

# Report: Design, Development, and Management of a Big Data Visualization Product Project

## Introduction

The field of data visualization has become increasingly important in the realm of big data analysis, particularly in contexts where clear communication of complex data is essential. As a data visualizer in a data science start-up company, the task at hand involves designing and developing a bespoke big data visualization product. The product is intended to serve junior government officials who have limited knowledge of data science or data visualization technologies. These officials need to install, deploy, and use the product to visualize a compelling data story for professional data-driven journalists working in public media. The success of this visualization product hinges on the ability to convey intricate data insights in a manner that is both accessible and impactful.

The project's design and development will be informed by a series of assumptions and requirements. The theme of the data and the purpose of the storytelling will be tailored to the selected audience's needs, drawing on the creator's individual interests and professional background. This report will critically discuss and evaluate the design stage of the data visualization product, serving as a blueprint for the proof-of-concept prototype to be delivered in Assignment Two.

## Data / Theme Selection

### Theme: Public Health and Safety

Public health and safety, particularly in the context of the COVID-19 pandemic, is an area of significant public and governmental interest. The chosen theme for this visualization project is therefore centred on COVID-19 data, including daily case numbers, vaccination rates, and hospital admissions over a three-month period. This focus is selected due to the still ongoing relevance of the pandemic, although it peaked in 2021. Additionally, to explore profound impact of the pandemic on public health policies and individual behaviours.

### Data Source:

- **Primary Source:** [data.gov.uk](https://data.gov.uk)
- **Specific Data Sets:** The datasets include daily COVID-19 statistics, vaccination data, and hospital admission rates.

### Pre-processing Requirements:

1. **Data Cleaning:** This involves handling missing values, identifying and correcting outliers, and resolving inconsistencies within the data to ensure accuracy and reliability.
2. **Data Transformation:** Aggregating daily data into weekly summaries to facilitate easier analysis and visualization, which helps in identifying trends over time.
3. **Data Integration:** Combining datasets from different sources to create a comprehensive view of the pandemic's progression, including linking case numbers with vaccination rates and hospital admissions.

The selection of this theme and data is driven by the need to provide actionable insights into the public health crisis. Visualizing this data can help government officials and journalists understand the effectiveness of interventions and communicate this information to the public effectively.

### Audience Requirements Analysis

The visualization product must satisfy both functional and non-functional requirements to effectively communicate the data narrative to the intended audience.

#### Functional Requirements:

1. **Interactive Dashboards:** The product should allow users to filter and explore data through interactive elements such as date range sliders, geographic filters, and category selectors.
2. **Visual Representations:** The visualizations should include line charts to show trends over time, bar charts for comparisons, and heat maps for geographic data visualization.
3. **Downloadable Reports:** Users should be able to export visualizations and the underlying data for further analysis and reporting, facilitating the dissemination of insights.

#### Non-Functional Requirements:

1. **Usability:** The interface should be intuitive and easy to navigate, particularly for users with limited data science knowledge. This involves clear labelling, straightforward navigation paths, and user-friendly design.
2. **Performance:** The application should have fast loading times and responsive interactions to ensure a smooth user experience, even with large datasets.
3. **Accuracy:** the dashboard interface will ensure that input commands lead to accurate results, thus each click produces the desired result

**Requirement Identification and Development:** The requirements were identified through stakeholder interviews and analysis of typical use cases. Junior government officials and

journalists were consulted to understand their needs and preferences. Functional requirements focus on the need for interactive and customizable data views, while non-functional requirements emphasize the importance of a user-friendly and accessible interface. Each requirement is justified based on its contribution to achieving the project's goal of clear and effective data communication.

## Data Storytelling and Process Design

As the big data in healthcare is increasing in volume, data visualization is becoming an important factor to aid users digest information easily. It is a well-known fact that the human brain processes visual information more quickly when compared to text. “A picture is worth a thousand words” is an often-cited statement. A compelling data story is crucial for making the visualizations meaningful and impactful. The narrative built around the COVID-19 data will highlight the progression of the pandemic, the effects of vaccination campaigns, and the strain on healthcare resources over the three-month period.

