Queen Mary
University of London

Group Project (GR11)

ECS7025p

Abhijeet Dhankhar, Arvind Ramchandra Jadhav, Ayushi Choudhury

Gargeya Sharma, Jayesh Gupta, Radhika Daithankar

Queen Mary University of London

Data Science Accelerator

- This program was designed by UK's **Government Digital Service (GDS).** (Data Science Accelerator, 2017) [1]
- It is a program which helps analysts from different public sectors to develop their data science skills.
- The DSA program provides a wide range of opportunities for the analysts like workshops, practical projects and mentoring.
- **Benefits:**
  - It provides a community of like-minded people.
  - It provides a practical experience on real-world projects.
  - Analysts make efficient decisions.

Public Sectors

Data Science Accelerator (cont'd)

- Participants from UK DSA program (Data Science Accelerator, 2017) [1] :
  - **Lisa Richardson** from the **Marine Management organization** simplified the process of analysing the satellite data.
  - **Chris Chin** from the **Ministry of defence** built a visualization tool to show how are the soldiers affected by different equipment loads.
  - **Gemma Coleman** from the **Department of Education** analysed the impact of Special Education Need (SEN) support.

- Other organizations using DSA:
  - **Marks & Spencer (M&S)** started the world's first Data Academy in retail. (Marks & Spencer) [2]
  - **Airbnb** has a program called as Data University that is designed to upskill their employees in data science and analytics. (Airbnb, 2017)[3]

Queen Mary University of London

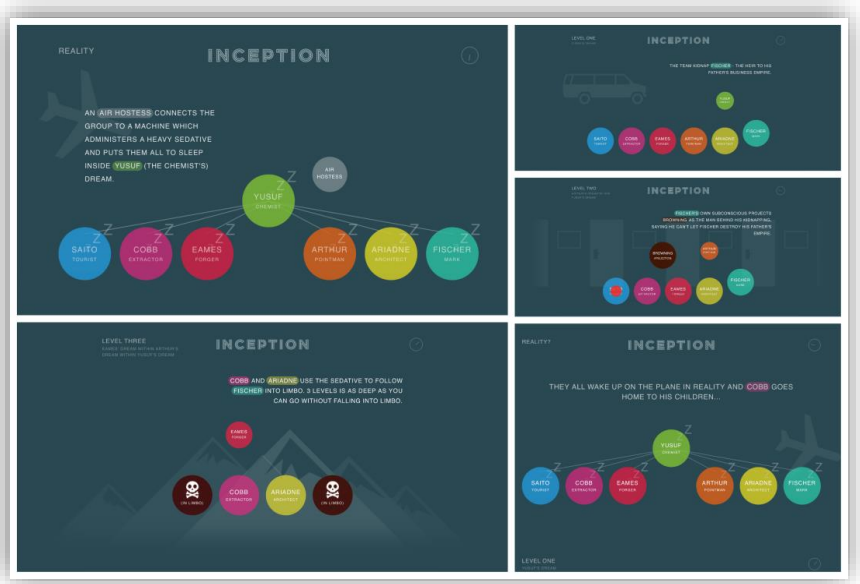
Data Science Visualization

- Data visualization is the process of putting information into a visual context, like a map or graph, to make it simpler for the human brain to grasp and draw conclusions from the data.
- Major objective of data visualization is to make it simpler to spot patterns, trends, and outliers in huge data sets.
- Data visualisation falls under the larger subject of data presentation architecture (DPA), which tries to search, locate, manipulate, format, and convey data as effectively as feasible.
- Data visualisation is one of the processes in the data science process, according to which data must be represented after it has been gathered, processed, and modelled in order to draw conclusions
- Data science visualisation initiatives can assist stakeholders in a variety of sectors, including business, government, and academia, in better comprehending and interpreting their data.

(TechTarget(2020))[4]

