BBC iPlayer

Data Cleaning & Behaviour Analysis (Python)

Project Overview

This project demonstrates an **end-to-end data cleaning, validation, and analysis workflow** using a **synthetic BBC iPlayer-style dataset**.

The focus is on **transforming unclean, inconsistent raw data into a reliable, analysis-ready dataset**, then extracting **insights into customer viewing behaviour** through exploratory visualisations.

Note: The dataset is fully synthetic and does not contain real BBC iPlayer data.

Objectives

- Clean and standardise messy real-world-style data
- Apply robust data validation techniques
- Handle missing values, duplicates, and outliers
- Prepare data for time-based and behavioural analysis
- Communicate methodology clearly and reproducibly

Dataset Summary

Input:
unclean_bbc_iplayer_dataset_5k.csv (~5,000 rows)

Key fields:

Column	Description
show_id	Unique row identifier
title	Programme title
genre	Programme genre
duration	Raw duration (inconsistent formats)
broadcast_date	Broadcast date (multiple formats)
rating	Viewer rating (mixed types)
views	View count (mixed types, missing values)

The dataset intentionally includes:

- Missing values (N/A, ?, empty strings)
- Inconsistent date and duration formats
- Mixed numeric and string types
- Duplicate records
- Outliers

Tools & Libraries

- **Python 3**
- **pandas**
- **NumPy**
- **Matplotlib**
- **Jupyter Notebook**

Data Cleaning & Preparation Pipeline (11 Steps)

1. Load & Inspect Raw Data

- Preserve an untouched copy of the raw dataset
- Inspect schema, data types, and anomalies

2. Normalise Missing Values

- Standardise placeholders ("N/A", "?", empty strings) to NaN
- Generate missing-value reports per column

3. Standardise Dates

- Convert broadcast_date to datetime
- Handle multiple date formats robustly
- Preserve raw date strings for traceability

4. Parse & Standardise Duration

- Convert inconsistent duration strings into numeric minutes
- Create duration_minutes
- Flag implausible values

