS(
                'New Non Project Hours'
                ,"ID", CONCATENATE("NP", 'New Non Project Hours'[_ccolID])
                ,"Work Type", "Non Project"
                ,"Action", 'New Non Project Hours'[Action]
                ,"Attachments", 'New Non Project Hours'[Attachments]
                ,"Created By", 'New Non Project Hours'[Created By.title]
                ,"Created Date", 'New Non Project Hours'[Created Date]
                ,"Description", 'New Non Project Hours'[Description]
                ,"Date Logged", 'New Non Project Hours'[Date Logged]
                ,"Project Name", "Non Project"
                ,"Modified Date", 'New Non Project Hours'[Modified]
                ,"Division/Group", 'New Non Project Hours'[Division / Group]
                ,"All", "All"
                ,"Team", 'New Non Project Hours'[Team]
                ,"Time Spent", 'New Non Project Hours'[Time Spent]
                ,"Minutes Spent", 'New Non Project Hours'[_ccolTimeSpentMinutes])
            ,SELECTCOLUMNS(
                'New Project Hours'
                ,"ID", CONCATENATE("P",'New Project Hours'[ID])
                ,"Work Type", "Project"
                ,"Action", 'New Project Hours'[Action]
                ,"Attachments", 'New Project Hours'[Attachments]
                ,"Created By", 'New Project Hours'[Created By.title]
                ,"Created Date", 'New Project Hours'[Created Date]
                ,"Description", 'New Project Hours'[Description]
                ,"Date Logged", 'New Project Hours'[Date Logged]
                ,"Project Name", 'New Project Hours'[Project Name]
                ,"Modified Date", 'New Project Hours'[Modified]
                ,"Division/Group", RELATED(Projects[Division / Group])
                ,"All", "All"
                ,"Team", 'New Project Hours'[Team]
                ,"Time Spent", 'New Project Hours'[Time Spent]
                ,"Minutes Spent", 'New Project Hours'[_ccolTimeSpentMinutes]))
```

ES2 Figure 1: Using the UNION function in DAX to combine data from two unrelated tables.

Figure 1 shows the DAX formula that I used in order to combine data from the "New Project Hours'' and "New Non Project Hour'' tables into visualisations seen in Figure 2,Figure 3,Figure 4,Figure 5,Figure 6 and Figure 7. I did this by using the SELECTCOLUMNS function which allows me to specify the columns needed in the output and I stated what the name of the columns should be in red as well as which table they have come from in blue. This DAX will be particularly useful for combining the project and non-project data which will allow the user to view a combined view of work hours whether this is project work or not, this saves the user the time it takes adding the data up from the two tables manually, therefore improving time efficiency. The CONCATENATE function allows you to combine values so in this context we are combining "NP" or "P" to the unique identifier in each table to distinguish between non-project and project.

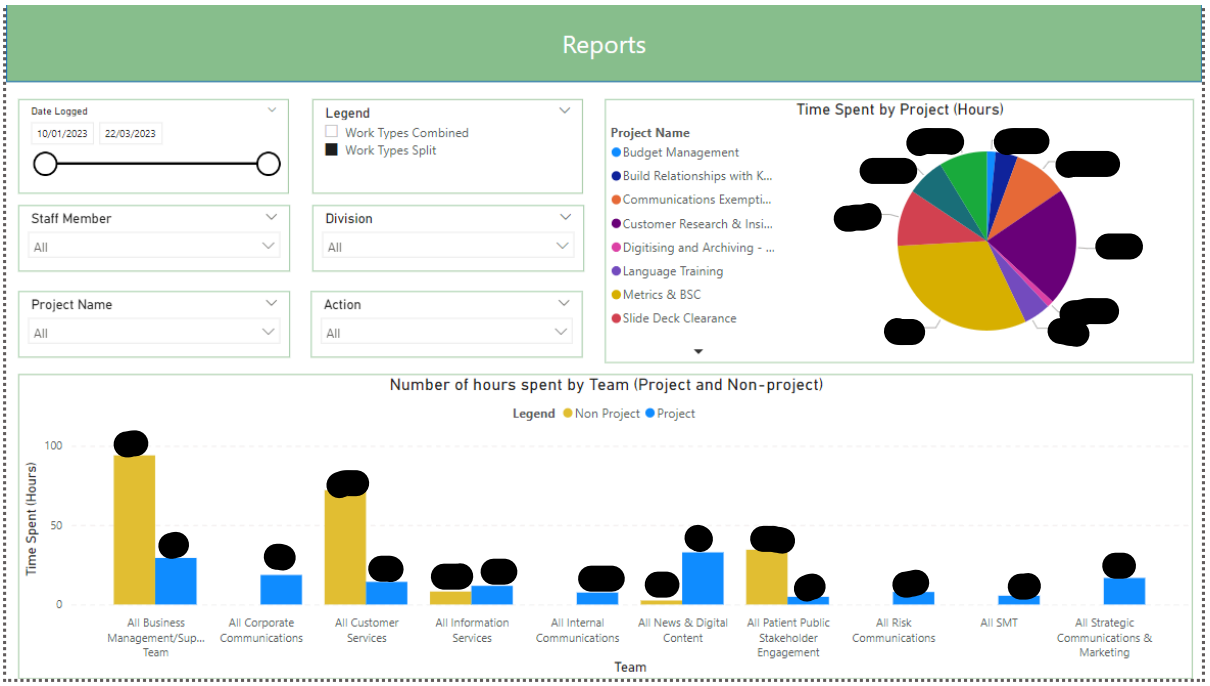
ES2 Figure 2: Result of UNION - the ability to combine project and non-project data into 1 visualisation - Work hours spent by division and action (Project and Non-Project)

Figure 2 shows us the number of hours spent by division and the number of hours spent by action in a comparative view between project and non-project hours. I created these visualisations using the Clustered column chart as it is quite easy to understand for users and it makes comparing categorical data incredibly easy. I included some filters on the left, this will provide extra use for the user as they will be able to filter by their specific need and because it is positioned on the left side of the report. It will usually be the first thing the user sees. There is also a date slider which will allow the user to filter by a specific date range. I created this page to provide an insight into how each division is balancing their work hours as well as to show what these work hours are being spent on as this can be incredibly useful for someone managing a team. This will allow the user to identify their team's focus and understand if their team is on track.



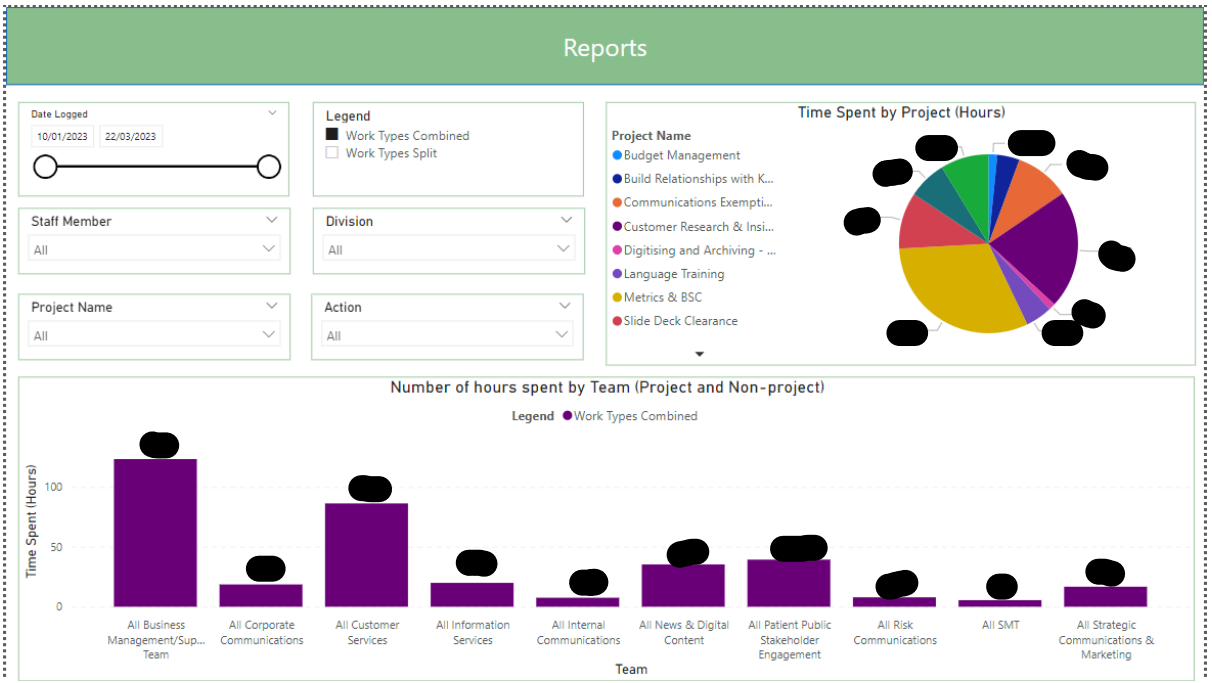
ES2 Figure 3: Result of UNION - the ability to combine project and non-project data into 1 visualisation - Work hours spent by division and action (Project and Non-Project) in combined view

Figure 3 shows the same data as evidence 2. However, the data in the two clustered column charts are combined which as I mentioned earlier will save the user some time in adding the numbers up to show a total figure of work hours spent.



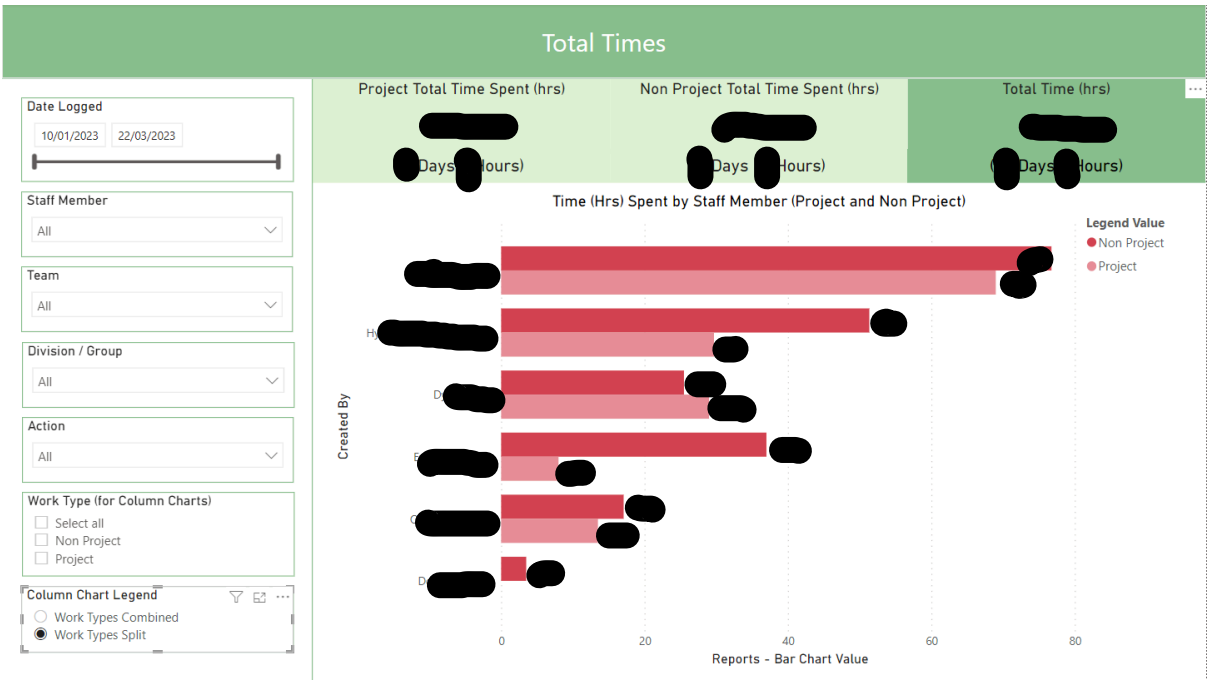
ES2 Figure 4: Result of UNION - the ability to combine project and non-project data into 1 visualisation - Work hours taken up by project and time spent by teams.

Figure 4 shows the time or hours that are taken up by projects that employees are working on as well as the number of hours spent by the team in a comparative view of the project and non-project categories. I created this page using a Pie chart as it is very easy to display data in a visual way but to make it more understandable I used data labels to show the percentage of time taken by each of the profits. I also used a Clustered column chart for the team's data as it would be easy to compare the project and non-project hours and it is easy to interpret. Like the previous pages in the report, I added numerous filters to give the user more options and use cases for this page. This would be useful as the user will easily be able to distinguish what projects are taking the majority of a team's time which can help in project time management.



ES2 Figure 5: Result of UNION - the ability to combine project and non-project data into 1 visualisation - Work hours taken up by the project and time spent by teams in a combined view.

Figure 5 shows the same views as Evidence 4 but displays the number of hours spent by the teams in a combined view to make it easier to find that insight.



ES2 Figure 6: Result of UNION - the ability to combine project and non-project data into 1 visualisation - Hours spent by the team members.

In Figure 6 we can see the total amount of horse spent on project work, non-project work as well as this breakdown of data between the employees. I created this page using some simple card visualisations as they will perfectly display single headline values and their positioning at the top of the page will make them even more easy to see for the stakeholder. I also used a column chart to display the data between the employees as it makes data comparison very easy for the viewer. It is also a very familiar type of visualisation so there won't be any difficulties understanding the visualisation. This would be incredibly useful as it will allow the stakeholder to see the breakdown of how each of their employees is balancing their time, this will allow them to instruct employees to spend more time on project work if there is an upcoming deadline or to spend more time on non-project work.

ES2 Figure 7: Result of UNION - the ability to combine project and non-project data into 1 visualisation - Hours spent by the team members in a combined view.

Figure 7 shows the same views as Evidence 6 but displays the number of hours spent by the staff members in a combined view to make it easier to find that insight.

ES2 - Resilient - viewing obstacles as challenges and learning from failure (B6)

There were two main challenges that arose when modifying/working on this report and this included combining the tables to allow for Project and Non-project to be used together in visualisations, another challenge was trying to improve the pages and the visualisations within them to provide the most utility to the stakeholder.

The biggest challenge was trying to combine the data from the “New project hours” and “New Non Project hours”. This was a challenge because of the way both tables were structured, trying to establish a relationship between these tables would result in a Many to Many relationship which is not the desired cardinality for these tables. I overcame this by doing a lot of research into the subject and recalling some knowledge about modelling. I understood that you could cancel out a Many to Many relationship by creating another table to connect the two other tables, offering two One to Many relationships. However, I wasn't sure about how to approach this. Only after undergoing research and a lot of trial and error did I come up with the solution of using DAX to solve the issue, more specifically using the UNION function to combine the columns from both tables (Can be seen in Figure 8 in Appendix).

The next challenge was improving the pages themselves and the visualisations to maximise the utility of the report. This proved to be a challenge as part of this included the combining of the two non-related tables, however, there was also an aspect of design and functionality which was difficult to improve on. I overcame this challenge by making small additions to the chart like adding data labels which would make the charts much easier to read as it saves the user time from having to hover over the visualisation to see the tooltip, I added a number of filters to provide the functionality of making specific searches and even fixed the old date range filter which the stakeholder could not get to work. Another small detail I added can be seen in Figure 6 and 7 whereby I have used a measure to convert the hours into days and remaining hours, its small quality of life changes like this which allow the stakeholder to gain more utility from the report.

Review

Output, Outcomes, and their Evaluations

The Output of this project is a report in which the stakeholder will be able to view the Project and Non-Project work hours both in a combined view and a separate view. This will help them gain insight on how different teams are balancing their work and this can help the stakeholder identify if enough time is being spent on projects or if there's not enough employee time devoted to Non-Project work. The predicted outcomes of this report is to be able to view the combined and split work hours on projects/non-project work, categorised by Division, action, Team and staff member, all of these predicted outcomes were fulfilled and were designed to provide the stakeholder with more insight on how time is being spent on multiple categorical levels.

Lessons learned | WWW & EBI

The main lesson I have learnt whilst doing this project is that there is a method of combining data from two unrelated tables without initially establishing a relationship and cardinality between them (As we can see in Evidence 1). This is a very useful lesson as there could be many different possible use cases for this in any future reports that I work on.

What went well?

The report is very functional and runs without any error which means it will be able to deliver the insights that I intended. I have made good use of filters which have only built upon the functionality and versatility of this report, ultimately improving on the utility the stakeholder will gain when using the report.

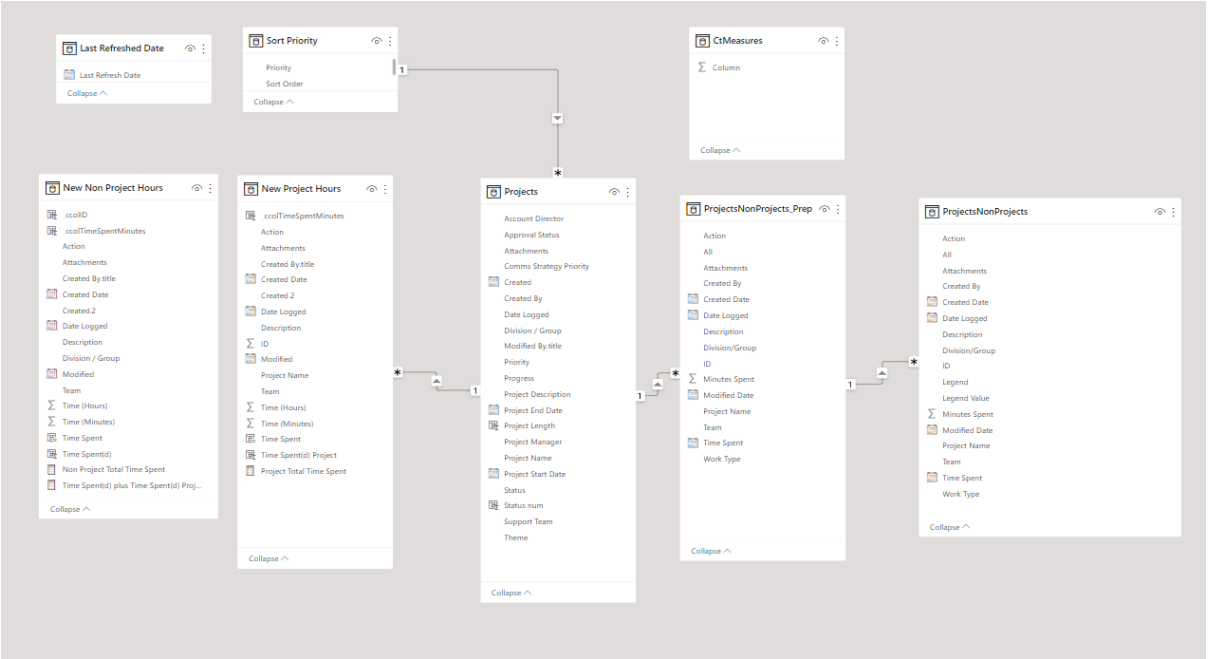
Even better if …

I was able to create visualisations that tell more of a story and provide information that is not present within the underlying dataset. Unfortunately, the dataset provided to build this report was limited in possibility from the beginning but I feel as though there is still room for me to improve and create on this report.

