FINDINGS REPORT

By SUBHANKAR BHARADWAJ

Key Insights of the Data:

- V_Gamma has the max avg crowd energy, followed by V_Delta, V_Alpha and V_Beta respectively
- Analysing the scatter plot between crowd energy and Ticket Price, we found that crowd energy has positive correlation with Ticket Price, although the correlation was quite less
- Similarly, the correlation between the Volume Level and Crowd Energy, as well as between Crowd Size and Crowd Energy was less(although it was positive).
- The Lead singer predicted that Tuesday was cursed, however the average crowd

energy for all the days were almost the samne. Even though Tuesday's average crowd energy was second lowest, however there is not much difference between the highest and the lowest.

- The average crowd energy was significantly higher during the night and evening as compared to afternoon and morning.
- There is no relationship between the band outfit and crowd energy because the average crowd energy for all three outfits were almost same.
- Average Crowd Energy were almost same for all weathers, however it is noteworthy that the lowest average crowd energy was located during Stormy weather
- For V_Gamma, as the ticket price increased, the crowd energy increased, but for rest of the venues, the crowd energy remained almost same for all ticket prices.

- V_Gamma and V_Beta had higher ticket prices as compared to the other two venues.
- It is clear that as the opener rating increases, the crowd energy increases, however the increase is not very high.
- Saturday and Friday are the days with highest average crowd energy.

MODEL SELECTION

Given the scope of my current knowledge, I focused on models that I have thoroughly studied: Linear Regression and Random Forest. While I am familiar with other algorithms like Logistic Regression and K-Nearest Neighbours, they were not the best fit for this specific problem:

- Why not Logistic Regression? Since we are predicting a continuous score rather than a category, Logistic Regression is unsuitable as it is designed for classification, not regression.

- Linear Regression as a Baseline**:** I started with Linear Regression because it is the fundamental approach for predicting numbers. It helps determine if there is a straight-line relationship between features like Crowd_Size or Ticket_Price and the resulting energy levels.
- Random Forest for Complexity**:** I chose Random Forest as my second model because the dataset contains many categorical variables .Random Forests are excellent at capturing non-linear patterns and interactions between these different variables that a simple line might miss.

After using both the algorithm, I evaluated their RMSE, MAE and R2 score

Since Random Forest had lower RMSE and MAE score, while higher R2 score, I went with Random Forest

HYPERPARAMETER TUNING

For model optimization, I performed manual hyperparameter tuning on the Random Forest model. I explored two key hyperparameters:

- n_estimators: [100, 200, 300] - the number of decision trees in the forest
- max_depth: [10, 20, 30] - the maximum depth of each tree

I used a simple train-test validation strategy where the model was trained on the training set and evaluated on the test set for each parameter combination. The combination that produced the lowest RMSE on the test set was selected as optimal.

Validation Strategy used:- Simple train-test split