ML MODULE - EDA

CODING WEEK 2025

AYUSH MISHRA
B.Tech (CSE'29)

Breaking Down J's Conclusions: Hits and Misses

Ticket Price

Ticket price **does not appear to strongly influence** crowd energy. Most energy values **cluster between 50–80** across all common price points ($40–$80). There is no clear evidence of a sweet spot at $62 or $80. So J's wrong here.

During analysis, I could not find any negative value, or outlier (Ex. $500). The **max value was around $120**. **V_Beta don't seem to care much for price**. **V_Gamma on the other hand show decrease in enthusiasm** with price increase! Also, **Compensated** Shows have **high energy**!

Moon Phase

I investigated the number of shows with **energy greater than 75**, and made a moon phase distribution.

J made a **blunder** in assuming that **Full Moon** is the best, it's the **worst** rather. Other than that, **moon phase doesn't really matter** a lot (Drummer wins this round!).

Day-Time of Week

I again made a distribution similar to that for Moon Phase. **Tuesdays are truly not ideal!** They consist of **just 9% of High Energy** shows. On the other hand, **weekends** account for **more than one-third** of enthusiastic crowd shows.

Goths i.e. V_Beta **come alive with a different spirit in Late Night shows**! J was right on spot here.

From the bar graph, I concluded that all the High Energy Shows have been Late Nighters in V_Beta.

Also, J rightly remembered **about a killer Afternoon Show!** I could find an **Afternoon Show** with crowd **energy of staggering 96**!

Weather

Weather **doesn't play a determining role** in energy of crowd. On finding the distribution, I could conclude that about **27 % of High Energy Shows have been during Rainy Weather** and **around 25 % during Clear Skies.** On a lighter note, **Crowd Energy drops** a bit during **Stormy Shows** for obvious reasons!

Band Outfit

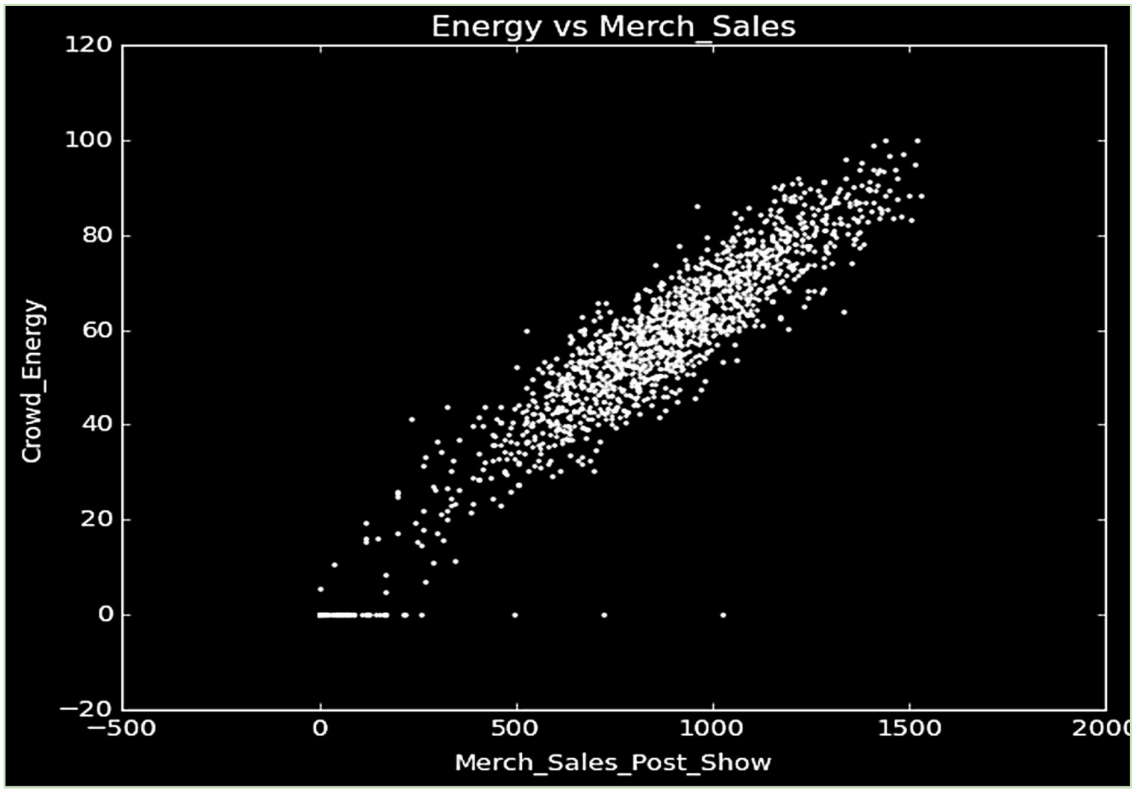
J's apprehension about proper band outfit have been answered! It turns out that whenever band dresses in **Leather outfit**, it has contributed to **nearly 45% of the high energy shows**. Also, Spandex may no longer be the go-to-choice of the band after this EDA! **Spandex outfit shows constituted just 20 % of great shows.**

Opener Rating

Drummer hasn't been exactly to the point, when he insisted on Opener Ratings. It turns out that **Opener Ratings don't really matter to the energy of Crowd significantly**. Afterall, the band needs to manage the whole show for great energy. Upon inspecting the scatter plots, V_Gamma i.e. **Snobs show somewhat dependence on Opener Rating**, other than that none!

Merch Sales Post Show

The Crowd_Energy is strongly related to how the sale of Merch is.



There is a strong linear dependency between the Crowd_Energy & Merch Sales. So, to ensure post show sales, focus on crowd energy!

Volume_Levels

There is a **lack of strong correlation between Volume_Levels and Energy of Crowd.**

Upon inspecting for each of the four venues individually, I could find that shows with a **large range of volumes** have been **conducted at V_Beta**, but with little to no variation in Crowd_Energy. For others, the scatter plot is very dense rectangle like. J was wrong on assuming bias of a venue to level of volume.

Crowd Size

The overall scatter plot of Crowd Energy vs Size, is uniform, nearly **circular in shape**. This suggests **lack of strong correlation** overall.

Exploring Venue Wise as well, there was lack of conclusive dependence of number of attendees and their energy. J's assumption of exponential or logarithmic variation at V_Delta, doesn't seem true.

Model Choice Justification

Since the data obtained after cleaning, is highly structured and tabular, I was confident of applying Decision Trees to the dataset. Among, Decision Trees, XGBoost is by-far the most enhanced technique. It gives the flexibility of choosing the best among multiple trees, while also working on the shortcomings of previous trees.

During Hyperparameter Tuning with GridSearchCV, I tuned

(a) n_estimators with values around 500 – 2000
(b) learning_rate from around 0.03 to 0.1
(c) max_depth with values like 2,4 among other
(d) reg_lambda with values 1,2.5,5. I had also tried larger and smaller sets, before settling on these.

Finally, I was able to generate maximum R2 score of 0.663 on the test_set, using the following values:

n_estimators: 500

learning_rate: 0.03

max_depth: 4

reg_lambda: 5