Data Science Visualization Project

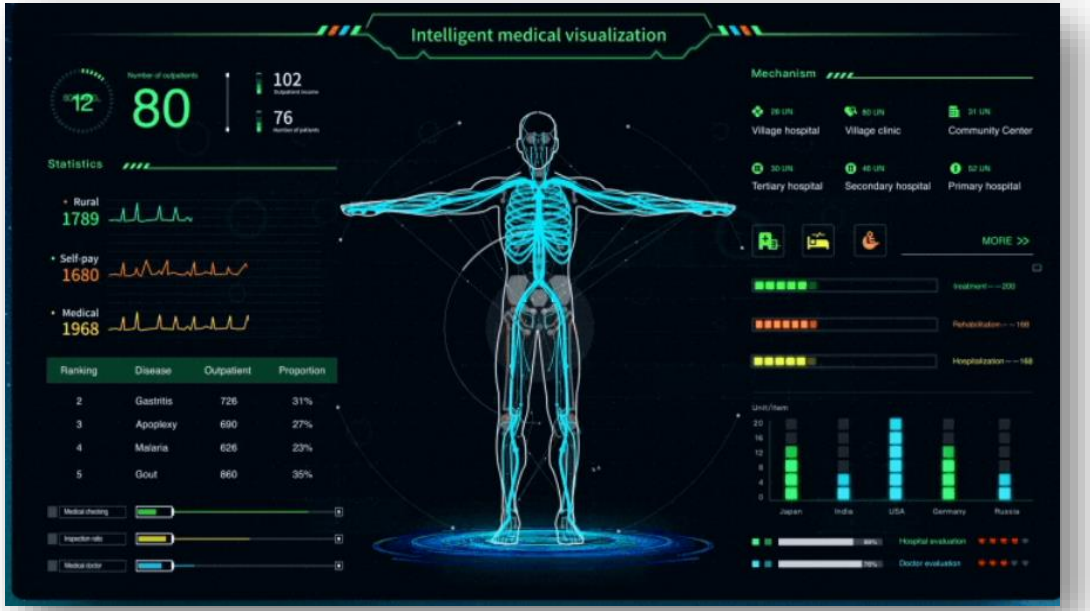
- The Data Science Visualization Accelerator programme was launched in 2021. (Data Science and Data Visualization Accelerator: GOV.UK)[5]
- It is for both analysts as well as for people with an interest in building their visualization skills.
- The objective of this project is to make visualising complex data sets easier and help to identify patterns and trends to make well informed, data-driven decisions.
- When collecting the data for a data visualization project, data controllers must ensure they have a valid legal basis for processing the data, which may include obtaining consent from individuals.
- They must also secure the individuals with clear and concise information about how their data will be processed, with whom it will be shared, and the purpose behind it.
- One of the best examples of data visualization is the Inception (Inception-explained)[6]

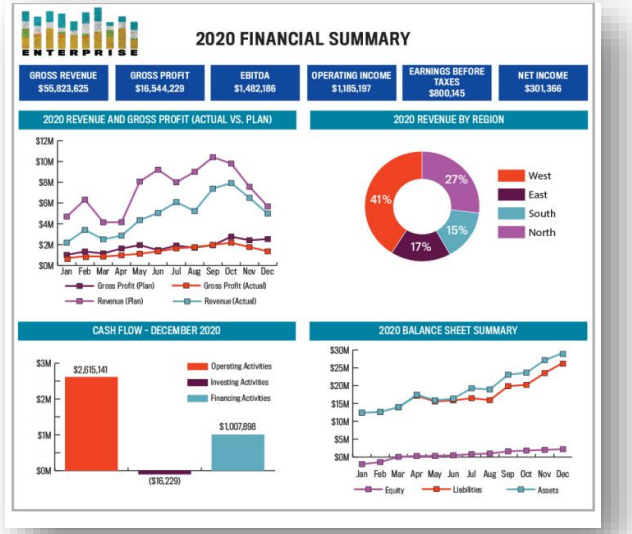


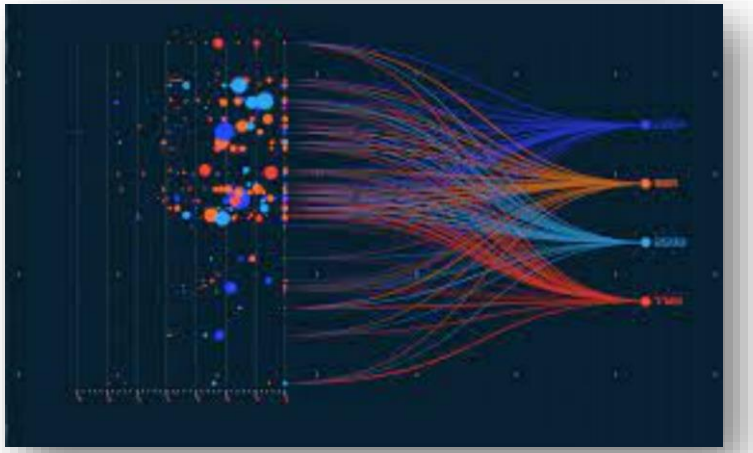
Data Science Visualization Uses

1. Healthcare 'Choropleth maps' and Dashboards
2. Scientific Visualizations
3. Finance/Investment uses 'Candlestick charts'

