MACQUARIE
University
SYDNEY·AUSTRALIA

BUSA3020 ADVANCED ANALYTICS TECHNIQUES

ASSIGNMENT #3: CLUSTERING

SEGMENTATION OF MUSIC AND MOVIE DATA USING

IBM SPSS STATISTICS 25

Name: Chris Jerylle Vargas Oidem

Student ID: 45476624

Due Date: Tuesday, 12 May 2020, 11:55pm

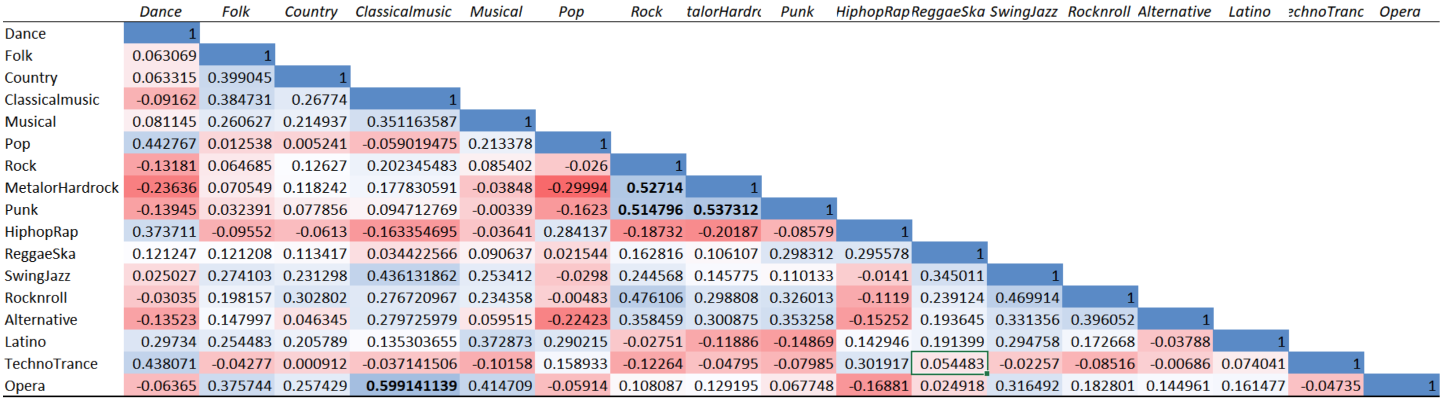
TABLE OF CONTENTS

1. Data Handling

This analysis reduces Music and Movie variables into fewer, more manageable, and interpretable factors. Resulting themes can be further used in profiling the respondents to gain better insight on their preferences.

Data with missing values were removed so that only those with complete observations can be included in the analysis. Unnecessary variables that are removed – *Music*, *Slow or Fast Songs*, and *Movies,* resulting to 17 Music and 11 Movie variables.

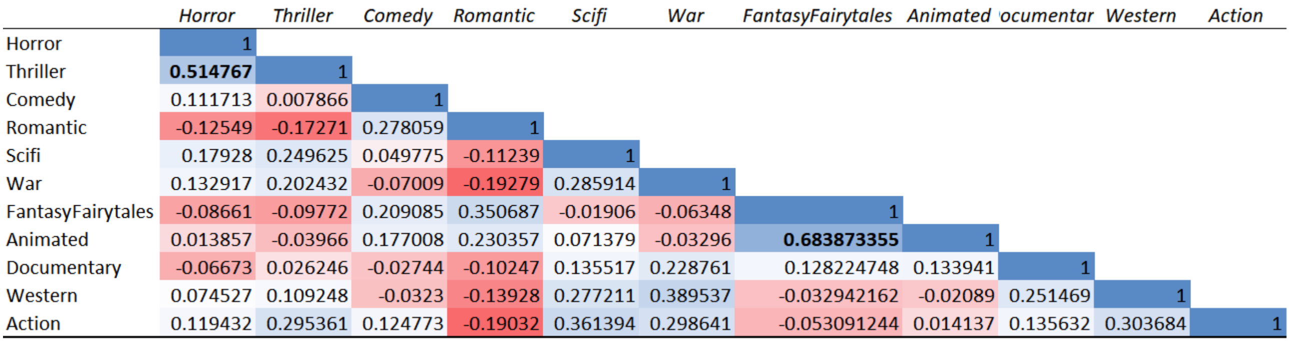
Correlation is first used among the complete cases.



