Mapping Museums:

Exploring Collections Data and KPI Trends

1. Introduction

1.1 Overview

National museum collection and footfall data offer valuable insight into cultural representation and public engagement. This project examines sculptures in British museums, identifying cultural, spatial and temporal patterns and providing evidence for historic and future visitor trends. By analysing sculptures on display at the V&A and the British Museum and profiling each item by attributes such as age, production place and material, it presents a broad view of how global cultures are represented in these collections. Data was sourced through several methods, including V&A records retrieved via API and British Museum records extracted as CSV.

Monthly and annual KPI data (footfall and income) collected by the Department for Digital, Culture, Media and Sport (DCMS) enables analysis of national museum footfall from 2004 to 2025. By tracking annual attendance, identifying seasonal patterns and developing predictive models, the project builds a clear picture of changing visitor behaviour to support data-driven planning and decision-making.

1.2 Background

The target audience includes DCMS and museum professionals who use data to understand public access to collections, cultural representation and museum management. The report provides insight into which cultures and regions are over or underrepresented in museum collections, highlights areas of poor or inconsistent documentation to support improvements in research and data management, and identifies short and long-term visitor trends to inform resource planning, exhibitions and engagement strategies. The project prioritised highly visual and accessible outputs to ensure that non-technical teams could use them easily and draw clear insights.

1.3 Research Questions

To guide the analysis and ensure a focused approach across the collections and visitor data, the project addressed the following research questions:

1. How are cultures, regions, materials and techniques represented across sculpture collections, and how have these changed over time?
2. How have visitor numbers to national museums changed over the past two decades?
3. What relationship can be observed between visitor attendance and museum income?
4. Can future museum footfall be forecast using time series models?

2. Data Collection

2.1 Required Information

The project required object-level metadata for sculptures, including production date, place, materials, techniques and display status. For visitor trends, monthly and annual footfall, income and KPI data were required. For the analytical stages, the team identified a small set of priority fields that would drive the core findings. These included production place, production date, material, technique and display status for the collections data, and annual footfall and income for the KPI data. Prioritising these fields provided a clear analytical direction and ensured that the subsequent cleaning workflows remained focused and consistent.

2.2 Available Information

The V&A provided API access to structured object data. The British Museum offered downloadable object records in CSV format. Tate collections and exhibition data were publicly accessible via GitHub and Figshare. DCMS published annual KPIs and monthly visitor statistics.

2.3 Data Sources

- V&A Collections API
- British Museum object CSV extracts
- Tate GitHub collections dataset
- Tate temporary exhibitions dataset from Figshare
- DCMS Key Performance Indicators dataset
- DCMS monthly footfall dataset (multiple sheets)

2.4 Data Collection Method

The V&A data was retrieved using structured API calls specifying object type and filtering conditions. British Museum and KPI datasets were downloaded as CSV or Excel files from their respective portals. Tate data was cloned directly from the GitHub repository. All sources were imported into Jupyter notebooks for cleaning. The API responses, CSVs and Excel documents were consolidated into pandas dataframes for transformation and integration.

3. Specifications And Design

A detailed architecture and workflow diagram, illustrating the full data pipeline from ingestion through cleaning, integration, analysis and modelling, is provided in the project's GitHub repository and can be found in the README file, under 'Project Workflow'.

3.1 Data Analysis Approach

3.1.1 Technical Requirements

The analysis was required to:

- Use Python with Pandas, NumPy and Matplotlib as core analytical tools.
- Incorporate at least one API as a data source.
- Clean, standardise and integrate datasets from multiple institutions.
- Develop at least one predictive model using an appropriate time series method.
- Maintain reproducibility through version control and documented workflows.

3.1.2 Non-technical Requirements

Alongside the technical work, the project needed to:

- Present insights in a clear and accessible way for museum and policy stakeholders.
- Handle cultural and historical information cautiously, avoiding overinterpretation.
- Retain full data provenance so that all cleaning steps remain auditable.
- Prioritise visual clarity to support decision-making for non-technical audiences, using formats that highlight patterns at a glance and reduce the need for specialist interpretation.

