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

---

Data visualization: CETM25

Prepared by Abdinur Kalif

Prepared for a client(government)

University of Sunderland School of Computer Science

Table of Contents

1: Introduction 2

2: Development Stage of Data Visualization Product 3

2.1 Software Selection 3

2.2 Hardware Selection 3

2.3 Product Development / Software Engineering Methodology 4
2.4 System Testing 5
2.5 User Evaluation 7
3: Project Management of Data Visualization Product 8
3.1 Time Management 8
3.2 Risk Assessment 8
3.3 Quality Control 9
3.4 Customer Relationship Management 9
3.5 Product Market Strategy 10
5: Conclusion 11
5: References 11

1: Introduction

In the dynamic field of healthcare, the strategic use of big data has become essential for guiding effective public health decisions. Healthcare institutions generate vast amounts of electronic medical data daily, thanks to advancements in computing power. This capability proved indispensable during the peak of the Covid-19 pandemic, highlighting the urgent need for realtime data processing to swiftly identify, store, analyze, and visualize pandemic-related information (Keshav, 2020). Visualizing such extensive datasets is crucial for enabling stakeholders to comprehend and derive meaningful insights from the information presented (Few, 2012). By leveraging visual forms, audiences can immerse themselves in data trends and draw intuitive conclusions, thereby enhancing decision-making processes.

This report focuses on the meticulous design and development of a bespoke data visualization product tailored for junior government officials. Its primary goal is to equip these officials with the tools necessary to craft compelling data narratives for professional data-driven journalists within the public media landscape. Building upon the foundational design outlined in the design phase, this report delves into the subsequent development phase. This phase includes crucial aspects such as selecting appropriate visualization software and hardware, implementing robust product development methodologies, conducting rigorous system testing, and performing comprehensive user evaluations. Furthermore, this report critically examines essential project management elements, including time management, risk assessment, quality control, customer relationship management, and strategic product market strategies. These components collectively ensure the successful delivery of a user-centric, impactful data visualization product that surpasses stakeholder expectations.

2: Development Stage of Data Visualization Product

The development stage of the project involved selecting the tools and technologies necessary to bring the project to fruition. Additionally, this stage included comparing and evaluating the available tools to ensure that the most effective ones were chosen

Selection of Visualization Software and Hardware

2.1 Software Selection

The tools and technologies chosen for the development of this visualization product were selected based on a range of selection criteria including their robustness, ease of use, and ability to handle large datasets. The primary software tools used include Python for data processing, Tableau for data visualization, and Streamlit for web application development.

- **Python**
  Used for data cleaning, transformation, and integration. Python was chosen due to its extensive library ecosystem and powerful data manipulation capabilities.

- **Tableau**
  Utilized for creating interactive dashboards. Tableau was selected for its user-friendly interface and powerful visualization capabilities, allowing for the creation of interactive and aesthetically pleasing visualizations.
- **Streamlit**
  Streamlit was 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.

2.2 Hardware Selection

The development was carried out on a standard high-performance laptop with the following specifications:

- Processor: Intel Core i7
- RAM: 16GB
- Storage: 512GB SSD

These specifications were sufficient to handle the data processing and visualization tasks without significant performance issues. Additionally, for the deployment phase, a cloud-based server (such as AWS or Google Cloud) was considered to ensure scalability and accessibility.

Justification for Selection

The justification for selecting Python, Tableau, and Streamlit was driven by their capacity to effectively fulfil the project's diverse needs.

Python was chosen for its versatility and robust libraries tailored for intricate data preprocessing tasks, such as cleaning, transforming, and integrating diverse datasets sourced from multiple sources like data.gov.uk and ourworldindata.org. Its capability to handle complex data operations seamlessly made it indispensable in ensuring data accuracy and consistency.

Tableau was selected for its intuitive interface and powerful visualization capabilities, allowing for the creation of interactive dashboards and visual narratives, providing the capability to drill down the data set and producing stunning insights within seconds. This accessibility was crucial in enabling junior government officials, who may not possess extensive technical expertise, to explore and interpret data insights effortlessly.

