SUMMARY

The key objective of this analysis was to predict **Crowd Energy** levels for future tour dates. The dataset presented significant challenges, including extreme outliers (energy levels > 5000), text-based 'dirty' raw data, negative crowd energy values and some complex, non-linear relationships between venue types and environmental factors.

MODEL CHOICE

Data Preprocessing Strategy

- Log-Transformation : The target variable (Crowd_Energy) was heavily right-skewed. Training on raw values caused the model to over-prioritize rare "super-events." We applied a log-transformation to normalize the distribution, ensuring the model learned typical crowd behaviors accurately.

- The "Boomerang" Technique: Predictions are generated in log-scale and mathematically reversed to produce the final real-world submission values.

- Text Cleaning: Features like Ticket_Price contained currency symbols ($), which were cleaned via regex and converted to floating-point numbers to allow for mathematical regression.

So XGBoost (Extreme Gradient Boosting) Regression was considered the best here, even over Linear Regression .

- Reasoning: Linear models failed to capture interaction effects. For example, a high ticket price is good for Venue_Gamma but bad or neutral for others. XGBoost's decision-tree architecture naturally learns these conditional rules ("If venue is gamma and price is high...") without manual feature engineering.

- Handling Non-Linearity**:** Linear regression failed to capture venue-specific "rules" (e.g., high prices boost energy at Venue C but lower it at Venue A). XGBoost's decision-tree architecture naturally learns these complex "if-then" interactions without manual equation adjustments.
- Robustness to Nulls**:** XGBoost handles missing data internally, providing a safety net for potential gaps in future test data.
- Performance on Structured Data**:** In our validation phase, XGBoost consistently outperformed linear baselines by effectively utilizing categorical features like Moon_Phase and Band_Outfit

VENUE CASE STUDY

1: Venue V_Alpha ("The Monastery")

- Theory: "The crowd here hates noise. High volume kills the vibe."
- Evidence: Regression analysis confirmed a negative correlation between Volume_Level and Crowd_Energy. As decibels increased, energy of the crowd dropped.
- Result: THEORY CORRECT
- Action: Cap volume levels for V_Alpha. Acoustic or "unplugged" sets must be prioritised.

2: Venue V_Beta ("The Vampire's Den")

- Theory: "They only wake up at night. The later we play, the better."
- Evidence: When temporal analysis was done, it showed a flat energy baseline for slots before 10 PM. Also an exponential spike occurred for start times after 11:00 PM, continuing into the early morning.
- Result: THEORY CORRECT
- Action: V_Beta slots should be scheduled as late as possible. Also opening acts, here, should be avoided.

3: Venue V_Gamma ("The VIP Lounge")

- Theory: "They are snobs. They only get excited if the ticket price is expensive."
- Evidence: Contrary to standard supply/demand, V_Gamma displayed a positive correlation between Ticket_Price and Crowd_Energy. Budget tickets resulted in lower engagement.
- Result: THEORY CORRECT
- Action: Maintain premium pricing strategies. Discounting tickets here damages the brand perception and crowd energy.

4: Venue V_Delta ("The Open Air Stage")

- Theory: "This place is at the mercy of the elements. If it rains, nobody dances."
- Evidence: Feature importance analysis for V_Delta highlighted Weather as the dominant predictor.
  - "Clear" or "Cloudy": High/Stable Energy.
  - "Rain" or "Snow": Massive drop in energy (near zero).
  - Note: This venue showed low sensitivity to Price or Band Outfit compared to the weather.
- Result: THEORY CORRECT
- Action: Implementation of a strict measures for overcasted/rainy weather. Minimize marketing spend if the forecast is poor, as performance quality cannot overcome the weather penalty here.

UNIVERSAL FACTORS

The model identified some universal patterns too in the data provided

1: The "Tuesday Curse"

- Theory: "Gigs on Tuesday are always dead."
- Finding: Box-plot analysis of Day_Name confirmed that Tuesday events had the lowest median energy and the lowest upper-quartile range of any weekday. The model assigned negative weight coefficients to the Day_Name_Tuesday feature.
- Result : THEORY CORRECT

2: Lunar Influence

- Here , the singer did not provide any statement/theory regarding this influence.
- Observation: Feature engineering regarding Moon_Phase suggested a statistically significant variance in energy during Full Moon phases compared to "New Moon" phases. This feature contributed to the model's accuracy, suggesting environmental/psychological factors play a subtle role.

CONCLUSION

Analysis confirms that engagement drivers are venue-specific rather than universal. The XGBoost model successfully encoded the non-linear variances between price-sensitive sectors Venue_Gamma and weather-dependent locations Venue_Delta. The resulting high accuracy scores validate the model's capability to handle complex feature interactions across diverse operational environments.