Bella Almeter, Rani Misra

Cluster Analysis (Taylor's Version)

Fall 2023: STA 3013, Multivariate Analysis

Dr. Anuradha Roy, Ph.D

Data: This data has been collected from Spotify's application programming interface on Kaggle. Spotify is a leading music streaming service that allows subscribers to listen to music from a variety of artists. The data collected is focused on Taylor Swift's discography and provides the name of her songs, albums, their release dates, track number, Spotify ID and URL. Additionally, it measures the acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, and valence. The popularity and duration of the tracks (in milliseconds) is also calculated in the data. The original data set contained 530 observations. We chose to analyze *Red (Taylor's Version)* as Swift herself describes it as her most varied album. In an interview, she described this album as "a real patchwork quilt of genres" (Meyers). Thus, during data cleaning, all tracks except those that are in *Red (Taylor's Version)* were removed. The "Taylor's Version" iteration was given preference to be kept compared to its original counterpart as this album has master ownership by Swift and contains additional songs compared to its original recording. Additionally, *State of Grace (Acoustic Version) (Taylor's Version)* was removed from the data set as it was a duplicate observation of *State of Grace (Taylor's Version)* with more acousticness and was skewing the data. After cleaning, 29 observations remain and the 10 continuous variables to be analyzed are described below:

- Acousticness is a confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence that the track is acoustic.
- Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

- Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- Instrumentalness predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- Liveness detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- Loudness describes the overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db. Based on Spotify's API, the closer a value is to 0, the louder the track is ("Loudness Normalization").
- Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

- Tempo describes the overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- Valence is a measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- Popularity measures the popularity of the song from 0 to 100.

Objective: Using the variables provided in the data, our aim is to separate each of Swift's songs on this album into categories that describe the genre of music it belongs to. We aim to categorize the songs on *Red (Taylor's Version)* into these genres.

Method: We will use cluster analysis to accomplish our goal. Specifically, we will use Ward's minimum variance approach as we are working with raw data rather than distances between observations.

Output:

Obs	name	album	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
1	State Of Grace (Taylor's Version)	Red (Taylor's Version)	0.000328	0.594	0.713	0	0.114	-5.314	0.0503	129.958	0.328	73
2	Red (Taylor's Version)	Red (Taylor's Version)	0.00108	0.516	0.777	1.62E-6	0.0761	-4.908	0.0375	125.047	0.408	81
3	Treacherous (Taylor's Version)	Red (Taylor's Version)	0.0344	0.645	0.593	0.000127	0.13	-6.506	0.0288	109.984	0.299	73
4	I Knew You Were Trouble (Taylor's Version)	Red (Taylor's Version)	0.0129	0.584	0.557	0	0.0576	-6.371	0.0342	154.008	0.767	81
5	All Too Well (Taylor's Version)	Red (Taylor's Version)	0.0171	0.44	0.528	0.00203	0.234	-7.809	0.0317	185.972	0.132	78
6	22 (Taylor's Version)	Red (Taylor's Version)	0.000443	0.642	0.695	0.0000102	0.0753	-5.62	0.0281	103.984	0.642	80
7	I Almost Do (Taylor's Version)	Red (Taylor's Version)	0.0167	0.511	0.559	0	0.113	-6.587	0.0264	145.88	0.248	72
8	We Are Never Ever Getting Back Together (Taylor's Version)	Red (Taylor's Version)	0.0317	0.567	0.686	1.86E-6	0.0732	-6.139	0.175	172.014	0.716	81
9	Stay Stay Stay (Taylor's Version)	Red (Taylor's Version)	0.0848	0.693	0.681	0	0.0768	-7.039	0.025	100.02	0.663	71
10	The Last Time (feat. Gary Lightbody of Snow Patrol) (Taylo	Red (Taylor's Version)	0.0399	0.502	0.534	0	0.0977	-5.954	0.0278	94.05	0.155	74
11	Holy Ground (Taylor's Version)	Red (Taylor's Version)	0.0288	0.622	0.809	0.00218	0.109	-5.623	0.0638	156.894	0.511	71
12	Sad Beautiful Tragic (Taylor's Version)	Red (Taylor's Version)	0.622	0.601	0.406	0.0000919	0.133	-11.827	0.0275	130.059	0.232	73
13	The Lucky One (Taylor's Version)	Red (Taylor's Version)	0.066	0.686	0.571	0	0.0608	-7.138	0.05	117.889	0.538	71
14	Everything Has Changed (feat. Ed Sheeran) (Taylor's Versio	Red (Taylor's Version)	0.271	0.498	0.61	0	0.223	-5.098	0.0363	79.918	0.474	77
15	Starlight (Taylor's Version)	Red (Taylor's Version)	0.00324	0.628	0.685	0	0.18	-5.864	0.0358	126.014	0.605	70
16	Begin Again (Taylor's Version)	Red (Taylor's Version)	0.075	0.519	0.527	0	0.132	-7.673	0.0274	78.915	0.267	73
17	The Moment I Knew (Taylor's Version)	Red (Taylor's Version)	0.0494	0.636	0.402	0	0.107	-7.855	0.031	125.952	0.208	70
18	Come Back...Be Here (Taylor's Version)	Red (Taylor's Version)	0.0158	0.46	0.632	0	0.0822	-6.031	0.0302	79.846	0.399	74
19	Girl At Home (Taylor's Version)	Red (Taylor's Version)	0.00955	0.691	0.736	0.0000188	0.101	-6.974	0.0326	125.089	0.612	68
20	Ronan (Taylor's Version)	Red (Taylor's Version)	0.661	0.623	0.279	0	0.193	-10.802	0.031	116.04	0.38	65
21	Better Man (Taylor's Version) (From The Vault)	Red (Taylor's Version)	0.214	0.473	0.579	0	0.0877	-5.824	0.0384	73.942	0.255	74
22	Nothing New (feat. Phoebe Bridgers) (Taylor's Version) (Fr	Red (Taylor's Version)	0.817	0.606	0.377	0	0.154	-9.455	0.0275	101.96	0.446	78
23	Babe (Taylor's Version) (From The Vault)	Red (Taylor's Version)	0.0538	0.584	0.743	2.83E-6	0.121	-7.075	0.0931	167.844	0.746	74
24	Message In A Bottle (Taylor's Version) (From The Vault)	Red (Taylor's Version)	0.00188	0.622	0.791	3.72E-6	0.083	-6.106	0.0535	115.915	0.494	75
25	I Bet You Think About Me (feat. Chris Stapleton) (Taylor's	Red (Taylor's Version)	0.167	0.391	0.715	0	0.183	-4.516	0.0495	149.654	0.473	77
26	Forever Winter (Taylor's Version) (From The Vault)	Red (Taylor's Version)	0.256	0.611	0.552	0	0.134	-5.828	0.031	116.012	0.41	70
27	Run (feat. Ed Sheeran) (Taylor's Version) (From The Vault)	Red (Taylor's Version)	0.817	0.61	0.488	0	0.312	-6.918	0.0293	125.039	0.443	72
28	The Very First Night (Taylor's Version) (From The Vault)	Red (Taylor's Version)	0.00115	0.678	0.733	0	0.104	-5.025	0.0281	121.009	0.581	77
29	All Too Well (10 Minute Version) (Taylor's Version) (From	Red (Taylor's Version)	0.274	0.631	0.518	0	0.088	-8.771	0.0303	93.023	0.205	87