(Knowledgehut(2023))[7]







Introduction: Problem

- GOV.UK receives over 20,000 pieces of feedback every month, which helps improve the site's content and services.
- However, a small percentage of this feedback is spam, which can be time-consuming to identify and remove manually.
- Hence the government decided to tackle this problem using Data Science and Machine Learning to automatically classify feedback submitted on the website as Spam or Not

(Reilly, F. (2022))[8]

Critical Analysis:

- From an Ethics perspective, the use of machine learning to automate spam detection could be seen as a positive step in protecting user data and privacy, as it reduces the need for human intervention and thereby minimizes the risk of human error or bias.

Developing the machine learning model

- A team of developers at GOV.UK used supervised machine learning to train a model to detect spam feedback.
- The team used a dataset of over 10,000 feedback items, of which around 8% were spam.
- They used Python programming language and Scikit-learn, an open-source machine learning library, to develop the model.
- The model uses natural language processing techniques to analyze the text of feedback and identify patterns that suggest spam.

(Reilly, F. (2022))[8]

Critical Analysis:

- The use of supervised machine learning to train the model is an effective approach, as it allows the model to learn from labelled data and improve its accuracy over time. The use of open-source frameworks makes it easier for other developers to replicate or build on this work in the future, Also The model created was open-sourced to the public **(Data Community Technical Documentation, 2022)**
- From an Ethics perspective, the use of natural language processing techniques raises questions about the use of personal data and privacy. However, the article does not provide enough information to determine whether the model is trained on personal data or not.

Evaluating the model

- The team tested the model on a separate dataset of feedback items, of which 10% were spam.
- The model achieved an accuracy rate of 98% in identifying spam feedback.
- An effort to understand and explain the behaviour of the model was also undertaken to understand and detect which features or inputs were valued more to detect potential biases

Reilly, F. (2022)[8]

Critical Analysis:

- The high accuracy rate of the model suggests that it is effective in detecting spam feedback. However, the article does not provide enough information to determine the precision and recall rates of the model, which are also important metrics in evaluating a machine learning model's performance.
- From an Ethics perspective, the high accuracy rate of the model could be seen as a positive step in protecting user data and privacy, as it reduces the risk of spam feedback being used to manipulate or influence user behaviour on the site.

Implementing the model

- The team integrated the model into the feedback system on GOV.UK.
- The model runs automatically on new feedback items and flags any that it identifies as spam.
- A human moderator reviews the flagged items to determine whether they are indeed spam and removes them if necessary.

Reilly, F. (2022)[8]

Critical Analysis:

- The integration of the model into the feedback system on GOV.UK is an effective way to automate spam detection and reduce the burden on human moderators. However, the article does not provide enough information to determine how the human moderators are trained to review the flagged items and avoid any biases.
- From an ethical perspective , the integration of the model into the feedback system could be seen as a positive step in protecting user data and privacy, as it reduces the need for human intervention and thereby minimizes the risk of human error or bias.

UK Data Ethics Framework Principles

Transparency
4/5

The model was created using open-source technologies and the model was also made available to the public. Methods and Outcomes for creation were highlighted

However recently the files and code were removed from access without any explanation

Accountability
3/5

The model was tested and analysed in both a mathematical and statistical manner and long-term improvements were highlighted. The public is invited to use and reach out for information and feedback

However, no scrutiny or oversight mechanism has been officially setup

Fairness
4/5

The model was studied from an explainability perspective to understand any biases. All feedback marked as spam is still reviewed by humans

However, it is not clear as to cases where the model is inaccurate are handled and how redressal is done for citizens

Action Points (UK Data Ethics Framework)

There are 5 Action Points mentioned in UK Data Ethics Framework. *(UK Data Ethics Framework)*[9] Specific data points guide different stages of project development and practical considerations. They are:

- 1. Define and understand the public benefit and user needs **(Clarity on needs specification)**
- 2. Involve diverse expertise **(Quality planning and gaining insights)**
- 3. Comply with the law **(Be systematic and sincere)**
- 4. Review the quality and limitations of the data **(Be realistic and analytical)**
- 5. Evaluate and consider wider policy implications **(Look at a bigger picture but with detail)**

Applied Action Points on the Case

1. The way they collected the feedback data from users through GOV.UK domain fulfils the first action as the feedback system and its collected data are there to understand **users' needs and ways, they can improve** their guidance to provide wider public benefit. (*Nightingale & Ansell, 2021*)[10]

2. Data generating pipeline was built with a team of a **data engineer, analyst, researcher support and some developers**, each contributing their expertise to result in a robust data infrastructure.