Changes I would make.

If I were to build this report from scratch again, I would like to include a more statistical side to this report as this would provide an even more in-depth insight into the work balance of the teams and divisions however, I fear this may not appeal to the stakeholder who I believe would have preferred more simplistic visualisations that specifically answer their questions.

Appendix



ES2 Figure 8: The new relationship diagram after the tables have been combined and then related.

Above is a screenshot of the modelling view for this report showing the bridging table that i had created to combine the Non-project and Project's data, If you focus on the "ProjectsNonProjects_Prep" table, this is the table which uses the UNION function to combine New Project Hours and New Non Project Hours.

EVIDENCE STATEMENT [3]

NAME Shaheer Hussain

TITLE **[Classified report for GP and patient data]**

PATHWAY **[DATA ANALYST]**

Scenario

Introduction & Business problem

The deputy director of technology and service operations at MHRA approached me to take on a project for an agency client. This project is to remain classified so I will not be able to disclose any information about the context of the project or work I completed but the main purpose of this project was to explore what insights I could provide with the dataset and the limitations of Power BI, however, I can mention that there were some areas of the dataset with a few thousand records and others with millions of records but I will censor my work and display a demonstration of what this project had required. In very simple terms the client wanted me to generate a report using the data that I was provided with and this was more of an exploratory project rather than one with a specific scope. This allowed me to experiment with different areas of data analysis like descriptive and predictive analysis which I have not yet had the chance to investigate yet.

Tasks

My role as a Data analyst means I am responsible for the sourcing of appropriate data, cleaning and transforming this data to be only what is needed, modelling relationships to relate data from different sources and visualising this data to give insights. However, focussing on this project, my role was to explore the data that was provided to me, create as many meaningful visualisations using this data and identify the limitations of Power BI as well as general problem-solving for issues that the dataset would pose. For this project, Power BI and SQL server management studio (SSMS) were the only tools used as the issues were all contained within Power BI and SSMS. Power BI was used to transform the data and visualise it to make it more understandable whereas SSMS was used to experiment with a feature the client wanted to implement. I chose to take an Agile methodology as I will be able to continuously improve the report and gather feedback from the stakeholder to make sure the insights I am providing are meaningful. This will also help me deliver a higher quality report and avoid any risks of developing a report that is not what the stakeholder wants

Activities

ES3 - Describes how they have ensured the true root cause of any problem is found and a solution is identified which prevents recurrence. (B5)

One aspect of this report was to allow the user to automatically “Sign in” to the report as a certain user, this would be done using a DAX function that takes the users email address and makes that into a sign in key. The user would then be able to view their metrics versus the rest of the market’s metrics, However I had identified a problem which was that the client wanted me to implement a security system which would not be possible to create due to DAX complications.

I located the root cause of this problem which was that the user wanted to implement this security system at the reporting layer. In other words they wanted a security system built into the Power BI report itself. After some research (into the topic of row level security and implementing a security system of this sort) and critical thinking I came up with the solution of implementing the security at the database layer. This would be the best solution to shift to as it would not be as exploitable as the security system I was creating in Power BI, it would also be more simple to create but more effective in practice, it would also mean I create rules to allow users to login as different users and only see their data versus the rest of the market. I had explained this concept to the stakeholder or end user and they were happy to proceed with this method.

ES3 - Explains the differences between Structured and Unstructured data (K5)

Differences Between Structured and Unstructured Data

Structured Data :

Definition : Structured data is well organised and formatted for searchability, within databases. This type of data follows a predefined model or schema.

Format : Typically stored in databases like SQL spreadsheets such as Excel and data warehouses.

Examples : Information like names, dates, addresses and numerical data arranged in tables.

Searchability : Simple to search using query languages like SQL.

Storage : Can be stored in tables with rows and columns facilitating querying and analysis.

Usage : Commonly found in systems, CRM systems and enterprise databases.

Unstructured Data :

Definition : Unstructured data lacks a predefined data model or schema. It often consists of text content along with multimedia elements.

Format : Includes documents like Word files, PDFs, emails, social media posts, images, videos and audio files.

Examples : Content such as research papers, medical reports, regulatory documents, social media discussions as multimedia content.

Searchability : More challenging to search and analyse due to its lack of structure : often necessitates tools, like natural language processing (NLP) and machine learning for extracting insights.

Data Storage : Information is kept in NoSQL databases, data lakes and content management systems.

Utilisation : It is applied in content management analysing customer sentiments and various types of data exploration.

Utilisation, within the Medicines and Healthcare products Regulatory Agency (MHRA) :

1. Drug Safety and Monitoring :

Organised Data :

Records of information, drug usage metrics and reports on events are stored in databases.

Purpose : Facilitates efficient monitoring and analysis of drug safety profiles identifies trends in effects. Ensures timely updates to regulatory guidelines.

Unorganized Data :

publications, social media posts, patient reviews and spontaneous reports on events.

Purpose : Utilising NLP and machine learning to scrutinise these sources for emerging safety issues that may not be evident from data

2. Compliance with Regulations and Documentation :

Organised Data :

Repositories containing information on approved medications, clinical trial findings and compliance checklists.

Purpose : Ensures that submissions adhere to standards while maintaining records of compliance status.

Unorganized Data :

Submissions from companies, research papers and regulatory communications.

Purpose : Extracts details through text mining to streamline the review process for evaluations.

3.. Research :

Organised Data :

Clinical trial data results from experiments as well as genetic information databases.In the realm of trials our focus is, on examining the outcomes and utilising data to guide decisions in drug development. We delve into types of data sources such as papers, research notes and logs from experiments.

Text analytics plays a role in extracting insights from scientific literature and historical research data. This process aids in forming hypotheses and fostering innovation within the field.

Moving on to public health surveillance structured data like health statistics, vaccination records and reports on disease incidence are components. These datasets are pivotal in tracking health trends and coordinating responses during health crises with precision.

On the other hand , unstructured data sources such as news reports, public health advisories and social media updates provide real time information for monitoring public health issues. By employing data mining techniques and sentiment analysis we can gauge reactions swiftly to inform policy decisions effectively.

When it comes to marketing authorization and licensing procedures structured data like application forms, product specifications and approval timelines are essential for the review process of medicines and healthcare products.

Unstructured data comprising supporting documents, product dossiers and expert testimonials play a role in decision making processes related to affairs. Leveraging document parsing tools enables us to extract information efficiently to support decisions seamlessly.

ES3 - Describes the fundamentals of data structures and database system design and explains how they are implemented and maintained.(K6)

Data structures consist of methods for arranging and storing data within a computer's memory or storage, aiming to facilitate effective retrieval and manipulation. They play a vital role in addressing intricate challenges and enhancing algorithmic efficiency.Whilst stacks, trees and hash tables were not used and are not commonly used in the MHRA, Arrays were used to store boolean values or even numbers used as flags for certain fields.

Some of the fundamentals of data structures are as follows:

- Arrays - These are sequential memory segments which are designed to contain elements of a uniform data type. Access to these elements is achieved through their respective indices.
- Stacks - This represents a linear arrangement of data where the Last-In-First-Out (LIFO) principle guides their behaviour. This means that elements are appended and removed at a singular extremity referred to as the "top."
- Trees - Trees are like family trees, starting with a main person (this is the root) and connecting to others (which are the children). Binary trees have at most two children per person. Balanced trees like AVL and red-black trees help find information quickly.
- Hash tables - Data structures that use a hash function to map keys to values. This allows for fast retrieval, insertion, and deletion of data.

Database systems are computer programs that handle storing, getting, and changing data. Designing a database system means making plans and structures to manage and use data well.AT MHRA, I make use of many database systems like Oracle sql developer and SSMS to gain access to development, testing or production environments, these databases will have a proper schema in AAS (Azure analysis services) where the data will also be normalised and there will be adequate security.

Some of the fundamentals of database system design include the following:

- Data models: These are representations of data as the relationships between the data, data models can come in different types like a relational model, network model (A network model is a method for structuring and organising data in a database. In this model, data is represented as nodes connected by relationships. Each node can contain information, and relationships define how nodes are related to each other.) or object-oriented model (way of designing and organising data in a database that mirrors the principles of object-oriented programming. In this model, data is stored in the form of objects, which are instances of classes.).
- Schema: A description of the structure of a database, this will often include tables, fields and the relationships between them.
- Normalisation: It involves reducing repetition and enhancing data reliability. This process includes dividing tables into smaller, interconnected ones to eliminate irregularities in data.

- Data integrity and security: Includes the implementation of strategies to uphold data precision, thwart unauthorised entry, and guarantee the confidentiality of data.

ES3 - Explains approaches to combining data from different sources to improve accuracy and / or efficiency and / or maximise benefits to the organisation and / or customer. (K10)

When the Medicines and Healthcare products Regulatory Agency (MHRA) gathers data from sources, it can enhance accuracy, efficiency and overall benefits. Below are some methods to achieve this :

1. Utilising Data Warehouses
Definition :
A repository for storing cleaned and organised data.
Advantages :
Enhanced Accuracy : Centralising data facilitates obtaining outcomes.
Efficiency : Simplifies the process of locating and analysing data swiftly.
Scalability : Capable of managing data from origins.

2. Implementing Data Lakes
Definition :
A storage system that preserves data in its state until required.
Advantages :
Flexibility : Accommodates types of data.
Cost Effective : Economical storage solution for datasets.
Advanced Analytics : Ideal for analyses and machine learning.

3. Leveraging APIs (Application Programming Interfaces)
Definition :
Tools enabling communication and real time data sharing among software applications.
Advantages :
Real Time Information : Ensures up to date data accessibility.
Interoperability : Facilitates collaboration between systems.
Automation : Reduces handling of data requirements.

Data Virtualization :

Definition :
Data virtualization refers to a method of utilising and managing data without relocating it from its source.

Advantages :
Speed : Enables access to data.
Cost Efficient : Reduces the necessity for storage space.
Unified Perspective : Provides a view of all data regardless of their locations.

ES3 - Describes the relevant tools or techniques used for working with the data systems architecture in their organisation. (S9)

Tools	This tool was used to…	Reason for selection

PowerBI	import the data, transform and cleanse the data, create visualisations and a dashboard to provide insights	This is a widely used tool within the agency and it is very intuitive which makes it easy for a beginner to create useful dashboards
Microsoft teams	Communication and meetings	Another heavily used tool in the MHRA but it's very effective when trying to get a message across very quickly which makes communication around the project very easy.

ES3 - Explains the principles of descriptive, predictive and prescriptive analytics and demonstrates how they have been applied within their own data analysis practice. (K14 S11)

Descriptive analytics

Descriptive analytics or analysis can be described as the process of using current and/or historical data to identify trends and relationships. It aims to answer questions like what the sales figures are for the past year or what the most popular products are in a catalogue, the keyword is “what”. Some of the characteristics of descriptive analysis include the summarisation of data to help understand patterns and trends that have occurred, the aggregation of data to generate statistics and the use of visualisation techniques like bar charts and line graphs.

I have used descriptive analysis in both Figure 1 and Figure 2, this is because I am using historical data points to show the relationship between two variables. In these two figures I have used time series analysis which means I am analysing the relationship of two or more variables over a given amount of time. This was used to just answer simple questions that the client may have about the data and to resolve the issue around actually understanding the data. I chose to use descriptive analytics as I was aiming to answer simple questions about the data with these visualisations.

I have demonstrated descriptive analysis in multiple projects at the MHRA, in this project I demonstrated descriptive analysis where I have used basic visualisations in the bar and line charts that I have used to show historical data, this can be seen in Figure X and Figure Y.

Predictive analytics

Predictive analytics or analysis can be described as the process of using current and past data to forecast future trends and events. Rather than addressing the “what” questions like descriptive analysis, this seeks to address the “what might happen” aspect. Some of the key characteristics of predictive analysis is that it uses historical data points to make a prediction about the future, predictive analysis may make use of different algorithms or statistical techniques like linear regression and some models will often provide the likelihood of different outcomes due to the uncertainty of the future, this can actually be seen when making use of a confidence level in a time series analysis diagram.

I have used predictive analysis in both Figure 3 and Figure 4 as I have used a linear regression technique and a parameter to give the user the ability to input a value for one variable and the visualisation will predict the output of the related variable based on the data in the chart shown in these figures. This was used to provide the client with an opportunity to explore the data in a predictive model but also to resolve any potential issues in

decision making as the client can use the model to predict values of a variable and make decisions accordingly. I chose to use predictive analytics here as the client may be interested in what may happen if the value of X variable were to undergo a specific change.

I have demonstrated predictive analysis in this project in Figure 3 and Figure 4 as i have used Linear regression and a parameter to predict the value of one variable based on the the value of the slider to the right of the page, this will allow the user to input one value and receive a prediction for the other variables value.

Prescriptive analytics

Prescriptive analysis can be described as the process of analysing data to determine an optimal course of action, this sort of analysis would be addressing the “what should we do” kind of questions by providing recommendations and actionable insights, this may make use of a combination of predictive models, historical data and optimisation techniques. Some of the key characteristics of prescriptive analysis is that it often includes optimisation techniques to identify the best course of action based on numerous objectives and constraints, it can also operate in real time which allows us to receive recommendations in dynamic situations.

ES3 - Critically evaluates the risks and benefits of predictive analytics (K14 S11) **[Distinction]**

As described earlier Predictive analytics or analysis can be described as the process of using current and past data to forecast future trends and events.

Benefits of predictive analytics include the following:

- Data driven insights leading to improved decision making- Predictive analytics utilises existing data points in order to uncover patterns and trends which is then used to forecast the future, this causes the forecasts to be data driven and can make the forecasts more accurate which can allow for more informed decision making.
- Preventative measures which could save on costs - Predictive analytic can identify potential issues before they occur, (e.g. equipment failure due to lack of maintenance or lack of inventory optimisation in a store) this allows a business to take action to prevent these issues and therefore saving them the costs of dealing with these issues.
- Mitigating risks like fraud - Predictive analytics can potentially be used to identify abnormalities in financial transactions which may indicate fraudulent activity however this would be more prevalent in finance and e-commerce.
- Improving efficiency and productivity with better resource allocation - The ability to predict future events or trends can grant businesses insight into what they should be allocating their resources towards, this can result in more efficient and effective allocation of resources, an example of this could be a retailer optimising their inventory so that they can meet demand when they previously would have failed due to a lack of insight on demand of their product.

Risks of predictive analytics include the following:

- Ethical implications in regards to bias and fairness - Blindly following forecasts created by a predictive model can lead to unfair outcomes as the predictive model may have biassed predictions due to the underlying dataset containing biases.
- Historical biases being present in the model - As mentioned earlier the dataset which a predictive model creates forecasts on may have underlying biases which can then make a reappearance in the outcome of the forecasts and this may result in inaccurate decision making.
- Data privacy and security risks - Predictive analytics would usually be carried out on larger datasets which can make these datasets more vulnerable to security breaches if the data is not stored properly. If any data in the dataset includes personal data then the risk of cyberattacks will also raise concern for data privacy issues.

- Unreliable forecasts - If the underlying dataset used for prediction is incomplete or contains missing / erroneous data then this may be reflected in the forecasted outcome which can then lead to inaccurate decision making.
- Over automation issues - Over relying on the insights provided by predictive models with little review by actual employees can result in decisions being made which may not be entirely appropriate or optimal for the business.

A scenario where this could be useful is regulatory decision making in drug approvals, the benefits of predictive analysis would play a significant role as we could have early detection of safety signals or potential adverse drug reactions that may not be immediately apparent, it can also accelerate decision making by predicting the safety profiles of different drug which are candidates for approval, These benefits can help the MHRA save a lot of money as not as much time is spent on decision making and safety signals can allow us to address situations before they escalate which makes predictive analysis more of a preventative measure,However, the risks can also play a significant role as there could be algorithm bias which can result in the predictive model making decisions that disproportionately favour certain demographic groups also there is always the risk that we become overly reliant on these predictive models which is neglective the importance of human expertise.

If MHRA heavily relies on predictive analysis for drug approval (as an example), The predictive model may be trained on historical data from clinical trials, real-world patient outcomes, and other relevant sources. However, a bias is inadvertently introduced into the model due to underrepresentation of certain demographic groups in the training data.

Concerns with biases :

The demographic bias :

The training data might not be representative of an entire population which will lead to less diversity in the data, If the model used at MHRA is predominantly trained on data from a specific medicine group, it may not generalise well to other groups, resulting in biassed predictions.

The data collection bias :

Historical data used for training the model may be biassed due to systematic errors or discriminatory practices in data collection. For example, if certain symptoms or side effects are more frequently reported in one medicine group than another due to healthcare disparities, the model may learn and perpetuate these biases.

Any fairness concerns with predictive analytics :

Impact on Underrepresented Groups:

If the predictive model is biassed against certain medicine groups, it may disproportionately impact the approval rates of drugs for those groups. This can lead to unequal access to medications and treatments which can increase the challenge for people to gain essential medical intervention.

MHRA being Transparent :

Lack of transparency in the predictive analytics process can make it challenging to identify and rectify biases. If the MHRA is not transparent about the data sources, variables, and methodologies used in the predictive model, it becomes difficult for external stakeholders to assess and challenge decisions.