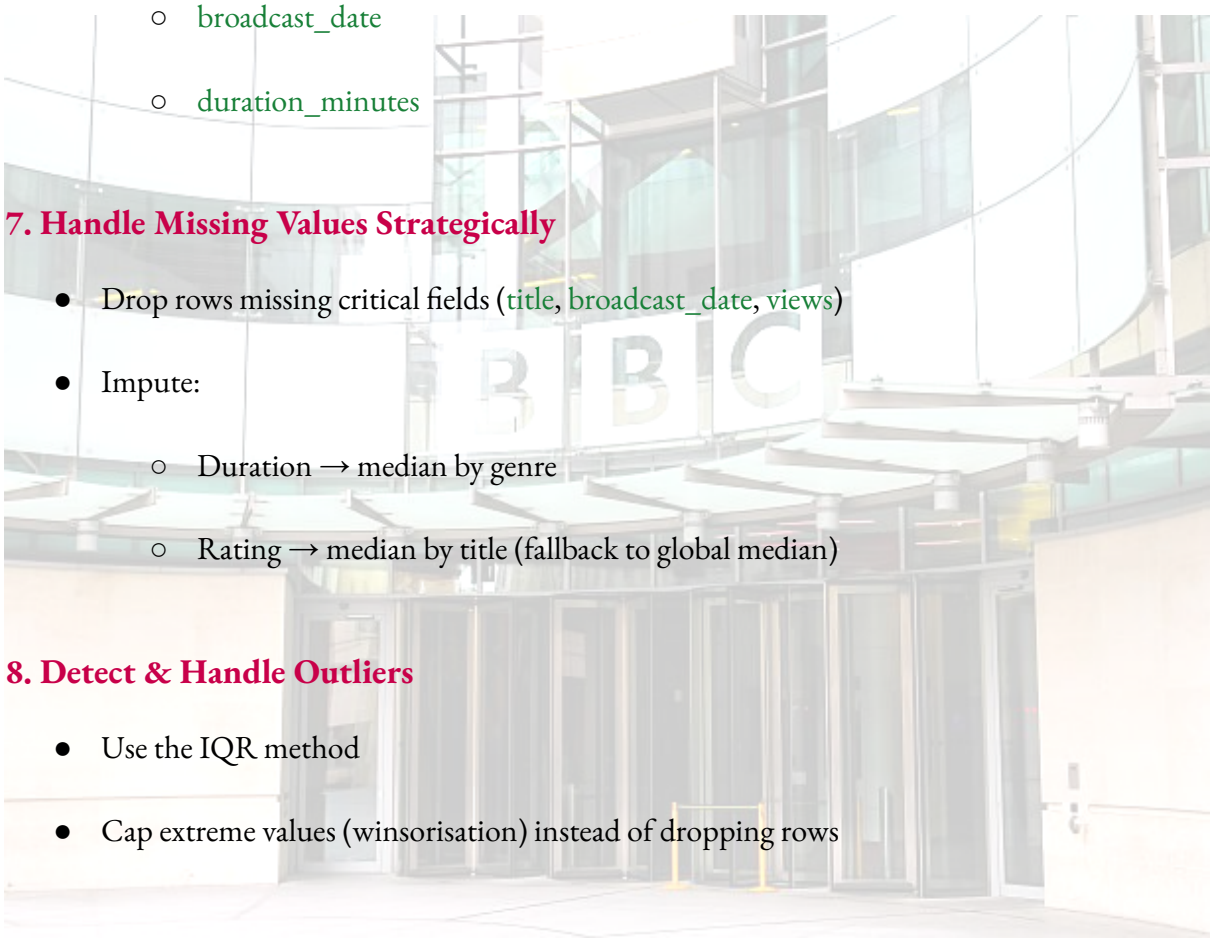
5. Clean Numeric Fields

- Convert rating and views to numeric
- Enforce domain rules:
    - Ratings $\in [1, 10]$
    - Views ≥ 0

- Retain raw values for auditability

6. Remove Duplicates

- Remove exact duplicates
- Remove logical duplicates using:
    - title
    - broadcast_date
    - duration_minutes



7. Handle Missing Values Strategically

- Drop rows missing critical fields (title, broadcast_date, views)
- Impute:
    - Duration → median by genre
    - Rating → median by title (fallback to global median)

8. Detect & Handle Outliers

- Use the IQR method
- Cap extreme values (winsorisation) instead of dropping rows

9. Data Validation

Automated checks ensure:

- No missing critical fields
- Correct data types
- Valid value ranges

- No remaining duplicates

10. Export Clean Dataset

Save analysis-ready data as:

bbc_iplayer_dataset_cleaned.csv

11. Behaviour Analysis & Visualisation

Generate insights including:

- Total views by genre
- Viewing trends over time
- Views vs ratings
- Heatmap: genre × day-of-week
- Top N shows by monthly views
- Optional session-style analysis (with user-level data)

Example Insights

- Certain genres show strong **weekend viewing patterns**
- Monthly top-performing titles highlight **seasonal trends**
- Ratings and views show **weak but interpretable correlation**
- Time-based aggregation enables release-impact analysis

📁 Project Structure

```
.
├── unclean_bbc_iplayer_dataset_5k.csv
├── bbc_iplayer_dataset_cleaned.csv
├── bbc_iplayer_analysis_notebook.ipynb
└── README.md
```

Why This Project Matters

This project demonstrates:

- Practical data cleaning skills
- Defensive programming with validation checks
- Realistic handling of messy data
- Clear analytical thinking
- Reproducible, well-documented workflows

It reflects challenges commonly encountered in **real-world analytics and data science roles**.

Advanced Behaviour Analysis (Post-Cleaning)

After producing a validated, analysis-ready dataset, the following analyses were conducted to extract **meaningful customer viewing behaviour patterns**.

These steps build directly on the cleaned data and demonstrate how preprocessing decisions enable reliable insights.

🔹 Heatmap: Genre × Day-of-Week Views

Objective:
Identify how viewing behaviour varies by genre across different days of the week.

Method:

- Extract day of week from broadcast_date
- Aggregate total views by:
    - genre
    - day_of_week
- Visualise results using a heatmap

Insights Enabled:

- Weekend vs weekday viewing differences
- Genre-specific viewing habits
- Identification of optimal release days per genre

Value:
Demonstrates time-based aggregation and categorical interaction analysis.

🔹 Session-Like Behaviour Analysis *(Optional Extension)*

Objective:
Explore deeper engagement patterns if user-level data is introduced.