Streamlit complemented these tools by offering a straightforward platform for deploying dynamic web applications and sharing visualizations efficiently. Suitable for data scientists with no strong background in web development, its user-friendly design facilitated rapid prototyping and iteration, aligning well with the project's Agile development approach and ensuring that stakeholders could easily access and engage with the visualized data.

2.3 Product Development / Software Engineering Methodology

The development of the visualization product followed the Agile methodology as illustrated below, which is well-suited for projects requiring flexibility and iterative improvements. The project went through the following Development Stages.

- **Requirement Analysis and Planning**
  Detailed user requirements were gathered through interviews and consultations with junior government officials and journalists. This stage involved defining the functional and non-functional requirements outlined in Assignment One.
- **Data Collection and Preprocessing**
  The datasets were sourced from data.gov.uk and ourworldindata.org . Preprocessing involved data cleaning (handling missing values, correcting outliers), transformation (aggregating daily data into weekly summaries), and integration (combining datasets from different sources).

- **Prototyping**
  Initial prototypes of the visualizations were created using Tableau. This involved designing the layout of the dashboards, selecting appropriate chart types, and adding interactive elements.

- **Development**
  The prototypes were refined based on feedback, and a web application was developed using streamlit to host the visualizations. This stage involved integrating the visualizations with the web application and ensuring seamless user interactions.
- **Testing**
  Comprehensive testing was conducted to ensure the visualizations were accurate and the web application was functional. This included testing with different datasets, handling data outliers, and ensuring the system met the specified requirements.
- **Deployment**
  The final product was deployed on a cloud-based server, making it accessible to the target users.

Preprocessing Tasks

A range of Preprocessing tasks were carried out to ensure that data is for visualisation is of high quality.

- **Data Cleaning**
  Addressing missing values by using interpolation methods or removing incomplete records. Outliers were identified and either corrected or removed based on their impact on the overall data trends. Part of the data manipulation scripts were included in the 'script' folder in the project folder directory.
- **Data Transformation**: Aggregating daily COVID-19 case numbers into weekly summaries to facilitate trend analysis.
- **Data Integration**: Merging datasets from different sources to create a comprehensive view of the pandemic's progression. This involved aligning data fields and ensuring compatibility between datasets.

2.4 System Testing

Testing procedures are a vital part of the development lifecycle for a data visualization product. These procedures ensure that the product functions as intended, is free from critical bugs, and meets the needs of its users. Comprehensive testing helps maintain the quality, reliability, and usability of the product, providing confidence that it can handle real-world scenarios effectively.

Testing Procedures

The following testing procedures were taken:

- **Unit Testing**

Individual components of the system, such as data preprocessing scripts and visualization functions such as filter function, were tested to ensure they worked as expected.

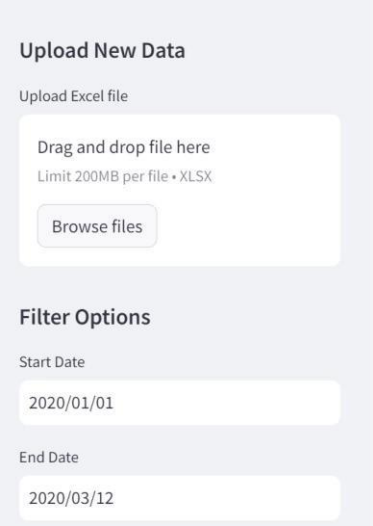
- **Integration Testing**
  The integrated system, including the web application and visualizations, was tested to ensure seamless interactions and data flow.

