REPORT of ASSIGNMENT 2: Factor models

Group 8:
Guus Bouwens (2701442)

15th of March 2024

1 Introduction

This report covers a factor analysis based on a dataset of songs to uncover underlying patterns within musical attributes. The dataset, "songlist.xlsx", comprises a selection of a total of 3090 songs from 50 distinct artists characterized by nine numeric variables: danceability, energy, loudness, acousticness, liveness, valence, tempo, popularity, and duration in seconds. Descriptive statistics for these variables can be found in Table 3. Each variable represents a distinct aspect of the music, providing a holistic perspective on the listening experience.

This analysis begins with the construction of an orthogonal factor model (OFM) utilizing three factors, followed by the application of the VARIMAX rotation. This step is pivotal in maximizing the interpretability of the factors, thereby enhancing our understanding of the data its structure. The subsequent sections report the parameters of the OFM, delve into the meaning behind the three factors, and interpret these in the context of practical application, aiming to find the constructs that these factors represent within the realm of music.

Further, the report delves into the communalities and specificities of the model, offering insights into the shared variance among variables and the uniqueness belonging to each one. This analysis is important to understand the extent to which individual song characteristics are explained by the factors.

The visual aspect of the investigation is addressed by plotting the first two factors and coloring each artist uniquely to show potential commonalities and disparities in the dataset. This shows the similarities and differences among artists but also highlights specific outliers and their distinct

musical signatures.

The objective of this report is not to just present statistical findings but to combine them into a study that is interesting for both academics and people within the music industry.

2 a) Orthogonal Factor Model Estimation with VARIMAX Rotation

In this section, we construct an orthogonal factor model (OFM), defined as below, utilizing a dataset comprised of various song attributes, each quantified by nine numerical metrics: danceability, energy, loudness, acousticness, liveness, valence, tempo, popularity, and duration in seconds.

$$\text{Orthogonal Factor Model: } X = \mu + \mathbf{QF} + \mathbf{U}$$

With restrictions: $\mathbb{E}[\mathbf{F}] = 0$, $\mathbb{E}[\mathbf{U}] = 0$, $\text{Var}[\mathbf{F}] = \mathbf{I}_m$, $\text{Var}[\mathbf{U}] = \text{diag}(\psi_1, \ldots, \psi_p)$, and $\mathbb{E}[\mathbf{F}\mathbf{U}'] = 0$.

Where X represents the observed variables, μ is the mean vector of these variables, indicating their average levels. The term $\mathbf{QF}$ involves $\mathbf{Q}$, the matrix of factor loadings that links observed variables to the underlying latent factors, and $\mathbf{F}$, the factor scores that represent these latent factors. The model assumes these factors are unobserved but influence the observed data. Finally, $\mathbf{U}$ denotes the unique factors or error terms, which capture the variance in observed variables not explained by the latent factors. The model sets several conditions to maintain orthogonality and independence: the expected values of $\mathbf{F}$ and $\mathbf{U}$ are zero, indicating they are centered around the origin; the variance of $\mathbf{F}$ is an identity matrix $\mathbf{I}_m$, ensuring factors are uncorrelated and standardized; the variance of $\mathbf{U}$ is a diagonal matrix, suggesting these unique factors are independent and vary across variables; and there is no covariance between $\mathbf{F}$ and $\mathbf{U}$, maintaining their orthogonality.

Our methodology starts with the standardization of these attributes to ensure a uniform scale, a crucial step for the accuracy of the subsequent factor analysis. It is mathematically represented as:

$$x = \frac{(z - \mu)}{\sigma}$$

where z is the original value, μ is the mean, and σ is the standard deviation of the variable. This standardization brings means and standard deviations close to 0 and 1, respectively, affirming the dataset's preparedness for further analysis (for proof, see Table 4 in the appendix).

The foundation of our factor model, a covariance matrix (Σ), detailed in Table 5, undergoes eigenvalue decomposition to unravel the dataset's underlying structure.

Σ, is computed as:

$$\Sigma = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

where $\mathbf{X}$ is the matrix of standardized data, and n is the number of observations.

And the eigenvalues (λ) and eigenvectors ($\mathbf{v}$) are computed as: $\Sigma\mathbf{v} = \lambda\mathbf{v}$. This decomposition pinpoints three predominant factors, dictated by the eigenvalues: 2.703, 1.874, and 1.089, suggesting significant dimensions within the musical attributes under examination. These factors were selected based on the Kaiser criterion, which favors factors with eigenvalues exceeding one, highlighting their importance in explaining variance more effectively than individual variables.

