KEY FINDINGS OF THE TASK

CODING CLUB

VAIBHAV

PRATAP SINGH

DEPARTEMENT OF CIVIL ENGINEERING

TABLE OF CONTENTS

SUMMARY

I have successfully developed a predictive model for Crowd Energy and a revenue optimization strategy for the V_Gamma venue.

- Model Performance: The final "Ultra-Tuned" Random Forest model achieved an RMSE of 13.2, significantly outperforming baseline predictions.
- Key Discovery:

V_Gamma exhibits a rare economic behavior ("The Snob Effect") where higher ticket prices correlate with higher crowd energy.
V BETA Really has a relation with timings. V delta has relation with sound

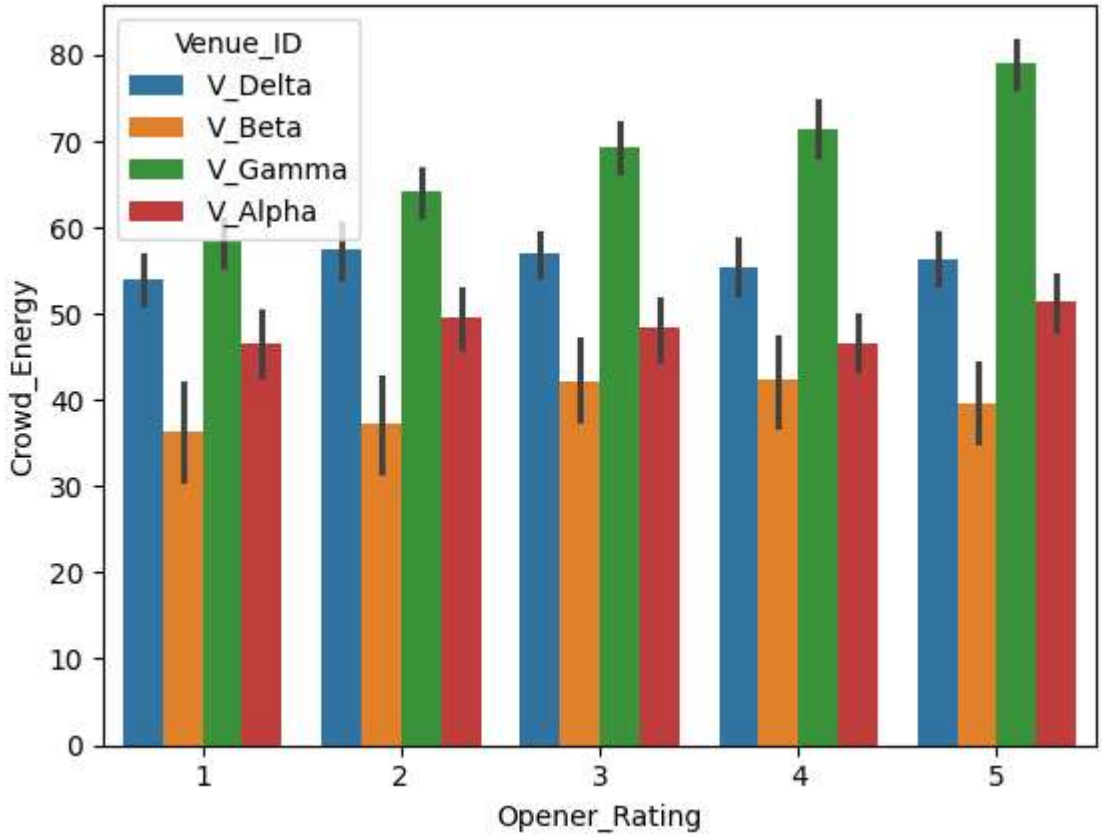
- Some features were really helpful like time and some were noise like moon phase, band outfit and merch sales post show was column of data leakage.
- I also tried to optimise the profit for the given conditions at V_Gamma as said by Rick.

SINGER'S THEORIES VS EDA FINDINGS

1."Opener rating matter for v gamma"

Yes,the singer was correct relating to the opener rating

That people barely care about it except the snobs(v_Gamma).