	Dance	*Folk*	*Country*	*Classicalmusic*	*Musical*	*Pop*	*Rock*	*talorHardrc*	*Punk*	*HiphopRap*	*ReggaeSka*	*SwingJazz*	*Rocknroll*	*Alternative*	*Latino*	*echnoTranc*	*Opera*
Dance	1																
Folk	0.063069	1															
Country	0.063315	0.399045	1														
Classicalmusic	-0.09162	0.384731	0.26774	1													
Musical	0.081145	0.260627	0.214937	0.351163587	1												
Pop	0.442767	0.012538	0.005241	-0.059019475	0.213378	1											
Rock	-0.13181	0.064685	0.12627	0.202345483	0.085402	-0.026	1										
MetalorHardrock	-0.23636	0.070549	0.118242	0.177830591	-0.03848	-0.29994	**0.52714**	1									
Punk	-0.13945	0.032391	0.077856	0.094712769	-0.00339	-0.1623	**0.514796**	**0.537312**	1								
HiphopRap	0.373711	-0.09552	-0.0613	-0.163354695	-0.03641	0.284137	-0.18732	-0.20187	-0.08579	1							
ReggaeSka	0.121247	0.121208	0.113417	0.034422566	0.090637	0.021544	0.162816	0.106107	0.298312	0.295578	1						
SwingJazz	0.025027	0.274103	0.231298	0.436131862	0.253412	-0.0298	0.244568	0.145775	0.110133	-0.0141	0.345011	1					
Rocknroll	-0.03035	0.198157	0.302802	0.276720967	0.234358	-0.00483	0.476106	0.298808	0.326013	-0.1119	0.239124	0.469914	1				
Alternative	-0.13523	0.147997	0.046345	0.279725979	0.059515	-0.22423	0.358459	0.300875	0.353258	-0.15252	0.193645	0.331356	0.396052	1			
Latino	0.29734	0.254483	0.205789	0.135303655	0.372873	0.290215	-0.02751	-0.11886	-0.14869	0.142946	0.191399	0.294758	0.172668	-0.03788	1		
TechnoTrance	0.438071	-0.04277	0.000912	-0.037141506	-0.10158	0.158933	-0.12264	-0.04795	-0.07985	0.301917	0.054483	-0.02257	-0.08516	-0.00686	0.074041	1	
Opera	-0.06365	0.375744	0.257429	**0.599141139**	0.414709	-0.05914	0.108087	0.129195	0.067748	-0.16881	0.024918	0.316492	0.182801	0.144961	0.161477	-0.04735	1

Table 1. Correlation of Music Variables

Table 1 shows that *Opera* looks highly correlated with *ClassicalMusic*, as well as *Rock* and *Metal*. *Pop* and *Metal* are negatively correlated.



	Horror	*Thriller*	*Comedy*	*Romantic*	*Scifi*	*War*	*FantasyFairytales*	*Animated*	*ocumentar*	*Western*	*Action*
Horror	1										
Thriller	**0.514767**	1									
Comedy	0.111713	0.007866	1								
Romantic	-0.12549	-0.17271	0.278059	1							
Scifi	0.17928	0.249625	0.049775	-0.11239	1						
War	0.132917	0.202432	-0.07009	-0.19279	0.285914	1					
FantasyFairytales	-0.08661	-0.09772	0.209085	0.350687	-0.01906	-0.06348	1				
Animated	0.013857	-0.03966	0.177008	0.230357	0.071379	-0.03296	**0.683873355**	1			
Documentary	-0.06673	0.026246	-0.02744	-0.10247	0.135517	0.228761	0.128224748	0.133941	1		
Western	0.074527	0.109248	-0.0323	-0.13928	0.277211	0.389537	-0.032942162	-0.02089	0.251469	1	
Action	0.119432	0.295361	0.124773	-0.19032	0.361394	0.298641	-0.053091244	0.014137	0.135632	0.303684	1

Table 2. Correlation of Movie Variables

Table 2 shows that *FantasyFairytales* and *Animated* are highly correlated, followed by *Thriller* and *Horror*. *War* and *Romantic* movies are negatively correlated.

The data has been reduced to fewer variables using Principal Component Analysis (PCA).

1.1 Music Preferences

17 Principal Components (PC) had been extracted for 17 variables:

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	**3.784**	22.257	22.257	3.784	22.257	22.257
2	**2.665**	15.675	37.933	2.665	15.675	37.933
3	**1.864**	10.962	48.895	1.864	10.962	48.895
4	**1.104**	6.495	**55.389**	1.104	6.495	55.389
5	1.033	6.075	61.465	1.033	6.075	61.465
6	.950	5.590	67.055	.950	5.590	67.055
7	.844	4.963	72.018	.844	4.963	72.018
8	.663	3.897	75.915	.663	3.897	75.915
9	.652	3.836	79.752	.652	3.836	79.752
10	.590	3.468	83.219	.590	3.468	83.219
11	.511	3.007	86.227	.511	3.007	86.227
12	.443	2.606	88.833	.443	2.606	88.833
13	.433	2.550	91.382	.433	2.550	91.382
14	.403	2.369	93.752	.403	2.369	93.752
15	.378	2.224	95.976	.378	2.224	95.976
16	.349	2.053	98.028	.349	2.053	98.028
17	.335	1.972	100.000	.335	1.972	100.000

Table 3. Percentage Variation Explained by Each Principal Component (Music)

For this report, the total variance explained by extracted factors should only explain 50% to 60%. Hence, the four PCs chosen account for 55.4% of the total variance and have eigenvalues that are greater than 1.