Subsequently, a VARIMAX rotation was applied to these selected factors to facilitate interpretability. It is calculated as

$$(\sum_{j=1}^{r}\left(\frac{1}{p}\sum_{i=1}^{p}\left(q_{ij}^4 - \frac{1}{p}\sum_{k=1}^{p}q_{kj}^2\right)^2\right),$$

where q is an element in the loading matrix, p is the number of variables, and r is the number of selected factors.

This rotation technique, which aims to maximize the variance of squared loadings within each factor, confirmed the orthogonality of the factors. The near-zero dot products between factor pairs validated the successful orthogonalization post-rotation. It was calculated as $\mathbf{r}_i \cdot \mathbf{r}_j \approx 0, \quad \text{for } i \neq j$ where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the rotated factor vectors.

In the next section, we will dive into the results of the rotation and their implications.

3 b) Interpretation of Orthogonal Factor Model Parameters and Factor Representations

The Orthogonal Factor Model (OFM) parameters, specifically the Factor Score Coefficient Matrix (Q) and the Unique Variances (Ψ), serve pivotal roles in interpreting the underlying structure of our dataset. The Q matrix translates the factor loadings and eigenvalues into coefficients that express how each original variable is represented in the factor space, while Ψ accounts for the variance unique to each variable, not captured by the model.

The Q matrix provides a refined perspective on the relationship between the original variables and the factors. By multiplying the rotated factors with the square roots of their corresponding eigenvalues, we obtain coefficients that signify the contribution of each variable to the factor scores. This mathematical process transforms abstract factor loadings into tangible coefficients, giving the computation of factor scores for each observation in the dataset. Factors range from -1 (negative) to 1 (positive association), with a value close to 0 indicating a weak association.

Table 1: The OFM parameters, the Factor Score Coefficient (Q) matrix, and the Unique variances (Ψ).

Variable	Factor1	Factor2	Factor3	Ψ
danceability	0.0552	-0.7903	0.1972	0.3338
energy	-0.9243	0.0327	-0.1022	0.1344
loudness	-0.8832	-0.1130	0.1321	0.1900
acousticness	0.8791	0.0014	0.0519	0.2249
liveness	0.0147	-0.0677	-0.7705	0.4019
valence	-0.3037	-0.7788	-0.0191	0.3012
tempo	-0.3604	0.1855	0.1278	0.8197
popularity	0.1708	-0.3191	0.5287	0.5898
duration_sec	-0.2024	0.6986	0.3607	0.3412

To interpret the parameters and their implications to the model, we analyze Table 1. Starting with the Factor Score Coefficients:

Factor 1 - "Calmness" or "Mellowness": Dominated by its strong negative association with 'en-

ergy' and 'loudness' (-0.9243 and -0.8832), Factor 1 captures the essence of songs characterized by their subdued intensity. This notion is backed up by the strong positive association with acousticness (0.8791) which also indicates peacefulness. This dimension reflects a calmness aspect, where lower energy levels suggest a preference for genres that prioritize ambiance over intensity, such as acoustic, ambient, or soft jazz. This factor's prominence, indicated by the highest eigenvalue (2.703), underscores its significant role in differentiating songs based on their energetic values, showing listeners' preferences for more easygoing songs.

Factor 2 - Upsetting "Non-Danceability" or "Rhythmic Complexity": Marked by negative loadings on 'danceability' and 'valence' (-0.7903 and -0.7788), Factor 2 delves into the rhythmic attributes of music that deter danceability. Songs that rank high on this factor will likely feature complex rhythms or loud sounds, making them less attractive to dance on. This factor might encompass genres like heavy metal, where the emphasis is on loud singing and guitar solos, or certain emotional branches of jazz that have rhythmic complexity. The eigenvalue of 1.874 for this factor indicates a substantial variance explained by these characteristics within the dataset.

Factor 3 - "Studio Production Quality": The strongest negative loading on 'liveness' (-0.7705) characterizes Factor 3, distinguishing between the live performance atmosphere and studio-produced clarity. This dimension reflects a preference for polished production over the raw energy of live recordings, aligning with genres that leverage studio technologies to enhance the listening experience. Pop, electronic, and highly produced rock music are probable genres, where studio craftsmanship plays a big role in shaping the final sound. This makes sense because the factor also has the strongest (positive) loading on popularity (0.5287), which those genres typically are. The eigenvalue of 1.089 signifies the relevance of production quality as a distinguishing factor within our music dataset.

