Cody Rorick

Final Project of MA4270 – Design and Analysis of Experiments

Relationship between total Spotify streams of a song and the key and the BPM of it

Section 1 Introduction

The number of Spotify streams a track gets is now one of the most important metrics to determine how successful it is deemed. Financial compensation, sponsorship deals, and music awards are tied to the number of Spotify streams that an artist gets on a song. Therefore, it is important for artists to know what variables play a role in total stream counts, and what will give them the best chance of maximizing their streams.

The streams variable contains the total number of streams on Spotify, this is used as the response variable in the analysis. The bpm (beats per minute) variable is a continuous quantitative variable that is used as a measure of song tempo. This variable was transformed into a categorical variable named bpm2 grouped into 5 levels with an approximately equal number of observations in each level using the Hmisc library in R. This was done to turn bpm into a discrete variable with levels for analysis purposes. The groupings of levels are ranges 65-97 bpm, 97-114 bpm, 114-129 bpm, 129-147 bpm, and 147-206 bpm. The key variable is the musical key of the song. The keys included in this data are B, C#, F, A, D, F#, G#, G, E, A#, D# which are 11 unique keys. There are 857 different songs across many different genres included in this dataset. The main purpose of this analysis is to discover if there is a relationship between the total number of streams that a song gets on Spotify and what key it's written in, the beats per minute of it and the interaction effect of the key and the beats per minute with it.

This dataset is observational data that was extracted from multiple data sources using the Python programming language according to its authors. This data was retrieved from Kaggle, a large online repository of various datasets.
(https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023)

Section 2 Statistical Methods

The R version 4.3.0 (Released 2023-04-21) (https://www.r-project.org/) was used in the analysis. The overall significance level was set as 0.05.

Because the number of observations in each treatment group are not identical the type III sum of squares will be used in order to investigate the ANOVA table. The Tukey method will be used to investigate the honest significant differences between any pairs of group means to see which pairwise differences exist.

Section 2.1 Exploratory Data Analysis

A box plot was obtained to see how total streams differed across the 11 keys and 5 BPM categories. An interaction plot was created to see if there were any interaction effects between the 11 treatment keys and 5 BPM categories on total Spotify streams.

Section 2.2 ANOVA

A two-way complete analysis of variance model was created. This model included the key the music was written in, the beats per minute of the music, and the interaction between these variables. The residual plots and the Q-Q plot of residuals were used to check model assumptions such as if there were any outliers, if there was independence, if there was equal variance, and if the normality assumption was met. The analysis of variance table was obtained to see if the total number of streams significantly differed across the treatment combinations of 11 musical key and 5 levels of beats per minute, and if there were any interaction effects. For significant effects, the simultaneous 95% confidence intervals for the pairwise comparisons of each term in the model were constructed to see which levels (or treatment combinations) have different effects.