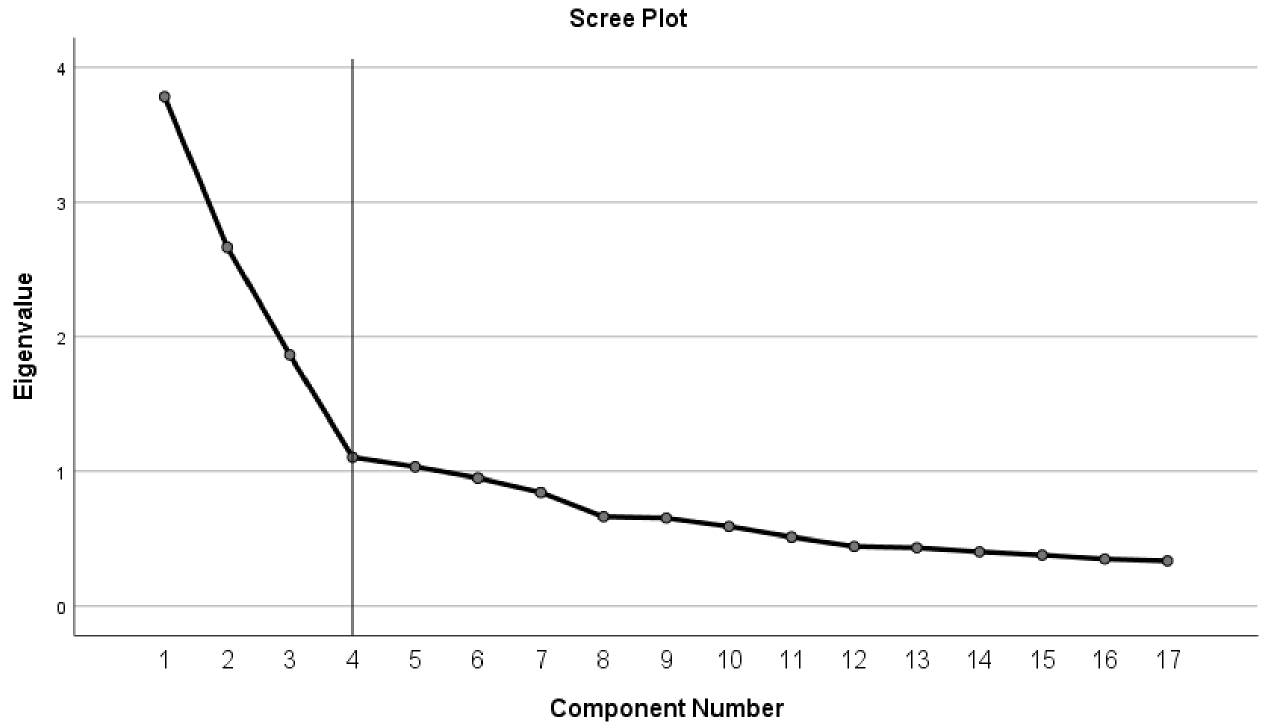
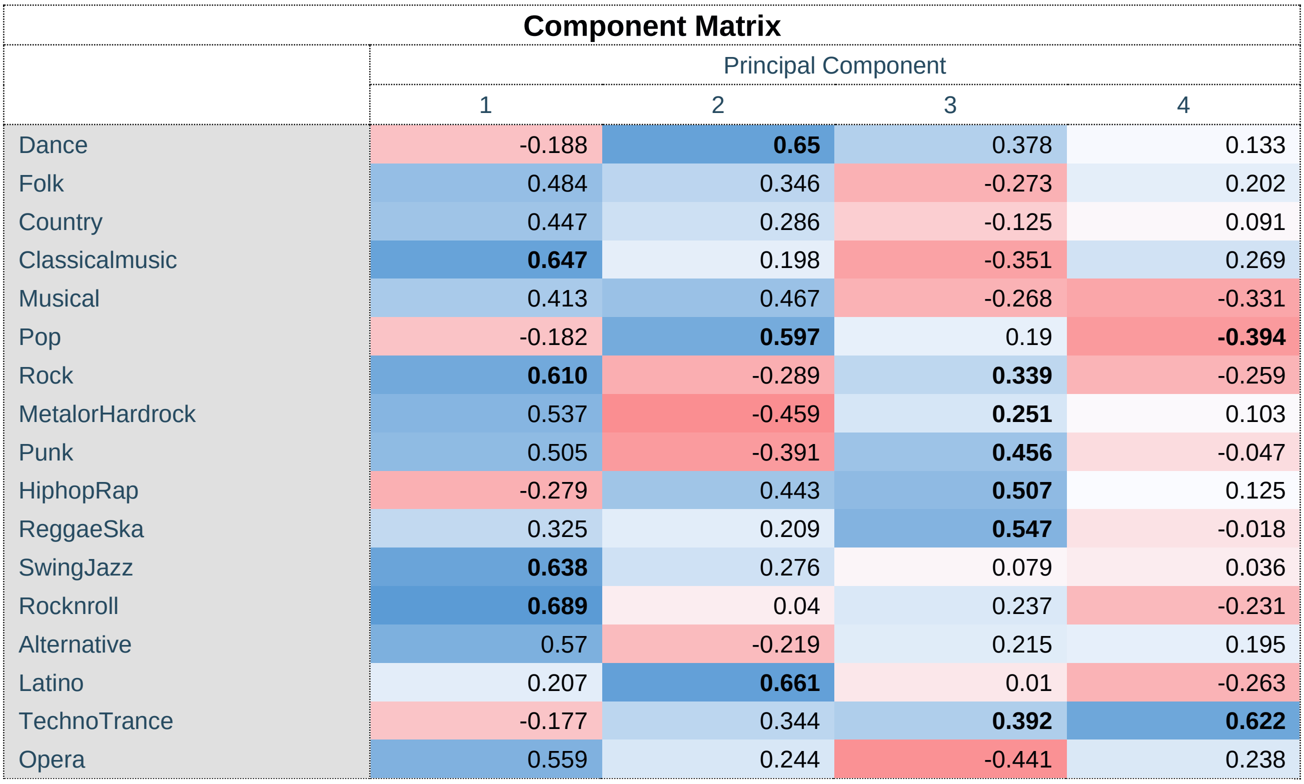


