Findings Report – Crowd Energy Prediction

Objective
The objective of this project was to predict Crowd Energy (0–100 scale) for future concerts using historical tour data while avoiding data leakage and ensuring real-world applicability.

Data Cleaning & Quality Issues
• Inconsistent date formats were standardized using datetime parsing.
• Ticket prices in multiple currencies (£, €, $) were converted to USD using provided exchange rates.
• Missing and zero-valued sensor readings were treated as invalid and imputed using median values.
• Extreme outliers caused by logging errors were removed using the IQR method.

Key Insights from EDA
• Crowd Energy varies significantly by venue, indicating venue-specific crowd behavior.
• Merchandise sales post-show show a strong positive correlation with Crowd Energy.
• Crowd size has a moderate positive relationship with energy levels.
• Other variables such as weather and opener rating show weaker influence.

Red Herrings Evaluation
Some singer hypotheses (e.g., moon phase, outfit) showed no strong evidence and were treated as noise rather than predictive features.

Modeling Approach
• A Linear Regression model was used for interpretability and robustness.
• Only features available before the show were used to avoid data leakage.
• Categorical variables were one-hot encoded with unseen categories handled safely.
• Validation results achieved MAE $\approx$ 11, indicating reasonable predictive performance.

Conclusion
The final model generalizes well, avoids post-event leakage, and produces realistic predictions. Venue and audience-related features are the strongest drivers of crowd energy, while several hypothesized factors were disproven by data.