Figure 1: Data Overview

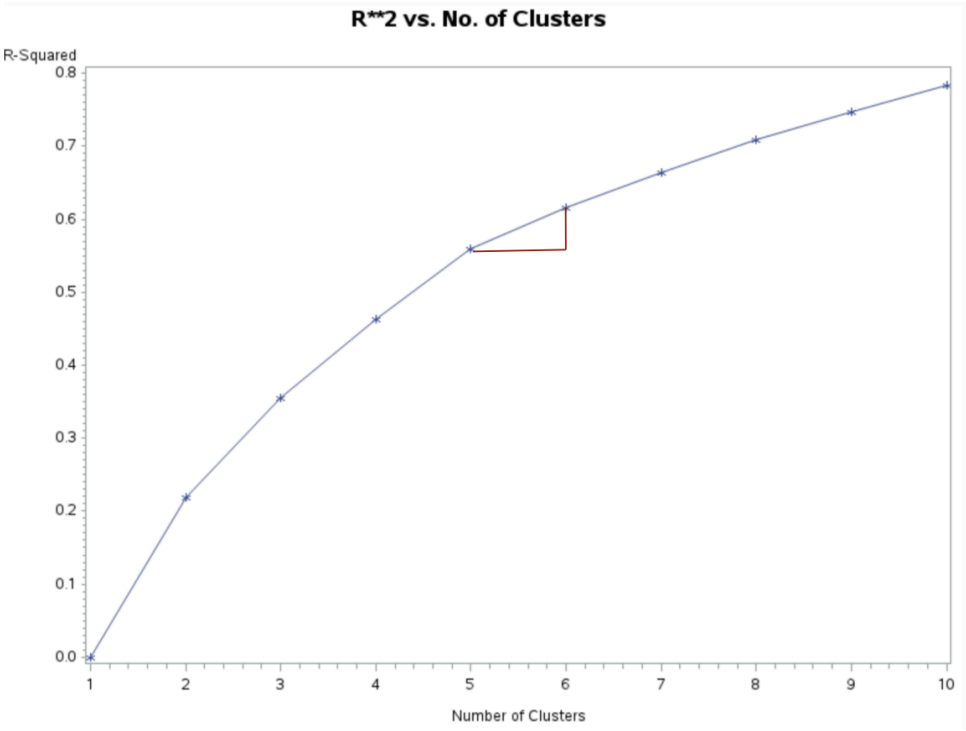

Figure 2: R^2 vs. No. of Clusters

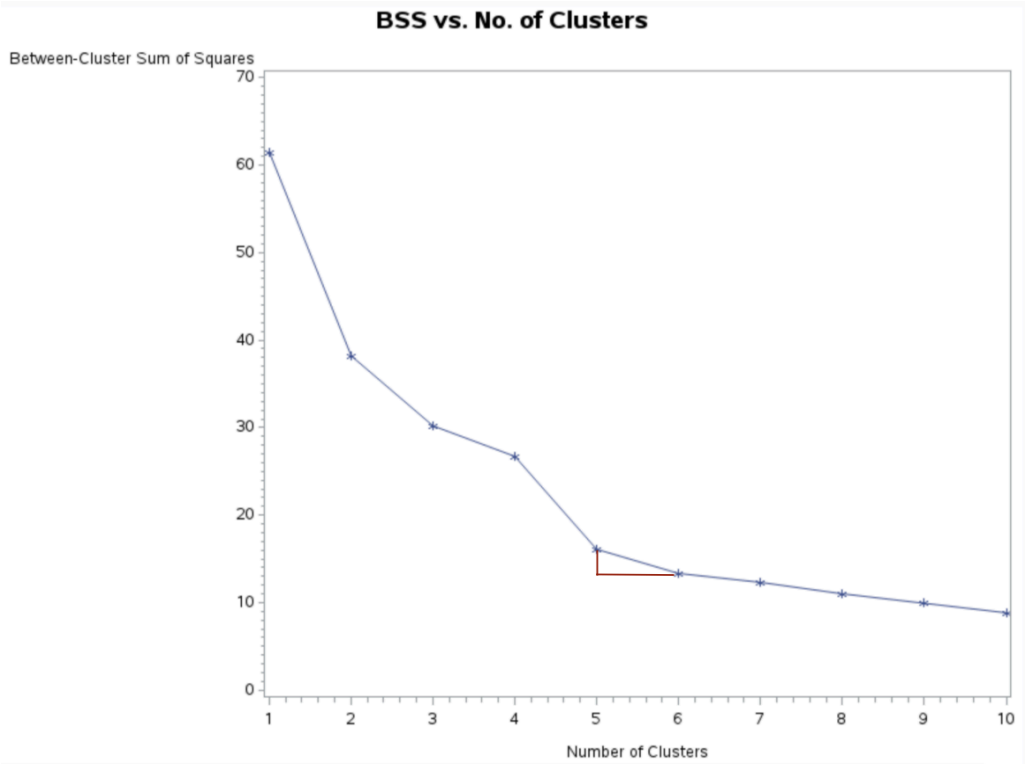

Figure 3: BSS vs. No. of Clusters

The graphs in Figure 2 and Figure 3 are used to estimate how many clusters are chosen in the analysis. In both visualizations, the points with the smallest slope, or where the line begins to flatten first, is chosen as the number of clusters. Flatness begins when the number of clusters is five. Therefore, five is the amount of clusters chosen.