Figure 1. Scree Plot for Music Preferences

The scree plot indicates 4 possible factors explaining Music Preferences as there are clear elbows at the first to fourth eigenvalues (indicating 1 to 3 dimensions). The plot flattens after the fourth eigenvalue.



Component Matrix

	Principal Component			
	1	2	3	4
Dance	-0.188	**0.65**	0.378	0.133
Folk	0.484	0.346	-0.273	0.202
Country	0.447	0.286	-0.125	0.091
Classicalmusic	**0.647**	0.198	-0.351	0.269
Musical	0.413	0.467	-0.268	-0.331
Pop	-0.182	**0.597**	0.19	**-0.394**
Rock	**0.610**	-0.289	**0.339**	-0.259
MetalorHardrock	0.537	-0.459	**0.251**	0.103
Punk	0.505	-0.391	**0.456**	-0.047
HiphopRap	-0.279	0.443	**0.507**	0.125
ReggaeSka	0.325	0.209	**0.547**	-0.018
SwingJazz	**0.638**	0.276	0.079	0.036
Rocknroll	**0.689**	0.04	0.237	-0.231
Alternative	0.57	-0.219	0.215	0.195
Latino	0.207	**0.661**	0.01	-0.263
TechnoTrance	-0.177	0.344	**0.392**	**0.622**
Opera	0.559	0.244	-0.441	0.238

Table 4. Component Matrix for Music Preferences

PC1 – Heavily reliant on *Classicalmusic*, *Rock*, *SwingJazz,* and *Rocknroll*

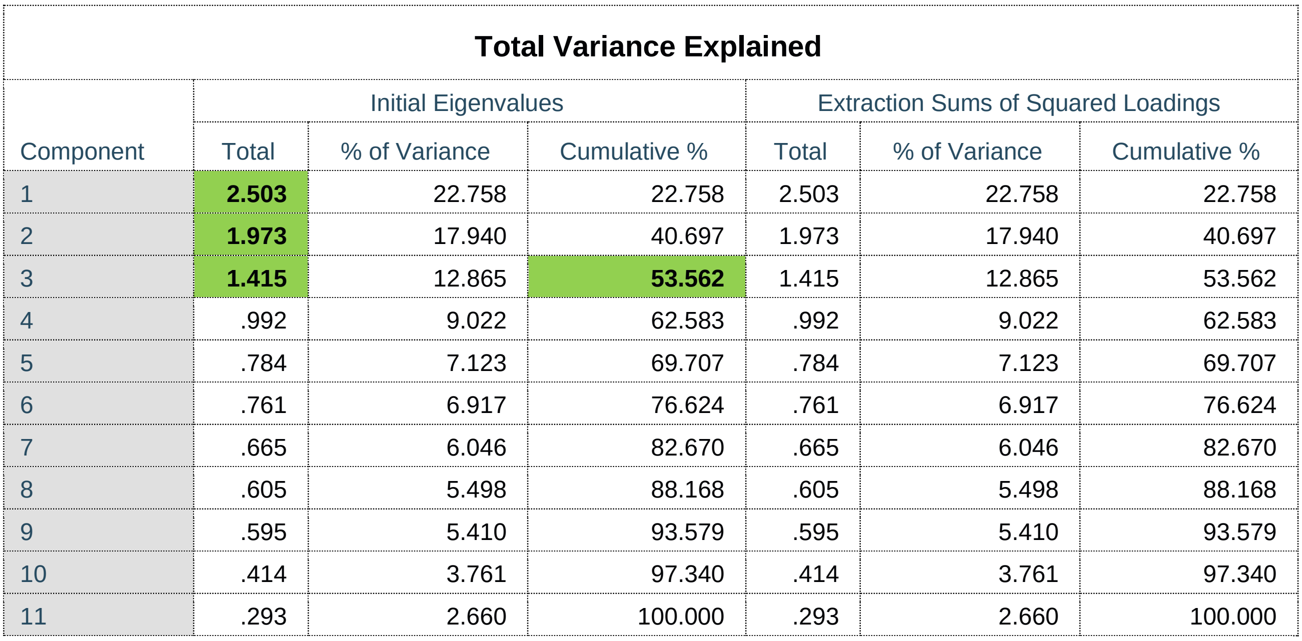
PC2 – The largest coefficients in the positive end are *Dance, Pop,* and *Latino* – a possible representation of up-beat music