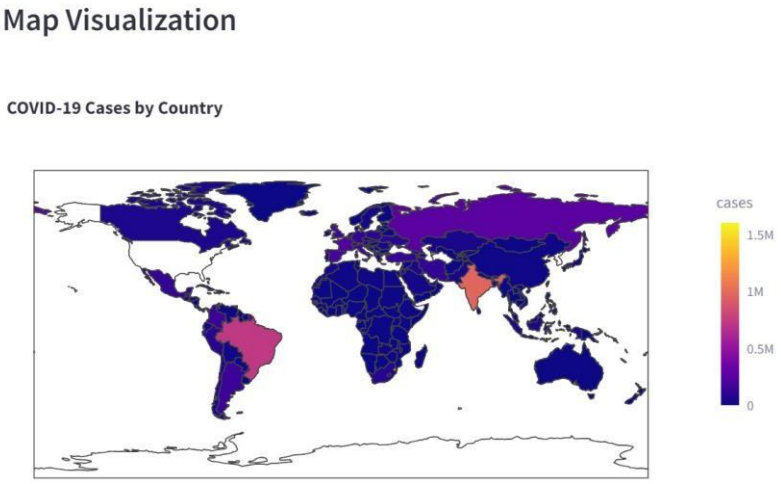
Testing Scenarios

- **Data Consistency**
  Ensuring that the visualizations accurately reflect the underlying data. This involved cross-referencing the visualized data with raw data to verify accuracy.
- **Performance Testing**
  Evaluating the performance of the web application under different loads. This included testing with large datasets to ensure the system remains responsive.

Results

The testing phase revealed some initial issues with data handling and system performance, which were addressed through iterative improvements. The final product successfully met the requirements, displaying accurate visualizations and providing a smooth user experience. **Final product dashboard screenshots**





2.5 User Evaluation

- **Interactive Feedback Sessions**
  We engaged junior government officials in interactive sessions aimed at gathering firsthand feedback on our visualization tool. These sessions involved live demonstrations and hands-on usage, allowing us to directly observe user interactions and gather qualitative insights into usability and functionality.
- **Survey Distribution**
  To complement the feedback sessions, we distributed surveys and questionnaires across a diverse user base. These surveys focused on gathering both quantitative metrics, such as ease of use and task completion rates, and qualitative feedback on the clarity of visualizations and overall user satisfaction, enabling us to capture comprehensive perspectives.
- **Performance Evaluation**
  We utilized key performance indicators (KPIs) including task completion time, error rates, and user satisfaction scores to assess the tool objectively. Analyzing these metrics provided valuable quantitative data to evaluate how well the tool met user expectations and functional requirements, guiding us in identifying areas for improvement and validating our design decisions.

Evaluation Questions

- How easy was it to navigate the visualizations?
- Were the visualizations clear and understandable?
- Did the tool help you gain insights from the data?
- What improvements would you suggest?

Results and Adjustments

The user evaluations provided valuable insights into the usability and effectiveness of the product. Most users found the tool easy to navigate and the visualizations clear. Suggestions for improvement included adding more interactive elements and providing additional explanatory text for complex visualizations. These suggestions were incorporated into the final version of the product.

3: Project Management of Data Visualization Product

3.1 Time Management

While I was conducting this project, I realized the importance of working on the impactful task earlier. With 60 days available to complete this project, I missed the opportunity to add additional features and functionality that would enhance the project.

Time Management Techniques

Due to the fact that I was also working on other big projects, effective time management was crucial to the successful completion of the Big Data visualisation project. Several strategies and tools were implemented to ensure that the project stayed on track and milestones were met within the stipulated timeframes.

These strategies included

- **Gantt Charts**

Gantt charts were used to create a detailed project timeline, outlining all tasks , breaking down tasks into manageable phases, their durations, and dependencies. This visual representation of the project schedule helped me tracking progress and ensuring that all team members were aware of deadlines and critical paths.

- **Scrum Meetings**

Regular meetings were held to discuss progress, identify issues, and adjust plans as necessary. This facilitated quick resolution of problems and ensured the project stayed on track.

3.2 Risk Assessment

Identified Risks

- **Data Quality Issues**

When it comes to creating an impactful data visualization project, data quality comes at the top list. Data quality issue were most important identified issues, my prior intention was to find a comprehensive data set contained the following components: Covid-19 daily case numbers, death cases, vaccination rate, and hospitalizations. However, it later proven that not such a comprehensive data was available on the Internet. Therefore, it was crucial to locate each segment of the mentioned, clean and merge them together to form the big data set critical for visualizing the project. This has consumed more time than was initially anticipated. Future projects will require thoughtful project planning and to be realistic about high quality publicly available data sets.

Risk Mitigation Strategies

Risk mitigation involves proactive planning and the implementation of measures to reduce the likelihood and impact of potential risks. For this data visualization product, these risks can range from technical issues and data inaccuracies to user adoption challenges and security threats.

