REPORT

VENUE ALPHA KEY FINDINGS:

The Scatter plot with the fitted regression line clearly indicates that the volume level has a significant effect on crowd energy at venue Alpha so, the singer was true about his theory at this venue.

Both afternoon and night shows occurred with the crowd energy in the range 40 - 60 so, the shows at this venue weren't either killer or a flop.

From the plots it has been observed that moon phase, band outfit, ticket price, weather, weekday and weekend are not affecting the crowd energy.

VENUE BETA KEY FINDINGS:

The goths at this venue don't have any noise limit. From the box plot it is clear that in the afternoon the crowd energy was in the range 0-20 whereas in the night it was in the range 50 - 65 which helps in concluding that the singer was right about his theory regarding timing.

As per the plots Moon phase, outfit, weather are affecting crowd energy.There is no price sensitivity at this venue.

It has been observed that weekday and weekend are having different ranges of crowd energy.

VENUE GAMMA KEY FINDINGS:

As per singers theory price is affecting the crowd energy which is agreeable as from the plot the crowd energy was increasing as the price was increasing.

Volume level, Moon phase, band outfit, weather, weekdays and weekend are not affecting crowd energy.

Afternoon shows are killer at this venue as per the plots.

VENUE DELTA KEY FINDINGS:

Noise level is affecting crowd energy as per the plots. Afternoon shows were neither killer nor a flop at this venue.

Moon phase, outfit, weekdays, weekend and price sensitivity are not affecting the crowd energy. Crowd energy range is almost same for rainy and stormy weather, it is also in the same range for remaining weather conditions.

MODEL CHOICE JUSTIFICATION:

Model used : RANDOM FOREST REGRESSOR

-> Crowd energy is influenced by non-linear interactions between time, venue and behavioural features.

-> Random Forest is an ensemble of decision trees, making it well-suited for capturing such non-linearities without requiring extensive feature scaling.

-> The model is robust to noise and outliers, which is important given the messy and inconsistent nature of the tour logs.

-> It handles mixed feature types(numerical and encoded categorical variables) effectively.

-> Compared to a single decision tree, Random forest reduces overfitting by averaging predictions across many trees.

Hence, Random Forest was selected as a strong baseline ensemble model for predicting Crowd Energy.

HYPERPARAMETER TUNING:

HYPER-PARAMETERS EXPLORED(Ranges & Values Tested):

The following hyper-parameters were tuned using grid search:

HYPERPARAMETER	VALUES TESTED	PURPOSE
n_estimators	[100 , 200]	Number of trees in the forest
max_depth	[5 , 10 , 15]	Controls maximum depth of each tree
min_samples_split	[2 , 5]	Minimum samples required to split a node

Reason for choosing these ranges :

-> Lower max_depth values prevent overfitting to noise.
-> Increasing n_estimators improves stability and reduces variance.
-> min_samples_split controls how aggressive tree splitting is, affecting generalisation.

VALIDATION STRATEGY USED DURING TUNING :

Validation method : 5-Fold Cross-Validation (KFold)

Implementation Details (From Code) :

-> The training data is split into 5 folds.
-> In each iteration :
4 Folds are used for training
1 Fold is used for validation.
-> The process is repeated until each fold has served as the validation Set once.

Cv = KFold (n_splits = 5, shuffle=True, random_state=42)

Evaluation metric :

RMSE (Root mean squared error), as it penalises larger prediction Errors more heavily.

FINAL HYPER-PARAMETER VALUES & REASONING :

After performing grid search, the optimal hyper parameters Obtained were :
n_estimators = 200

max_depth = 5
min_samples_split = 5

Reasoning :
-> n_estimators = 200 improved prediction stability compared to 100 trees.
-> max_depth = 5 prevented trees from becoming overly complex and memorising noise.
-> min_samples = 5 encouraged more conservative splits, improving generalisation.

COMPARISON : TUNED MODEL vs DEFAULT PARAMETERS

Model Version	RMSE (Cross Validated)	Observation
Default Random Forest	Higher RMSE	Overfits due to deeper trees and aggressive splits
Tuned Random Forest	15.5287	Improved generalisation and stability

Conclusion :
-> Hyper parameter tuning led to a clear improvement in predictive performance.
-> The tuned model generalises better than the default Random Forest.