PC3 – There is some comparative disparity between fast (*Rock, MetalorHardrock, Punk, HiphopRap, ReggaeSka, TechnoTrance* – positive coefficients) and slow music (*Folk, Country, Classicalmusic, Musical, Opera* – negative coefficients)

PC4 – Contrasts *TechnoTrance* at the positive end of the scale and *Pop* at the negative end

1.2 Movie Preferences

11 Principal Components (PC) had been extracted for 11 variables:



Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	**2.503**	22.758	22.758	2.503	22.758	22.758
2	**1.973**	17.940	40.697	1.973	17.940	40.697
3	**1.415**	12.865	**53.562**	1.415	12.865	53.562
4	.992	9.022	62.583	.992	9.022	62.583
5	.784	7.123	69.707	.784	7.123	69.707
6	.761	6.917	76.624	.761	6.917	76.624
7	.665	6.046	82.670	.665	6.046	82.670
8	.605	5.498	88.168	.605	5.498	88.168
9	.595	5.410	93.579	.595	5.410	93.579
10	.414	3.761	97.340	.414	3.761	97.340
11	.293	2.660	100.000	.293	2.660	100.000

Table 5. Percentage Variation Explained by Each Principal Component (Movie)

There are three eigenvalues that are greater than 1. The three PCs accounts for 53.6% and are chosen to adequately summarise the data for Movie Preferences.

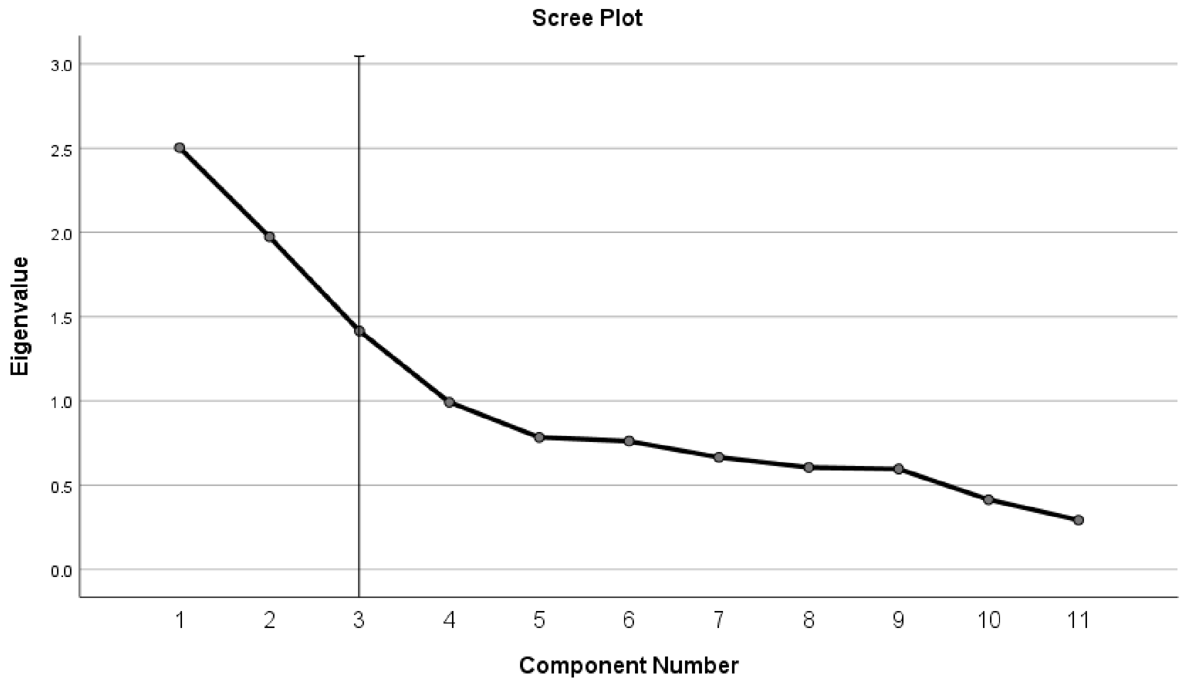
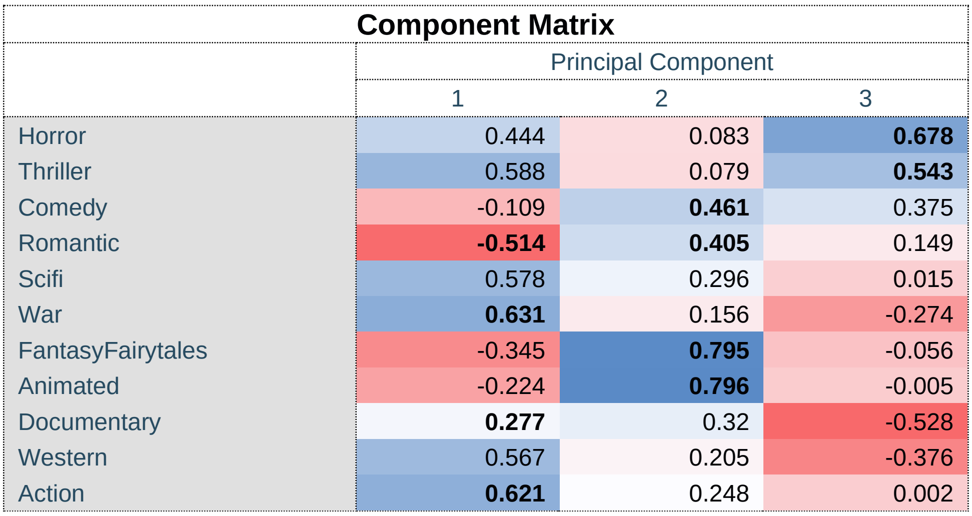


