Energy in Music

Introduction

My name is Eric Lin, a 3rd year Computer Science student at Rutgers, New Brunswick. I have taken data science courses ever since my interest in data started during my first year at university. Naturally, DATA101 at Rutgers falls into my line of interest. Besides data science and programming, I am a musician. I have been playing violin and piano since I was three years old. And even now that I am much older, I still love making and playing music for myself as well as others. So a lot of my personal projects usually have something to do with music. In this case, I decided to look at trends in YouTube song analytics.

The Dataset

This dataset is a collection of analytics by Salvatore Rastelli from popular songs in 2023 that can be found on both YouTube and Spotify. The dataset includes basic categorical variables such as track name, artist, album type, etc., and basic numerical variables such as YouTube views, Spotify streams, likes, comments, etc. However, it also includes some special numerical variables such as danceability, loudness, speechiness, valence, and more. These numerical variables are defined by the author of the dataset and values between 0 and 1. In my exploration, I will mainly be focusing on 3 select variables: YouTube views, Energy – the perceptual measure of intensity and activity in a music track – and Licensed – whether or not the YouTube music video is officially licensed and claimed by a YouTube partner.

Cleaning Data

Before anything else, it is always important to clean the data you are planning to use. In my case, I am observing the relationships between musical energy, YouTube views, and licensed content. I noticed with this dataset, in all three categories there are missing or empty values. Before I start making any observations, I clean up those empty values.

```
songs$Views[is.na(songs$Views)] <- 0
songs$Energy[is.na(songs$Energy)] <- 0
```

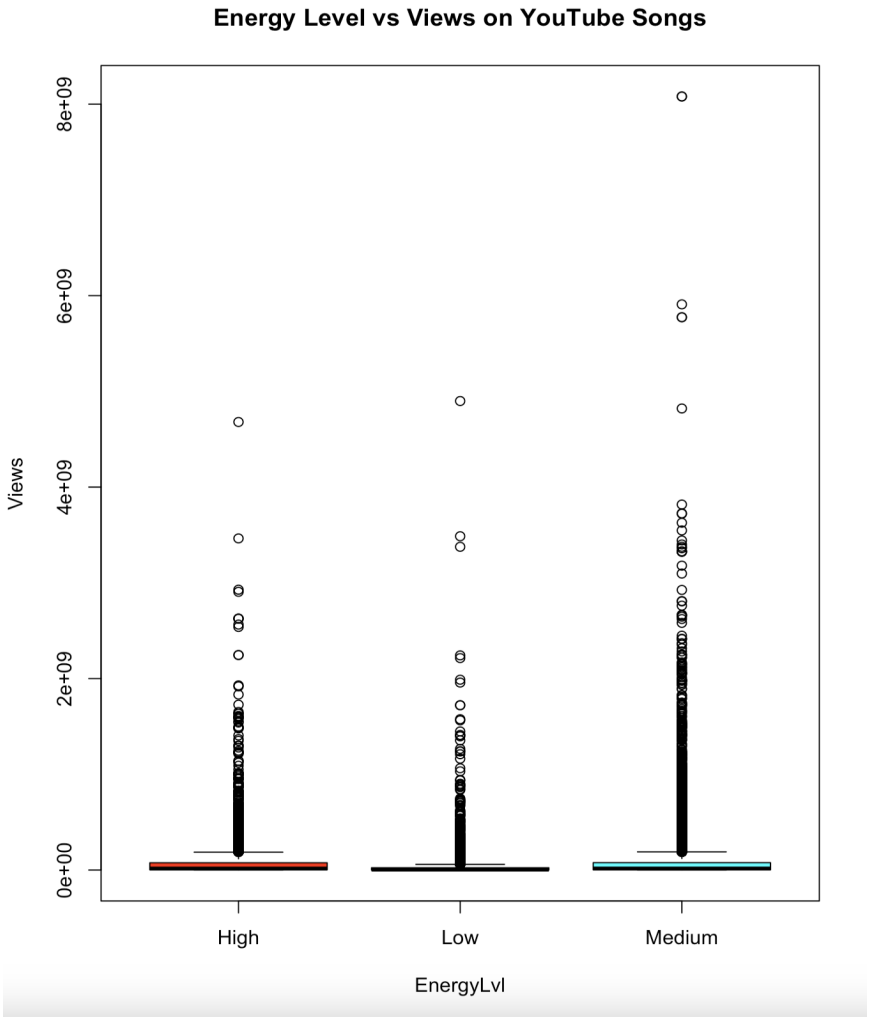
```
songs_yt <- songs_yt[songs_yt$Licensed == 'True' | songs_yt$Licensed == 'False',]
```