3. Not enough information but, when they **shared** their data with their colleagues for their **collaborative expertise and support**, it was in an **anonymized** format, which means abstracting all the personal information of the feedbackers; following the project's compliance with GDPR and DPA 2018. (*Legislation and codes of practice for use of data* (2020))[11] (Reilly, F. (2022))[8]

4. Use of multiple metrics for the model's performance evaluation while experimenting with different modelling techniques to get the best results. Data is also available for open-source dev **(Point 4.6)** [1] to cover data-specific limitations mentioned in *Nightingale & Ansell, 2021.* [10] (Rao et al., 2021)[12]

5. In this case, they have stated under '**Using user feedback to improve GOV.UK**' section of the *Nightingale & Ansell, 2021* [10] that they will be improving GOV.UK domain and its services in terms of presentation of guidance to reduce confusion over what people should do currently versus what to expect in the future, as new changes come from the user feedback.

Queen Mary University of London

Abhijeet's Self-Reflection

1. The UK government collects and analyzes vast amounts of data from various sources to improve service delivery and policy outcomes.
2. Data science techniques, such as machine learning and natural language processing, can be used to identify patterns and insights in large datasets.
3. The UK government is using data science techniques to reduce spam comments on Gov.uk, improving the quality of feedback and allowing officials to focus on genuine feedback.
4. Data science techniques can provide valuable insights into citizen behavior and preferences, helping governments to design policies and services that better meet the needs of citizens.
5. Data science techniques can also help automate some tasks, freeing up resources and allowing government officials to focus on higher-value work.

Queen Mary University of London

Arvind's Self-Reflection

1. It helped me improve my teamwork skills.
2. It enhanced my critical thinking.
3. I gained knowledge about UK ethics and law.
4. I learned about machine learning and data analysis (through the case study).
5. By meeting deadlines and prioritizing tasks, I have improved my time management skills.
6. Developed leadership skills such as delegation, communication, and conflict resolution.
7. By working on a group project, I have gained experience collaborating with others and learning how to work towards a common goal.
8. Working with others from different backgrounds and with different experiences exposed me to new perspectives and ideas.

Ayushi's Self-Reflection

1. The best aspect about working together as a group over the past few weeks was the excellent team dynamics we as a group shared. Firstly, we divided the task so everybody had equal amounts of contribution. Each member completed their task smoothly and were always available if someone else in the group needed any support in their part of the work. This certainly helped in enhancing my teamwork qualities, and it was a beneficial experience.

2. My task was to work on the question 2 with Radhika, which gave me an insight into data visualisation as a conept and the equivalent accelerator programme which is organised by the Office for National Statistics (ONS), Data Science Campus. This contributed towards my technical skill set, which I am sure I would be able to implement in my future endeavours.

3. I have always been a bit weak when it comes to communicating with people. So, this project really helped me work on improving it as well. Planning on how to get tasks done, how to go ahead with the presentation, all these really enhance my communication skills. Hence, overall this project was a fun and educating experience.

Queen Mary University of London

Gargeya's Self-Reflection

1. Designing has always been an intrinsic skill of mine. Through this presentation, I am happy I was able to use my skills and learn a few more techniques along the way to uplift the quality of our team's work as elegant, clear and professional. The world is largely governed by how something is presented (**perceived**) before what it actually is. Content-wise, I contributed in simplifying the case and especially solving the 5th question.

2. It has been a long time since I worked with a team this big to accumulate a single work. Towards the end, I can confidently say that teamwork teaches you a lot, not just all the positive aspects of working with people but the adjustment to their workflow, patience, people and work management, altruism and harnessing each individual's strength into the best result possible. In my opinion, that's what leadership is all about: not just you but the people around you. Overall, I am glad to have gotten this experience for my personal growth.

3. My motive to take on this module was to understand what the world is doing about the legal, and managerial side of the development of artificial intelligence. While doing my own research for each of the questions I found a lot of answers that I was looking for. I am satisfied with the efforts taken until now by the EU and UK, but a lot of jurisdiction-based decision needs to be discussed and established before the mass usage of these intelligent agents among us. The case study was a good narrative for validating my understanding of the current ethical situation in the domain of A.I.