The ethical implications of predictive analytics:

The patient trust and public perception (very important to MHRA):

Biassed predictions and unfair outcomes can break public trust in the regulatory process.This distrust can have broader implications for public health and the credibility of the MHRA.

access to medication :

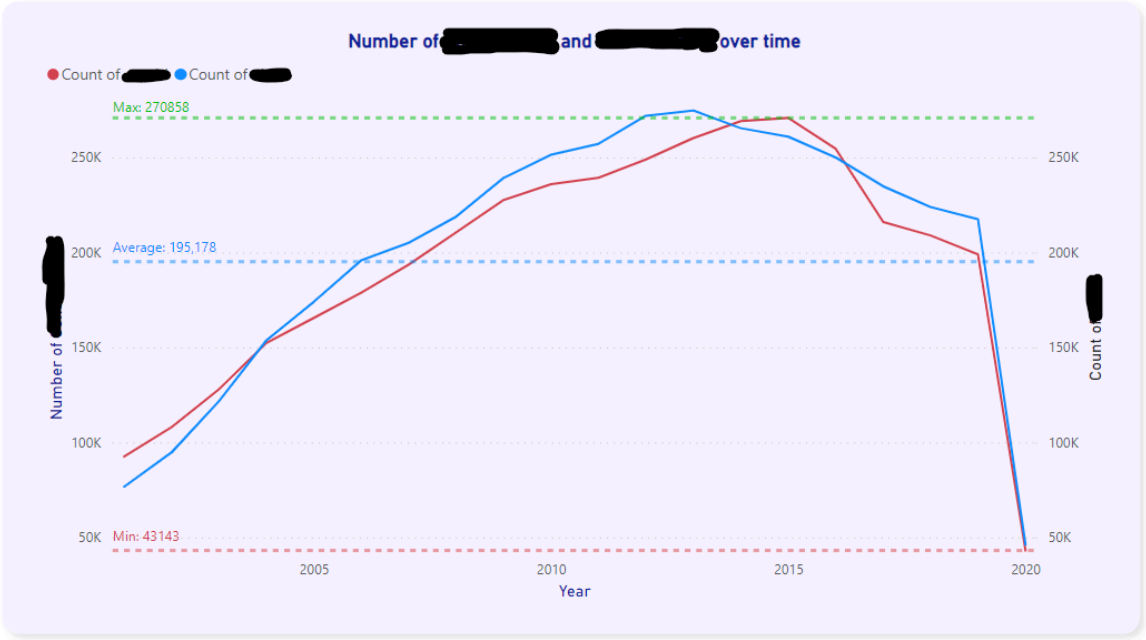
Biassed models can result in medications that are more effective or safer for certain groups being disproportionately rejected. This can lead to inequitable access to treatments, impacting the health outcomes of specific populations and perpetuating health disparities.

In conclusion, Predictive analytics can be an incredible tool to use for resource allocation and improved decision making. However, it does show some flaws which can point to inaccurate forecasts. This means that predictive analytics should not be the only type of data analysis used for decision making as it may result in many issues like automation bias (described as just accepting every insight from the predictive models and making decisions upon this).

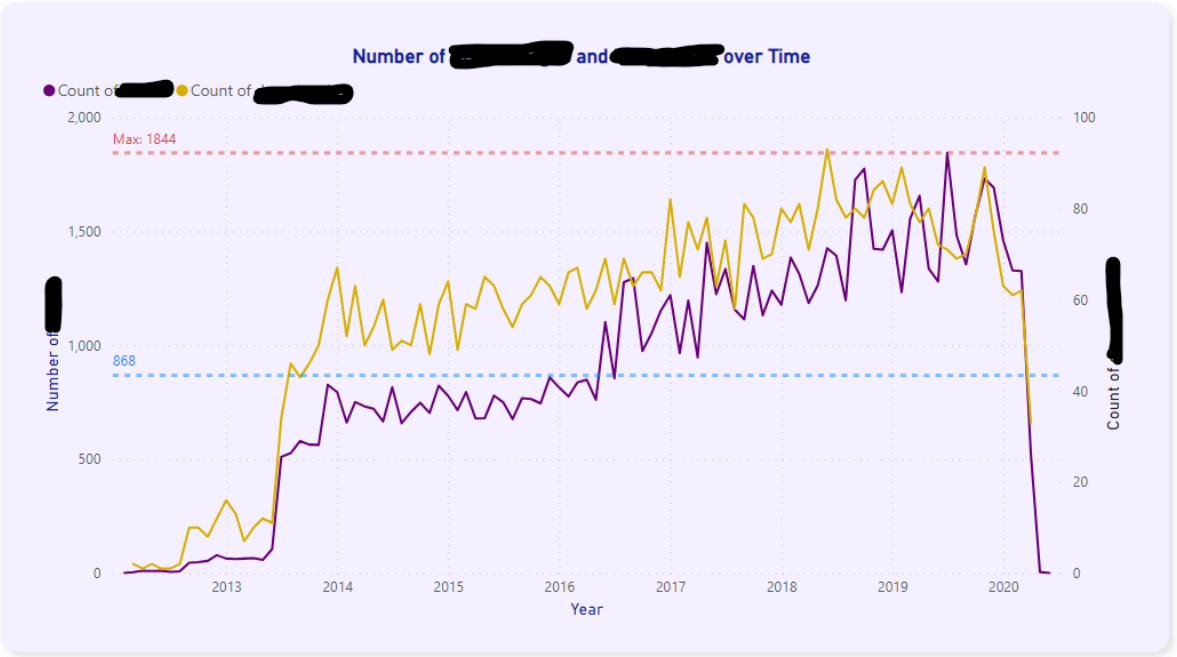
ES3 - Explains how they have applied analytical techniques for data mining and time series forecasting and other modelling techniques (S13)

Time series analysis involves the examination of patterns within data that is recorded sequentially over time. In other words it's a way of studying the characteristics of the response variable with respect to time as the independent variable. This data is gathered at regular intervals, forming a time-ordered series of observations. In simple terms time series analysis is a technique which uses historical and current data points to make predictions over a period of time. The time series analysis technique I used for this project allowed me to view the trend of variables and the correlation between them which will be very important to the business user.

This report that I worked on includes highly classified information which I will not disclose however, I applied time series forecasting by taking highly important variables and using line charts to compare them over a specific period of time, this would give the end user the insight of how these variable have changed over time and if there is a correlation between them. This can be demonstrated by Figure 1 and Figure 2. The difference between these two applications is that Figure 1 is just broken down by the “Year” metric whereas the chart in Figure 2 is broken down by the “Year” and the “Month” metric which gives us a more in-depth look at how the data points have varied over time.



ES3 Figure 1: Time series analysis technique by Year



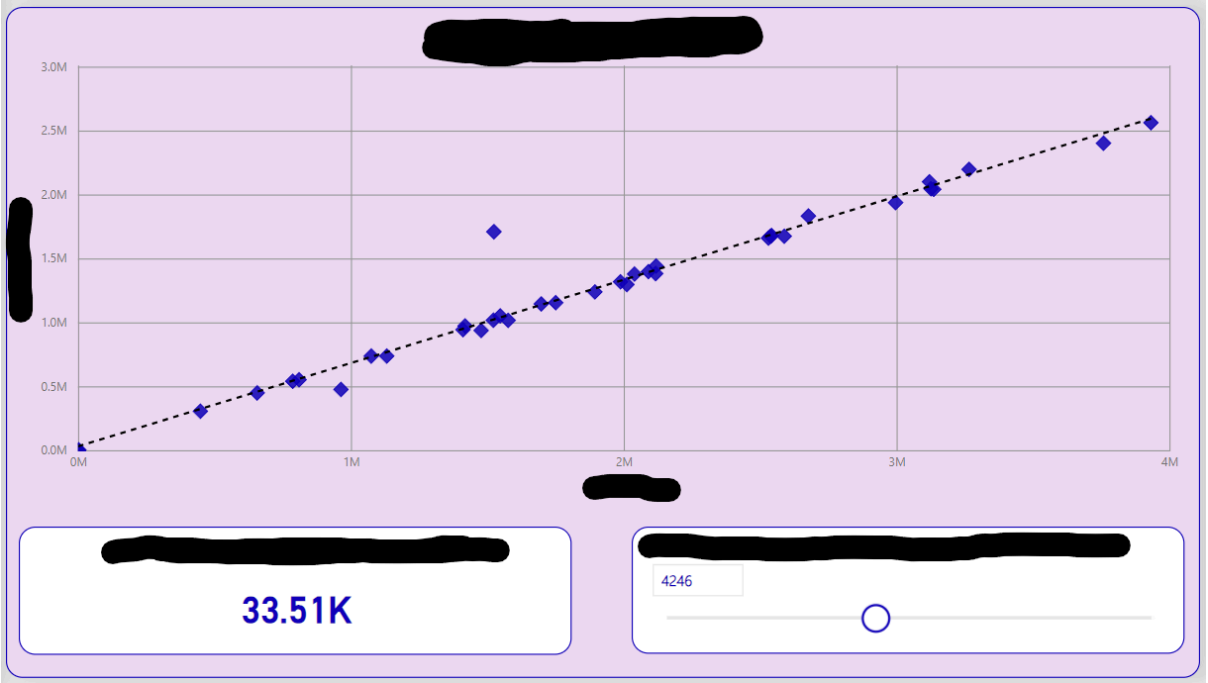
ES3 Figure 2: Time series analysis technique by Year

Linear regression is a statistical technique employed to establish a connection between a dependent variable and one or more independent variables by fitting a linear equation to the available data. The main goal of linear regression is to identify the optimal straight line, known as the line of best fit or regression line. This effectively describes the variability in the dependent variable in relation to changes in the independent variables.

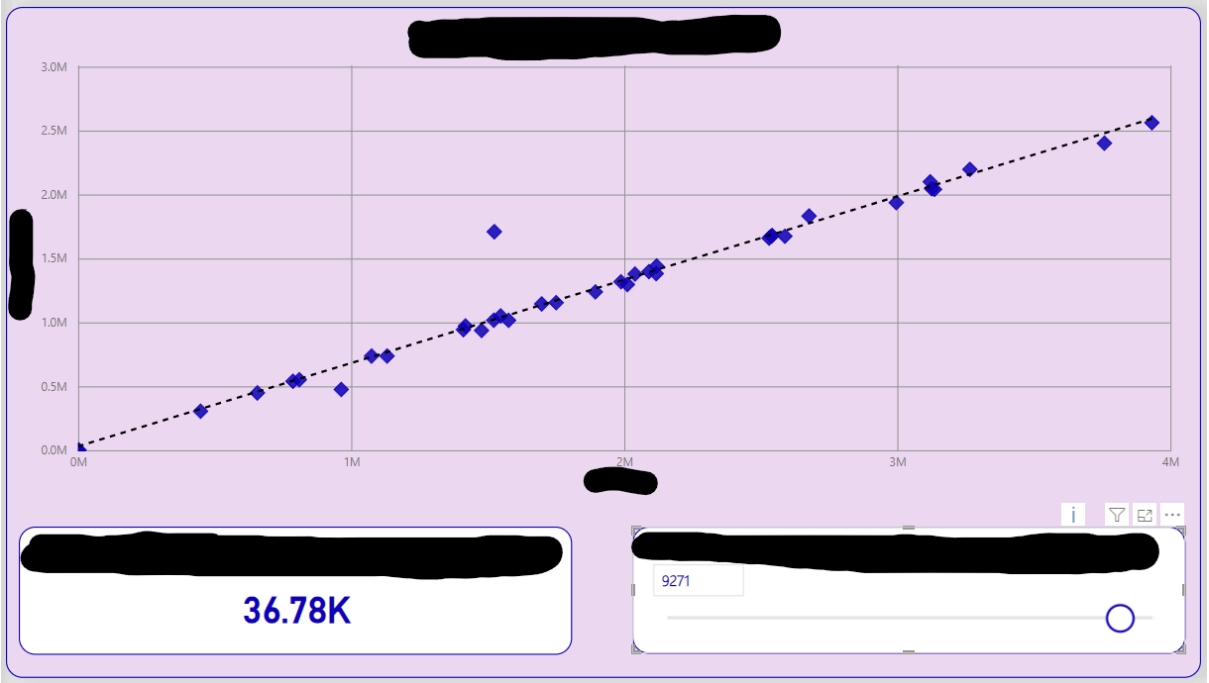
I used Linear regression to allow the end user to make predictions for one variable based upon changes to another variable, as you can see in Figure 3 and Figure 4 there is a slider which the user can manipulate and

this will affect the prediction or outcome on the left card. This uses data points from the graph above to predict the relationship between the two variables, in Figure 5 you can see the DAX formula I had used to generate the prediction card on the bottom left, this uses in built DAX functions like LINESTX (This is a function which calculates a straight line that best fits the given data) to do the majority of the work, the LINEST function itself just generates a bunch of useful values with the given data like slope, intercept, degrees of freedom etc however our variable “_slope” and “_Intercept" seek to just select the slope and intercept out of the bunch of values as that is all we will need for this linear regression.LINEST and LINESTX are two DAX functions that calculate a linear regression by using the Least Squares method. Both functions return multiple values, represented in a table that has a single row and one column for each of the values returned.

In Figure 3 and 4 we can see a graph with two small panels at the bottom of the screen, the big graph in the middle of the page is the data which we are using for linear regression, we have to choose two variables which have a relationship or correlation in order to do this, the panel on the bottom right is a slider which allows the user to have an input value for one of the variables, using the data in the graph above the report will predict the future relationship between the two variables to predict a value or output based on the users input, this predicted value is then outputted on the small panel in the bottom left corner of the pages.This application of linear regression does not solve a immediate problem, rather it seeks to allow the end user to experiment with the predictive model and see what values they may get.



ES3 Figure 3: Linear regression analysis technique



ES3 Figure 4: Linear regression analysis technique (2)

```
1 [redacted] =
2 VAR _linest = LINESTX[redacted]
3 var _Slope = SELECTCOLUMNS(_linest, [Slope1])
4 VAR _intercept = SELECTCOLUMNS(_linest, [intercept])
5 RETURN
6 _intercept + _Slope * [redacted]
```

ES3 Figure 5: Linear regression DAX formula

ES3 - Identifies areas of work where they adapted to changing contexts within the scope of a project, direction of the organisation or Data Analyst role (B7).

At a certain point during this project the stakeholder had asked me to alias a table in the dataset and use it as a control table (this just means that I will duplicate a table and use this for testing). This was done so that I could work on a security system that the deputy director had asked for. In simple terms, this security system would allow the user to “Log in” as a user and view their metrics against the rest of the market. The security system would hide the “rest of market” metrics broken down by the rest of the users in the market. However, after days of development and testing I came to the conclusion that a security system like this should not be implemented at the reporting layer but rather it should be implemented at the database level. To adapt to this I had to change the scope of the project to utilise a database management to implement the security system in question, instead of continuing to develop the security system on Power BI, I opted to load all of the data into SQL server management studio and use a method called row level security to only allow a user to have access to their own metrics.

Review

Output, Outcomes, and their Evaluations

The Output of this project is a report in which the stakeholder will be able to view the meaningful insights which will enable them to understand their data more. The stakeholder will be able to understand the relationships between certain variables when looking at these insights, this can be incredibly important as the dataset includes GP and patient data, so there will be a lot of important trends to consider. The data used for this project was incredibly detailed, for context some tables had millions of rows of data so this report outcome is a much clearer understanding of the very granular dataset.

Lessons learned | WWW & EBI

The main lesson that I learned with this project is that we can use time series analysis to compare multiple variables to see how they have changed over time and to identify any trends between these variables. This will enable the stakeholder to make more informed decisions as they will further understand the relationships between the variables or data. Taking an agile approach worked well as I was able to take on more and more feedback from meeting with the stakeholder and showing off my progression.

What went well?

The report delivers very clear insights and has an intuitive design. It makes use of buttons and bookmarks to allow users to switch between views on a single page within the report, This allows the user to view more insights without further complicating navigation between these insights. The stakeholder was happy with the high level insights provided which he was easily able to pick out stories within the data and this was the main objective of the project.

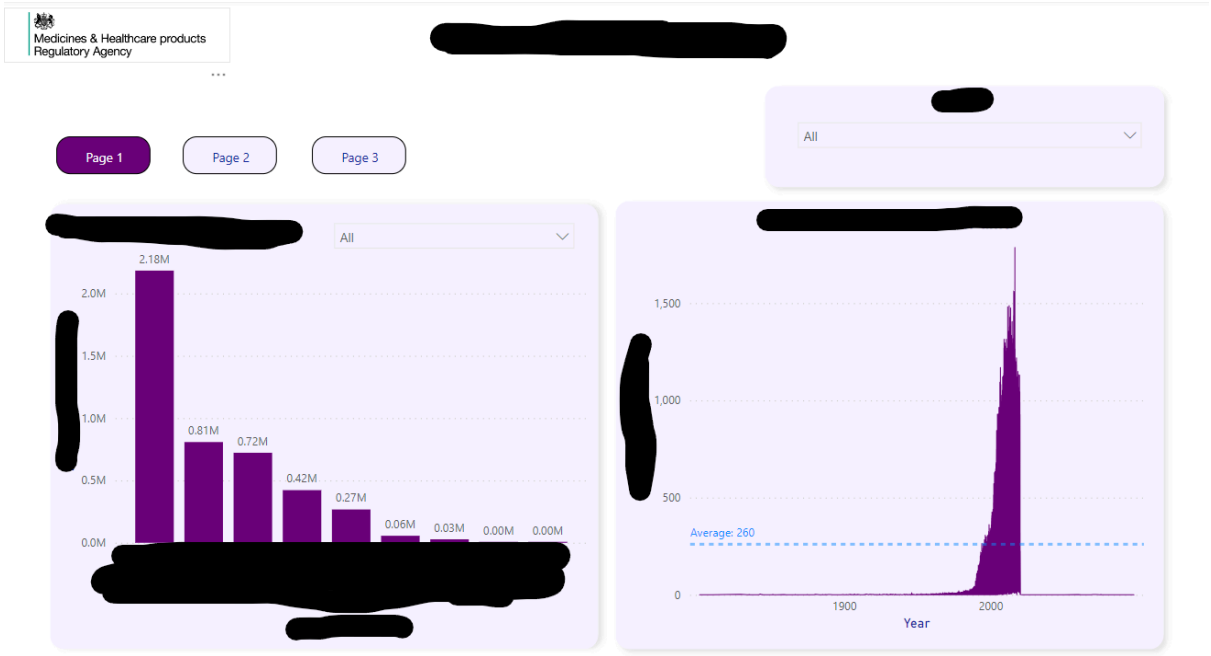
Even better if …

There was a clear vision of the project and what exactly the stakeholder wanted to understand about the data, this would have given me clear goals to work towards and achieve by a timeline. A lack of clear vision leaves me in a position of thinking about what insights would be useful to the stakeholder whereas if I had been given a list of visualisations that would be needed then I could get to developing much sooner.