Preprocessing

I also want to categorize my numerical energy variable if I want to see how energy affects views or licenseship. I can do that by breaking energy into three energy level categories. First, I find the mean of energy, then one standard deviation away from the mean on both sides. I categorize this as average energy or "Medium." Anything above this threshold or in other words, any above one standard deviation away, I can categorize as above average or "High." Likewise, anything below this threshold will be categorized as below average of "Low."

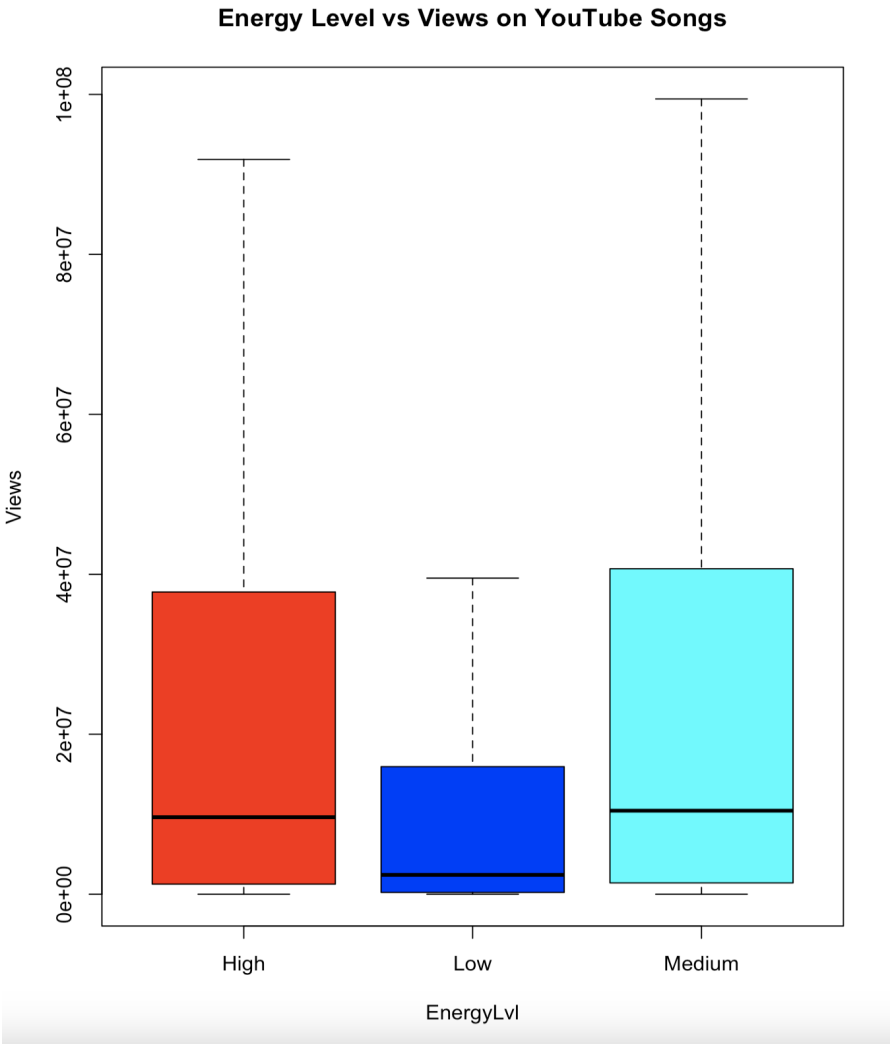
```
songs_yt$EnergyLvl <- ifelse(songs_yt$Energy >= mean + sd(songs$Energy), "High",
                      ifelse(songs_yt$Energy >= mean - sd(songs$Energy), "Medium",
                      "Low"))
```

Removing outliers is also a big part of preprocessing. Using a boxplot, I plotted Energy Levels against YouTube views and found more than a few major outliers.