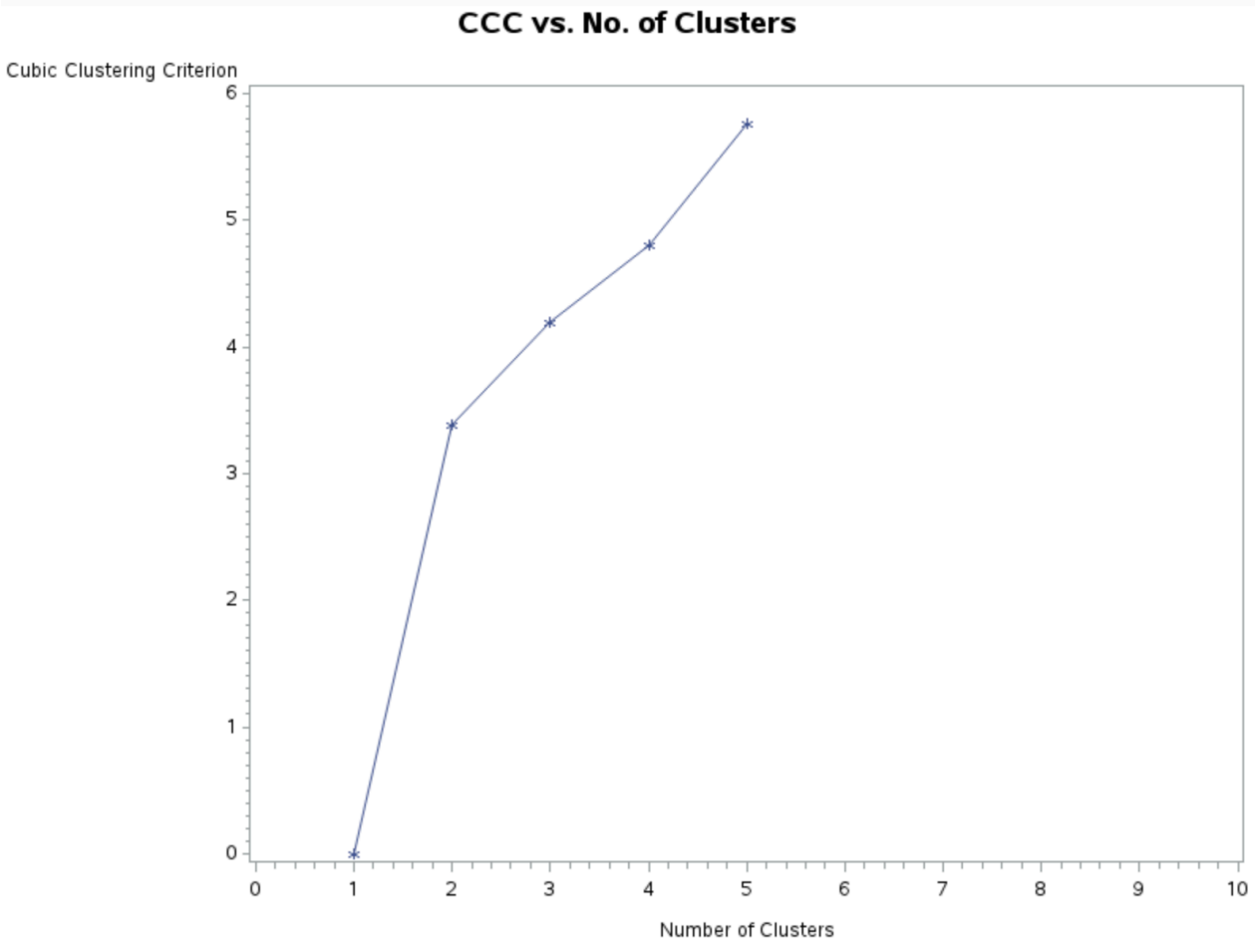

Figure 4: CCC vs. No. of Clusters

Typically, the CCC vs. No. of Clusters is also considered when determining the number of clusters. However, Figure 4 does not show the values for all clusters as SAS will not compute the calculation if the number of clusters exceeds 20% of the number of observations. 20% of 29 observations is 5.8. As 5 has a marginal difference compared to 5.8, the CCC vs. No. of Clusters graph is not used to determine our number of clusters.