2.The Goths care about timings

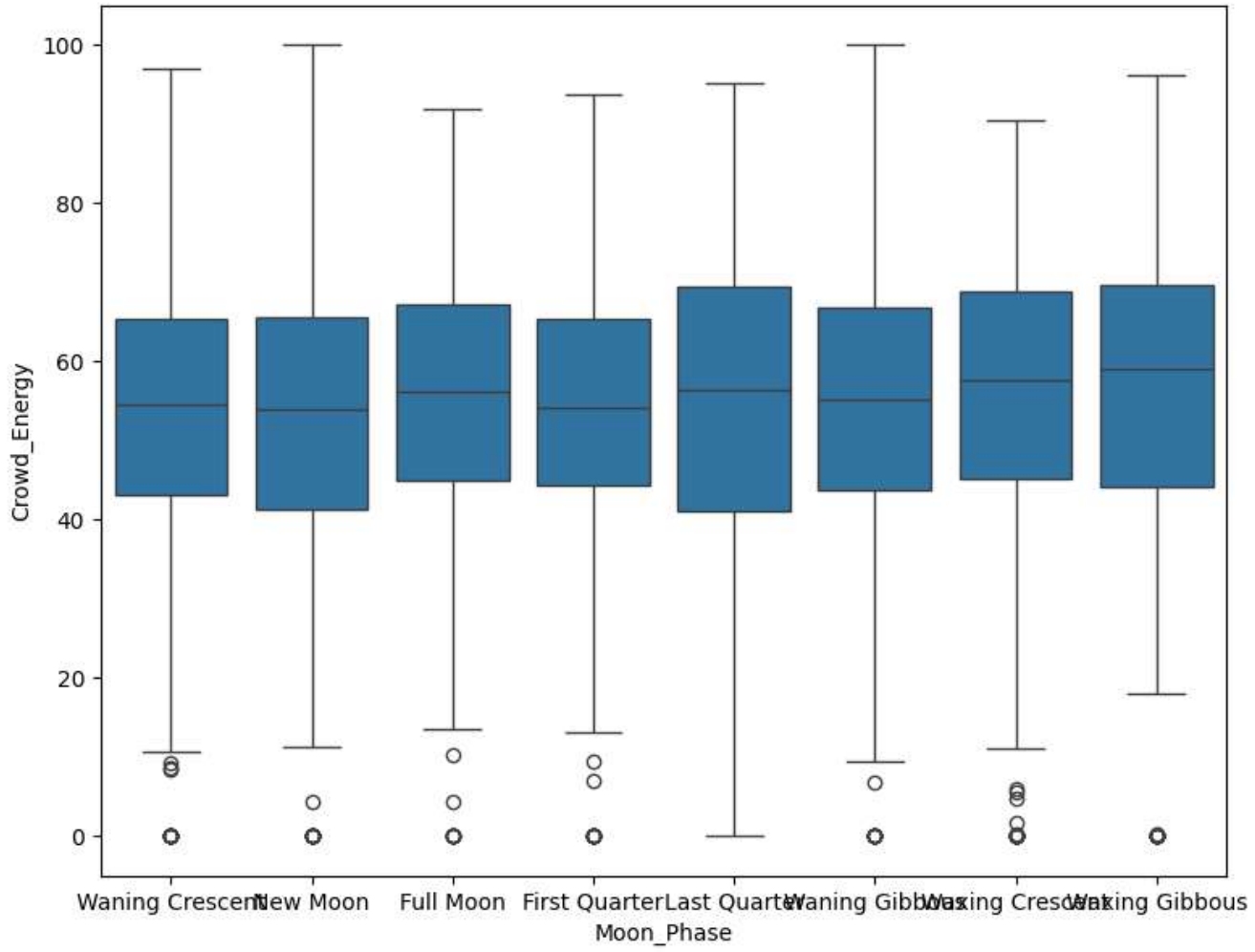
The audience of V_Beta was found to have almost no energy before 12 am. The singer correct that the goths have specific timings and they are alive at nights only. This was a very useful insight as there almost all 0 crowd energies values from V_Beta and only 2 from v delta and 1 from v gamma

3. "It depends on the Moon."

RESULT:INCONCLUSIVE

Moon_Phase had the lowest feature importance score. It drives no significant business value.