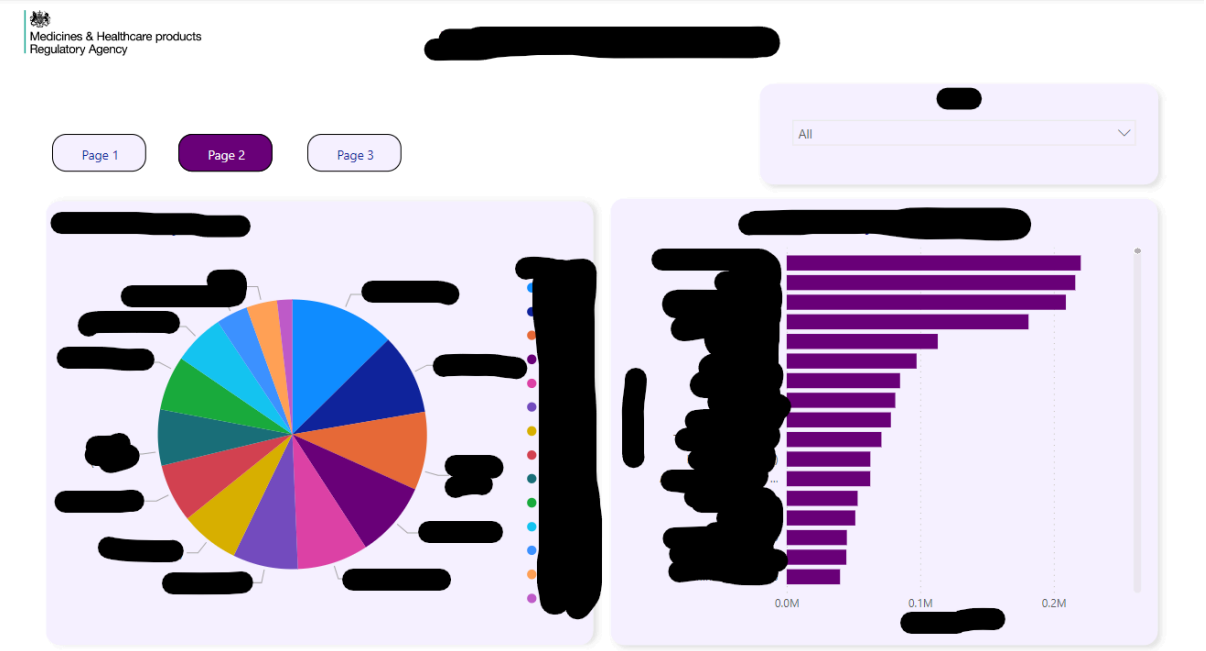
Changes I would make.

If I were to build this report from scratch again, I would like to include a more statistical side to this report by making use of some Python within the Power BI application as this could aid me in delivering very meaningful insights especially when considering the size of the dataset.

Appendix



ES3 Figure x: Ability to switch between views on one page



ES3 Figure y: Ability to switch between views on one page (2)

EVIDENCE STATEMENT [4]

NAME	Shaheer Hussain
TITLE	EVIDENCE STATEMENT **[4]** - **[Analysis for New GP dataset]**
PATHWAY	**[DATA ANALYST]**

Scenario

Introduction & Business Problem:

The deputy director of technology and service operations at MHRA approached me to take on a project for an agency client who is an observational and interventional research organisation. This Project is quite similar to my work shown in Evidence Statement 3 as I worked with the same kind of GP dataset however this project will be unique as I have been tasked with taking a more statistical approach, not only that but my method of providing visualisations is much different in this project as I make use of Python. The purpose of this project is to explore further the GP data from a more statistical stance and the goal of this is to explore what kind of statistics we can draw from the data using the tools we have, by completing this kind of project/exercise we are creating the opportunity to assess our analytical skills but also identifying anywhere where we may have skill gaps, this will allow us to address these gaps in the future, this naturally makes this project quite important as it will allow apprentices like me to gradually improve my skill set to provide more analysis in the future.. I proposed that I had some experience with statistics in Python due to a statistics module I completed with Ada College in my first year and we agreed upon this being the route we should choose considering we had looked at other options like power BI already, another reason that I had chosen Python is simply because it is a very powerful and versatile programming language so it will be more than capable for some simple statistics visualisations. Again this project included files with thousands or tens of thousands of rows and it was more of an exploratory project rather than a project with a single focus.

Tasks

My role as a Data analyst means I am responsible for the sourcing of appropriate data (In this project I was provided with a couple of CSV files which had a standard layout with no unique characteristics), cleaning and transforming this data to be only what is needed, modelling relationships to relate data from different sources and visualising this data to give insights. However, focussing on this project, my role was to explore the data that was provided to me and create as many meaningful statistical visualisations like distribution and bar graphs using Python to identify what information or story I could generate using the data, I used Python libraries such as Pandas to read files such as csv and matplotlib to create visualisations. I did not conduct any in-depth preliminary analysis however when working with Python and the Pandas library, I make use of the ".info()" function which displays all of the columns in a dataset as well as other important information like the data type of these columns (Pandas is a Python library used for working with datasets. It allows us to analyse big data and make conclusions based on statistical theories. Pandas can also clean messy data sets, and make them readable and relevant). For this project, I chose to use Python as I have had some experience with this language in a statistics context and it is a very powerful general programming language. I chose to take an Agile methodology as this will result in a higher quality product that will provide higher satisfaction to the end-user or stakeholder however, there are limitations to an agile methodology approach like the fact that it takes more time to develop an end product and it can easily get sidetracked as more contact with stakeholders can increase the chances of stakeholders bringing in additional requirements which will affect the development of the original project. This will also help me deliver higher quality insight and avoid any risks of developing insights that are not what the stakeholder is asking for.

Table X: Key Stakeholders

Name/Organisation/Team e.g. Finance team	Role in the project e.g. Budget oversight	Specific interest in project e.g. ensuring project is on budget
Shaheer Hussain	Data analyst	creating insightful visualisation
Deputy director	Overseeing progress	ensuring project is on track
Client	communication	ensuring they get what they want out of project

Table X: Risks & Constraints

Type of Risk / Constraint e.g. Limited budget	Impact or affect on project e.g. Budget cannot be exceeded	Risk reduction / avoidance e.g. Regular checks on project spending

Constraint - Time	There may not be enough time to fully explore the capabilities of visualising with python	Include more colleagues within the project to fast track the exploration process
Constraint - Skill	I may not have the experience to create the most complex but insightful visualisations	Upskilling so that i am able to deliver more insightful visualisations

ES4 - Describes the relevant tools or techniques used for working with the data systems architecture in their organisation. (S9)

Table X: Tools

Tools	This tool was used to…	Reason for selection
Python	create visualisations	This is an incredibly powerful programming language which has a lot of support online with access to a multitude of libraries.

Activities

ES4 - (K1 B1) Explains how current, relevant legislation impacts on the safe use of data and how their role contributes to a productive, safe, and secure working environment.(K1 B1)

In the context of data analysis within the MHRA, several pieces of legislation impact the safe use of data. Some of these may include:

the General Data Protection Regulation which is a data protection regulation that applies to the processing of personal data in the EU, this includes the UK. The regulation establishes principles for lawful and transparent processing of personal data and places obligations on organisations that handle such data. As a data analyst at MHRA, I need to ensure that I am complying with GDPR by handling data responsibly, respecting individuals' privacy rights, and implementing security measures to protect the data.

Pharmacovigilance Legislation: Given the nature of the MHRA's responsibilities in monitoring the safety of medicines, there may be specific pharmacovigilance regulations that impact the collection, analysis, and reporting of adverse drug reactions and safety data. My role as a data analyst may involve contributing to these processes while adhering to relevant legislation. I contribute to the Pharmacovigilance process by creating medicines reports using both Power BI and Microstrategy which is a BI tool that is similar to Power BI.The data used for Pharmacovigilance is typically stored in Azure analysis services or other in-house systems like HALO or Sentinel. With Power BI we used to point the report to the correct AAS environment(there is a development, testing and production environment) and directly report off of this data however we have now moved these reports to Microstrategy.

Cybersecurity and Information Security Regulations: Ensuring the security of data is crucial in any regulatory environment. My role as a data analyst may include implementing and following cybersecurity and information security best practices to safeguard sensitive information from unauthorised access or breaches.Some examples of measures I take to follow these regulations include Using data purely for lawful purposes, not granting access to data or reports to people who should not have access to these materials and even making sure I do not store any data on removable storage devices to eliminate the possibility of losing these devices and causing a security breach.

To contribute to a productive, safe, and secure working environment within the MHRA, my responsibilities as a data analyst may include data governance which would be Implementing and adhering to data governance policies to ensure the quality, integrity, and confidentiality of data, an example of this would be the Data protection policy in the MHRA, this holds policies like “3.1 Lawful bases for processing” which states “Before processing personal data, the Agency will establish at least one lawful basis for processing from GDPR Article 6(1) ” and policies such as “3.2 Special category personal data conditions of processing” which states “We can only process special category personal data lawfully if we can also satisfy at least one condition from GDPR Article 9(2).”.Another one of my important responsibilities would be data security: this could be fulfilled by employing encryption, access controls, and other security measures to protect data from unauthorised access or breaches.

Compliance Monitoring is also another important responsibility: Regularly monitoring and ensuring compliance with relevant data protection and regulatory requirements, this is incredibly important because if we stop monitoring reports for compliance and data protection it could be the case that there are new vulnerabilities that appear in the future, this could be caused by several things like a change to the structure of the data, in any case, we need to continuously monitor to make sure we are complying with sat protection regulations so potential vulnerabilities do not become a threat or are not taken advantage off.Ethical Considerations should also be considered by adhering to ethical standards in data analysis, especially when dealing with sensitive health-related information. For example, when working as a data analyst for the MHRA, I adhere to the civil service code, which includes the following core values:

- ‘integrity’ is putting the obligations of public service above your own personal interests
- ‘honesty’ is being truthful and open
- ‘objectivity’ is basing your advice and decisions on rigorous analysis of the evidence
- ‘impartiality’ is acting solely according to the merits of the case and serving equally well governments of different political persuasions

Each one of these core values has standards of behaviours, for example with “Honesty”, it is expected that I must “use resources only for the authorised public purposes for which they are provided”(Civil Service, 2015) and must not “be influenced by improper pressures from others or the prospect of personal gain”(Civil Service, 2015)

Finally Collaborating with other professionals would ensure that data practices align with regulatory requirements.There are many examples of this in any project I get involved in as I take a more agile approach to work which means I'm in constant communication with stakeholders to discuss current progress on a project, improvements or corrections.A specific example in this project would be

communicating with the deputy director of technology about updates about implementations in the project.

ES4 - Describes how they have appropriately adapted their activities to meet minor, unexpected changes at work. (B2)

At the Medicines and Healthcare products Regulatory Agency (MHRA) my focus has mainly been on using PowerBI for analysis. This method has proven to be very effective for our reporting and decision making processes, allowing us to present metrics and trends in a user friendly way.Recently the deputy director of technology and service operations approached me with a project for one of our agency clients. This project required an approach to data analysis, which was different from my usual descriptive methods. Moreover the visualisations for this project needed to be created using Python rather than PowerBI so naturally it was much more difficult for me as I find PowerBI more intuitive and easy for creating visualisations.

To meet the stakeholders requirements effectively I took the following steps :

Skill Enhancement : I immediately started enhancing my skills in statistical analysis by revisiting concepts like hypothesis testing, regression analysis and predictive modelling. It was essential to ensure that I could apply the techniques to analyse the clients data effectively.

Python Proficiency : Although I was already familiar with Python I delved deeper into using libraries such as pandas, numpy, matplotlib and seaborn for manipulating and visualising data. Additionally I explored libraries, like scikit learn to support statistical analysis.

Tools : In order to shift from using PowerBI to Python for creating visualisations I crafted a Python script to delve into the dataset and developed visual representations based on that data.

Progress : I took an approach to my development process. By crafting versions of models and visuals and seeking input from the deputy director and the client I continuously refined my work to meet their expectations.

Clear Communication : To simplify insights for the client I prioritised clear and succinct communication. I produced reports that outlined the methodology, findings and implications in a to understand manner.

ES4 - (K7) Explains the principles of user experience and domain context for data analytics.

User experience (UX) and domain context are crucial considerations in the field of data analytics, as they significantly impact the effectiveness and usability of analytics solutions. Let's explore the principles of user experience and domain context in the context of data analytics:

User Experience (UX) Principles for Data Analytics:

User-Centred Design (UCD):
UCD involves designing products or systems with the end user in mind. In data analytics, this means understanding the needs, preferences, and goals of the users who will be interacting with the analytics tools.An example of this at MHRA would be understanding the needs of business users when creating the Safety Connect medicines reports.

Usability:
Analytics tools should be easy to navigate and use effectively. Users should be able to perform tasks intuitively, without unnecessary complexity as this will make it harder for the users to adapt and learn to a tool. This includes providing clear navigation, simple interfaces, and straightforward interactions.For example with any of the Power BI reports I created and logged in my previous evidence statements (e.g. the Communications Dashboard in Evidence statement 2) I have created a very simple interface so it is easy to understand for the end users and the interactions you can have with the report is straightforward and apparent so that the user knows exactly what they can do with the report as soon as they start using it. If I use the communications dashboard as an example, we can see that the user is able to filter the visualisations on date, legend, staff member, division and team.

Accessibility:
It's incredibly important to ensure that analytics tools are accessible to users with diverse abilities. This involves considering factors such as colour contrast, text size, and compatibility with assistive technologies. This would allow more people in the MHRA to make use of analytics tools.

Feedback and Guidance:
Providing real-time feedback to users as they interact with data is quite crucial as it will aid with upskilling users and improving their skills, this could include tooltips, error messages, or visual cues that help users understand the implications of their actions and guide them toward effective analysis.An example of this could be seen in my first evidence statement where I created the DIGG Training app report, I had used a lot of tooltips in that report to provide the end user with a lot more information that would help with usability.

Responsive Design:
Include interfaces that adapt to different devices and screen sizes as responsive design ensures that users can access and interact with data analytics tools seamlessly across various platforms, this also reduces any probability of users on certain platforms having issues with tools.

Consistency:
Maintaining consistency in the design elements, terminology, and interactions throughout the analytics platform. Consistency enhances predictability and reduces cognitive load for users.

Performance Optimization:
Optimising the performance of analytics tools to ensure quick response times and efficient data processing. Slow and unresponsive tools can frustrate users and hinder their analytical processes.

Domain Context Principles for Data Analytics:

Understanding User Roles:
It is important to identify and understand the different roles within the domain that will interact with the analytics platform. This could include analysts, data scientists and other stakeholders. Understanding the different roles allows you to tailor the user experience to meet the specific needs of each role.

Domain-Specific Language:
Using terminology and language that are familiar to users in the specific domain. This will reduce the learning curve and make the analytics tools more accessible to users who may not have a technical background. In the MHRA the business users could be a perfect example of this.

Contextual Relevance:
Presenting data in a contextually relevant manner. Understanding the business or industry context in which the analytics will be applied and ensure that the insights derived align with the goals and challenges of that context.

Integration with Workflows:
Integrate analytics tools seamlessly into existing workflows. This helps users incorporate data-driven insights into their daily tasks without disrupting their established processes.In the MHRA we have a number of reports

created in Power BI and Microstrategy, these reports allow end users to efficiently access data driven insights and easy to understand information which helps in the final decision making, with our new implementation of Microstrategy, I alongside the Data Science team are testing whether we can hand Microstrategy as a self service tool to end users where they will be able to create reports that they need by themselves, we are doing this by creating reports that they need now and observing whether the tool meets the requirements of the end users.

Customization and Flexibility:
Allow users to customise their analytics dashboards and reports to suit their specific needs. Providing flexibility in visualisation and analysis options accommodates diverse requirements within the domain.

Security and Compliance:
Ensuring that data analytics solutions comply with security and regulatory requirements within the domain. This includes safeguarding sensitive information and maintaining data privacy standards.

Collaboration Support:
Facilitating collaboration among users by incorporating features that allow for easy sharing of insights, collaborative analysis, and communication within the analytics platform.

ES4 - Describes impact on user experience and domain context on data analysis (S5)

Impact on User Experience in Data Analysis

Comprehensibility:
Impact: Good UX design ensures that data visualisations are easy to understand. Clear, well-designed visuals help users quickly grasp key insights without confusion.
Example: Using intuitive bar charts with clear labels and a consistent colour scheme in Power BI makes it easier for MHRA employees to understand the visualisation.

Efficiency:
Impact: Efficient UX allows users to find and interpret relevant data quickly, which is crucial in a regulatory environment where timely decisions can have significant consequences.
Example: Interactive elements like filters and drill-down options in Power BI enable users to swiftly navigate to the specific data they need, improving decision-making speed However I find this more difficult to account for in Python since I am producing visualisations themselves and not a dashboard.

Engagement:
Impact: A well-designed user experience keeps users engaged and encourages deeper exploration of data, leading to more thorough analysis.
Example: Interactive dashboards in Power BI allow users to explore the data in ways they may not have been able to do before, they can filter in various ways, allowing them to tap into different layers of data granularity.

Accessibility:
Impact: Ensuring accessibility in UX design means that all users, regardless of their abilities or technical expertise, can effectively interact with and understand the data.
Example: Using accessible colour schemes and ensuring that visualisations are readable on various devices ensures that all MHRA staff can benefit from the data analysis.

Accuracy and Compliance:

Impact: In a regulatory environment like the MHRA, accuracy and compliance are critical. The domain context dictates stringent data validation and reporting standards.
Example: Ensuring that work hour data is accurate and compliant with MHRA's data protection policies is essential for maintaining trust and regulatory adherence.