The other OFM parameter, Unique variances, represents the proportion of variance for each variable that remains unexplained by the factor model. Essentially, Ψ offers insight into the signatures of each variable, showing which variable its variability is not accounted for by the extracted factors. Lower values of unique variances indicate that the factors successfully capture a substantial portion of the variable's variance, whereas higher values highlight the presence of specific aspects of the variable not well explained by the model. This makes sense because the second and third variables are associated with the most influential Factor 1, and they have the lowest unique variance. Conversely, the tempo (seventh) variable has no real influence on any of the factors and has the highest unique variance, indicating it is badly explained by the OFM.

The relationship between Q, and Ψ paints a comprehensive picture of the dataset's dimensional structure. The Q matrix helps to identify which variables significantly contribute to each factor and quantifies this relationship. Meanwhile, Ψ assesses the model's explanatory power, indicating the variance of each variable. This holistic view gives us a deeper understanding of how each variable interacts within the factor model framework.

4 c) Communality and Specificity in the Factor Model: Implications and Interpretations

In our orthogonal factor analysis with $r = 3$ factors, communality measures the variance of each variable explained by the model, while specificity identifies the variance unique to the variable, not captured by the factors.

For any given variable i, its communality (h_i^2) is determined by the sum of squared loadings of that variable on all factors:

$$h_i^2 = \sum_{j=1}^{r} (a_{ij})^2$$

where a_{ij} denotes the element in the Q matrix of the i^{th} variable on the j^{th} factor, and r represents the total number of factors.

Specificity (u_i^2) is computed as $u_i^2 = diag(\Psi)$. One can also calculate them as $u_i^2 = 1 - h_i^2$, assuming that all variables are standardized, as discussed in section 2.

The results in Table 2 show that certain variables have high communalities, suggesting a big portion of their variance is well-explained by the three extracted factors. This highlights a strong association of these variables with the factors. Conversely, variables with high specificities reveal segments of the dataset containing unique information not fully captured by the factors. This could imply the existence of attributes inherent to these variables, needing the inclusion of more factors for a comprehensive understanding of their variance.

These values together are very similar to the unique variances in the last section. Because the second and third variables are associated with the most influential Factor 1, and they have the lowest specificities also. This is of course because the specificities are equal to the diagonal of the unique variances. Conversely, the tempo (seventh) variable has no real influence on any of the factors and has the lowest communality, indicating it is badly explained by the OFM.

Table 2: Communalities and Specificities for the variables of X.

Variable	Communality	Specificity
danceability	0.6665	0.3335
energy	0.8659	0.1341
loudness	0.8103	0.1897
acousticness	0.7755	0.2245
liveness	0.5984	0.4016
valence	0.6992	0.3008
tempo	0.1807	0.8193
popularity	0.4105	0.5895
duration_sec	0.6591	0.3409

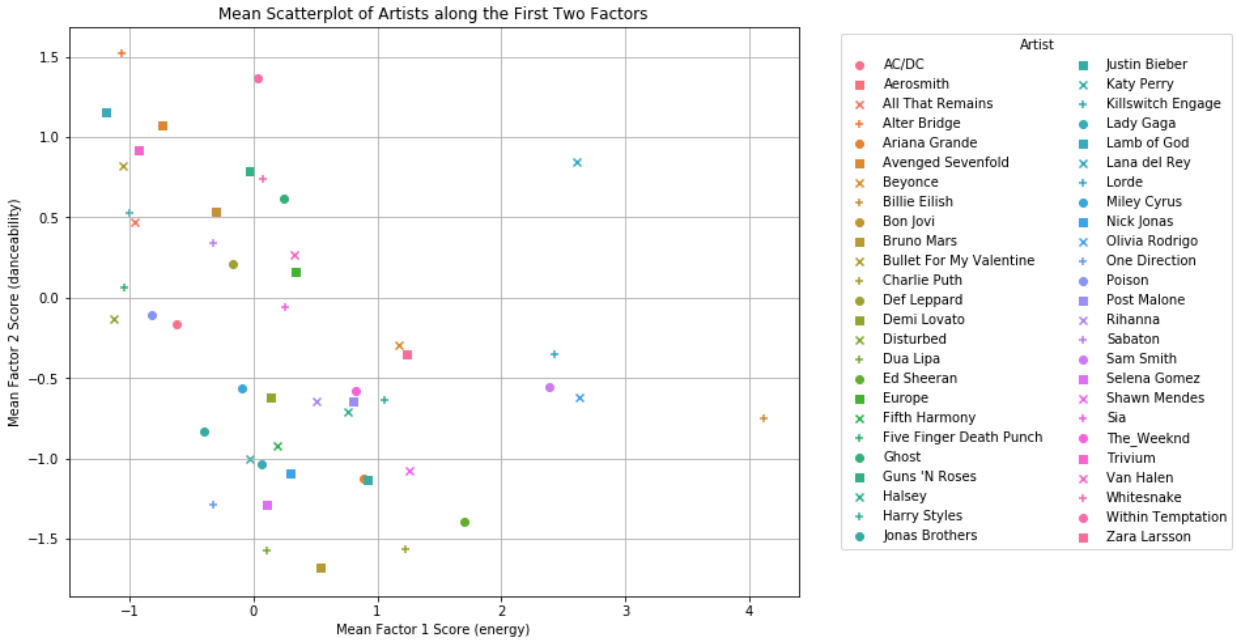
5 d) Analysis of Artist Positioning Based on Energy and Danceability Factors