Findings #1: Permutation Testing

Upon removing outliers, we can make a better observation about energy levels of music on the number of views it can obtain on YouTube. From the new plot below, we can clearly see that the higher Energy Levels have much higher views than the lower Energy Level.



Using the tapply() function, we can also see the specific view count means of each Energy Level.

```
> tapply(songs_yt$Views, songs_yt$EnergyLvl, mean)
    High      Low   Medium
27450154 15335097 28241351
```

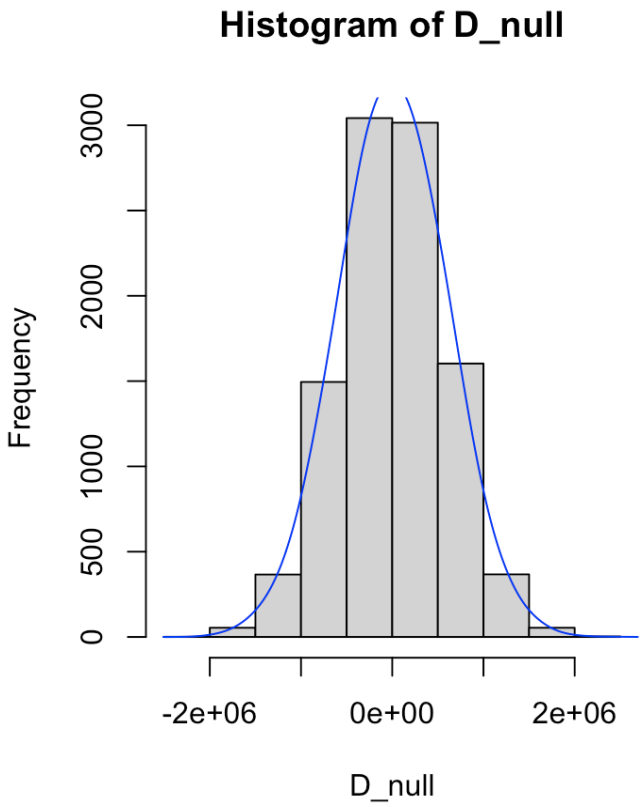
From these differences, I came up with two hypotheses. And when you are conducting multiple hypothesis tests simultaneously, you will need to correct for the significance level using Bonferroni Coefficient. In this case, there are only two hypotheses we are testing, so we simply need to divide the standard significance level (0.05) by 2 which results in a new significance level of 0.025. With this significance level in mind, I can test the following hypotheses:

1.)
Null Hypothesis: Songs with medium energy levels have the same view counts as songs with lower energy levels.

Alternative Hypothesis: Songs with medium energy levels have higher view counts as songs with lower energy levels.

To test this hypothesis, I ran the following permutation test:

```
> permutation_test(songs_yt, 'EnergyLvl', 'Views', 10000, 'Low', 'Medium')
[1] 0
```