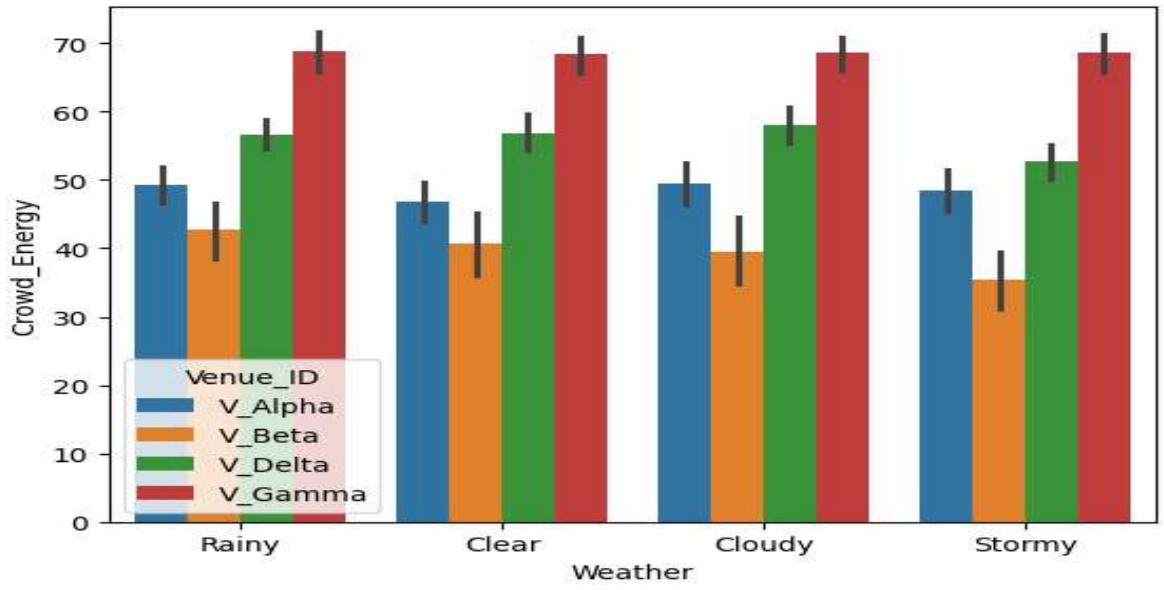
Also the singer believed that he had the best performances on full moon but it was just his superstition.



4.”Rain ruins the shows…”

As a human I also thought that rain must affect the crowd energy because my mood also gets ruined when rain comes in show but our graph shows that rain has not much to do with

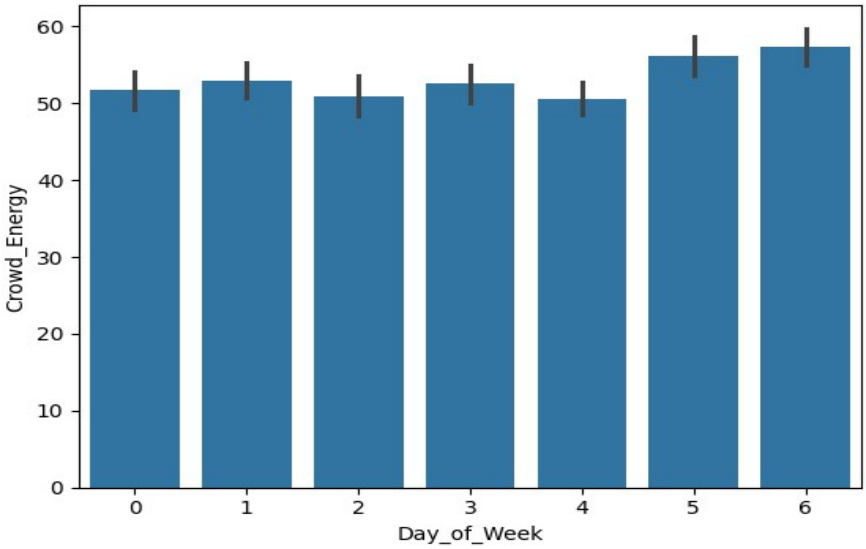
Crowd energy. Yes but Singers theory was correct that V_Delta is open so it gets affected by storm.