In this section, we visualize artists on a spectrum of the first 2 factors found in the previous sections. We discuss the visualization and discern patterns from the observations.

The mean scatterplot in Figure 1 of artists along the first two factors reveals distinctive positions that artists occupy in relation to the musical attributes of energy and danceability. The individual values of the artist can be found in Table 6. A scatterplot showing all the songs can be found in Figure 2, it is ommitted here because showing more than 3000 datapoints is not very insightfull when comparing 50 artists.

For Factor 1 (inverse of energy), artists like Billie Eilish, Olivia Rodrigo, and Lana Del Rey exhibit lower energy, positioning themselves as outliers on the higher end of Factor 1, suggesting a tendency towards more mellow, subdued music. On the other hand, artists such as Disturbed, Lamb of God, and Trivium showcase higher energy levels, clustering on the lower end of this factor, which aligns with the intensity often found in rock and metal genres.

Figure 1: Mean Distribution of Artists on Orthogonal Factors Representing Energy and Danceability. Each point represents the average position of an artist's songs in the space defined by the first two factors—Factor 1 (mean energy level) and Factor 2 (mean danceability)—derived from OFM. The plot illustrates the relative positioning of artists within these two dimensions, providing insights into the underlying structure of musical attributes in the dataset.



Factor 2 (inverse of danceability) delineates the degree to which the music is suitable for dancing. Artists like Alter Bridge and Within Temptation are placed higher on this factor, indicating a less danceable quality to heavy metal songs. Conversely, artists such as Bruno Mars, Dua Lipa, and Charlie Puth score lower, suggesting their pop songs feature more rhythmic and danceable elements.

Some patterns found during this analysis in the music streaming industry, by Figure 1:

1. The central (biggest) cluster, where artists like Ed Sheeran, Justin Bieber, and Harry Styles are found, may indicate a balance between energy and danceability, a characteristic of pop music that straddles these attributes. It is not for nothing that the pop genre is short for

popular music, as it maximizes appeal, as evidenced by the biggest cluster in this study.

2. Artists who exhibit both lower energy and lower danceability (high on both factors), like Lorde, might represent genres with more lyrical emphasis, such as indie or alternative styles.

3. Notably, the lower-left quadrant, which would represent high energy and danceability, is sparsely populated (this is especially notable when looking at all songs in Figure 2), suggesting that songs generally do not exhibit both high calmness and high danceability—a reasonable expectation given the negative loading of both factors. This goes together with the first pattern that we addressed.

4. Low-energy songs that do not make you want to dance are the least popular, as the upper-right quadrant is almost empty.

This analysis provides an insightful representation of where artists stand in the energy-danceability spectrum. Music industry stakeholders could utilize this information for artist branding, genre classification, and tailoring music recommendations to listener preferences.

6 Conclusion

In this project, we dived into a dataset of songs to understand the patterns that define musical preferences. Starting with the construction of a factor model using three distinct factors, the goal was to simplify the complex interactions among various musical features into something more straightforward.

The factor analysis process revealed three main themes: 1. Calmness: This theme emerged from songs with lower energy and loudness levels, indicating a preference for genres that are more subdued. 2. Non-Danceability: This factor identified songs that aren't typical dance tracks, possibly due to their intense sounds or dynamic flow. 3. Studio Production Quality: This theme focuses on the production aspect, distinguishing studio-produced tracks from live recordings.