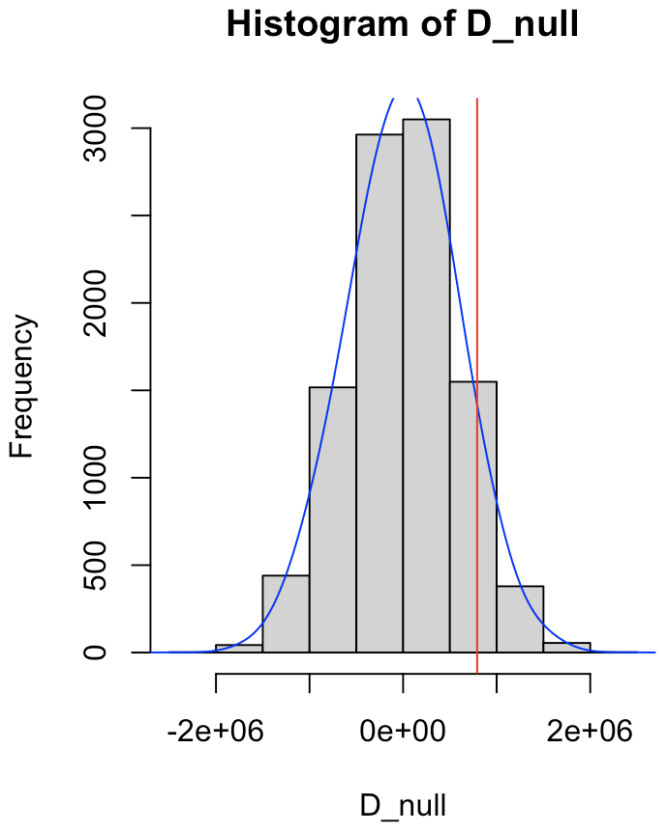
The resulting p-value was a 0. While that is practically impossible, it just means that our p-value is an extremely small number close to zero. Because our p-value is less than 0.025, that indicates enough evidence to fully reject our original null hypothesis stated above.

2.)
Null Hypothesis: Songs with medium energy levels have the same view counts as songs with high energy levels.

Alternative Hypothesis: Songs with medium energy levels have higher view counts as songs with high energy levels.

To test this hypothesis, I ran the following permutation test:

```
> permutation_test(songs_yt, 'EnergyLvl', 'Views', 10000, 'High', 'Medium')
[1] 0.0889
```



The resulting p-value was 0.0889. While it is close, our p-value is greater than 0.025, which indicates there is not enough evidence to reject our original null hypothesis stated above.