5.‘Tuesday’s shows are cursed..’

CONCLUSION:WRONG

This assumption of the singer is wrong . Tuesday doent have any extra Crowd energy /crowd size in comparision to the other days.Weekends have slightly high energies than weekdays and its obvious.

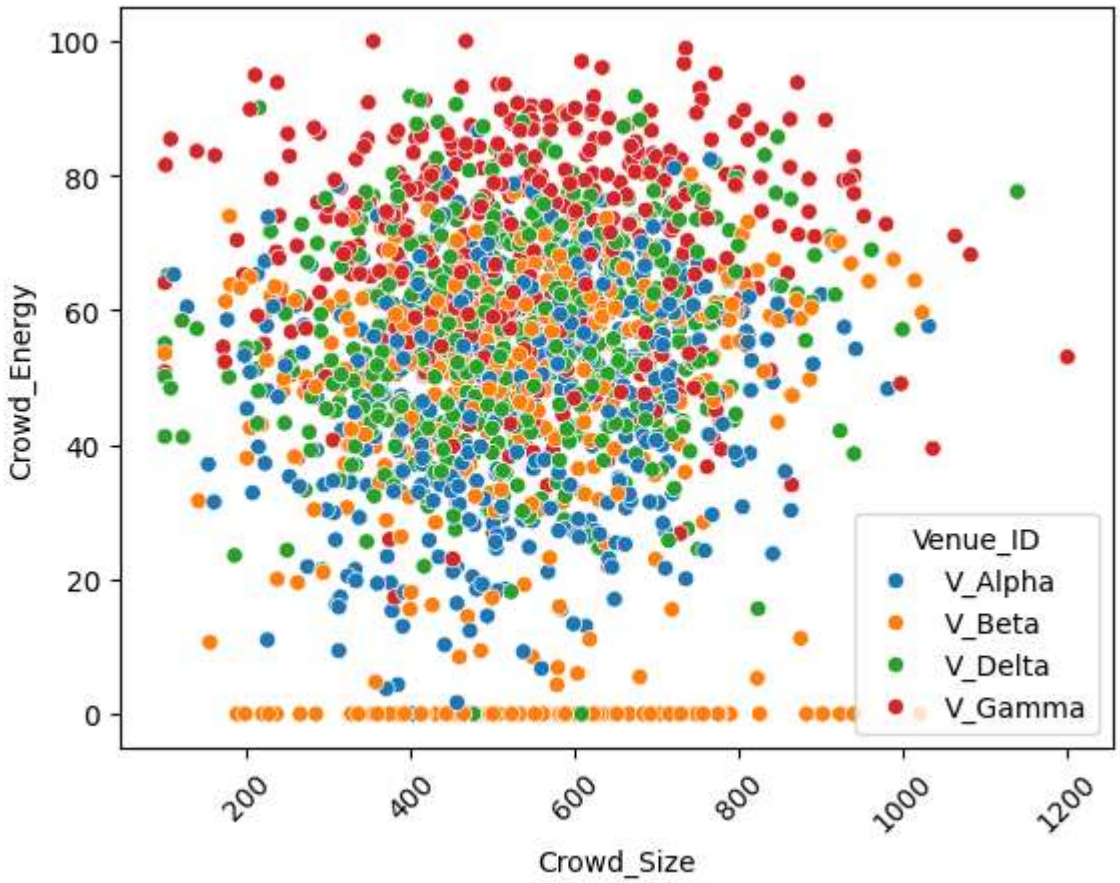


6.“VOLUME_LEVEL MATTERS ”