The report also shed light on how well the model captured the essence of the music. While certain attributes like danceability and energy were well-explained, there were unique aspects of some songs that the model didn't fully capture, like tempo. This insight is crucial, as it suggests that while the model provides a solid foundation, there's room to include more factors or refine the

approach to better understand the full spectrum of musical attributes.

These findings are more than just academic; they have real-world applications, particularly in developing music recommendation systems and understanding artist positioning in the music industry. By mapping artists and songs along these dimensions, we can get a clearer picture of their unique sounds. This information is invaluable for marketing, discovering new trends, and predicting the next big hit.

In summary, this project was an exploration of the music industry through a data science lens. The ability to quantify and analyze these factors opens up new possibilities, showing that even something as subjective as music can be understood in a structured way.

7 Appendix

7.1 Introduction

Table 3: Descriptive statistics for the variables in the songlist.xlsx file.

	danceability	energy	loudness	acousticness	liveness	valence	tempo	popularity	duration_sec
count	3090	3090	3090	3090	3090	3090	3090	3090	3090
mean	49.759	75.128	-5.932	14.191	20.960	39.758	123.905	50.224	241.805
std	14.785	22.107	2.743	24.681	16.904	20.938	30.577	16.553	66.185
min	7.320	0.826	-26.383	0.000	1.930	2.720	60.269	0.000	16.000
25%	39.600	60.925	-7.061	0.033	9.910	23.500	99.999	38.000	203.013
50%	49.900	81.500	-5.297	1.340	14.100	37.250	121.001	49.000	231.267
75%	59.100	94.500	-4.116	15.100	28.875	53.275	143.867	62.000	269.543
max	96.800	99.800	-1.347	99.400	99.600	96.900	235.998	96.000	939.139

Table 4: The means and standard deviations after standardisation for the variables in the songlist.xlsx file. They should be close to 0 and 1, respectively.

Variable	Mean	Standard Deviation
danceability	-3.491633e-16	1.000162
energy	2.580460e-16	1.000162
loudness	-6.654152e-17	1.000162
acousticness	-2.076728e-16	1.000162
liveness	-2.461533e-16	1.000162
valence	4.720424e-16	1.000162
tempo	-4.504847e-16	1.000162
popularity	6.500374e-16	1.000162
duration_sec	3.365521e-16	1.000162

Table 5: Covariance matrix for the variables in the standardized songlist.xlsx file.

	danceability	energy	loudness	acousticness	liveness	valence	tempo	popularity	duration_sec
danceability	1.000	-0.194	0.017	0.091	-0.188	0.463	-0.200	0.398	-0.285
energy	-0.194	1.000	0.727	-0.774	0.158	0.157	0.227	-0.290	0.149
loudness	0.017	0.727	1.000	-0.613	-0.014	0.201	0.167	-0.062	0.078
acousticness	0.091	-0.774	-0.613	1.000	-0.111	-0.141	-0.191	0.232	-0.146
liveness	-0.188	0.158	-0.014	-0.111	1.000	-0.047	0.003	-0.212	-0.003
valence	0.463	0.157	0.201	-0.141	-0.047	1.000	0.050	0.149	-0.232
tempo	-0.200	0.227	0.167	-0.191	0.003	0.050	1.000	-0.045	0.025
popularity	0.398	-0.290	-0.062	0.232	-0.212	0.149	-0.045	1.000	-0.092
duration_sec	-0.285	0.149	0.078	-0.146	-0.003	-0.232	0.025	-0.092	1.000

7.2 a)

7.3 d)