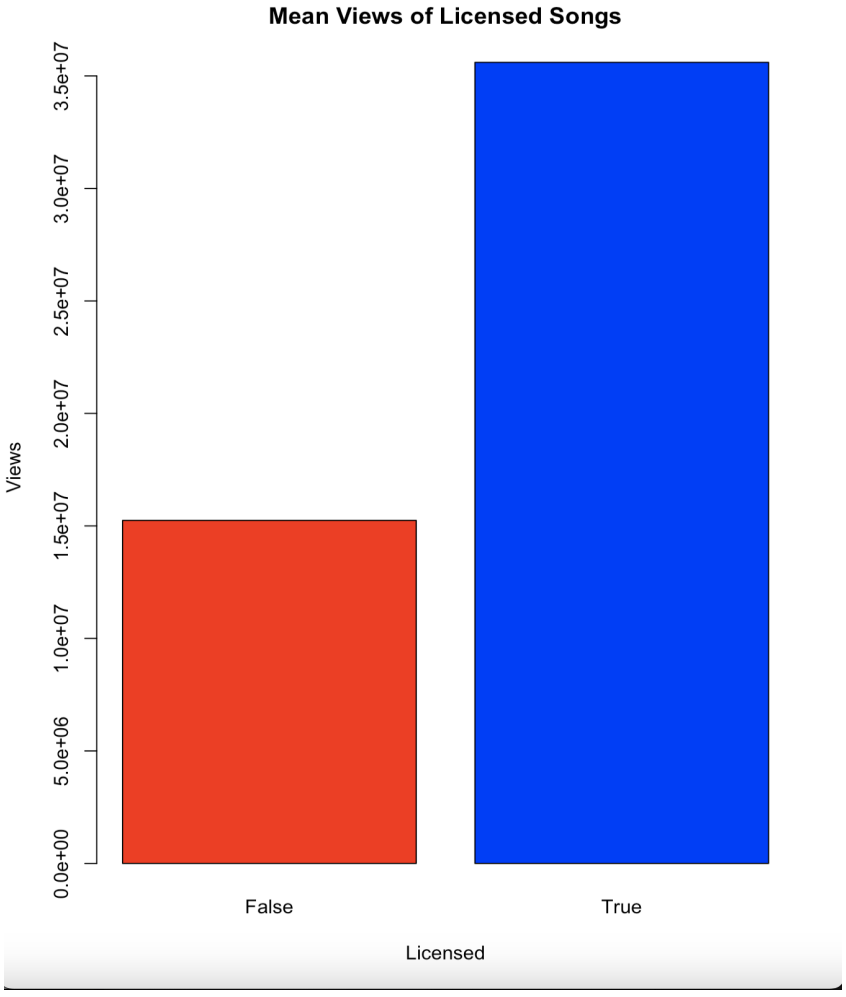
Conclusion

We can conclude that while there is likely a relationship between higher views in medium vs low energy levels in music, there is nothing that we can definitively say about higher views and high energy levels. All in all, this is expected and somewhat makes sense when you think about how

popular music is trending nowadays. While some people enjoy slower “chill” songs, a majority of people who use social media tend to be attracted toward music with faster beats and higher energy. It gives them that kick and excitement you do not often have on a boring never-changing day.

Findings #2: Bayesian Odds

Calculating Bayesian Odds allows us to see the likelihood and probability of an event occurring given conditions. In this case, I wanted to know if songs by licensed artists or companies garnered more views. From our previous finding, I learned that medium energy music had the highest mean of views and it definitively obtained more YouTube views than low energy music. So I came up with the following Bayesian Odds Task: What are the odds that a song with medium energy and above average views is an officially licensed music video?



In the graph above, I plotted the mean views of Licensed vs Not-Licensed YouTube songs with medium energy levels. As expected, we can see that Licensed songs obtain more average views.

Using Bayesian Odds, I calculated the following:

```
> Prior <- nrow(songs_yt[songs_yt$Licensed == 'True',]) / nrow(songs_yt)
> Prior
[1] 0.6653366
> PriorOdds <- Prior / (1 - Prior)
> PriorOdds
[1] 1.988077
> TruePositive <- nrow(songs_yt[songs_yt$EnergyLvl == 'Medium' & songs_yt$Views > avg & songs_yt$Licensed == 'True',]) / nrow(songs_yt[songs_yt$Licensed == 'True',])
> TruePositive
[1] 0.2870117
> FalsePositive <- nrow(songs_yt[songs_yt$EnergyLvl == 'Medium' & songs_yt$Views > avg & songs_yt$Licensed != 'True',]) / nrow(songs_yt[songs_yt$Licensed != 'True',])
> FalsePositive
[1] 0.1127576
> LikelihoodRatio <- TruePositive / FalsePositive
> LikelihoodRatio
[1] 2.545386
> PosteriorOdds <- LikelihoodRatio * PriorOdds
> PosteriorOdds
[1] 5.060423
> Posterior <- round(PosteriorOdds / ( 1 + PosteriorOdds), 2)
> Posterior
[1] 0.83
```

Obtaining a posterior of 0.83 indicates that there's an 83% chance that a song with medium energy levels and above average views is by a licensed creator or company.

Possible Flaws

When going about finding the qualities that affect view count or licenseship of music, I only took into consideration at most two variables. This is obviously not how the real world works and the dataset contains numerous complex special variables that could be features in predicting view count of YouTube videos.

Why Do These Findings Matter?

It's important to find out and identify why things are the way they are for any field, especially ones you are passionate about. Music is everywhere and the industry is ever-growing. Music might not be a viable career right now for most, but it is crucial to identity trends. That way, we can improve on what type of product we produce and for what audience – not just for music tracks online. Besides, data is fun to play around and humans naturally like to put complex things into numbers to help better contextualize and understand them. Whether it be music or the next biggest discovery in the world, we should always strive to identity what and how things are happening and make conclusions based on that.