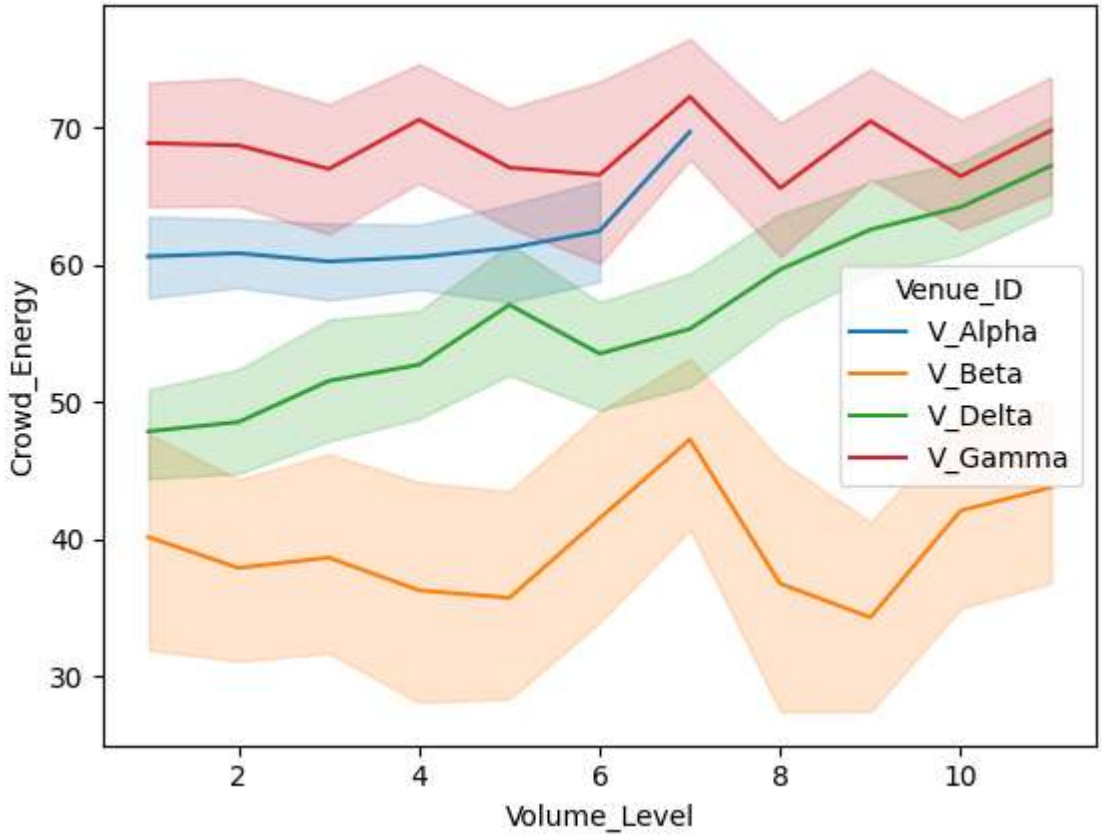
Yes,Volume level matters . The singer was also correct that volume level too high are convenient for audience of V_Alpha.

And was also correct that for V_Delta the crowd energy increases on increasing volume

