REPORTS:

Analysis of Volume vs. Crowd Energy (Post-Cleaning) After excluding the anomalous data points where **Crowd Energy** was recorded as 1000—an impossible value on a 0–100 scale found in **V_Gamma** and **V_Delta**—the remaining data reveals distinct acoustic behaviors across venues. **V_Alpha** exhibits the clearest trend, showing a strong positive correlation where increasing volume reliably boosts crowd energy, though this effect appears to plateau at higher levels. In contrast, **V_Beta** displays high variance with no discernible pattern, indicating that for this specific venue, volume is not a primary driver of crowd engagement and other factors likely dominate. Meanwhile, once the extreme outliers are removed from **V_Gamma** and **V_Delta**, the data mostly clusters within realistic energy ranges, although **V_Delta** still contains a likely volume entry error near 100 that should be addressed. This variation strongly suggests that the impact of volume is venue-dependent, justifying the use of Venue_ID as a critical feature in the machine learning model to capture these local differences.

The singer's theories against noise limits in V_Alpha is proven here.

Moon Phase affecting crowd energy: The graphs plotted for moonphase vs crowd energy disproves the singer's intuition that full moon leads to more crowd energy.

Price sensitivity: The graphs plotted show that the price sensitivity is low in V Gamma and V delta with the curves almost being a flatline plateau. The people of V_Alpha and V_Beta were more sensitive to prices than the others.

Weekends: Weekends USUALLY have higher energy than the weekdays.

Showtime influence at the goths: The plotted graphs show that that V_Beta/gothic nightclub has showtime preferences late night as compared to other venues which are mostly flat or plateau.

Weather: The plotted boxplots with identical medians and similar spread shows that weather doesn't directly affect crowd energy disproving the singer.

1. Model Selection Justification

Why Random Forest outperformed Linear Regression:

- **Non-Linear Relationships:** Linear Regression assumes straight-line relationships (e.g., "Higher Volume = Higher Energy"). However, our EDA showed complex behaviors, such as the "Volume Plateau" at Venue Alpha and the inverted U-shape of the Revenue Curve. Random Forest naturally captures these non-linear patterns and thresholds without requiring manual equation adjustments.
- **Complex Interactions:** The dataset involves complex interactions between categorical features (e.g., Venue_ID) and numerical features (e.g., Volume_Level). Random Forest automatically learns these interaction effects (e.g., "High volume works at Alpha but fails at Beta"), whereas Linear Regression would require complex manual feature engineering (interaction terms) to achieve the same result.

Why Random Forest outperformed XGBoost:

- **Small Sample Size:** The training dataset contains approximately 1,200 rows. XGBoost typically thrives on massive datasets ($10,000+$ rows) where it can iteratively correct subtle errors. On smaller datasets, XGBoost's aggressive error-correction often leads to overfitting. Random Forest's "Bagging" (Bootstrap Aggregating) approach is inherently more stable for smaller sample sizes as it builds independent trees and averages them.
- **Noise Tolerance:** The target variable, Crowd_Energy, is highly subjective and noisy (influenced by human behavior). XGBoost tends to "chase the noise" by trying to minimize every residual error, leading to higher variance on unseen data. Random Forest smooths out this noise by averaging the predictions of hundreds of decorrelated trees, resulting in a more generalized and robust model for this specific problem.

2. Hyperparameter Configuration Justification

The following hyperparameters were selected for the GridSearchCV to balance model complexity with generalization:

- **n_estimators [50,100, 200]:**
  - *Role:* The number of trees in the forest.
  - *Justification:* We tested up to 200 trees to ensure sufficient "voting power" to stabilize predictions and reduce variance. Beyond this point, the accuracy gains typically diminish while computational cost increases ("diminishing returns").
- **max_depth [10, 20, None]:**
  - *Role:* Controls how deep each tree can grow.
  - *Justification:* Allowing trees to grow infinitely (None) often leads to overfitting, where the model memorizes specific training rows. We introduced limits (10, 20) to force the trees to learn broader, more generalizable patterns rather than memorizing every unique outlier in the training data.

- **min_samples_split [2, 5]:**
    - *Role:* The minimum number of samples required to split an internal node.
    - *Justification:* Increasing this value to 5 acts as a regularizer. It prevents the model from creating specific rules for tiny groups of people (e.g., "If volume is 8.2 and it's raining, energy is 95"), ensuring that every decision branch is based on a statistically significant chunk of data.

Cross validation strategy has been used during tuning.

ADVICE FOR MORE CROWD ENERGY:

Acoustic management must be tailored to the location: while increasing volume at Venue V_Alpha directly drives higher crowd energy and should be encouraged, the same tactic at Venue V_Beta yields no benefit, suggesting resources there should be redirected to other performance elements. Financially, a significant opportunity exists at Venue V_Gamma, where the audience is price-insensitive regarding their enjoyment; this allows for ticket price increases without negatively impacting the show's atmosphere. Operationally, the business can reduce costs by booking cheaper, off-peak time slots and reducing weather-related insurance, as the data confirms that neither the time of day nor adverse weather conditions have any significant negative impact on the crowd's energy levels.

Plotting a graph for which factor affects the crowd energy most according to the model it is the HOUR so focus should be given to find which timings ensure greater energy.

(USE OF AI HAS BEEN USED TO GENERATE THE ADVICE BASED ON THE MODEL DUE TO LACK OF TIME)