Jayesh's Self-Reflection

1. Our Data Ethics group project was a fantastic learning opportunity, and I had the pleasure of working with an amazing team. One of the most important lessons learned was the value of collaboration and effective teamwork. Our team was able to complete the project with exceptional results thanks to active participation and open communication.

2. As a data analyst, this project assisted me in developing critical skills for success, such as data visualisation, interpretation, and analysis. The project also provided me with a thorough understanding of Data Ethics principles and how to apply them in real-world scenarios.

3. We evaluated a case study as a group using the UK Government's Data Ethics framework to identify potential ethical issues. We then worked together to devise practical solutions to these concerns, demonstrating our problem-solving and teamwork abilities.

4. Overall, I am grateful for this project because it allowed me to hone important data analysts skills such as teamwork, communication, and problem-solving, as well as gain a better understanding of Data Ethics. Working with such an amazing team made the experience even more rewarding, and I'm excited to put these skills to use in future projects.

Radhika's Self-Reflection

1. Over the past few days, I have had the opportunity to work with Ayushi, Gargeya, Jayesh, Arvind, and Abhijeet on a group assignment focused on "Pharmacies, people and ports: the Data Science Accelerator". The purpose of this assignment was to develop our knowledge and ethics in the data science accelerator program. I worked with Ayushi on question no. 2. I was responsible for answering the definition of data visualization and its uses.

2. Working with my team members was a great experience. And we had a good time working together. Thank you for this opportunity. To assign tasks and roles, and set deadlines for each question, we scheduled meetings. Ensuring that we understood every section and successfully contribute to the project's success.

3. My experience working on this project was good. I found this task engaging and informative. I believe that working with The Data Accelerator Program has given me better knowledge.

Queen Mary University of London

References

1. *Madeline Lasko. (2017). "Pharmacies, people and ports: the Data Science Accelerator". GOV.UK. Available at: https://dataingovernment.blog.gov.uk/2017/08/11/pharmacies-people-and-ports-the-data-science-accelerator/ (Accessed: March 15, 2023).*
2. *"Launching the world's first data science & AI academy in retail". Marks & Spencer. Available at : https://corporate.marksandspencer.com/launching-worlds-first-data-science-ai-academy-retail (Accessed: March 15, 2023).*
3. *Jeff Feng, Erin Coffman & Elena Grewal. (2017). "How Airbnb Democratizes Data Science With Data University". Airbnb. Available at : https://medium.com/airbnb-engineering/how-airbnb-democratizes-data-science-with-data-university-3eccc71e073a (Accessed: March 15, 2023).*
4. Brush, K. and Burns, E. (2022) *What is data visualization and why is it important?*, *Business Analytics*. TechTarget. Available at: https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization (Accessed: March 23, 2023).
5. *Introduction to the data science and Data Visualisation Accelerator Programmes* (no date) *GOV.UK*. Available at: https://www.gov.uk/government/publications/data-science-accelerator-programme/introduction-to-the-data-science-accelerator-programme (Accessed: March 23, 2023).
6. *Inception explained* (no date) *Inception Explained*. Available at: https://www.inception-explained.com/ (Accessed: March 23, 2023).
7. Narang, M. (2023) *Most interesting data visualization projects in 2023*, *KnowledgeHut*. Knowledgehut. Available at: https://www.knowledgehut.com/blog/business-intelligence-and-visualization/data-visualization-projects (Accessed: March 23, 2023).
8. Reilly, F. (2022) *How we are using machine learning to detect GOV.UK feedback spam*, *Data in government*. Available at: https://dataingovernment.blog.gov.uk/2022/10/03/how-we-are-using-machine-learning-to-detect-gov-uk-feedback-spam/ (Accessed: March 15, 2023).
9. *Data Ethics Framework* (2018) *GOV.UK*. Available at: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020 (Accessed: March 15, 2023).
10. Nightingale, V. and Ansell, I. (2021) *Working with user feedback during COVID-19*, *Inside GOVUK*. Available at: https://insidegovuk.blog.gov.uk/2021/08/16/working-with-user-feedback-during-covid-19/ (Accessed: March 15, 2023).
11. *Data Ethics Framework: Legislation and codes of practice for use of data* (2020) *GOV.UK*. Available at: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-legislation-and-codes-of-practice-for-use-of-data (Accessed: March 15, 2023).
12. Rao, S., Verma, A.K. and Bhatia, T. (2021) "A review on Social Spam Detection: Challenges, open issues, and Future Directions," *Expert Systems with Applications*, 186, p. 115742. Available at: https://doi.org/10.1016/j.eswa.2021.115742.