4. Implementation And Execution

With the project requirements established, work began with preprocessing across all datasets to create consistent inputs. Column names and values were standardised, duplicates resolved and datasets merged to support reliable analysis. Once cleaned, the data enabled a range of techniques, including choropleth mapping, heatmaps and initial machine learning models. Collections analysis focused on production dates, places, materials and techniques, while footfall and KPI analysis centred on visitor numbers and income. For predictive modelling, the data was aggregated to total UK visitors to produce a clearer and more interpretable time series. The full implementation process is documented in the Jupyter notebooks stored within the project repository.

4.1 Team Roles

Across the project, each team member was assigned clear roles and responsibilities to ensure all components of the project were completed efficiently, with accountability maintained throughout. A breakdown of overall roles can be found below:

Name	Roles & Responsibilities
Angie	● **Data cleansing:** cleansed materials and techniques in collections data ● **Data visualisation:** timelines for materials and techniques
Diana	● **Data cleansing:** V&A collections data for locations and time periods ● **Documentation:** authored report section for timelines for materials and techniques (Angie's work)
Jenny	● **Data cleansing:** footfall data; standardising cell values and column names; handling missing values ● **Data visualisation:** heatmap and line charts for annual visitor numbers; polar chart for average monthly footfall ● **Predictive modelling:** Prophet time-series model forecasting five years of British Museum footfall ● **Documentation:** Slide deck and report editor
Jennifer	● **Data cleansing:** footfall data; pivot and melting data from financial year to calendar year ● **Data visualisation:** time series multi-plots and ML model predictions; mermaid flowchart ● **Predictive modelling:** multiple model testing; cross-validation; feature engineering; MAE, RMSE and % error rate ● **Documentation:** GitHub set up
Jo-Ann	● **Data cleansing and integration:** Tate collections and temporary exhibitions data ● **Data visualisation:** horizontal bar chart showing Tate temporary exhibitions over time by medium; updating colour scheme to ML visualisations
Nicky	● **Project Manager:** Created and managed project plan, monitored progress, ensured overall delivery ● **API Integration:** Retrieved and parsed V&S sculpture records using the museum's public API ● **Data cleansing and integration:** BM and V&A collections data and DCMS KPIs dataset ● **Data visualisation:** choropleth map for BM, V&A and combined showing distribution of sculpture origins; stacked bar/line graph comparing National Museum footfall vs income ● **Documentation:** Project README author, slide deck and report editor

4.2 Project Approach

Although the project was originally planned using a waterfall structure, in practice it developed into a hybrid approach incorporating several agile elements. The fixed deadline and limited weekly availability required a clear sequence of phases and defined responsibilities, broken down into smaller tasks with regular check-ins to adjust methods and resourcing. To maintain alignment across sub-groups, the team used shared integration checkpoints to ensure consistent naming conventions, data structures and cleaning decisions across the BM, V&A, Tate and KPI datasets. This allowed the team to retain flexibility while still working within an overall linear framework. The high-level project plan can be found as a Gantt chart in the project repository's README, under "Project Timeline".

As the project progressed, the scope evolved in a more iterative way. Once the core collections datasets were underway, the team identified capacity to incorporate the DCMS footfall and KPI datasets and other sources that had initially been set aside. The predictive modelling was also developed iteratively, refining the approach as cleaned data became available and intermediate outputs revealed new opportunities or constraints. These additions were incorporated only after the team reassessed feasibility and confirmed that expanding the scope could be achieved without affecting the quality or timely delivery of the core collections analysis.

The result was a practical blend of waterfall and agile principles, guided by a detailed roadmap that supported task allocation, milestone tracking and structured planning. Regular check-ins acted as sprint reviews, enabling the team to surface progress, adjust priorities and resolve blockers. Git branching and pull requests supported collaborative working and informal code review, and the overall scope remained adaptable as further datasets and opportunities were identified.