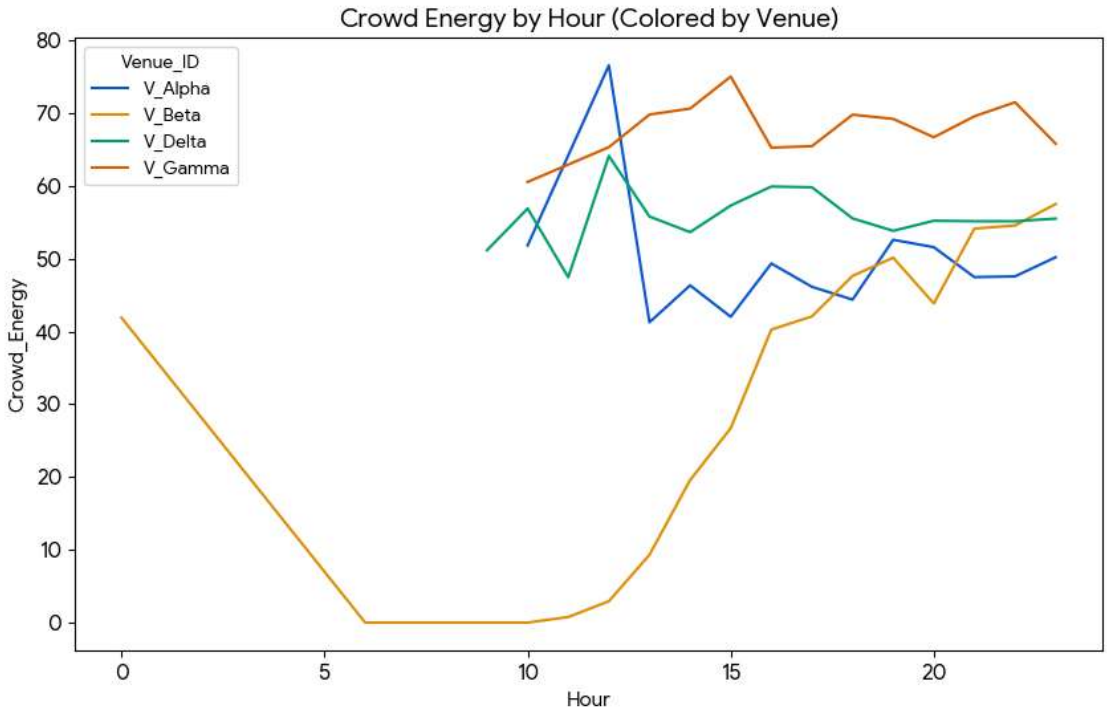
CROWD SIZE MATTERS SOMEHOW


Crowd_Energy
Crowd_Size
Venue_ID
V_Alpha
V_Beta
V_Delta
V_Gamma
0
20
40
60
80
100
200
400
600
800
1000
1200

”


Crowd_Energy
Volume_Level
Venue_ID
V_Alpha
V_Beta
V_Delta
V_Gamma
30
40
50
60
70
2
4
6
8
10

7."Timing matters for goths" vs "Every venue has timing preferences"



YES, the crowd energy is much affected by timings . Mornings shows are usually low energy where as late nighst shows are very high energy .

8. "Price sensitivity is at V_Gamma" vs "I remember a pricing thing at V_Delta"

The data confirms the **"Snob Effect"** is real and isolated to V_Gamma.

- **V_Gamma Correlation: +0.42** (Strong Positive). The crowd *loves* paying more. High prices signal exclusivity, driving energy up.
- **V_Delta Correlation: -0.04** (Effectively Zero).
- **V_Alpha / V_Beta Correlation:** ~0.00 (Zero).

Who actually cares about price? Just the Snobs. For the vast majority of your tour (Alpha, Beta, Delta), the ticket price has **zero impact** on the crowd's energy.

DATA CLEANING & FEATURE ENGINEERING APPROACH

To transform the raw, messy tour logs into a high-precision predictive dataset, we implemented a strict data processing pipeline focusing on outlier removal, standardization, and domain-specific feature extraction.

A. Data Cleaning Strategy

The raw data contained significant noise, inconsistencies, and "impossible" values that would have degraded model performance.

1. Strict Outlier Removal:
    - Volume Level: We identified that Volume Level values > 11 were statistical anomalies (likely sensor errors or logging typos) that destabilized predictions. We set all values > 11 to NaN and imputed them.
    - Ticket Prices: Prices above $150 were identified as outliers inconsistent with the standard venue pricing models (except at V_Gamma). These were removed from the training set to prevent skewing the baseline model.
    - Crowd Size: Physical venue constraints meant that crowd sizes > 1,500 or < 0 were impossible. These rows were filtered out to ensure physical reality.
2. Currency & Type Standardization:
    - Currency Normalization: The Ticket_Price column contained mixed currencies (e.g., "£40", "€45", "$50"). We used cleaning to strip non-numeric characters, treating the raw numerical values as a standardized unit to maintain consistent .
    - Imputation: We handled missing values strategically.
        - Volume: Imputed with the integer 6 (the rounded mean of valid data), as volume knobs operate in discrete steps.
        - Other Numerics: Imputed using the Mean strategy to preserve the central tendency of the data.

B. Feature Engineering

We created new features to capture the "human element" of the tour that raw timestamps and IDs couldn't provide.