Decision-Making Support:
Impact: Data analysis tailored to the domain context supports informed decision-making by providing insights that are specific to the operational needs of the organisation.
Example: Visualisations showing imbalances in work hours can inform management decisions on resource reallocation or process improvements within the MHRA.

User-Centric Design: I start by understanding the specific needs and preferences of my users, conducting user interviews or surveys to gather insights.

Data Accuracy and Validation: I ensure that my data sources are reliable and that my Power BI reports or python scripts include validation checks to maintain data integrity.

Customization and Personalization:I tailor my visualisations to meet the needs of different user groups within the MHRA.I use Power BI's customization features to create relevant and focused reports.

Continuous Feedback and Improvement:I regularly gather feedback from users and iterate on my designs to continuously improve the user experience and relevance of my data analysis.

ES4 - (K13 S10) Explains and applies the principles of statistics for analysing datasets

Statistical analysis is the process of collecting and analysing large volumes of data to identify trends and develop valuable insights. In the professional world, statistical analysts take raw data and find correlations between variables to reveal patterns and trends to relevant stakeholders. Here are some fundamental principles of statistics for analysing datasets:

Descriptive Statistics:

- Mean - This is the sum of all values divided by the number of observations.
- Mode - This is the most frequently occurring value in a dataset.
- Median: This is the middle value in a dataset when it is sorted.

Probability:

- Probability Distribution which describes the likelihood of different outcomes in a sample space.
- Random Variables - Variables that can take on different values with certain probabilities. (There can be discrete random variables where a variable can take on a countable number of distinct values and continuous random variables where a variable can take on any value within its range and its possible values form a continuous interval)

Hypothesis Testing:

- Null Hypothesis (H0) - A statement of no effect or no difference.
- Alternative Hypothesis (H1) - The statement that researchers are testing.
- Significance Level (Alpha) - This is the threshold used to determine statistical significance.

Statistical Inference:

- Confidence Intervals - A range of values used to estimate the true value of a population parameter.
- P-Value - The probability of obtaining results as extreme as the observed results, assuming the null hypothesis is true.

Regression Analysis:

- Linear Regression - Modelling the relationship between a dependent variable and one or more independent variables. (A form of statistics I have shown in Evidence statement 3)
- Correlation: Measures the strength and direction of a linear relationship between two variables.

Statistical Distributions:

- Normal Distribution - A symmetrical bell-shaped distribution often seen in nature.
- Skewness and Kurtosis - Measures of the shape of a distribution.

Statistical Software:

- Use of statistical software (e.g., R, Python with libraries like NumPy, SciPy, and Pandas, or dedicated statistical packages) to perform analyses efficiently.

Data Visualisation:

- Graphical representation of data through charts, graphs, and plots to better understand patterns and trends. Visualising data is crucial for gaining insights, communicating findings, and making informed decisions. Effective data visualisation enhances the interpretability of complex datasets, making it a valuable tool in the fields of statistics, data analysis, and decision-making.

At the MHRA, I have demonstrated regression analysis, which was displayed in evidence statement 3, this was done to give the end user a chance to predict one value based upon the value of another variable, this model would use historical data points making it more data-driven. I have also displayed statistical distributions and data visualisation in both Figure 1 and Figure 2. In both Figures, I have used Python as it is a very capable and versatile programming language. On top of this, I have had a little experience using Python for statistics due to my Statistics module at Ada College. In Figure 1, I have shown the weight distribution of patients in the dataset, from this we can infer that there is a normal distribution due to the bell curve which means the data near the mean are more frequent in occurrence than data far from the mean. In Figure 2, I am showing the number of male and female patients in a small part of the data, from this visualisation we can infer that there were more males than females in this small subset of data.In this project, there were no clear requirements in terms of providing specific statistical visualisations as this was more like an exercise to see what statistical visualisations i could generate using the datasets that were provided.

ES4 - Explains the principles of descriptive, predictive and prescriptive analytics and demonstrates how they have been applied within their own data analysis practice. (K14 S11)

Descriptive Analytics :

Descriptive analytics involves looking at data to understand what happened previously. This type of analytics focuses on combining data and using mining techniques to gain insights into performance.

Potential applications in MHRA :

- Adverse Event Reporting : By gathering and analysing reports of drug reactions MHRA can spot patterns and trends helping to pinpoint drugs or devices that need investigation.
- Drug Approval Processes : Examining data from drug trials to understand approval timelines, common challenges faced and areas for improvement in the approval process.
- Compliance Monitoring : Descriptive analytics can be used to track compliance rates among manufacturers and identify compliance issues.

Example :

The MHRA could utilise analytics to study data from the Yellow Card Scheme, which collects details on suspected drug reactions. By summarising this information they can spot trends like the reported side effects, which medications have the highest number of reports and any demographic patterns, in the reports.

I have demonstrated descriptive analytics in all of my evidence statements which includes this statement, this is because I am making use of historical data to create a summary or overview and understand what has already happened and this can be seen in Figure 1, Figure 2, Figure 3, Figure 4 and Figure 6.

Predictive Analytics :

Predictive analytics involves using models and machine learning methods to foresee events based on past data. Its goal is to anticipate outcomes.

Potential applications in MHRA :

- Risk Forecasting : Anticipating which medications or medical devices are more prone to causing reactions by analysing data and other relevant factors.
- Demand Estimation : Predicting the demand for medicines and medical devices to ensure there is supply and proper allocation of resources.
- Trend Projections : Foreseeing trends in healthcare like the rise of illnesses or changes in patient demographics to enhance regulatory readiness.

Example :
Through machine learning techniques, the MHRA could potentially predict which approved drugs might carry risks of adverse reactions by studying data from prior drug approvals, clinical trials and demographic details. This predictive approach can assist in prioritising market surveillance efforts.

At this point in time there is not a lot of predictive analysis that is used within the MHRA but it is something the Data Science team is looking at for the future however, I have demonstrated some predictive analysis in my 3rd evidence statement which was on a confidential GP dataset.

Prescriptive Analytics :

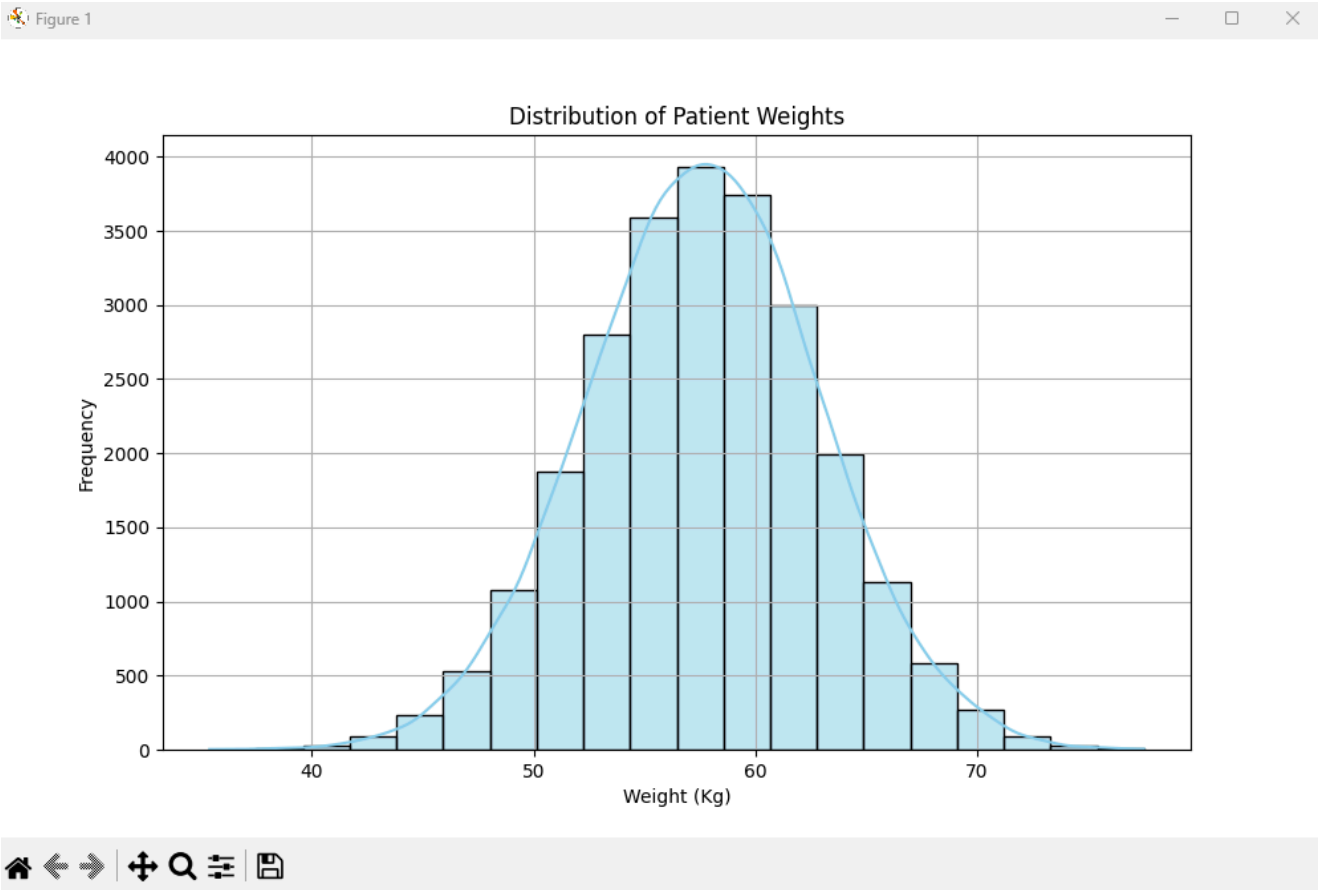
Prescriptive analytics takes it a step further by suggesting actions to achieve desired results. It utilises optimization and simulation algorithms to offer guidance on outcomes and decision pathways.

Potential application in MHRA :

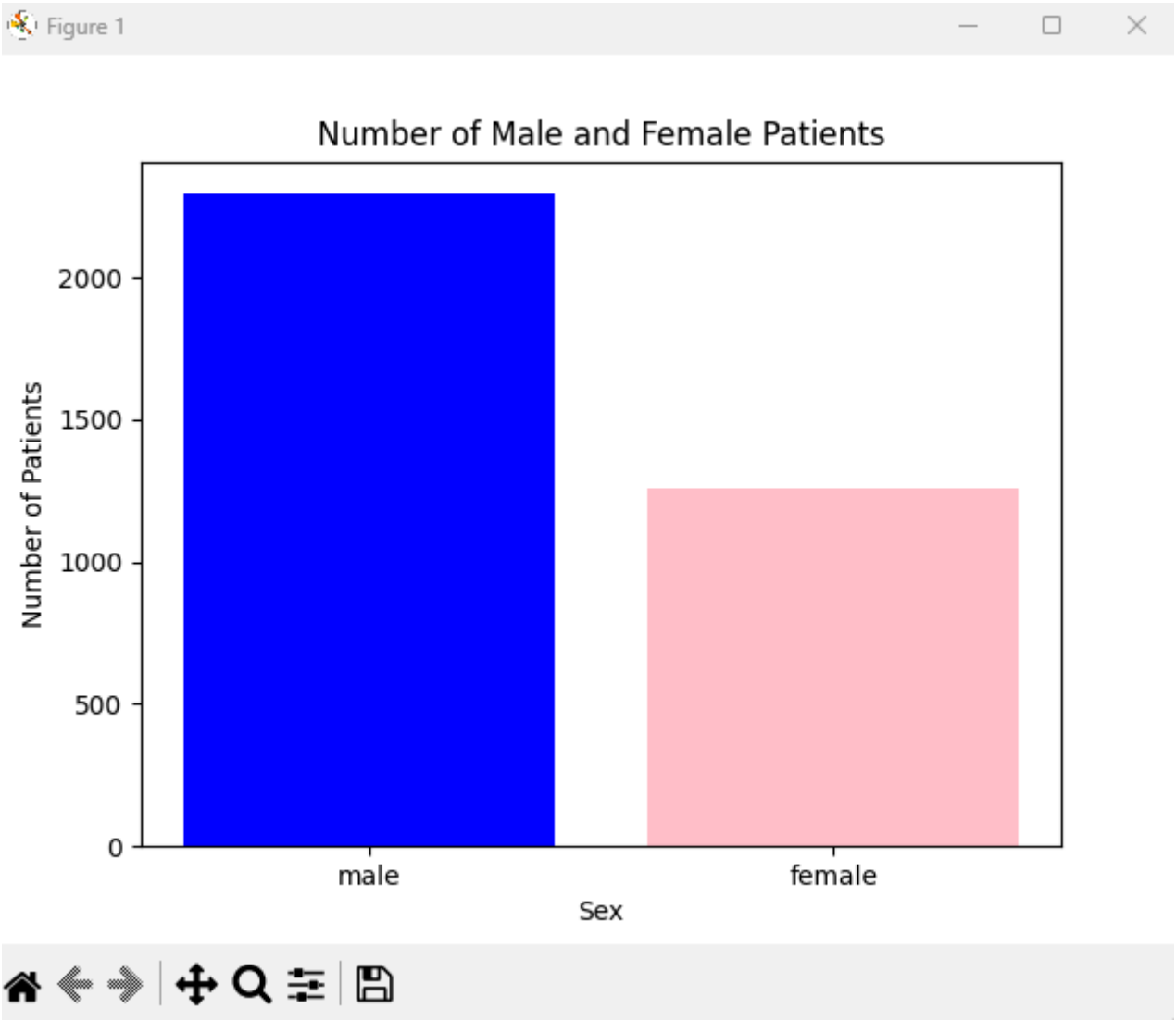
- Support for Regulatory Decisions : Providing advice on the regulatory steps to take in particular scenarios like recalling a medical device or approving a new drug.
- Resource Management : Efficiently allocating resources for inspections, audits and compliance monitoring to enhance effectiveness and productivity.
- Policy Formulation : Suggesting changes in policies based on insights to improve regulations concerning drug safety and efficacy.

Example :
If predictive analytics show a probability of reactions with a certain medication, prescriptive analytics might recommend actions such as increased monitoring, additional labelling requirements or even temporarily halting approval of the drug until further testing is done.

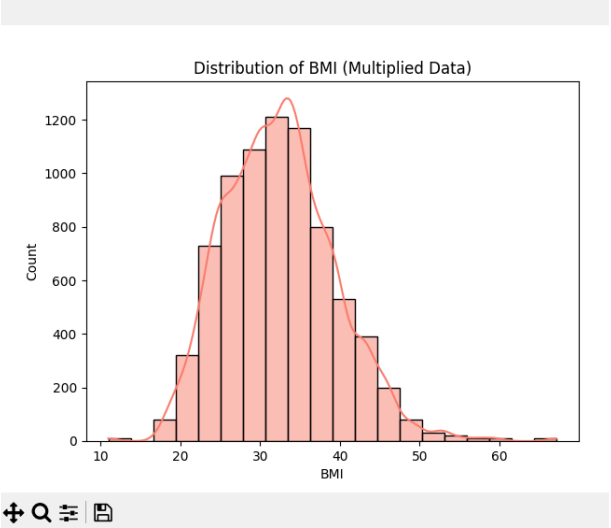
At this point in time there is not a lot of prescriptive analysis that is used within the MHRA but it is something the Data Science team is looking at for the future.



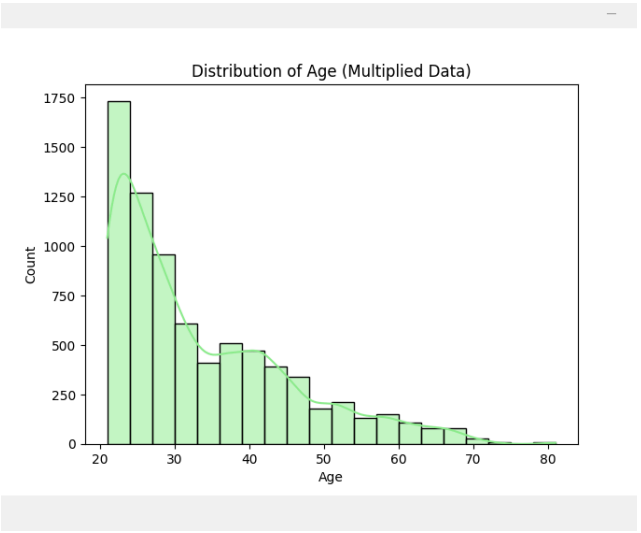
ES4 Figure 1: Statistical analysis showing the distribution of age amongst the patients in the dataset. (Different set of data than Figures 2 and 3). I believe in a normal situation the weight of patients would not have a normal distribution, however this is what the data (that was provided to me) shows and showing the data to the stakeholder gave me confirmation that this is what they would have expected.I chose this visualisation as it can provide clarity around the frequency of data within a range of values.



ES4 Figure 2: Statistical analysis showing the distribution of patients by Gender. (Different set of data than Figures 1 and 3) This is a simple bar graph showing the number of male patients compared to female patients in one of the datasets given to me for this project. I chose this visualisation as it is quite simple and easy to understand whilst providing insightful comparative data.



ES4 Figure 3: Statistical distribution of BMI and Age for patients. (Different set of data than Figures 1 and 2) This is a distribution graph which shows the distribution of BMI in patients and I believe this displays a positive or right skew so most of the extreme values are on the left side of the mean. I chose this visualisation as it can provide clarity around the frequency of data within a range of values.