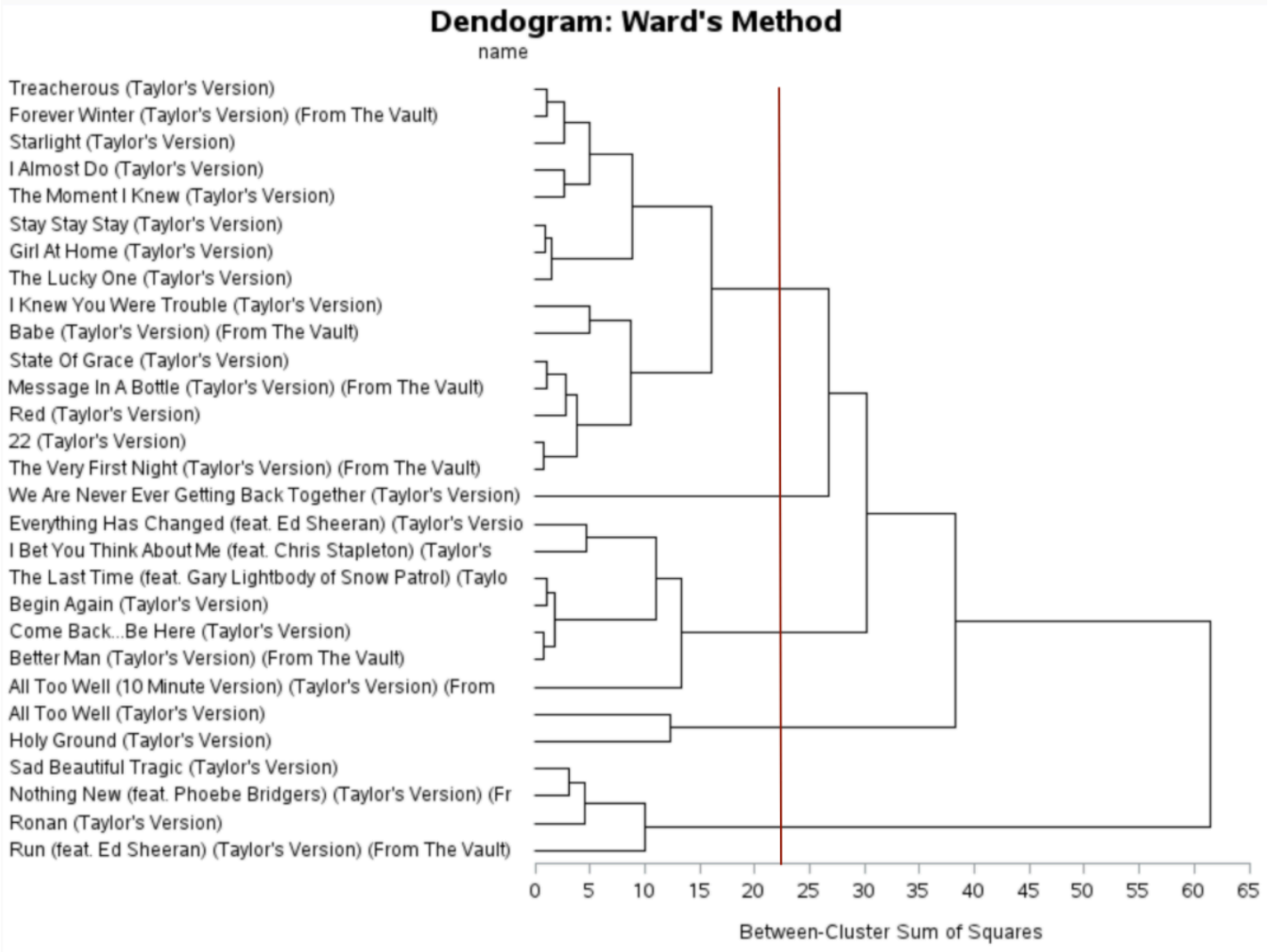

Figure 5: Dendogram using Ward's Method

A vertical line is drawn through the dendogram where maximum distances between clusters are found. The horizontal intersections with this line determine the amount of clusters chosen. As the red vertical line in Figure 5 intersects with five vertical lines, five clusters are also chosen through this method.