Additional Fields (Optional):

- user_id
- device
- timestamp

Potential Analyses:

- Session length (number of shows per user per day)
- Binge-watching behaviour
- Device-based consumption patterns
- Peak viewing hours

Insights Enabled:

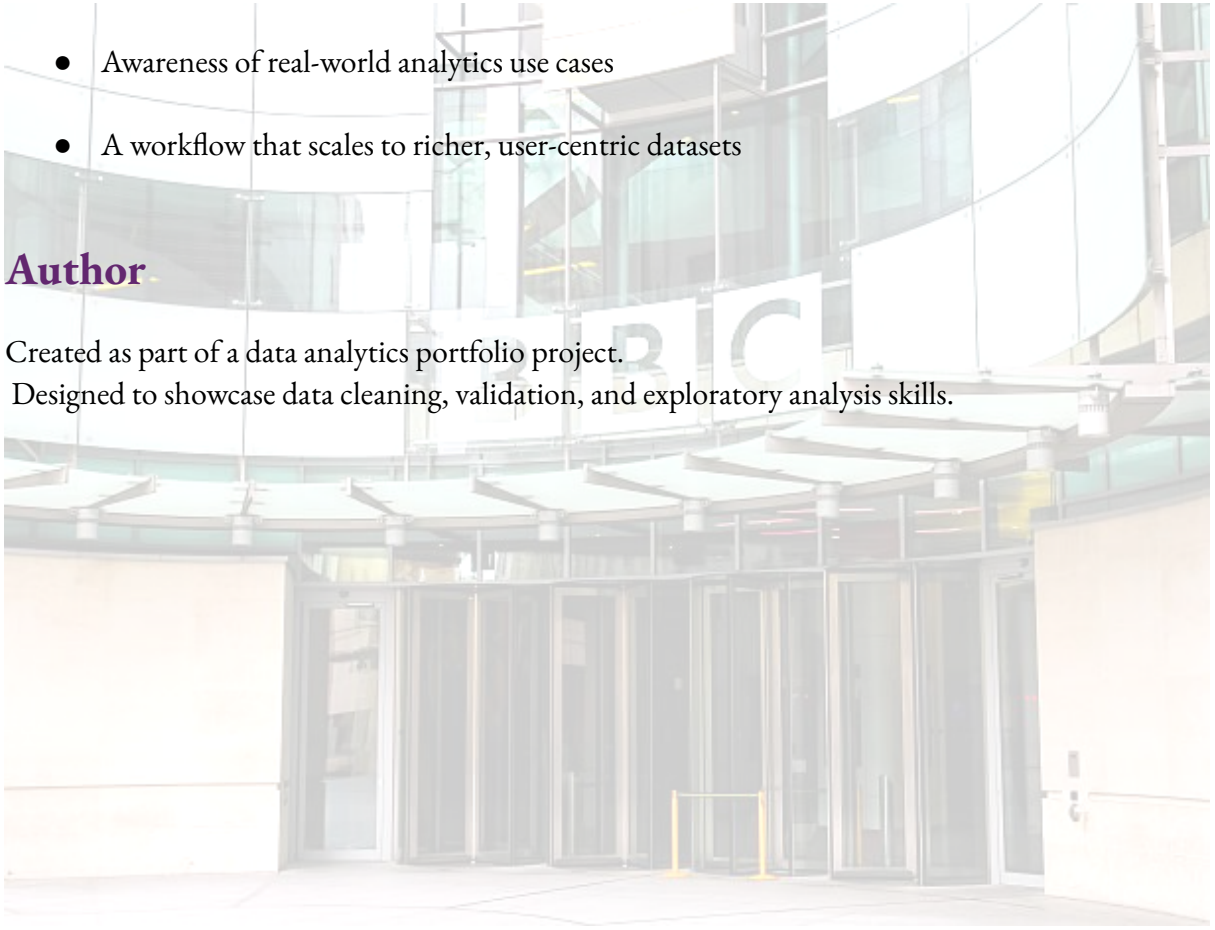
- User engagement intensity
- Preferred viewing devices
- Daily and hourly viewing trends

Value:
Demonstrates scalability of the pipeline from content-level analysis to user-level behavioural analytics.

Project Outcome

By extending the analysis beyond data cleaning, this project demonstrates:

- Strong data preprocessing foundations
- Ability to translate clean data into actionable insights
- Awareness of real-world analytics use cases
- A workflow that scales to richer, user-centric datasets



Author

Created as part of a data analytics portfolio project.
Designed to showcase data cleaning, validation, and exploratory analysis skills.