Figure 2. Scree Plot for Movie Preferences

In the scree plot above, there are clear elbows at the first, second, and third eigenvalues (indicating 1 to 2 dimensions).



Component Matrix

	Principal Component		
	1	2	3
Horror	0.444	0.083	**0.678**
Thriller	0.588	0.079	**0.543**
Comedy	-0.109	**0.461**	0.375
Romantic	**-0.514**	**0.405**	0.149
Scifi	0.578	0.296	0.015
War	**0.631**	0.156	-0.274
FantasyFairytales	-0.345	**0.795**	-0.056
Animated	-0.224	**0.796**	-0.005
Documentary	**0.277**	0.32	-0.528
Western	0.567	0.205	-0.376
Action	**0.621**	0.248	0.002

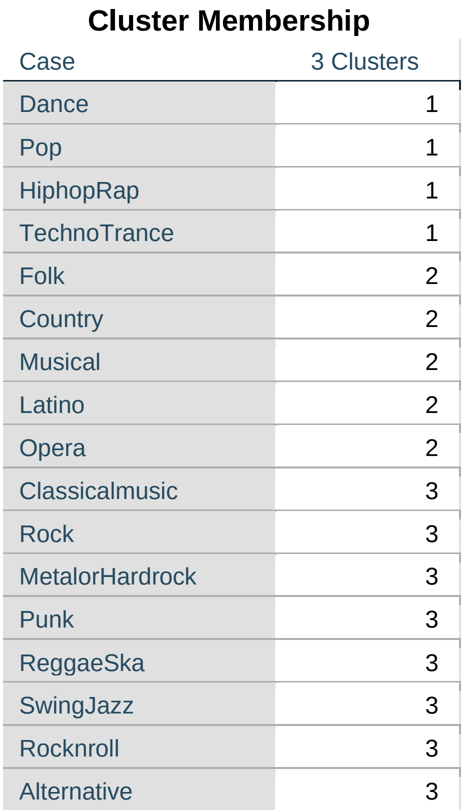
Table 6. Component Matrix for Movie Preferences

PC1 – Contrasts *War* and *Action* movies in the positive end, and *Romantic* in the negative end of the scale
PC2 –The largest coefficients are *FantasyFairytales* and *Animated* (.795 and .796 respectively) which are substantially higher than the rest, followed by *Comedy* and *Romantic* (.461 and .405)
PC3 – Accentuates dominance of *Horror* and *Thriller* movies with coefficients .678 and .543, and *Documentary* at the negative end

2. Clustering Methods

2.1 Agglomerative Hierarchical Clustering

Music Clustering



Cluster Membership

Case	3 Clusters
Dance	1
Pop	1
HiphopRap	1
TechnoTrance	1
Folk	2
Country	2
Musical	2
Latino	2
Opera	2
Classicalmusic	3
Rock	3
MetalorHardrock	3
Punk	3
ReggaeSka	3
SwingJazz	3
Rocknroll	3
Alternative	3