4.3 Tools and Libraries

Pandas and NumPy were used for data cleaning, transformation and integration across the collections, KPI and footfall datasets. Visualisations were produced with matplotlib and seaborn, including choropleth maps, heatmaps and combined bar and line charts. XGBoost and statsmodels supported the development and comparison of predictive models for visitor trends. Work was carried out in Jupyter notebooks, which provided an auditable environment for documenting cleaning steps, exploratory analysis and model development. The project notebooks detailing each working process can be found within the project repository.

Git and GitHub enabled version control and collaborative working across the sub-teams. Shared documentation platforms such as Google Docs, Sheets and Slides, together with communication tools including Slack and Zoom, supported coordination and consistent project management.

As a quick side project, Tableau was also used to produce an interactive dashboard designed for non-technical audiences, presenting key metrics from the collections and footfall datasets in a front-end format. The dashboard supports quick exploration of trends and complements the static visualisations included in this report.

Dashboard link: https://public.tableau.com/app/profile/nicky.cross/viz/MuseumDataDashboard/Dashboard1

The codebase was organised into separate Jupyter notebooks for data cleaning, integration, exploratory analysis, visualisation and modelling. Functions were grouped logically, with consistent naming conventions and inline comments describing each transformation. Intermediate datasets were saved at key stages to support reproducibility, and markdown cells were used throughout to explain purpose, assumptions and decisions.

4.4.1 Achievements and Successes

The team successfully cleaned, standardised and integrated several complex datasets, producing reliable inputs that supported clear and high-impact visualisations. A consistent visual style and defined colour palette helped make the findings accessible for non-technical audiences, meeting a core requirement of the project. The collections work produced interpretable timelines and geographic distributions, while the KPI and footfall analysis provided both long-term patterns and seasonal insights.

Predictive modelling, initially considered a stretch target, became a substantive element of the project. The footfall data was reshaped into a coherent time series, multiple models were tested, and XGBoost was refined through feature engineering and cross-validation. While not designed for precise operational forecasting, the modelling demonstrated the feasibility of estimating broad future visitor volumes.

Effective project management and collaborative working underpinned these achievements. Clear role allocation, regular communication and a hybrid waterfall–agile approach supported steady progress and allowed the team to expand the scope once capacity was identified. Git branching and pull requests facilitated collaborative coding and ensured consistent standards across cleaning, integration, visualisation and modelling.

4.4.2 Challenges

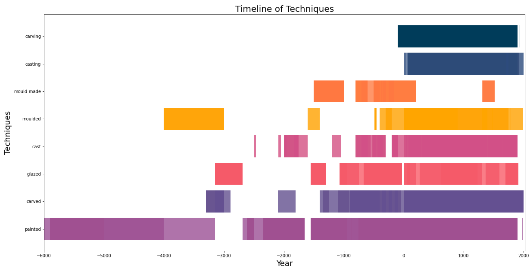
This project faced several data cleaning challenges across all datasets, largely due to inconsistent values, complex historical information and non-standardised formatting. Collections data required substantial work to standardise dates, places, materials and techniques. Historically complex or inconsistently recorded place names were normalised and mapped to modern countries using staged rules, with fallback fields used cautiously. Materials and techniques presented similar issues, with multi-value and inconsistent entries complicating categorisation. To ensure reliability, visualisations were restricted to objects with a single clearly defined material or technique, and all original fields were retained for auditability.

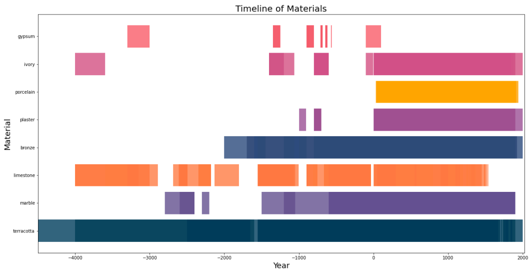
Within the KPI dataset, key tasks included standardising year formats, aligning museum names, correcting malformed or mis-scaled values and recalculating annual totals where data was incomplete. The footfall dataset required further restructuring, as monthly data spanned both financial-year and calendar-year formats. Pivoting, reshaping and aggregation were carried out to create a consistent annual structure and reduce null values.