### Story Elements:

1. **Pandemic Progression:** Visualizing the rise and fall of COVID-19 cases over time, identifying peaks and troughs in the data, and correlating these with public health interventions.
2. **Vaccination Impact:** Demonstrating the relationship between vaccination rates and the number of new cases, showcasing the effectiveness of vaccination efforts in controlling the spread of the virus.
3. **Healthcare Strain:** Highlighting hospital admissions and capacity, illustrating how the healthcare system is coping with the influx of patients during peak periods.

**Impact:** The visualization aims to provide insights into the effectiveness of public health measures and inform policy decisions. By clearly communicating key facts, such as the reduction in case numbers and hospitalizations due to vaccinations, the visualization supports the continued importance of public health interventions and encourages compliance with health guidelines.

The storytelling aspect of the visualization will help contextualize the data, making it easier for users to understand the broader implications. For example, linking spikes in cases with delays in vaccination rollout or lifting of restrictions can provide valuable lessons for future policymaking.

## Software Architecture Design

### Tools and Technologies:

1. **Data Processing:** Python, using libraries such as Pandas and NumPy, will be employed for data cleaning, transformation, and integration. These libraries are chosen for their robustness and ease of use in handling large datasets.
2. **Visualization:** Tableau will be used to create interactive and visually appealing dashboards. Tableau is selected for its powerful visualization capabilities, ease of use, and ability to handle complex data with interactive features.
3. **Deployment:** Streamlit is an open-source Python library that simplifies the process of creating and sharing interactive web applications for data science and machine learning. It is designed to be intuitive and easy to use, requiring minimal web development knowledge.

### Rationale for Tool Selection:

- **Python:** Python is a versatile programming language with extensive libraries for data manipulation and analysis, making it ideal for processing large datasets efficiently.
- **Tableau:** Tableau is renowned for its user-friendly interface and powerful visualization tools, enabling the creation of interactive and aesthetically pleasing dashboards that can convey complex data insights effectively.
- **Streamlit:** Streamlit will be used to develop a web application to host and share the visualizations. Streamlit is very intuitive and requires minimal web development knowledge. It is designed to be easy for data scientists and analysts who are more familiar with Python than web development. This makes Streamlit suitable for rapidly building and deploying interactive data applications and dashboards, allowing for quick iteration and sharing of insights.

### Product Case Use Specifications

The visualization product is designed to be user-friendly and accessible, enabling junior government officials to effectively use the system to gain insights from the data.

1. **Interaction:** Users will interact with the dashboards to filter data by date, region, or other relevant parameters. Interactive elements such as sliders and drop-down menus will allow users to customize their views.
2. **Analysis:** Users can analyse trends, compare data points, and generate reports based on their findings. The interactive nature of the dashboards will facilitate in-depth exploration of the data.
3. **Reporting:** Visualizations and underlying data can be exported for further use in reports, presentations, and decision-making processes. This feature ensures that the insights gained from the visualizations can be effectively communicated to a broader audience.

## Conclusion

This report outlines the plan and progress for designing a big data visualization product focused on COVID-19 data. The chosen theme of public health and safety is crucial for informing public policy and communicating important information to the public. By adhering to both functional and non-functional requirements, the visualization product aims to be intuitive, informative, and accessible. The next phase will involve developing a proof-of-concept prototype to be delivered in Assignment Two. The detailed design and thoughtful selection of tools and technologies set a solid foundation for creating a valuable and impactful data visualization product.

## References

1. Few, S., 2013. *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring*. Analytics Press.
2. Zhang, Y., Reynolds, M., Lugmayr, A., Damjanov, K. and Hassan, G.M., 2022. A Visual Data Storytelling Framework. *Informatics*, 9(4), p.73.
3. McCandless, D., 2014. *Knowledge is Beautiful*. HarperCollins UK.
4. Kirk, A., 2019. *Data Visualisation: A Handbook for Data Driven Design*. Sage Publications.
5. Yau, N., 2013. *Data Points: Visualization That Means Something*. Wiley.
6. Ware, C., 2013. *Information Visualization: Perception for Design*. Elsevier.
7. Microsoft Corporation, n.d. Power BI. [Online] Available at: <https://powerbi.microsoft.com/en-us/data-visualization-vs-storytelling/> [Accessed 20 February 2023].