Table 7. Cluster Membership for Music Preference Clusters

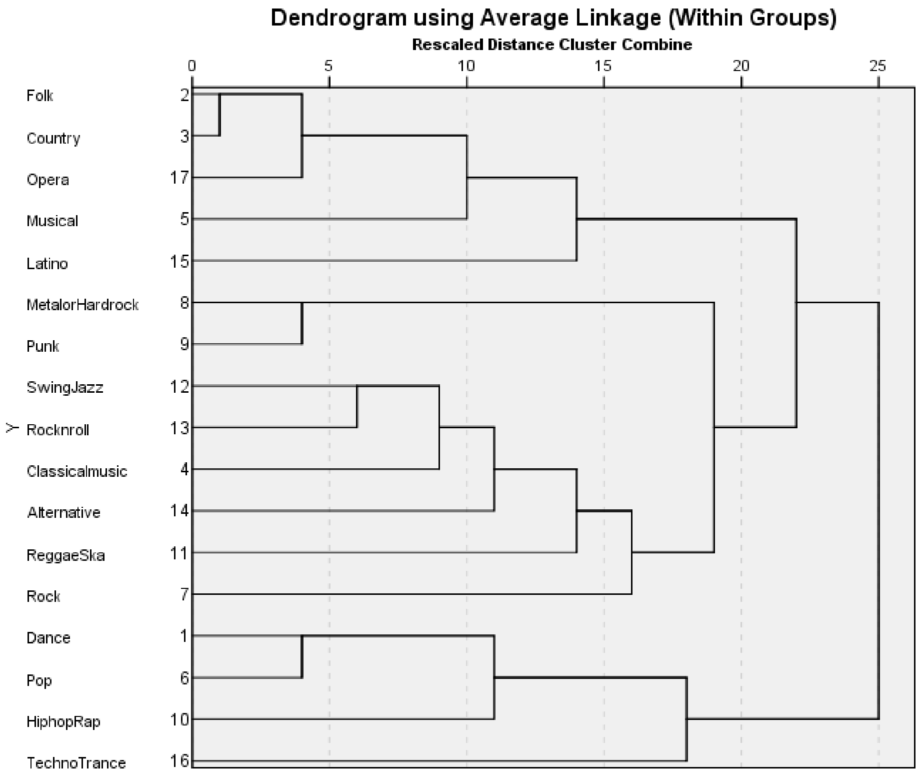


Figure 3. Dendrogram for Music Preferences (Clusters saved by all Variables)

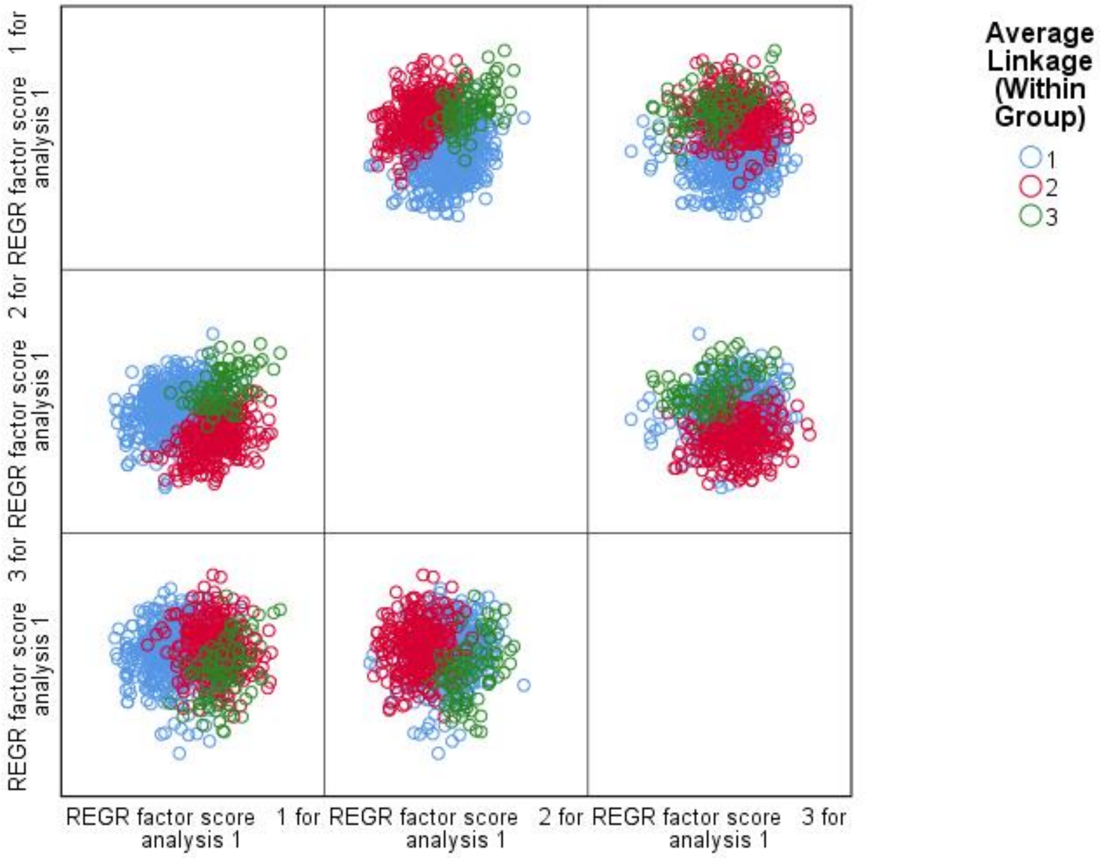


Figure 4. Hierarchical Music Clustering Scatterplot Matrix of PC1, PC2, and PC3 (Clusters saved by all Cases)