5-Clusters solution: Ward's Minimum Variance Clustering

CLUSTER=1

Obs	name	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity	CLUSNAME
1	22 (Taylor's Version)	0.000443	0.642	0.695	0.0000102	0.0753	-5.62	0.0281	103.984	0.642	80	CL5
2	The Very First Night (Taylor's Version) (From The Vault)	0.00115	0.678	0.733	0	0.104	-5.025	0.0281	121.009	0.581	77	CL5
3	Stay Stay Stay (Taylor's Version)	0.0848	0.693	0.681	0	0.0768	-7.039	0.025	100.02	0.663	71	CL5
4	Girl At Home (Taylor's Version)	0.00955	0.691	0.736	0.0000188	0.101	-6.974	0.0326	125.089	0.612	68	CL5
5	Treacherous (Taylor's Version)	0.0344	0.645	0.593	0.000127	0.13	-6.506	0.0288	109.984	0.299	73	CL5
6	Forever Winter (Taylor's Version) (From The Vault)	0.256	0.611	0.552	0	0.134	-5.828	0.031	116.012	0.41	70	CL5
7	State Of Grace (Taylor's Version)	0.000328	0.594	0.713	0	0.114	-5.314	0.0503	129.958	0.328	73	CL5
8	Message In A Bottle (Taylor's Version) (From The Vault)	0.00188	0.622	0.791	3.72E-6	0.083	-6.106	0.0535	115.915	0.494	75	CL5
9	The Lucky One (Taylor's Version)	0.066	0.686	0.571	0	0.0608	-7.138	0.05	117.889	0.538	71	CL5
10	I Almost Do (Taylor's Version)	0.0167	0.511	0.559	0	0.113	-6.587	0.0264	145.88	0.248	72	CL5
11	The Moment I Knew (Taylor's Version)	0.0494	0.636	0.402	0	0.107	-7.855	0.031	125.952	0.208	70	CL5
12	Starlight (Taylor's Version)	0.00324	0.628	0.685	0	0.18	-5.864	0.0358	126.014	0.605	70	CL5
13	Red (Taylor's Version)	0.00108	0.516	0.777	1.62E-6	0.0761	-4.908	0.0375	125.047	0.408	81	CL5
14	I Knew You Were Trouble (Taylor's Version)	0.0129	0.584	0.557	0	0.0576	-6.371	0.0342	154.008	0.767	81	CL5
15	Babe (Taylor's Version) (From The Vault)	0.0538	0.584	0.743	2.83E-6	0.121	-7.075	0.0931	167.844	0.746	74	CL5

CLUSTER=2

Obs	name	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity	CLUSNAME
16	Come Back...Be Here (Taylor's Version)	0.0158	0.46	0.632	0	0.0822	-6.031	0.0302	79.846	0.399	74	CL6
17	Better Man (Taylor's Version) (From The Vault)	0.214	0.473	0.579	0	0.0877	-5.824	0.0384	73.942	0.255	74	CL6
18	The Last Time (feat. Gary Lightbody of Snow Patrol) (Taylo	0.0399	0.502	0.534	0	0.0977	-5.954	0.0278	94.05	0.155	74	CL6
19	Begin Again (Taylor's Version)	0.075	0.519	0.527	0	0.132	-7.673	0.0274	78.915	0.267	73	CL6
20	Everything Has Changed (feat. Ed Sheeran) (Taylor's Versio	0.271	0.498	0.61	0	0.223	-5.098	0.0363	79.918	0.474	77	CL6
21	I Bet You Think About Me (feat. Chris Stapleton) (Taylor's	0.167	0.391	0.715	0	0.183	-4.516	0.0495	149.654	0.473	77	CL6
22	All Too Well (10 Minute Version) (Taylor's Version) (From	0.274	0.631	0.518	0	0.088	-8.771	0.0303	93.023	0.205	87	CL6

CLUSTER=3

Obs	name	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity	CLUSNAME
23	Sad Beautiful Tragic (Taylor's Version)	0.622	0.601	0.406	0.0000919	0.133	-11.827	0.0275	130.059	0.232	73	CL9
24	Nothing New (feat. Phoebe Bridgers) (Taylor's Version) (Fr	0.817	0.606	0.377	0	0.154	-9.455	0.0275	101.96	0.446	78	CL9
25	Ronan (Taylor's Version)	0.661	0.623	0.279	0	0.193	-10.802	0.031	116.04	0.38	65	CL9
26	Run (feat. Ed Sheeran) (Taylor's Version) (From The Vault)	0.817	0.61	0.488	0	0.312	-6.918	0.0293	125.039	0.443	72	CL9