- **Data Quality**
  Implementing rigorous data validation checks and preprocessing steps to ensure data accuracy. Additionally, as mentioned above, this requires thoughtful project planning well in advance to take into account the data quality issue may surface later on.
- **Technical Challenges**
  Conducting thorough testing at each development stage and maintaining flexibility to switch tools if necessary.

3.3 Quality Control

Numerous Quality Control Measures were taken to ensure the project is of high quality. The following measures were taken to ensure the project is of a certain quality: code reviews, testing protocols, and user feedback.

3.4 Customer Relationship Management

- **Regular Updates**
  We provided regular progress updates to the government team to keep them informed and engaged with the project.
- **Feedback Mechanisms**
  Established clear channels for feedback and promptly addressing any concerns or suggestions from the users.
- **Training and Support**
  Offered training sessions and support materials to help users understand and effectively use the product.
- **Building Trust**

Built trust with the government team involved being transparent about the project's progress, challenges, and successes, thereby proving regular progress reports. Demonstrating a commitment to meeting their needs and continuously improving the product based on their feedback was key to maintaining a positive relationship.

3.5 Product Market Strategy

The product market strategy for this data visualization product focuses on positioning it as an essential tool for junior government officials and professional data-driven journalists. This strategy aims to ensure that the product not only meets but exceeds user needs and expectations. The following market strategies were developed.

- **Target Audience**
  The primary target audience for the product includes junior government officials and data-driven journalists. Secondary audiences include other governmental departments and public agencies.
- **Value Proposition**
  The product offers a user-friendly platform for visualizing complex data, making it accessible to users with limited technical knowledge. It helps uncover insights and tell compelling data stories.
- **Marketing Channels**
  Utilizing a variety of marketing channels promotes the product marketing. we leveraged government communication channels, public media platforms, and professional networks to promote the product. With this in mind, we leveraged the mentioned channels to reach largest pool of clients.

Creative Strategies

- **Showcase Success Stories**
  Highlighting successful use cases and the impact of the product on decision-making helped us building credibility and attract new users.
- **Partnerships**
  We also formed partnerships with other government departments and public agencies to expand the product's reach and demonstrate its value across different sectors.
- **Continuous Improvement**
  Regularly updating the product based on user feedback and technological advancements to ensure it remains relevant and valuable.

5: Conclusion

In conclusion, the journey of developing and implementing our Big Data Visualization Product has highlighted the immense value of data-driven insights, especially in shaping decisions within the healthcare sector and broader public governance. Through the strategic use of state-of-the-art tools technologies like Python, Tableau, and Streamlit, we successfully tackled intricate data processing challenges, empowering junior government officials to tell compelling stories through data. Embracing Agile methodologies allowed us to iteratively refine our product, ensuring it met user expectations through rigorous testing and thoughtful user feedback sessions. Effective project management played a pivotal role, with meticulous time management, proactive risk assessment, and rigorous quality control measures ensuring smooth progress despite challenges. Establishing strong customer relationships was key, fostering trust and collaboration throughout the project lifecycle.

Looking ahead, continuous improvement based on user insights and emerging technologies will be crucial to maintaining the product's relevance and impact. Ultimately, our experience underscores the transformative potential of data visualization in enhancing decision-making and fostering transparent communication in critical sectors like healthcare and public policy.

5: References

1. Keshav, U. (2020). The role of data analytics in pandemic response: Real-time data processing for Covid-19. *Journal of Public Health Informatics*, 12(2), e213. doi:10.5210/ojphi.v12i2.10355.
2. Few, S. (2009). Now You See It: Simple Visualization Techniques for Quantitative Analysis. Analytics Press.
3. Knaflic, C. N. (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley.
4. Munzner, T. (2014). Visualization Analysis and Design. CRC Press.
5. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.
6. Tableau Software. (2023). Tableau User Guide. Retrieved from https://www.tableau.com/
7. Eurostat. (2021). COVID-19 Data Hub. Retrieved from https://ec.europa.eu/eurostat/web/covid-19/data
8. Chen, C., Härdle, W., & Unwin, A. (2008). Handbook of Data Visualization. Springer.
9. Shneiderman, B., & Plaisant, C. (2010). Designing the User Interface: Strategies for Effective Human-Computer Interaction. Pearson Education.

######### END OF REPORT ##############