In the machine learning stage, the main challenge was working with a dataset containing initially few usable features. Feature engineering improved performance, but structural breaks such as COVID introduced volatility. The resulting model is suitable for broad long-term volume estimates, though further development would be needed for more precise month-by-month forecasting.

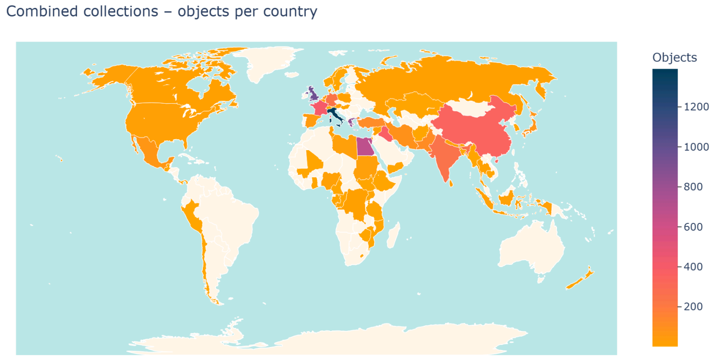
5. Results Reporting

5.1 Collections Data - British Museum & V&A Museum

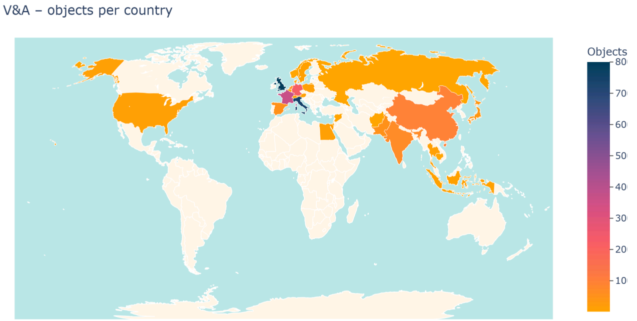


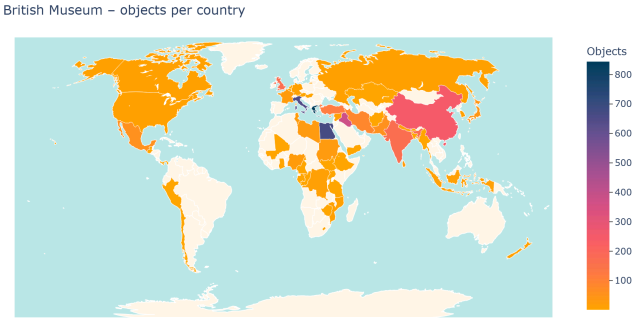


These gantt-style timelines show that key techniques such as carving and casting, and materials like terracotta, marble and bronze, have been used across many centuries, while others appear only in shorter periods as technologies and artistic preferences changed. Increasing density in later periods reflects both higher production and improved documentation, highlighting the longevity of major sculptural traditions and the emergence and decline of others over time.



The choropleth combining data from the British Museum and the V&A shows a predominantly Eurocentric distribution, with the highest concentrations of sculptures linked to Italy, the UK and Greece, alongside a substantial contribution from Egypt.