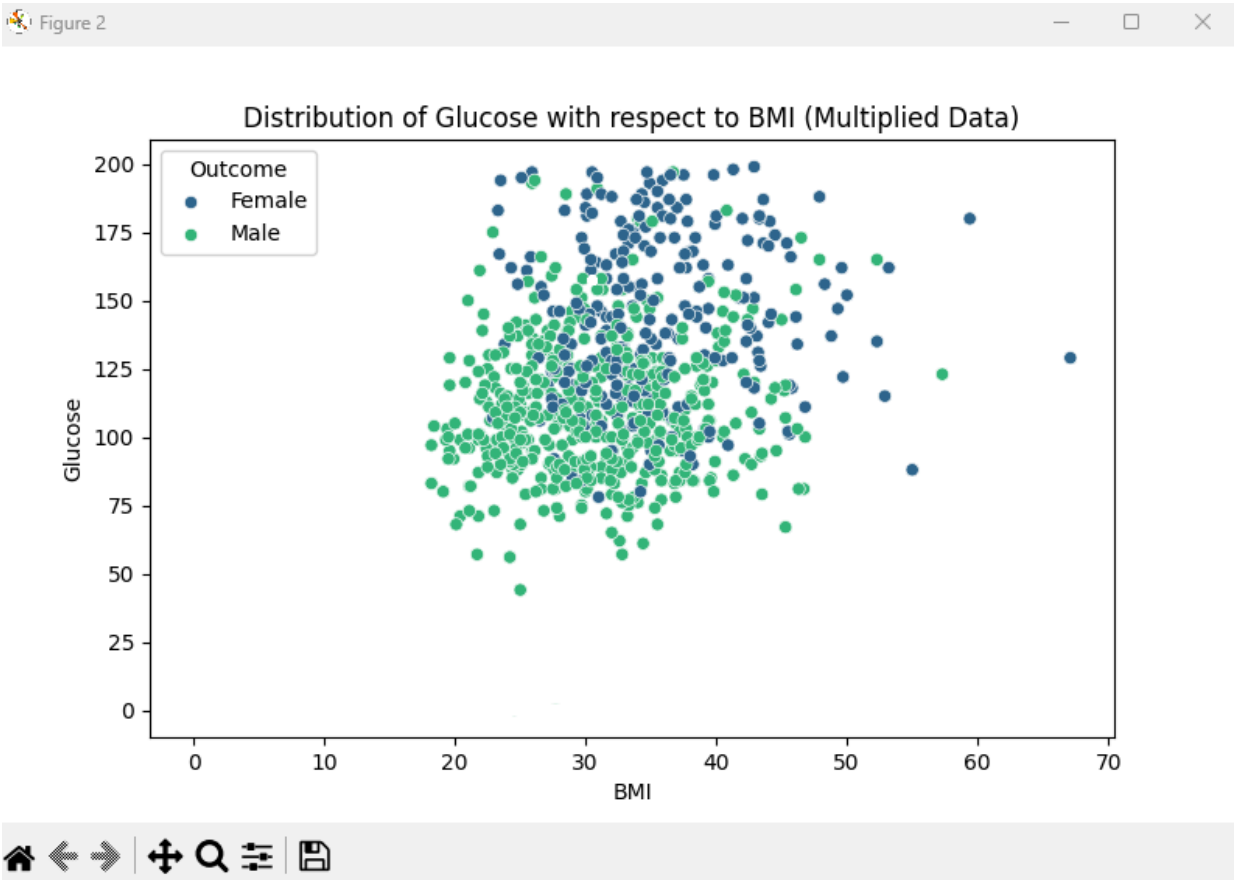


ES4 Figure 4: Statistical distribution of Age of Patients. (Different set of data than Figures 1,2 and 3) This is another distribution graph which shows the positive or right-skewed distribution of patients' age in one of the datasets, it is important to note that the data this graph is presenting is different to the data that Figure 3's graph is showing.I chose this visualisation as it can provide clarity around the frequency of data within a range of values.

```
# Here I am trying to Visualise the BMI of some Patients in a distribution graph
plt.figure(figsize=(8, 5)) #Here I am creating a new figure for the plot with the specific size :
sns.histplot(df_multiplied['BMI'], bins=20, kde=True, color='salmon') #Here i am making use of sea
plt.title('Distribution of BMI (Multiplied Data)') #I have created a title for the visualisation h
plt.show() #This function displays the visualisation so we can see its output.

# Here I am trying to Visualise the Age of some patients in a distribution graph
plt.figure(figsize=(8, 5)) # Here I am creating a new figure for the plot with the specific size
sns.histplot(df_multiplied['Age'], bins=20, kde=True, color='lightgreen') #Here i am making use o
plt.title('Distribution of Age (Multiplied Data)')
plt.show()
```

ES4 Figure 5: Python code for Figure 3 and 4 showing the way that I created these visualisations using matplotlib and seaborn. I have also included some comments in my code as good practice, this helps others understand what is going on in my code.



ES4: Distribution of glucose levels with respect to BMI in both Male and Female patients who have diabetes, this is showing that there is some correlation between these two variables and that higher glucose levels can varies between different BMI levels, this could potentially provide insight into how body weight can influence blood glucose levels.I chose to use this visualisation as it makes identifying the outliers very easy and it allows me to easily compare variables.

Review

Output, Outcomes, and their Evaluations

The Output of this project is a set of statistical visualisations which help provide insights to the stakeholder about the dataset, the impact of these insights would be great as it gives the deputy director a sense of the capability in terms of visualisations using methods other than power bi and we are also able to answer questions the client may have about their data. The stakeholder will be able to understand key information like what the distribution of the “X “ variable looks like within the dataset. The data used for this project was incredibly detailed, for context some tables had hundreds to thousands of rows of data so this report's outcome is a much clearer understanding of the very granular dataset.

Lessons learned | WWW & EBI

The main lesson that I learned with this project is that I can utilise Python skills to create visualisations the same way I would do in Power BI. Using Python provides a lot of options in terms of constructing a visualisation as you have to code the different aspects from the title to the colour of data bars. Taking an Agile approach worked well as I was able to take on more and more feedback from meeting with the stakeholders and showing off my progression by presenting the new visualisations I was able to create since the last meetings and any improvements on the old visualisations. The feedback was provided to me during the meeting which means this was verbal feedback given to me for me to improve on my approach towards this project and improve my outcome.

What went well?

The visualisations deliver very clear insights. The stakeholders were happy with the information provided by the visualisations as it gave them more context and understanding of certain areas of the dataset.

Even better if …

There was more data to work with, this would give me more leeway to provide more statistical insights. it would have also been better if I could use more methods of providing statistical insights to the stakeholders using the dataset. This is because there could be more variables to consider in statistical analysis, this would allow me to make more interesting and insightful comparisons as well as identify insightful trends which may be useful for the stakeholder

Changes I would make.

If I were to do this again, I would seek to provide more complex visualisations. However, at this point, I am only able to provide very basic statistical visualisations as this is what was covered in my statistics course and I feel more ready to undertake more work at MHRA regarding providing statistical analysis whether the tool I use is Power BI or Python.

Appendix

Civil Service (2015). *The Civil Service Code*. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/civil-service-code/the-civil-service-code.

EVIDENCE STATEMENT [5]

NAME	Shaheer Hussain
TITLE	EVIDENCE STATEMENT **[5]** - **[Communications dashboard for requests]**
PATHWAY	**[DATA ANALYST]**

***Some evidence of this project cannot be shown due to confidentiality and/or intellectual property restrictions. The use of a red asterisk (*) will highlight where there is detail missing that cannot be included for these reasons.**

Scenario

Introduction & Business Problem:

An employee from the communications department at the MHRA and the head of technology operations had approached me to take on a task which was to do some quick analysis of some requests, the business problem was that there was a lack of data visibility on these workflow and data requests and the stakeholders needed a tool to easily view their data in an understandable manner so they could make more data informed decisions based upon those insights.(I have been asked to keep these requests censored in my evidence statement and to not give important information about these requests in this document, this has been done to protect potential sensitive data and workflows, also since I have been asked to keep the data confidential I have chosen to respect this request from the stakeholder.) that the MHRA receives about some of the work it does and this was the purpose of this project, the other purpose of this project was to build the employees trust in the tools I use to provide insights and deeper understanding of data, by providing some analysis into the data provided. I can create a good impression for business intelligence tools like Power BI (It's important to create a good impression for business intelligence tools like Power BI as it will encourage employees in other divisions within the MHRA to adopt the tool for more data visibility and insights, this can only result in upskilling employees and allowing them to understand their data more effectively and efficiently which will result in a boost of employee productivity across divisions), this project did not include a very large dataset as the requests that I provided analysis on aren't as abundant as the data in some of the other projects I have worked on.

Tasks

My role as a Data analyst means I am responsible for the sourcing of appropriate data (In this project I was provided with a couple of CSV files which had a standard layout with no unique characteristics), cleaning and transforming this data to be only what is needed, modelling relationships to relate data from different sources and visualising this data to give insights. However, focussing on this project, my role was to explore the data that was provided to me and create as many meaningful visualisations to help the end user understand more about the data and make a more informed decision. For this project, I used Power BI as it was more than sufficient for this small task as I just had to provide some basic analysis on the data provided to me without any sort of complexity in the analysis. I chose to take an Agile methodology as this will result in a higher quality product that will provide higher satisfaction to the end-user or stakeholder (The stakeholder in this case are some of the employees in the Comms department/division within the MHRA) however, there are limitations to an agile methodology approach like the fact that it takes more time to develop an end product and

it can easily get sidetracked as more contact with stakeholders can increase the chances of stakeholders bringing in additional requirements which will affect the development of the original project. This will also help me deliver higher quality insight and avoid any risks of developing insights that are not what the stakeholder is asking for. Agile methodology is more appropriate as it allows me to break the project down in more manageable sprints and it also accounts for adaptability, this means that if there is a change in the requirements given by stakeholders then this can be accounted for, also it just allows for more collaboration with stakeholders which makes the stakeholders feel more involved and results in a better product.

Develop a Prototype Dashboard:

Specific: Create a prototype dashboard in Power BI that displays key metrics and insights from the data.

Measurable: Prototype should include at least 3 key visualisations that stakeholders can review.

Achievable: Utilise available data and Power BI capabilities.

Relevant: Demonstrates the potential of Power BI for data visibility.

Time-bound: Deliver within a week for the initial insights

Stakeholder Review and Feedback Sessions:

Specific: Conduct weekly review sessions with stakeholders to gather feedback and make necessary adjustments.

Measurable: Document feedback from each session and track changes made.

Achievable: Schedule sessions in advance to ensure stakeholder availability.

Relevant: Ensures the tool meets stakeholder needs and builds trust.

Time-bound: Hold sessions every week throughout the project duration.

Table 1: Risks & Constraints

Type of Risk / Constraint e.g. Limited budget	Impact or affect on project e.g. Budget cannot be exceeded	Risk reduction / avoidance e.g. Regular checks on project spending
Limited time	I had a time constraint of 2 weeks to come up with a resolution to the stakeholders lack of data visibility and a decision for tracking the telephone requests	Made effective use of my time by planning out how I would carry out each step in the process. I made sure to communicate effectively

Limited options	There only a limited amount of options we could consider to keep track of the telephone requests which meant there wasn't much we could do to actually effectively track the telephone requests whilst complying with data protection regulations	Used Miro board to brainstorm ideas with the stakeholder to make sure we could generate as many possible options for tracking telephone requests, this allows us to evaluate our limited options and choose one based on what we thought was best.

Activities

ES5 - Explains approaches to combining data from different sources to improve accuracy and / or efficiency and / or maximise benefits to the organisation and / or customer. (K10)

Merging information from origins can greatly improve the precision, effectiveness and advantages for an Agency, like the Medicines and Healthcare products Regulatory Agency (MHRA). Below are some strategies for accomplishing this :

1 - Techniques for Integrating Data
A - Utilising Data Warehousing :

Approach - Gather data from sources into a database.
Advantages - Improves data accessibility facilitates analysis and maintains data coherence.
Application Scenario = Combining clinical trial data, event reports and regulatory details for analysis.

B - ETL (Extract, Transform, Load) :

Approach - Extracting data from sources transforming it into a format and loading it into a target system.
Advantages - Ensures that information from sources is standardised and prepared for analysis.
Application Scenario - Standardising data from firms, healthcare providers and research projects for assessment.

2 - Data Fusion
A - Integration of Multisource Data :

Approach - Merging information, from origins to generate consistent, accurate and valuable insights.
Advantages - Improves decision making by offering a view.
Application Scenario - Combining electronic health records, prescription data and patient feedback to monitor drug safety.

3 - Advanced Data Analysis Techniques

A - Machine Learning and Artificial Intelligence :

Approach - Implement machine learning algorithms to analyse and merge data from sources.

Benefits - Improves analytics, pattern recognition and decision making processes.

Use Case - Utilising AI to forecast reactions to drugs by combining data from trials, patient records and social media platforms.

B - Natural Language Processing (NLP) :

Approach - Employ NLP methods to extract and integrate information from sources such as research papers and medical notes.

Benefits - Enables the extraction of insights from written content.

Use Case - Examining research publications and medical notes to discover drug interactions and potential side effects.

4 - Collaborative Platforms and Data Exchange

A - Data Sharing Partnerships :

Approach - Establish agreements with entities for sharing data.

Benefits - Enhances access to a range of data sources.

Use Case - Partnering with bodies and pharmaceutical companies to exchange information on drug safety.

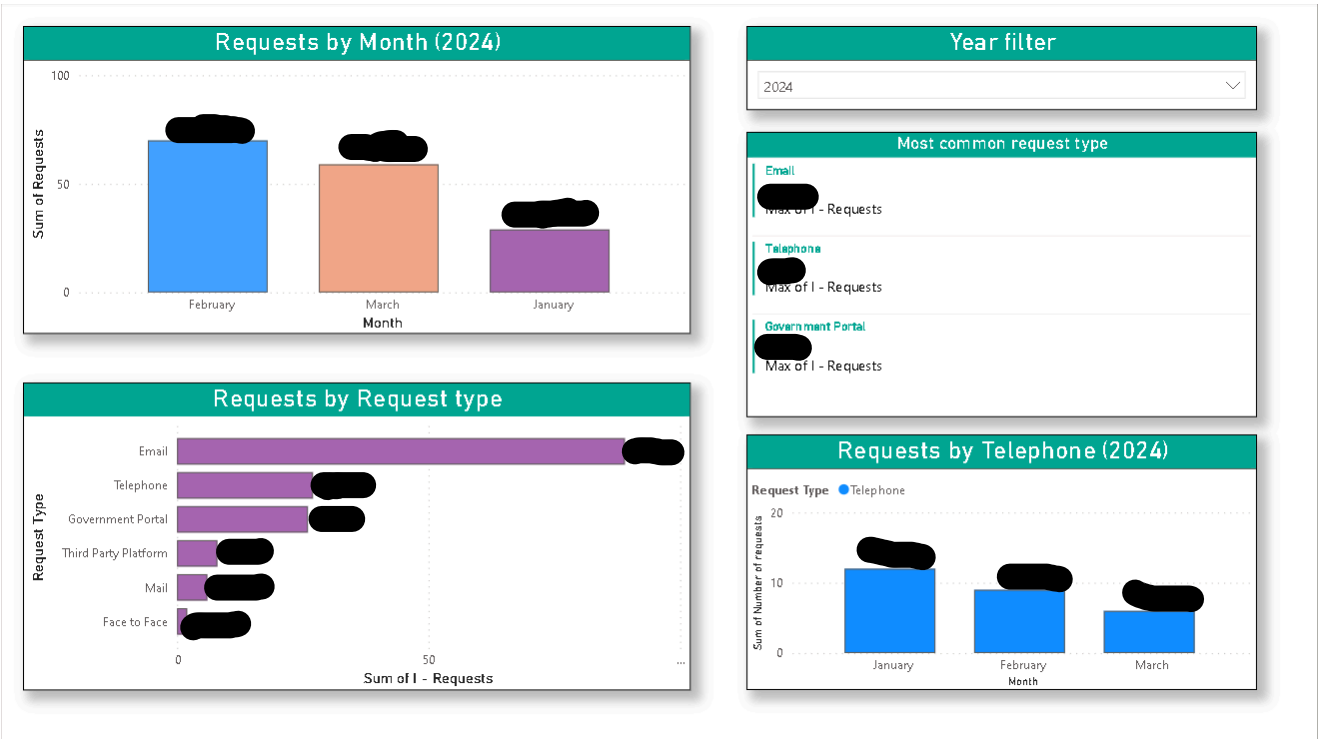
B - Interoperable Systems :

Approach - Develop systems that can seamlessly share data with systems.

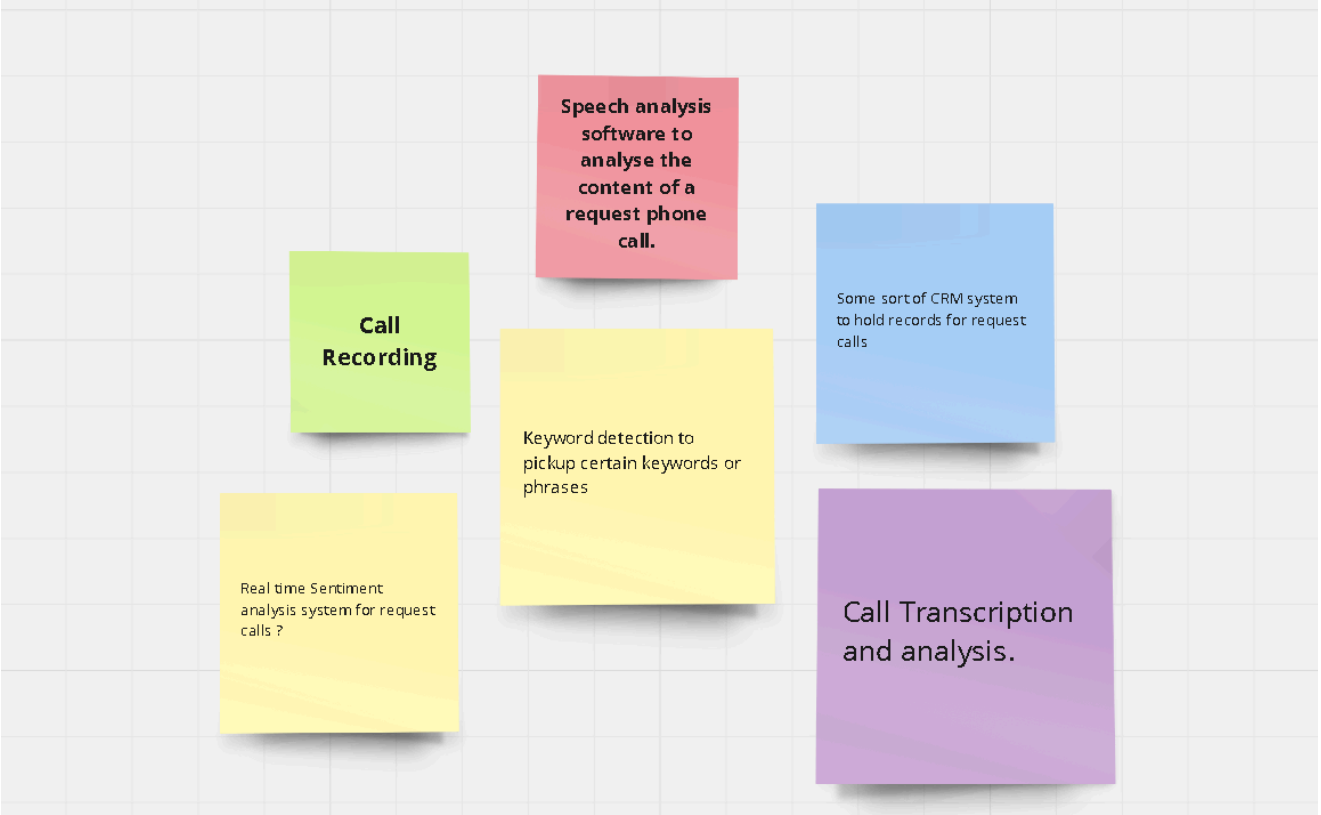
Benefits - Improves data flow efficiency and minimises integration complexities.