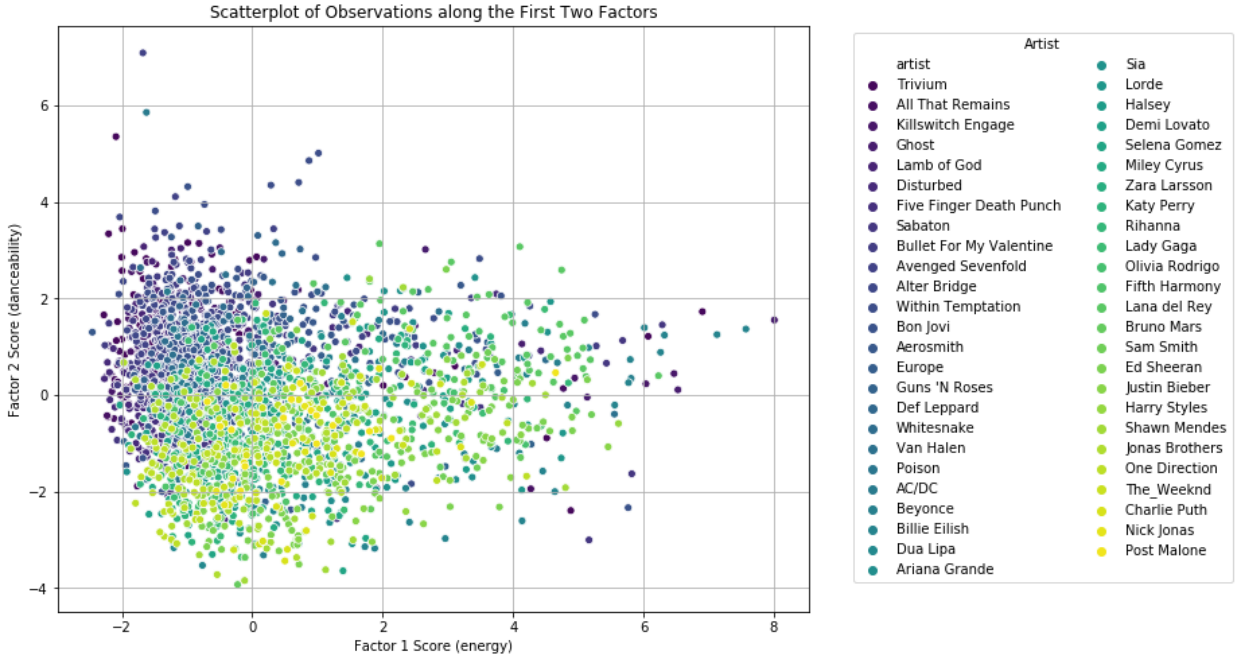


Figure 2: Distribution of Artists on Orthogonal Factors Representing Energy and Danceability. Each point represents the position of an artist's song in the space defined by the first two factors—Factor 1 (energy level) and Factor 2 (danceability)—derived from an orthogonal factor analysis. The plot illustrates the relative positioning of artists within these two dimensions, providing insights into the underlying structure of musical attributes in the dataset.

Table 6: Observations from figure 1.

artist	Factor1	Factor2
AC/DC	-0.623812	-0.162670
Aerosmith	-0.306823	0.532423
All That Remains	-0.957305	0.470631
Alter Bridge	-1.062780	1.518019
Ariana Grande	0.886371	-1.122492
Avenged Sevenfold	-0.739375	1.072003
Beyonce	1.172632	-0.293560
Billie Eilish	4.119681	-0.753919
Bon Jovi	-0.303400	0.536330
Bruno Mars	0.536336	-1.674376
Bullet For My Valentine	-1.054447	0.823257
Charlie Puth	1.222056	-1.567111
Def Leppard	-0.165611	0.207751
Demi Lovato	0.137702	-0.616893
Disturbed	-1.132185	-0.134029
Dua Lipa	0.105628	-1.575599
Ed Sheeran	1.703751	-1.390816
Europe	0.342814	0.159513
Fifth Harmony	0.186463	-0.919677
Five Finger Death Punch	-1.046954	0.060253
Ghost	0.243965	0.620807
Guns ’N Roses	-0.029239	0.787616
Halsey	0.757257	-0.711889
Harry Styles	1.059748	-0.632899
Jonas Brothers	-0.401152	-0.828615
Justin Bieber	0.924184	-1.134395
Katy Perry	-0.028377	-1.002966

Table 7: Table 6 continued.

artist	Factor1	Factor2
Killswitch Engage	-1.002886	0.527546
Lady Gaga	0.061753	-1.031749
Lamb of God	-1.192713	1.152867
Lana del Rey	2.614655	0.844708
Lorde	2.427345	-0.353600
Miley Cyrus	-0.094278	-0.566446
Nick Jonas	0.295458	-1.090745
Olivia Rodrigo	2.631333	-0.617711
One Direction	-0.320969	-1.283495
Poison	-0.823196	-0.110925
Post Malone	0.801256	-0.648195
Rihanna	0.512196	-0.640556
Sabaton	-0.327446	0.342522
Sam Smith	2.389272	-0.550723
Selena Gomez	0.102295	-1.285196
Shawn Mendes	1.255945	-1.071771
Sia	0.252248	-0.059600
The_Weeknd	0.823182	-0.575684
Trivium	-0.926381	0.916823
Van Halen	0.334107	0.269830
Whitesnake	0.071594	0.740650
Within Temptation	0.032387	1.366553
Zara Larsson	1.239788	-0.351051