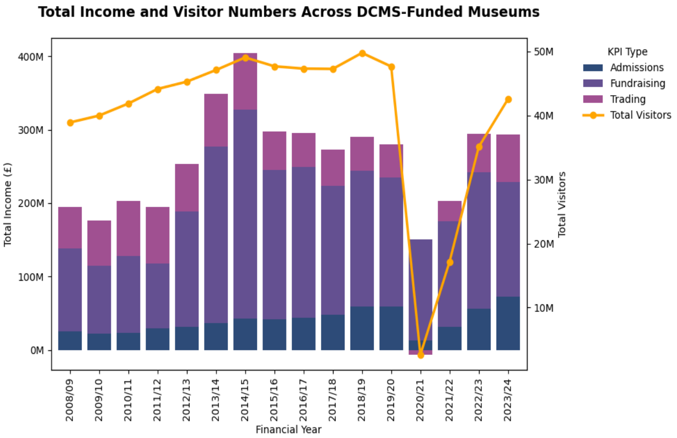




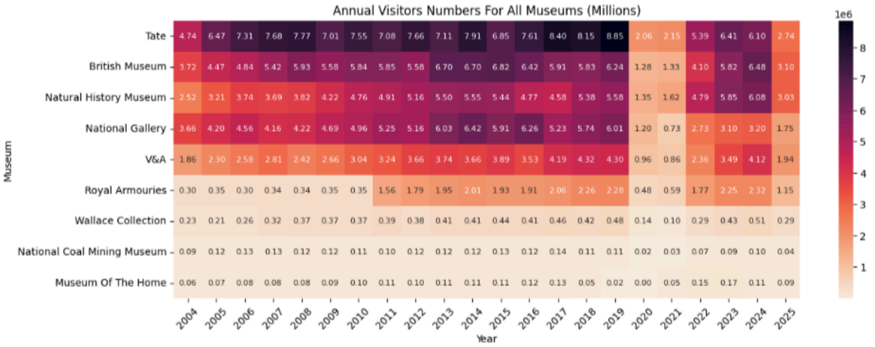
Examined separately, the two datasets reveal clearer distinctions in their profiles. The British Museum dataset presents a wider global spread but retains its highest densities in Italy, Greece and Egypt, reflecting the institution's longstanding strengths in classical and ancient collections. This concentration aligns with the Museum's historic collecting priorities and the scholarly focus that has shaped its holdings. By contrast, the V&A dataset is more strongly European in character, with notable concentrations in the UK, Italy and France, consistent with its focus on European art and design. Taken together, these patterns illustrate both the shared geographical emphases and the differing collecting trajectories of the two institutions.

5.2 KPI Data (Footfall & Income)

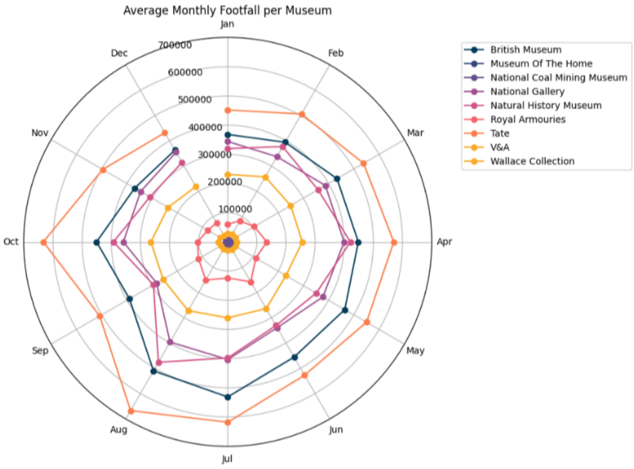
Analysis of key performance indicators shows a clear relationship between visitor numbers and income across DCMS-sponsored museums. Both peak in 2014-15, a pattern that may reflect First World War centenary activity, though this cannot be confirmed from the dataset alone. Income then declines despite relatively stable footfall. The sharp falls during the COVID-19 pandemic reflect closures and restrictions, with losses concentrated in admissions and trading income while fundraising remains comparatively stable. By 2023-24, both footfall and income have largely returned to pre-pandemic levels, indicating a strong sector-wide recovery.



5.3 KPI Data (Footfall)



The heatmap above shows a steady increase in footfall from 2004 - 2019 across all museums, with the Tate leading in footfall figures. There is a sharp decline over COVID. Some museums have been able to reach or exceed pre-COVID attendance figures.