Use Case - Building health information systems of exchanging data with hospital databases, research organisations and regulatory authorities.

ES5 - Explains the principles of descriptive, predictive and prescriptive analytics and demonstrates how they have been applied within their own data analysis practice. (K14 S11)



ES5 Figure 1 - Dashboard showing the Requests received so far in 2024 with a slight focus on the requests received by telephone.

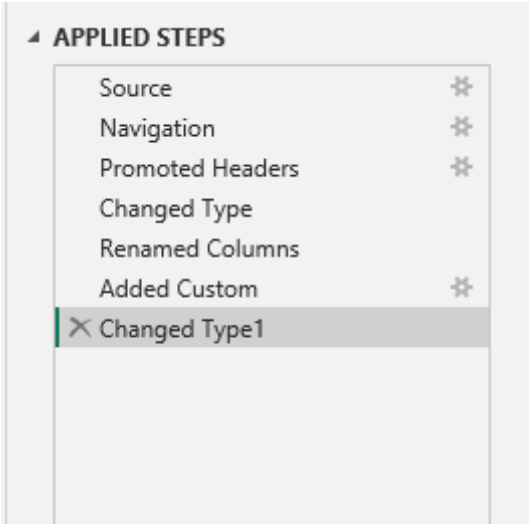


ES5 Figure 2 - Miro Board of Ideas collected by me and stakeholders regarding keeping a record of request phone-calls.

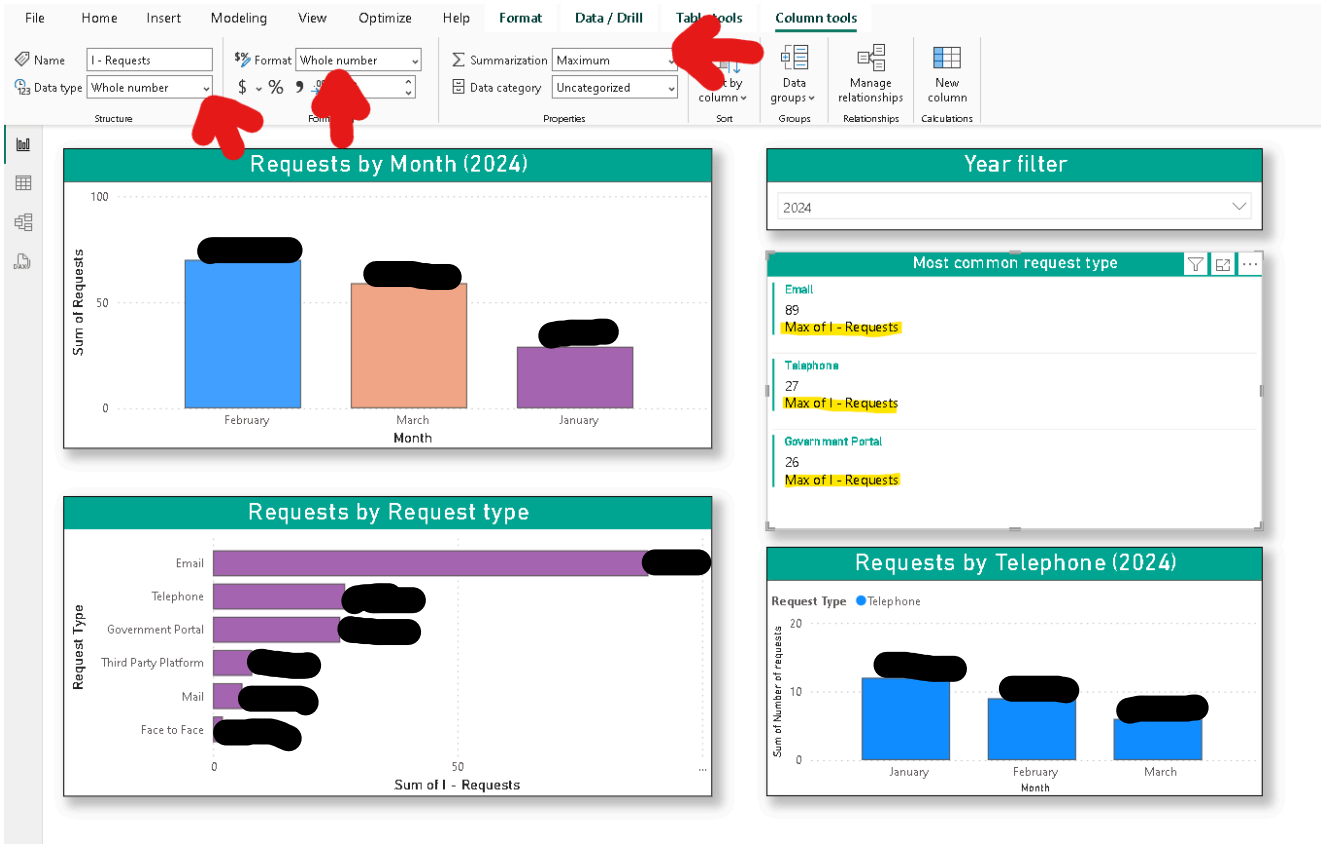
Potential solutions	Benefits	Risks
Call Recording	- Provides an accurate record of conversations for dispute resolution and compliance purposes.	- Privacy concerns around recording without consent may violate regulations. - Storage and security: Recorded calls must be securely stored to prevent unauthorised access or data breaches. - Will require manual call monitoring in case phone calls need to be re-reviewed which will require more time
Speech Analysis	- Enables automated analysis of call content for actionable insights. - Identifies trends, patterns, and keywords relevant to business objectives.	- Speech analysis algorithms may misinterpret accents, background noise, or complex language, leading to inaccurate results so it could become inaccurate. - - It will be quite resource-intensive Requires significant computational resources for

		real-time analysis.
Keyword detection	- Automates the identification of specific keywords or phrases during calls. - Facilitates quick identification of relevant topics or issues.	- There may be false positives as it may flag irrelevant keywords or miss important ones, leading to inaccurate analysis. - There is limited scope as it only captures predefined keywords, potentially overlooking nuanced conversation elements.
Sentiment analysis	- Provides insights into customer satisfaction levels and emotional responses. - It helps identify potential issues or opportunities for improvement.	- There could be Subjectivity as Sentiment analysis may not accurately capture the nuances of tone or emotion in speech, leading to misinterpretations. - Language barriers: May struggle with non-standard language or cultural nuances.
Call Transcriptions	- Converts spoken conversations into text format for easier review and analysis. - It enables keyword search and indexing for efficient call retrieval.	- Transcription errors may occur, especially with accents, background noise, or fast-paced conversation. - There may be Data privacy concerns as Transcribed conversations must be handled securely to protect sensitive information.

ES5 Figure 3 - Benefits and Risks analysis on potential options regarding keeping a record for request phone calls.



ES5 Figure 4 – Transformation steps applied to the dataset in order to prepare it for visualisation



ES5 Figure 5 – Summarisation used on key visuals and data types selected

Prescriptive analytics

Prescriptive analysis can be described as the process of analysing data to determine an optimal course of action, this sort of analysis would be addressing the “what should we do” kind of questions by providing recommendations and actionable insights, this may make use of a combination of predictive models, historical data and optimisation techniques. Some of the key characteristics of prescriptive analysis is that it often includes optimisation techniques to identify the best course of action based on numerous objectives and constraints, it can also operate in real time which allows us to receive recommendations in dynamic situations.

I have demonstrated Prescriptive analysis in Figure 1, Figure 2 and Figure 3 as I have used my visualisations or analysis of the data to find out that one of the most common forms of requests type is telephone which means people may call in to make a request from the MHRA. Since other methods like emails are easy to trace back to and telephones aren't exactly easy to track, I used this information to make a suggestion and answer the “What should we do ?” question. I made a few suggestions but among them was one suggestion that we need to track phone calls with keywords, so when a keyword is used in a phone call then a meaning is already assigned to that keyword and there can be a record created in place of the phone call or some sort of similar system to be implemented, this way, when telephone calls are taken, there is a record created and this record uses keywords to assign a semantic to this request, this will make this workflow regarding request management much more efficient as you would not need to go through call transcriptions or audio recordings which can be much more time consuming, especially when there is a lot of records to go through. The downside to this is that this sort of system can be expensive to set up and there will be considerations around

privacy. My method of optimisation within this prescriptive analysis was decision analysis, Decision analysis is a systematic, quantitative, and visual approach to making complex decisions. It involves using structured methodologies to evaluate and compare different options based on their potential risks and benefits, with the goal of selecting the most advantageous course of action. The decision analysis technique I used was SWOT analysis, SWOT analysis is a strategic planning tool used to identify and analyse the internal and external factors that can impact an organisation, project, or individual decision. The acronym SWOT stands for Strengths, Weaknesses, Opportunities, and Threats. However, there are also other techniques in decision analysis like Cost-Benefit analysis, Multi-Criteria decision analysis and even risk assessments.

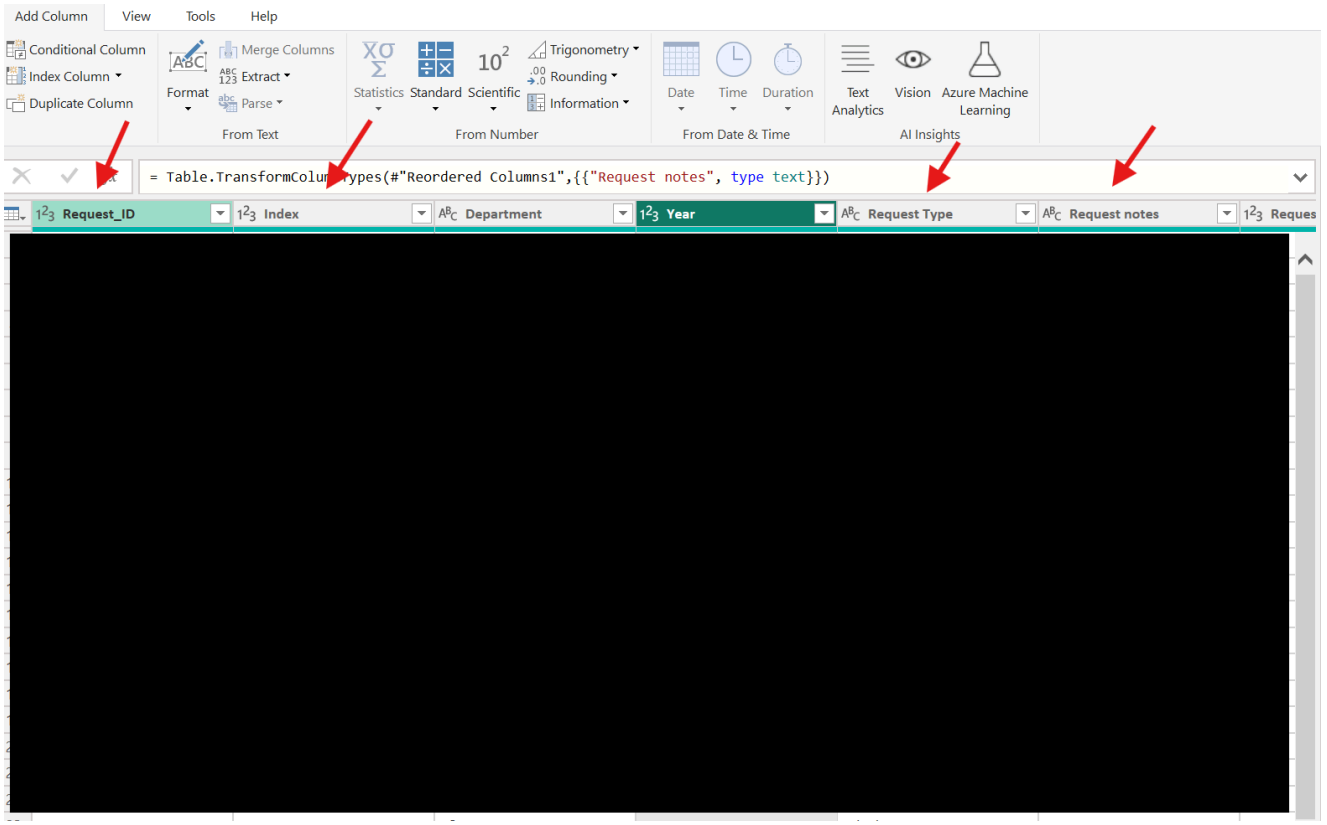
Predictive analytics leverages statistical techniques, machine learning algorithms, and data mining to analyse current and historical data, allowing organisations to make informed predictions about future events. This approach can significantly enhance decision-making processes by providing insights into potential future scenarios so naturally it has a lot of potential applications within the MHRA, In this scenario, it could be used to forecast the number of requests in future years by the type of request, for example I could predict the number of telephone requests in 2025 based on the historical data of telephone request frequency in the previous years, this could provide the stakeholders with insight like the fact that they may have more telephone requests and so they need to make sure they have a system in place to readily access information about these telephone requests.

ES5 - Describes the relevant tools or techniques used for working with the data systems architecture in their organisation. (S9)

Table 2: Tools

Tools	This tool was used to…	Reason for selection
Power BI	Load and analyse the data, creating some visualisations to provide quick insights for the stakeholders to make more data driven decisions	Power BI is a MHRA approved tool and it's quite capable for handling small datasets like the one used in this project, since I have some experience using Power BI, I would be more efficient using this tool and I would understand how to get the tasks done.
Miro board	Give myself and the stakeholders the ability to share our ideas in regards to keeping a record of telephone requests	This is a tool I had used quite frequently at ADA so I was familiar with how easy it is to use and its focus on collaborative work so I thought it would be the perfect tool to collate our ideas.
Microsoft Teams	Hold meeting to discuss the project, more specifically any updates, progress and next steps	An heavily used tool within the MHRA for communication and its very effective in the regard of getting information to people very quickly.

ES5 - Demonstrates data analysis activities involving the collation and interpretation of qualitative and quantitative data and displays results using visual representations. (S14)



ES5 Figure 6 - PowerBI Power Query view of dataset , this is only a section of the dataset.

I have demonstrated the collation of qualitative and quantitative data in a visual representation in Figure 5 as I have created a visualisation for the requests by type, quantitative data is present in my visuals where you can see the counts of requests by month or the count of requests by the request type, this is qualitative as it can be quantified, counted, measured or it can be assigned a numerical value, On the other hand I have also shown the use of qualitative data when transforming the data, this can be shown in figure 6 where you can see the input of notes for records of “Telephone” request type, as for the collation of data, you can see that I have included an index which will be used to quantify a request type as well as the individual request notes which have been collated into the same query / table.

ES5 - Critically evaluates the risks and benefits of predictive analytics (K14 S11) [**Distinction]**

Aspect	Risk	Benefit
Data privacy	- Risk of unauthorised access or data breaches. - Potential for misuse of sensitive health data. - Concerns about patient consent and autonomy.	- Improved patient outcomes through personalised medicine. - Early identification of adverse drug reactions. - Enhancing public health interventions.
Bias and Fairness	- Potential for algorithmic bias leading to unequal treatment of certain demographic groups. - Reinforcing existing disparities in healthcare. - Lack of transparency in algorithm decision-making.	- Enhanced objectivity in decision-making. - Ability to identify and mitigate biases in healthcare delivery. - Targeted interventions to address health disparities.
Accuracy and Reliability	- Inaccurate predictions leading to inappropriate interventions or resource allocation. - Reliance on incomplete or biassed datasets.	- Early detection of potential safety issues with medications or medical devices. - More efficient allocation of resources based on predictive insights. - Timely intervention in public health emergencies.
Regulatory Compliance	- Compliance challenges due to the evolving regulatory landscape. - Potential conflicts with existing regulations regarding data privacy and patient rights.	- Streamlined regulatory processes through automation and data-driven decision-making. - Improved monitoring of safety and efficacy of healthcare products. -Facilitation of evidence-based policymaking and regulation.
Stakeholder Trust	- Erosion of trust due to perceived or actual misuse of predictive analytics. - Lack of transparency in how predictions are generated and utilised.	- Increased transparency in healthcare decision-making processes. - Building trust through accountable and ethical use of predictive analytics. - Enhanced collaboration between regulators, healthcare providers, and the public.

Resource Allocation	- Misallocation of resources based on flawed predictions. - Over-reliance on predictive models at the expense of clinical judgement.	- Optimal allocation of resources to areas with the highest risk or need. - Improved efficiency and cost-effectiveness in healthcare delivery. - Better preparation for public health crises and emergencies.

ES5 Figure 7 – Table showing Critical evaluation of Predictive analysis in the MHRA

One vital aspect is the need for collaboration for implementation. Implementing analytics in healthcare requires cooperation among regulators, healthcare providers, data experts and various stakeholders. Effective communication and coordination among these groups are crucial to ensure that predictive models are accurate, reliable and ethically utilised. The MHRA must invest in creating teams that bring together expertise from fields like data analysis, pharmacology, epidemiology and regulatory science through partnerships with industry leaders and academic institutions; this is what the Data science team do to an extent.