CLUSTER=4

Obs	name	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity	CLUSNAME
27	All Too Well (Taylor's Version)	0.0171	0.44	0.528	0.00203	0.234	-7.809	0.0317	185.972	0.132	78	CL7
28	Holy Ground (Taylor's Version)	0.0288	0.622	0.809	0.00218	0.109	-5.623	0.0638	156.894	0.511	71	CL7

CLUSTER=5

Obs	name	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity	CLUSNAME
29	We Are Never Ever Getting Back Together (Taylor's Version)	0.0317	0.567	0.686	1.86E-6	0.0732	-6.139	0.175	172.014	0.716	81	We Are Never Ever Getting Back Together (Taylor's Version)

Figure 6: Clusters

cluster	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
1	0.0394447	0.6214	0.6525333	0.00001095	0.10224	-6.2806667	0.0390267	125.6403333	0.5032667	73.7333333
2	0.1509571	0.4962857	0.5878571	0.00000000	0.1276571	-6.2667143	0.0342714	92.764	0.3182857	76.5714286
3	0.72925	0.61	0.3875	0.00002298	0.198	-9.7505	0.028825	118.2745	0.37525	72
4	0.02295	0.531	0.6685	0.00210500	0.1715	-6.716	0.04775	171.433	0.3215	74.5
5	0.0317	0.567	0.686	0.00000186	0.0732	-6.139	0.175	172.014	0.716	81

Figure 7: Comparisons of Means for Variables Between Clusters

The five clusters displayed in Figure 6 represent the five main genres on this album; Pop, Country Pop, Alternative, Ballads, and Dance Pop.

Cluster 1, Pop, contains Swift's energetic and upbeat pop songs. They are overall more "happy" sounding. They have the highest means for danceability (0.6214) and valence (0.5032). The high mean for valence further proves that this cluster has an overall positive theme as high valence is an indicator of happiness. While the mean for energy (0.6525) is not the highest among all clusters, it is still significantly large. High danceability and energy contribute to the idea that these songs are very cheerful.

Cluster 2, Country Pop, contains songs that are slower and more acoustic. They have the slowest tempo (92.764) and second highest means for acousticness (0.1509), which are common traits in country music. This genre also contains the lowest means for valence (0.3182), as it contains sadder, more gloomy-sounding songs than you would see in pop music.

Cluster 3, Alternative, contains songs that are sad and melancholic. They contain the highest values for acousticness (0.7292) and liveness (0.198). Tracks in this genre rely on acoustic instruments and have a much more intimacy which can cause the significant means of acousticness and liveness. They have the lowest value for energy (0.3875), most likely due to their depressive nature. This cluster also has low means for valence (0.3752) , which can contribute to the negative nature of the songs.

Cluster 4, Ballads, consists of narrative, fast paced songs. This cluster contains the highest means for instrumentalness (0.0021), a common trait in many ballads, and has the fastest tempo (171.433). It also has the second highest mean for speechiness (0.0477), most likely due to its storytelling nature.

Cluster 5, Dance Pop, contains only one song which was the most successful single on the original recording. It contains the highest means for energy (0.686), speechiness (0.175), tempo (172.014), and valence (0.716). It also has the most significant mean for loudness (-6.139). The combination of multiple variables being the most significant could have been a reason for its success, as it is the only cluster on the entire album with the highest mean for popularity.

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall CLUSTER Effect
H = Type III SSCP Matrix for CLUSTER
E = Error SSCP Matrix

S=4 M=2.5 N=6.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00000558	34.20	40	58.734	<.0001
Pillai's Trace	3.55468690	14.37	40	72	<.0001
Hotelling-Lawley Trace	373.27114670	129.60	40	31.291	<.0001
Roy's Greatest Root	346.76781573	624.18	10	18	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Figure 8: MANOVA Test Criteria

Since all four MANOVA tests are statistically significant with p-values that are less than 0.0001 at a 0.05 significance level, it is assumed that the cluster means are significantly separated.