Section 3 Results

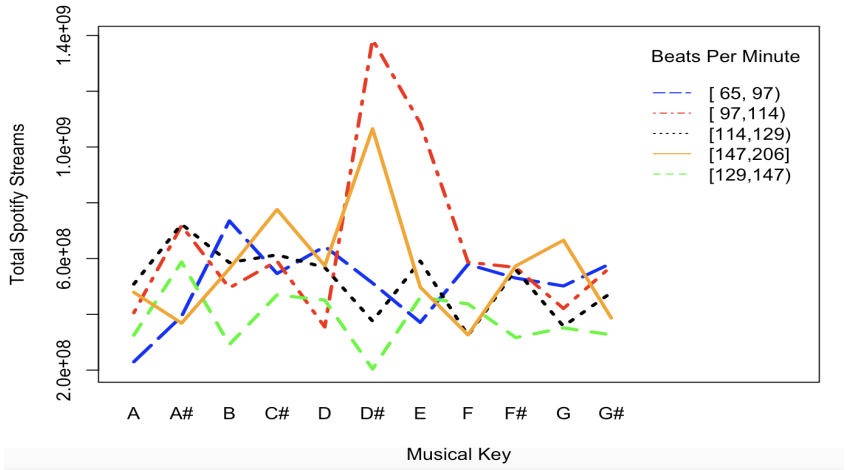

Figure 1 Interaction plot between bpm and musical key on total Spotify streams

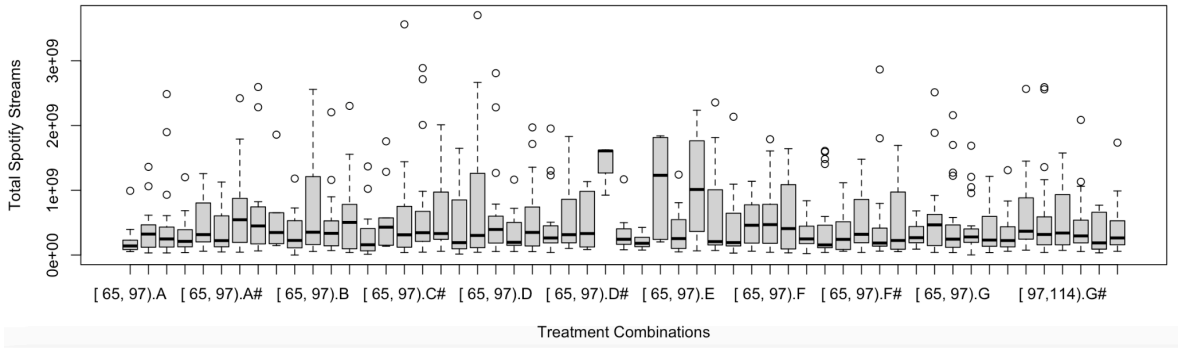

Figure 2 Treatment Combinations vs Total Spotify Streams

Table 1 Analysis of Variance Table utilizing type III sum of squares

Variable	Degrees of Freedom	Sum of Squares	Mean Square	F Value	P Value
bpm2	4	5.1394e+18	1.28485e+18	3.9862	0.003288
key	10	4.3756e+18	4.3756e+17	1.3575	0.1956
bpm2:key	40	1.4181e+19	3.54525e+17	1.0999	0.312037
Residuals	802	2.5850e+20	3.2232e+17	-	-