Additionally dealing with bias and fairness is an issue in predictive analytics. Biases present in models can worsen healthcare inequalities and amplify disparities in access and results. The MHRA needs to prioritise fairness and transparency when developing, validating and deploying algorithms to reduce these risks. This involves monitoring and evaluating models to detect and address biases that may stem from skewed or incomplete data sets or inherent algorithm limitations.

Moreover, incorporating analytics into decision making processes requires a thoughtful balance between automation and human judgement. While predictive models can boost efficiency and accuracy in recognising safety signals and assessing risks they should work alongside, not replace, clinical expertise and regulatory judgement. The MHRA should equip regulators with the tools and training to interpret and contextualise insights fostering a culture of evidence-based decision making and ongoing learning.

To sum up, while predictive analytics presents opportunities for enhancing oversight and improving public health outcomes, successfully integrating it into the MHRA's operations demands careful attention to regulatory, ethical and technical factors. By addressing these challenges, the MHRA can leverage the transformative potential of predictive analytics to fulfil its mission of guaranteeing the safety, effectiveness and quality of healthcare products for everyone.

ES5 - Compares and contrasts visual data representation approaches and how they aid understanding by stakeholders (S14) **[Distinction]**

Comparing and contrasting visual data representation approaches involves evaluating different methods of presenting data visually, considering their strengths and weaknesses in aiding stakeholder understanding. Here's an explanation of how this visualisations / representation approached compare with each other:

Bar Charts vs. Pie Charts:

Bar Charts:

Strengths: Bar charts excel in comparing categorical data, such as different types of requests like phone calls, emails, and in-person visits. They provide a clear visual representation of the quantity of each category and allow for easy comparison between them.

Weaknesses: Bar charts may become cluttered or difficult to read if there are too many categories or if the values vary widely in magnitude.

An example of this can be seen in Figure 5 as it has been used to show requests by month and requests be telephone, This is a commonly used visualisation in MHRA because of its simplicity and the fact that all if not most business users (Stakeholders) will be able to easily understand them

Pie Charts:

Strengths: Pie charts offer a visual depiction of each category's share of the whole, making it easy to see the proportion of phone calls, emails, and in-person visits relative to each other. They are particularly effective for displaying data with a small number of categories.

Weaknesses: Pie charts can be less effective when there are too many categories or when the differences in proportions are subtle, as it may be challenging to accurately compare the sizes of different slices.

Pie charts aren't used as much in the MHRA as it generally isn't needed as much as other visualisations like the bar charts or column charts however this visualisation has been used in my 2nd, 3rd and 4th evidence statement, it has been used to show the values of different categories in a dataset.

Line Charts vs. Area Charts:

Line Charts:

Strengths: Line charts are ideal for tracking trends over time, such as the volume of requests received each month. They emphasise changes in values over time and are particularly effective for illustrating patterns or trends.

Weaknesses: Line charts may not clearly convey the total magnitude of values, as they focus primarily on the trend or pattern of change over time.

Line charts are also used quite a lot in the MHRA and can even be seen in my 3rd evidence statement where I have used a line chart for time series analysis, it is commonly used for a lot of power bi reports as it is very easy to understand and work with on power bi.

Area Charts:

Strengths: Area charts, like line charts, show trends over time but also illustrate the overall magnitude of values. They provide a visual representation of both the change and the total volume of requests, offering a more comprehensive view.

Weaknesses: Area charts can sometimes make it difficult to accurately assess the exact values at specific points in time, particularly if the areas overlap or if there are significant fluctuations in the data.

Area charts are not used very frequently within the MHRA and I don't believe I have made much use of area charts in Power BI reports within the MHRA.

Heatmaps vs. Scatter Plots:

Heatmaps:

Strengths: Heatmaps are useful for identifying patterns and trends in large datasets, such as the frequency of certain keywords in phone conversations. They use colour gradients to visually represent the density of data points, making it easy to spot areas of high or low activity.

Weaknesses: Heatmaps may oversimplify complex data or obscure individual data points, making it challenging to identify specific details or outliers within the dataset.

Scatter Plots:

Strengths: Scatter plots are effective for revealing relationships between variables, such as the relationship between keyword frequencies and the types of requests received. They plot individual data points on a two-dimensional graph, allowing for a visual assessment of correlations or patterns.

Weaknesses: Scatter plots may become cluttered or difficult to interpret if there are too many data points or if the relationship between variables is not clear.

I have used scatter plots a lot of times within the MHRA and this can be seen in my 4th evidence statement where I was asked to conduct some statistical analysis on a dataset.

Enhancing Stakeholders Understanding

Clarity and Simplicity:

Visuals should simplify complex data for stakeholders without causing confusion. Clear and straightforward visuals are preferred as they make data interpretation easier. By using intuitive designs and avoiding unnecessary clutter, stakeholders can quickly grasp key insights without needing extensive explanations. For example, employing clean layouts, clear labelling, and concise annotations can enhance clarity and simplify the understanding of visual data representations.

Decision-Making Relevance:

It's essential to select visualisations that offer insights directly relevant to stakeholder’s decision-making processes. For instance, if the objective is to track trends over time, employing line charts or area charts may provide more insightful observations compared to bar charts. By aligning visualisations with stakeholder’s specific analytical needs and strategic objectives, decision-makers can derive actionable insights that inform their decision-making processes effectively.

Interactivity and Customization:

Integrating interactive elements such as drill-down options or filters can enhance stakeholder engagement and comprehension. Interactive dashboards empower stakeholders to explore data from various perspectives, enabling them to uncover tailored insights aligned with their specific requirements. By allowing stakeholders

to interact with the data dynamically, interactive features facilitate deeper exploration and understanding of complex datasets, leading to more informed decision-making.

Accuracy and Precision:

The accuracy and precision of visual representations are paramount for stakeholders to make informed decisions confidently. It's crucial to select visualisations that faithfully portray the underlying data without distortions or uncertainties. By ensuring that visualisations accurately communicate the core data and avoid misinterpretations or misleading impressions, stakeholders can trust the insights derived from the visual representations to guide their decision-making processes effectively.

Decision and Implementation

Feedback and Iteration:

Seeking input from stakeholders regarding the effectiveness of visual representations is crucial for refining dashboard design. By soliciting feedback, you can identify areas where improvements are needed to enhance clarity and relevance. Stakeholder's insights can help pinpoint potential issues, such as unclear labels, confusing layouts, or missing key information. Iterating based on this feedback allows you to make necessary tweaks to the dashboard design, ensuring that it effectively aids stakeholder's understanding and aligns with their needs and preferences.

Training and Support:

Providing guidance and resources to stakeholders is essential for ensuring they can interpret and utilise visualisations effectively. This may involve offering explanations of different chart types, clarifying the meanings of specific visual elements, and assisting stakeholders in navigating dashboard features such as filters or interactive elements. Additionally, providing additional learning materials such as tutorials, documentation, or training sessions can further support stakeholders in building their skills and confidence in interpreting visual data representations.

Integration with Workflow:

Seamless integration of visual data representations into stakeholder's existing workflow and decision-making procedures is key to maximising their impact. This involves ensuring that the dashboard is easily accessible and user-friendly for all stakeholders, regardless of their technical proficiency. Providing clear instructions on how to access and use the dashboard, as well as offering ongoing support as needed, can help streamline integration efforts. Moreover, motivating stakeholders to incorporate data-driven insights into their planning and day-to-day operations fosters a culture of data-driven decision-making, where insights derived from visualisations inform strategic initiatives and operational activities. By embedding visual data representations into stakeholder's workflow, organisations can leverage data more effectively to drive performance and achieve their objectives.

ES5 - Explains the ethical aspects associated with the collation and use of data and justifies why this is important. (K15)

Collating data from phone calls using keyword detection can raise some ethical considerations to light and these are important to address. One ethical consideration is Privacy as people will expect their calls to be private and collecting data without any of their consent will be in violation of this expectation so it could be a good solution to include an automated voice explaining to callers that their call may be recorded and ask if their consent is given.

Another consideration is transparency, it would be good to let callers know that the call will be recorded for data storage purposes, who will have access to this data and for how long they will have this data for, this is important as it maintains trust whilst allowing us to collect data for our intended purposes.

Another ethical consideration will have to be data security, The data from any telephone calls will have to be stored using very robust security measures to protect this information from unauthorised access, hacking or any misuse. A breach of data can have incredibly serious consequences regarding privacy and misuse of data.

Why does this matter?

It's crucial to follow guidelines when gathering data to uphold trust, between people and businesses. When individuals believe their data is handled responsibly, they are inclined to interact with products and services that depend on data collection creating advantages for everyone involved. Moreover, sticking to standards enables organisations to steer off legal and reputational challenges linked to unethical data handling. In essence emphasising ethics, in data collection guarantees that technology is employed responsibly for the good of society.

ES5 - Evaluates the benefits and risks inherent in combining data from different sources (K10). **[Distinction]**

Combining data from data sources can bring about advantages and drawbacks depending on the approach and the situation in which it is utilised.

Advantages:

Deeper Understanding: Merging datasets can offer a holistic perspective on a topic enabling deeper insights and informed decision making. For example, combining sales data with customer feedback may uncover patterns that would go unnoticed when analysing each dataset in isolation.

Enhanced Precision: Pooling data from sources aids in validating and checking information resulting in improved accuracy. By comparing data points across sources, inconsistencies or inaccuracies can be. Rectified.

Enhanced Predictive Capability: Integrating data from sources can bolster models. For instance, merging weather data with sales figures could assist retailers in forecasting demand accurately by considering how weather impacts consumer behaviour.

Discovering Synergies: Combining datasets might unveil synergies or connections that were previously overlooked fostering innovation and identifying prospects. For example, merging healthcare records with information could reveal trends in disease prevalence among demographics.

Cost Effectiveness: Utilising existing data, from sources of gathering new data can be a more economical approach.
It's important to consider the significance of using datasets or relying on just one, for analysis or decision making.

An example of this can be seen in my 2nd evidence statement where I used PowerBI to combine Project and Non-project hours to create a more holistic view of how employees were spending their time at work.This also

improved data precision as I could use the project and Non-project data to compare against each other to cross-validate the data.

Potential Risks:

1. Data Quality Concerns: When merging data from sources there's a possibility of encountering quality issues like inconsistencies, errors or biases due to differences in formats, standards or completeness levels.

2. Privacy Issues: Combining data from sources may pose risks to privacy rights especially if sensitive personal information is involved. It's essential to adhere to data protection regulations to address this concern effectively.

3. Security Vulnerabilities: Integrating data from sources can expose the dataset to a risk of security breaches. Safeguarding the integrated dataset against access and cyber threats becomes more complex.

4. Legal and Compliance Challenges: Consolidating data from sources might present compliance challenges particularly when dealing with various regulations or contractual obligations. Ensuring adherence to laws on data protection and intellectual property rights is crucial.

5. Contextual Loss: Bringing together datasets may result in losing details or nuances present in individual datasets potentially leading to misinterpretation or oversimplification of complex phenomena.

Depending on Sources: Counting on information from sources creates a reliance on their accessibility, trustworthiness and consistency. Any alterations or interruptions in the data origins might impact the credibility and utility of the dataset.
In general, although merging data from origins can reveal perspectives and openings it is crucial to thoroughly evaluate and handle the related risks to safeguard the credibility, confidentiality and protection of the merged dataset.

When combining the project and non-project data there were some format issues simply due to the way that some employees format their excel workbooks, this means I need the use of Power query to transform the data in a way that allows me to effectively join the two datasets.

ES5 - Describes how they have ensured the true root cause of any problem is found and a solution is identified which prevents recurrence. (B5)

I have ensured that I found the root cause of the problem by conducting some basic descriptive analysis to provide some insights into how many requests the stakeholders were receiving by both request type and by each month, this would allow me and the stakeholders to understand how many requests were coming in by telephone and the stakeholders did not have an effective system to record the information from telephone requests and easily access this information, it was from here that I decided to conduct some prescriptive analysis and I created a miro board for me and the stakeholders to share our ideas on how we could create a system to hold the telephone request information and we decided to choose a keyword detection system as this would allow us to automate the records of these requests to some extent and it would allow us to access this information very easily, even allowing us to present this information in BI tools like PowerBI. This would ensure that the issue of data visibility would not be a reoccurring dilemma.

ES5 - Identifies areas of work where they adapted to changing contexts within the scope of a project, direction of the organisation or Data Analyst role (B7).

I had changed the context within the scope of the project when delivering my PowerBI report to the stakeholders to provide them with more insights into their data, Since the level of telephone requests were much higher than expected it forced us to discuss how to effectively create a record of these requests and this resulted in me creating a meeting for me and the stakeholders to collaborate and put some ideas together in regards to creating this information database for telephone requests. This was the shift from simply providing some descriptive analysis to conducting prescriptive analysis as I went from providing insights to the stakeholders to utilising those insights to determine an optimal course of action.

Review

Output, Outcomes, and their Evaluations

The Output of this project is a report in which the stakeholder will be able to view the meaningful insights which will enable them to understand their data more. The stakeholders will be able to understand my point of view and what I believe we should do in this situation which is to start tracking phone calls via keywords. The data used for this project was not incredibly detailed as it really is just the number of requests by request type and date. However, it is suitable as simple insights is all we need to make a more informed decision on what the communications team should do next. In addition to this, we also have a range of other ideas we could pursue in order to achieve our goal of having a more organised record of requests that come in.Overall the requests from the stakeholders were met as we were able to make a decision on how to record the telephone requests and this was because I was able to provide data driven insights on the requests.

Lessons learned | WWW & EBI

The main lesson that I learned with this project is that we can use very simple visualisations to carry out very effective prescriptive analysis that can then potentially improve the efficiency of operations. Taking an Agile approach worked well as I was able to take on more and more feedback from meeting with the stakeholder and showing off my progression.

What went well?

The report delivers very clear insights and highlights the telephone as being a clear focus on approaching the records for these requests in the MHRA. This was helpful in allowing the team to make a decision about these phone calls. The stakeholder was happy with the high-level l insights provided which he was easily able to pick out stories within the data and this was an objective of the project.

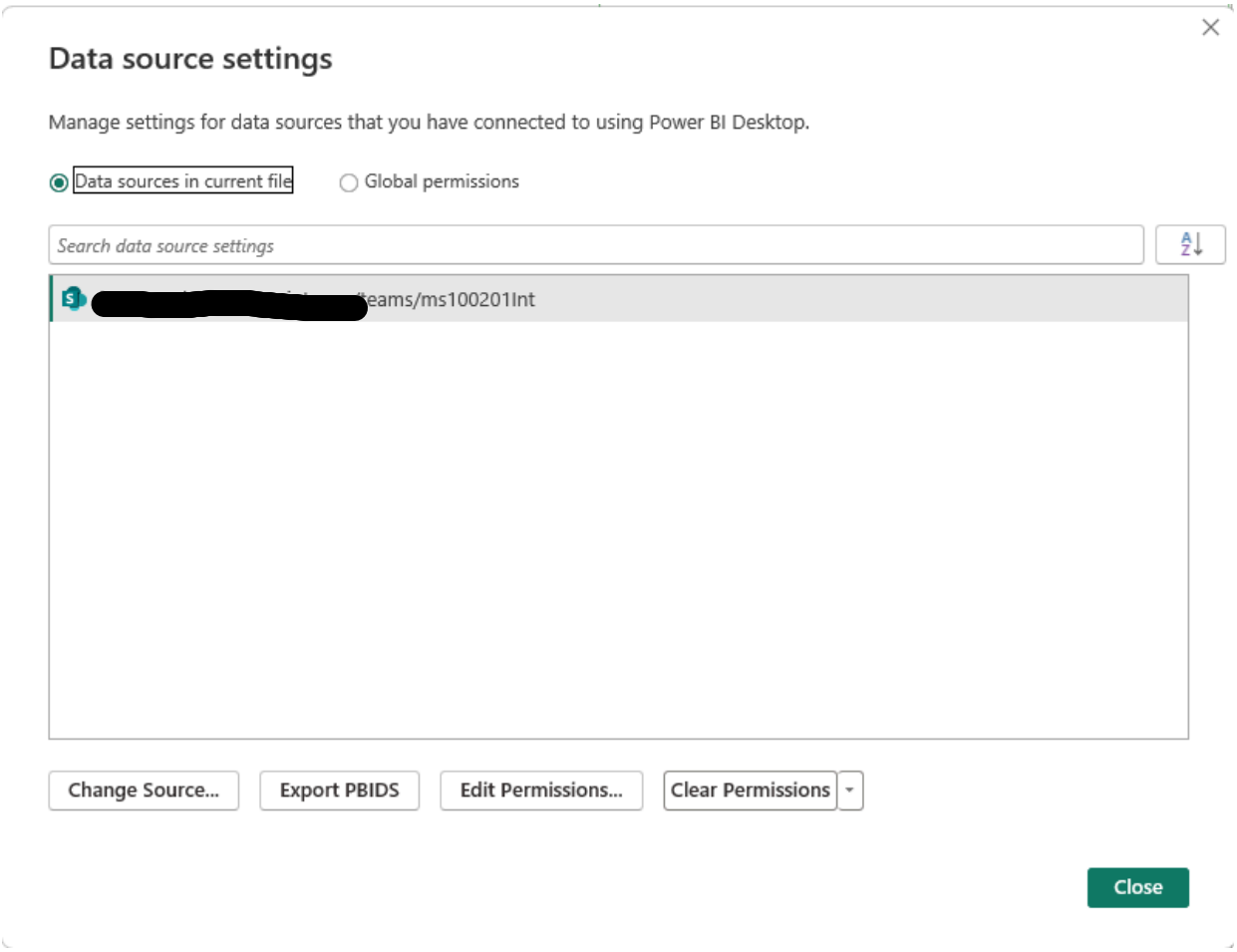
Even better if …

There was a longer timescale for us to brainstorm more potential solutions around tracking the key bits of information around these telephone calls and potentially identify any other areas where we could apply a similar solution to solve a problem.

Changes I would make.

If I were to build this report from scratch again, I would not change much as I believe I had delivered the insights needed to make a decision, if there wasn't any telephone calls or very little telephone requests then implementing a keyword or language processing system would not be worth the time to develop and implement as it would not be of as much use, Overall I think i made the best decision as I was able to deliver insights that we could use to make more informed decisions.

Appendix



ES5 Figure 8 : Data source for Power BI report