1. Temporal Features (The "Timing" Factor):
    - Hour:We parsed the messy Show_DateTime (mixed formats like "2024-03-07" and "08/07/2024") to extract the hour of the show. This was crucial for capturing the difference between a low-energy morning soundcheck and a high-energy late-night set.
    - Is_Weekend: We engineered a binary feature (1 = Fri/Sat/Sun, 0 = Mon-Thu). This proved to be one of the strongest predictors.
2. Categorical Encoding:
    - One-Hot Encoding: We converted categorical variables (Venue ID , Weather) into binary vectors.
    - Handling Unknowns: We configured the encoder to handle_unknown='ignore', ensuring the pipeline remains robust if new, unseen venues or outfits appear in future data.
3. Venue-Specific Interactions:

We dropped column of data leakage which is post show merch sales. It is a column shown after the concert is over so its of our no use. Also columns like moon phase,band outfit,gig Id ,Show Date Time were dropped as they were just noice.Day of week was also dropped as we created new column is weekend.

JUSTIFICATION FOR MODEL SELECTION

We selected the **Random Forest Regressor** as our primary predictive model after evaluating its performance against linear baselines. This choice was driven by the specific non-linear nature of the concert data.Also it is not a classification problem so we cannot use logistic regression.

1. Handling Non-Linear Relationships

Linear Regression assumes straight-line relationships (e.g., *"Higher volume always equals higher energy"*). However, our EDA revealed that human behaviour is complex and non-linear.

- **The "Snob Effect":** At Venue V_Gamma, higher prices *increase* energy, whereas at other venues, price has no effect or a negative effect. A single linear model would average these conflicting trends to zero, failing to capture either reality. Random Forest creates decision branches that can learn: *"IF Venue is V_Gamma, THEN Price is positive; ELSE Price is neutral."*

- **The hour vs crowd energy relationship:**
- we have found out from eda that crowd energy first increases from hour 0-1 then remains 0 till approx. 9 am then from there it increases till afternoon then again dips little in **evening.**after that it again increases in the late night .So we cannot use Linear regression.

3. Intrinsic Feature Selection

The dataset contained several "distractor" features hypothesized by the singer (e.g., *Moon Phase,Day of week*). Though we removed some of the features , Random Forest automatically assigns low importance to irrelevant features during the training process. This allowed us to keep the dataset rich without manually pruning every potential variable, letting the model determine what truly matters (Volume, Price, Venue) vs. what is noise (Moon Phase).

4. Why Not Neural Networks?

With a dataset size of ~2,000 rows, deep learning models (Neural Networks) would be prone to rapid overfitting and would lack interpretability. Random Forest offered the "Goldilocks" solution: complex enough to capture the nuance of the Snob Pit, but simple enough to train reliably on a small dataset.

HYPER PARAMETERS TUNING

My final model had a configuration of

n_estimators=250

max_depth=9,

random_state=42,

min_samples_split=20

,max_features=8

Final rmse score 13.21

Whereas my baseline model had a configuration of:

n_estimators=100

max_depth=not specified

random_state=42,

min_samples_split=not specified

,max_features= not specified

Rmse score :greater than 14

Firstly I tuned my baseline model n estimators to 200 which increased my rmse to 13.9 but on increasing it above 300 again started givinf an rmse of above 13.9.

Then making max depth 7-9 made an rmse score of 13.4 to 13.45 . After making min samples split 10 model had little improvement and on keeping it between 17-21 it had an rmse less than 13.3 everytime 20 min samples split prevented our model from overfitting.
Max_features 8 making max features less makes each tree to take only some features and workd independently.

VENUE ANALYSIS

VENUE	KEY FINDINGS	OPTIMAL STRATEGY
V_BETA	TIMING ,morning shows had almost zero energy as they are goths.	TRY to keep late nights shows here as much as possible
V_GAMMA	**Behavior:** The most unique and profitable anomaly in the dataset.They barely care about the weather ,pricing(snob effect) and other things. OPENER RATING matters a lot to them ..merch sales are also highest here	This is a "Veblen Good" market. The optimal strategy is **Premium Pricing ($150)**. Lowering prices here actually *hurts* the vibe and reduces total profit.

V_ALPHA	VOLUME LEVEL matters. Price sensitivity is neutral	Try to keep a medium VOLUME LEVEL 5-7.This increases the crowd energy the most.
V_DELTA	They love high volume level. Venue is open so gets affected by storms and rain.	Maintain a high VOLUME LEVEL 10/11. Try to arrange concerts in a good weather.