Findings Report

Venue Wise Report

V_Alpha:

- *Volume limit. Volume doesn't go above 7.*
- *However, there are many shows at max volume.*
- *There was a killer afternoon show.*
- *Ticket price or Opener rating didn't matter much.*

V_Beta:

- *Comes alive at night. Afternoon and morning shows are flop.*
- *There are free shows here.*

V_Gamma:

- *Ticket price is the most important factor here.*
- *Higher ticket price leads to higher crowd energy.*
- *Opener Rating also matters here. Higher opener rating results in more energetic crowd.*

V_Delta:

- *Higher volume leads to slightly better shows on average.*
- *Ticket price range has been narrower here than the others.*

Validating Singer's Theories

Tuesday Shows are CURSED!

There is no clear evidence of Tuesday shows having any significant deviation than any other day. However, weekend shows ARE better by a bit.

I am CONVINCED our best shows happen during full moons.

No, the data suggests that there is no clear indication that moon phases affect the shows at all. The drummer is right here!

Actually, speaking of socks - does outfit color matter?

It is clear from the visualizations that the outfit colour or type does not matter significantly.

Also, I've been tracking the weather. Rain definitely sucks. But does it actually affect energy or does it just affect MY mood and I play worse

Weather doesn't affect the crowd energy much. Probably just affects his mood. However, V_Delta has had some great shows during clear days.

The packed shows were INSANE. The empty shows were dead. (V_Delta)

There is a slight relationship between the crowd size and the crowd energy. The effect is more pronounced at V_Alpha though.

1. **"Volume matters" vs "Crowd size matters more"**

Volume definitely matters more, especially for certain venues.

2. **"Weekends are always better" vs "That Tuesday in London was fire"**

Weekends are ON AVERAGE better, but not always.

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

Only snobs care about the price (V_Gamma).

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

Timing only really matters for the goths (V_Beta)

5. **"Moon phase is real" vs "The drummer says I'm crazy"**

The drummer is definitely right. There is no relatioin between moon phases and the crowd's energy.

6. **"Weather affects energy" vs "Weather affects MY mood"**

The latter is true. Weather does not seem to have much of a correlation with crowd energy.

Model Choice Justification

Due to the large number of features and the different characteristics of each venue, a linear regression will fail to learn the different characteristics. A ***decision-tree based model*** *would work best here. Hence we can use a* ***Random Forest Regressor****.*

However, a Random Forest Regressor fails to produce a decent R^2 score, of around 0.06, indicating that it was unsuccessful in picking up the trends.

From the graphs and visualizations, it is clear that it is a very noisy dataset, where trends are overshadowed by the variations. To overcome this, we can use ***Gradient Boosting****, which allows tree-based models to pick up trends through the noise. Therefore, I have decided to use an* ***AdaBoost Regressor model****.*

Features Used

- *Venue_ID*
- *Volume_Level*
- *Ticket_Price*
- *Opener_Rating*
- *Is_night_show*
- *Is_weekend*

Hyperparameter Tuning

Parameters explored for AdaBoostRegressor:

- ***n_estimators**: [200, 500, 1000, 1200, 1500, 2000]*
  This is the maximum number of trees in the sequence (forest).
- ***learning_rate**: [0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0]*
  This controls the weights applied to each tree (regressor) at each training iteration.
- ***loss**: ["linear", "square", "exponential"]*
  This is the loss function used when updating the weights after each boosting iteration.

*Validation Strategy used: **k-fold cross-validation** with **R^2 score** as metric.*

Final Hyperparameter values:

```
n_estimators = 500

learning_rate = 0.01

loss = exponential
```

Comparison (before vs after tuning):

Before: `Mean R2: 0.4202944177832036, Mean RMSE: 16.223554543926767`

After: `Mean R2: 0.449703034335129, Mean RMSE: 15.796460194910976`