Conclusion: Our aim to cluster the songs on this album was accomplished by analyzing the R^2 vs. No. of Clusters (Figure 2), BSS vs. No. of Clusters (Figure 3) and Dendogram (Figure 5), which yielded that choosing five clusters is most appropriate. Thus, we interpreted the five clusters as five genres on *Red (Taylor's Version)*. Using means for all variables per cluster, we described them as Pop, Country Pop, Alternative, Ballads, and Dance Pop.

Acknowledgements:

Priester, J. (2023, November 7). Taylor Swift Spotify Dataset (Version 12). Kaggle www.kaggle.com/datasets/jarredpriester/taylor-swift-spotify-dataset

Meyers, Seth. “Taylor Swift Explains Why She’s Re-Recording Her Albums.” *YouTube*, NBC, New York City, New York, 12 Nov. 2021.

“Loudness Normalization.” *Spotify*, Spotify, support.spotify.com/us/artists/article/loudness-normalization/. Accessed 16 Nov. 2023.

Appendix:

```
PROC IMPORT OUT=tss
   DATAFILE="/home/u63069202/Multivariate Analysis/Project/taylor_swift_spotify.csv"
   DBMS=CSV REPLACE; GETNAMES=YES; RUN;
data tss; set tss;

drop var1 release_date track_number id uri duration_ms;
if album in ('Taylor Swift',  'Live From Clear Channel Stripped', 'Fearless', 'Fearless
(International Vers', 'Fearless Platinum Edition', 'Speak Now (Deluxe Edition)', 'Speak Now',
'Speak Now World Tour Live', 'Red',  'Red (Deluxe Edition)', '1989', '1989 (Deluxe Edition)',
'reputation', 'reputation Stadium Tour Surprise', 'Lover', 'folklore',  'folklore: the long pond studio
s',  'folklore (deluxe version)',  'evermore', 'evermore (deluxe version)',  'Midnights', 'Midnights
(3am Edition)', 'Midnights (The Til Dawn Edition)', "Fearless (Taylor's Version)", "Speak Now
(Taylor's Version)", "1989 (Taylor's Version)", "1989 (Taylor's Version) [Deluxe]") then delete;
if name in ("State Of Grace (Acoustic Version) (Taylor's Version)") then delete;
run; proc print data=tss;

proc cluster data=tss method=average noeigen nonorm out=tree1;
id name;
var acousticness danceability energy instrumentalness liveness loudness speechiness
tempo valence popularity; run;
```

```
proc tree data=tree1  out=tss_out nclusters=3 horizontal;
copy acousticness danceability energy instrumentalness liveness loudness speechiness tempo
valence popularity; run;

proc sort data = tss_out; by cluster; run;
proc print data=tss_out; by cluster; title "5-Clusters solution: Average Linkage Clustering"; run;

proc cluster data=tss  method=ward standard noprint noeigen nonorm out=tree2;
id name;
var acousticness danceability energy instrumentalness liveness loudness speechiness tempo
valence popularity; run;

proc sort data=tree2; by _ncl_;

data graph2; set tree2; if _ncl_<=10; proc print data=graph2; var _ncl_ _ccc_ ;

symbol1 i=join value=star;

proc gplot data=graph2; plot _rsq_*_ncl_ ; title "R**2 vs. No. of Clusters"; run;
proc gplot data=graph2; plot _height_*_ncl_ ; title "BSS vs. No. of Clusters";
Proc gplot data=tree2; plot _ccc_*_ncl_/haxis=0 to 10 by 1 ; title "CCC vs. No. of Clusters"; run;
```

```
proc tree data=tree2 out=newdata nclusters=3 horizontal;
id name;
copy acousticness danceability energy instrumentalness liveness loudness speechiness tempo
valence popularity; title "Dendogram: Ward's Method"; run;

proc means data=newdata;
var acousticness danceability energy instrumentalness liveness loudness speechiness tempo
valence popularity; by cluster; run;

PROC IMPORT OUT=clustermeans
   DATAFILE="/home/u63069202/Multivariate Analysis/Project/clustermeans.xlsx"
   DBMS=XLSX
   REPLACE;
   GETNAMES=YES;
RUN;

proc print data=clustermeans;

proc glm data=newdata; class cluster;
model acousticness danceability energy instrumentalness liveness loudness speechiness tempo
valence popularity =cluster/nouni; manova h=cluster; run;
```