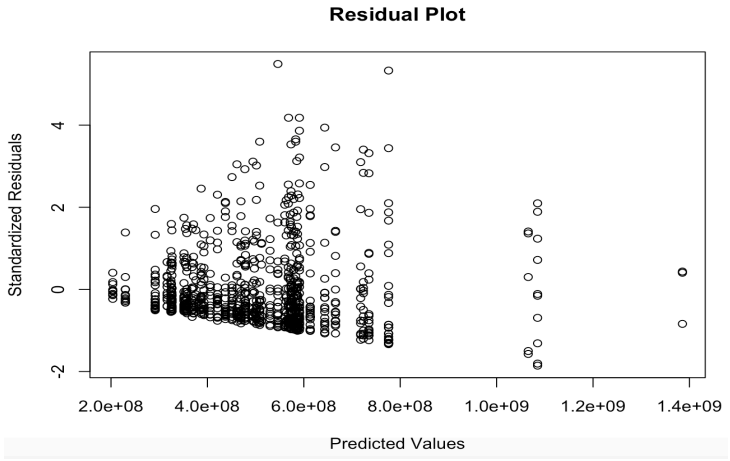

Figure 3 Residual Plot

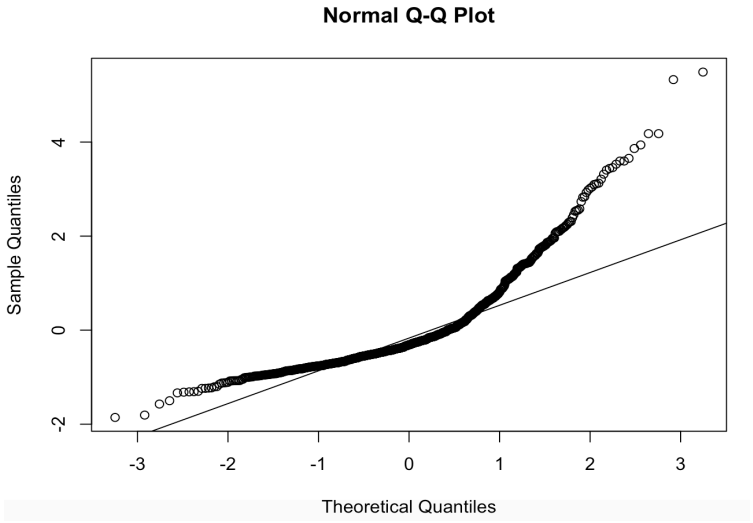

Figure 4 Normal Q-Q Plot

```
 contrast               estimate       SE  df  lower.CL  upper.CL t.ratio p.value
 [ 65, 97) - [ 97,114) -1.42e+08 68472350 802 -3.29e+08  45596810  -2.068  0.2350
 [ 65, 97) - [114,129) -5.54e+06 62930445 802 -1.78e+08 166511886  -0.088  1.0000
 [ 65, 97) - [129,147)  1.27e+08 64246914 802 -4.82e+07 303136338   1.984  0.2746
 [ 65, 97) - [147,206] -5.93e+07 65299782 802 -2.38e+08 119265895  -0.908  0.8940
 [ 97,114) - [114,129)  1.36e+08 69278969 802 -5.33e+07 325474847   1.964  0.2847
 [ 97,114) - [129,147)  2.69e+08 70476951 802  7.64e+07 461775356   3.818  0.0014
 [ 97,114) - [147,206]  8.23e+07 71438059 802 -1.13e+08 277654040   1.153  0.7782
 [114,129) - [129,147)  1.33e+08 65105904 802 -4.50e+07 311024342   2.043  0.2464
 [114,129) - [147,206] -5.37e+07 66145101 802 -2.35e+08 127116520  -0.812  0.9269
 [129,147) - [147,206] -1.87e+08 67398809 802 -3.71e+08  -2481086  -2.771  0.0452

Results are averaged over the levels of: key
Confidence level used: 0.95
Conf-level adjustment: tukey method for comparing a family of 5 estimates
P value adjustment: tukey method for comparing a family of 5 estimates
```

Figure 5 Contrasts Between BPM

A pairwise analysis was done on the beats per minute of the music to see which ranges varied from each other in terms of total Spotify streams. It was shown at the 0.05 significance level that music written with bpm in the range of 97 to 114 achieved a statistically significant amount of streams greater than music written with bpm in the range of 129 to 147. It was also shown that music written with bpm in the range of 147 to 206 achieved a statistically significant amount of streams greater than music written with bpm in the range of 129 to 147.

The normality assumption does not appear to be met, the normal curve shows that the data skews high at both the low end and the high end of the theoretical quantiles. The equal variance assumption does not appear to be met either because as the predicted values get larger, the variance appears to fan outwards with those predicted values. The data also appears to contain outliers, as there is shown to be songs with standardized residuals greater than 4 standard deviations away from the predicted values.

Section 4 Discussion

The outcomes of this analysis have been insightful into the investigation of musical keys and bpm impact on the total number of Spotify streams. The anova table showed that the hypothesis of bpm having no impact on the total number of Spotify streams should be rejected, as the p-value was significantly lower than the alpha level assigned for significance at the beginning of the experiment. In this analysis, the hypotheses that musical keys are negligible and the interaction effects between musical keys and bpm are negligible on the total number of Spotify streams fail to be rejected. This does not mean that the key and bpm/key interaction do not have an impact on total Spotify streams, but this analysis just failed to find a significant effect.

*The results obtained from this analysis should be looked at with scrutiny, as the assumptions for Anova testing such as normality, equivalent variance, and lack of outliers are suspect.