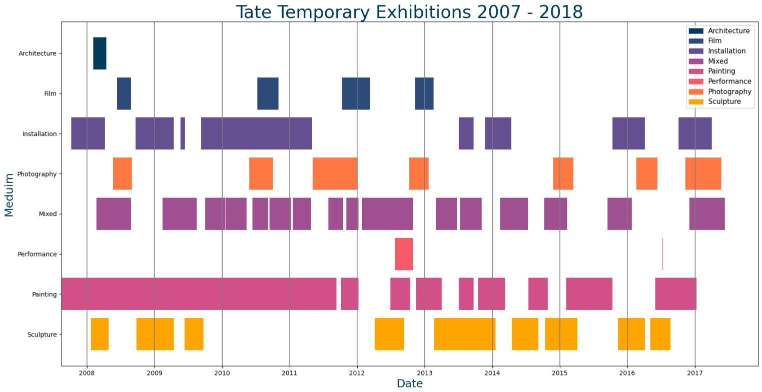


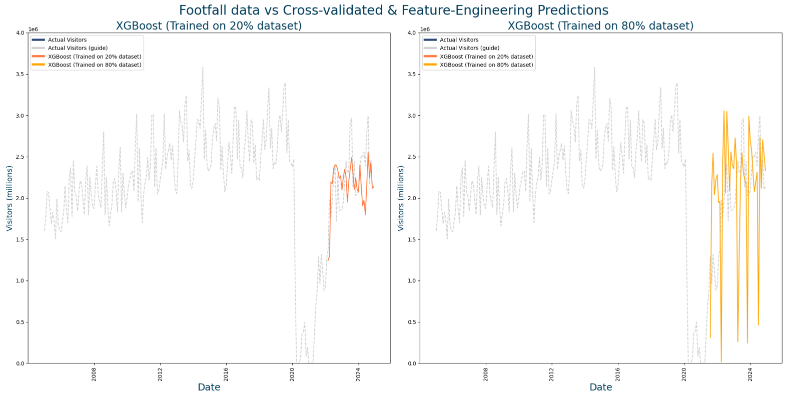
High footfall coincides with summer months, low footfall around winter months, therefore, more resource can be put towards high footfall months, with low footfall months ideal for renovations or gallery changes.

This makes planning for staffing, maintenance, special exhibitions etc. easier and more effective by anticipating visitor trends in advance.

5.4 Tate Exhibitions Data

Exhibition data between 2007 and 2018 indicates varying durations across media categories, with sculpture-themed exhibitions appearing intermittently and absent between 2010 and 2012. The medium labels suggest that sculpture was often incorporated into broader ‘mixed’ exhibitions, indicating a continued presence in the programme even when not the primary focus.





5.5 Machine Learning

Both time series represent a prediction of the final 20% of the data. A line chart was chosen to clearly show the viewer the accuracy of the predictions versus presenting decimals, demonstrating the mathematical accuracy. Future recursive predictions could provide insight into future visitor numbers allowing DCMS to plan accordingly.

6. Conclusion

This project demonstrates the value of integrating museum datasets to generate insight into cultural representation and public engagement. The collections analysis revealed that sculptures are predominantly associated with Europe, particularly Italy, the UK, Greece and Egypt, with the British Museum showing a wider global spread and the V&A retaining a more Eurocentric focus. Materials and techniques exhibit long-term continuity, reflecting enduring sculptural traditions and institutional collecting histories.

Visitor numbers rose steadily from 2004 to 2019 before falling sharply during COVID-19, with most museums returning to or exceeding pre-pandemic levels by 2023–24. Seasonal patterns remain stable, and income broadly tracks attendance, although the dataset does not allow for a stronger causal interpretation. The modelling showed that future footfall can be forecast at a general level, with XGBoost performing best among the models tested. Structural breaks and limited features, however, reduce month-to-month reliability, indicating the need for further feature engineering for operational forecasting.

Overall, the project demonstrates that once cleaned and standardised, museum data can support clearer cross-museum comparison, inform planning and strengthen evidence-based decision-making. Further work could expand the modelling through additional features, including seasonality terms and flags for structural shocks such as COVID-19. Broader analysis across more museums or object types, together with improved metadata standardisation and richer documentation of materials, techniques and provenance, would enhance analytical depth. Incorporating qualitative factors such as exhibition programming or marketing activity may also improve explanatory power in future models.

Taken together, these findings demonstrate that the project met its core objectives by analysing representation within the sculpture collections, identifying long and short-term visitor trends, clarifying the relationship between attendance and income, and assessing the potential for future footfall forecasting.