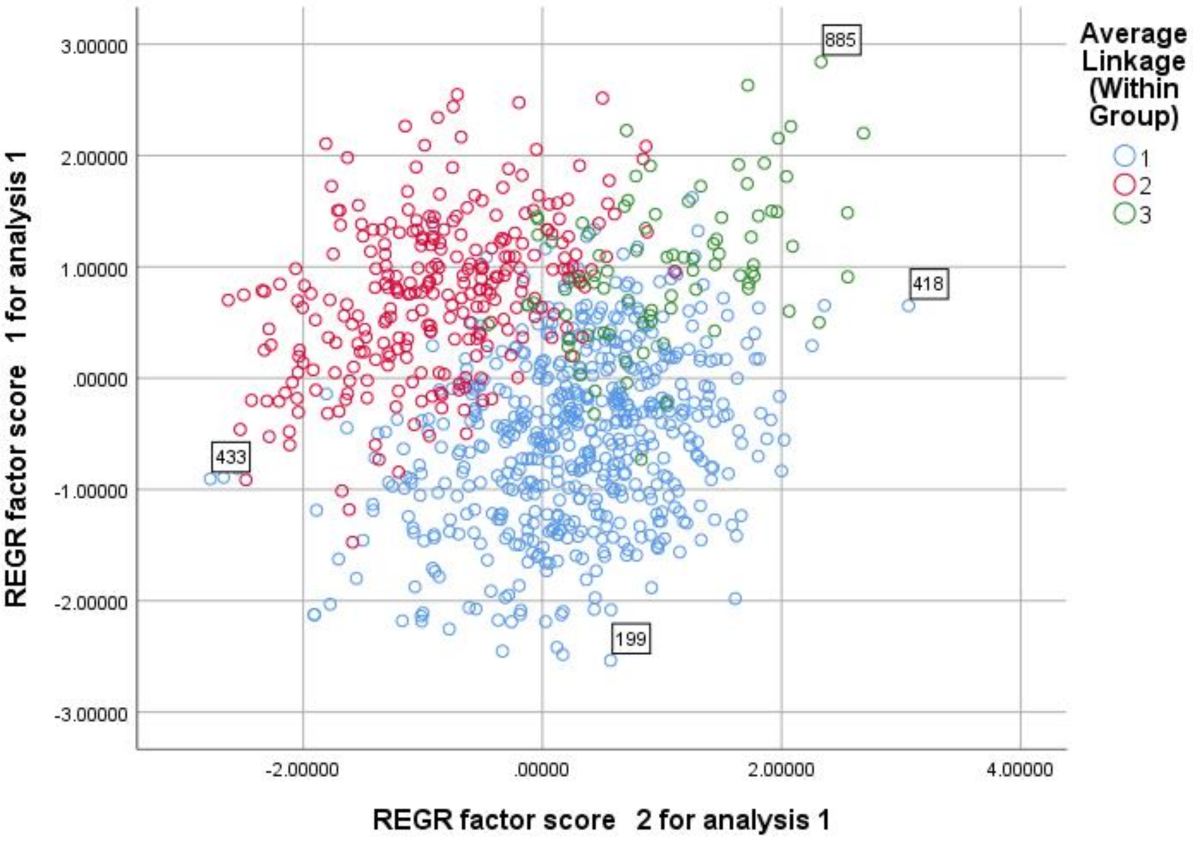
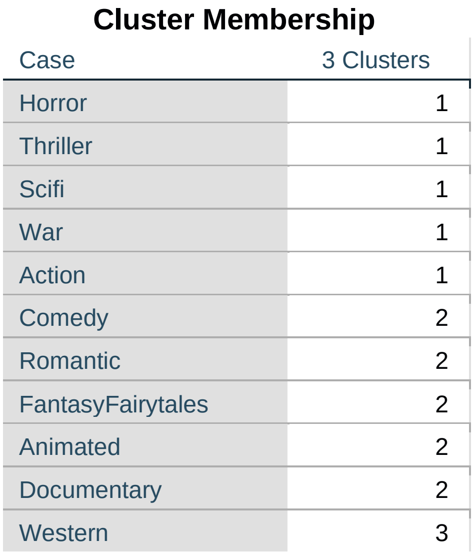


Figure 5. Hierarchical Music Clustering Scatterplot of PC1 and PC2

Movie Clustering



Cluster Membership

Case	3 Clusters
Horror	1
Thriller	1
Scifi	1
War	1
Action	1
Comedy	2
Romantic	2
FantasyFairytales	2
Animated	2
Documentary	2
Western	3

Table 8. Cluster Membership for Movie Preference Clusters

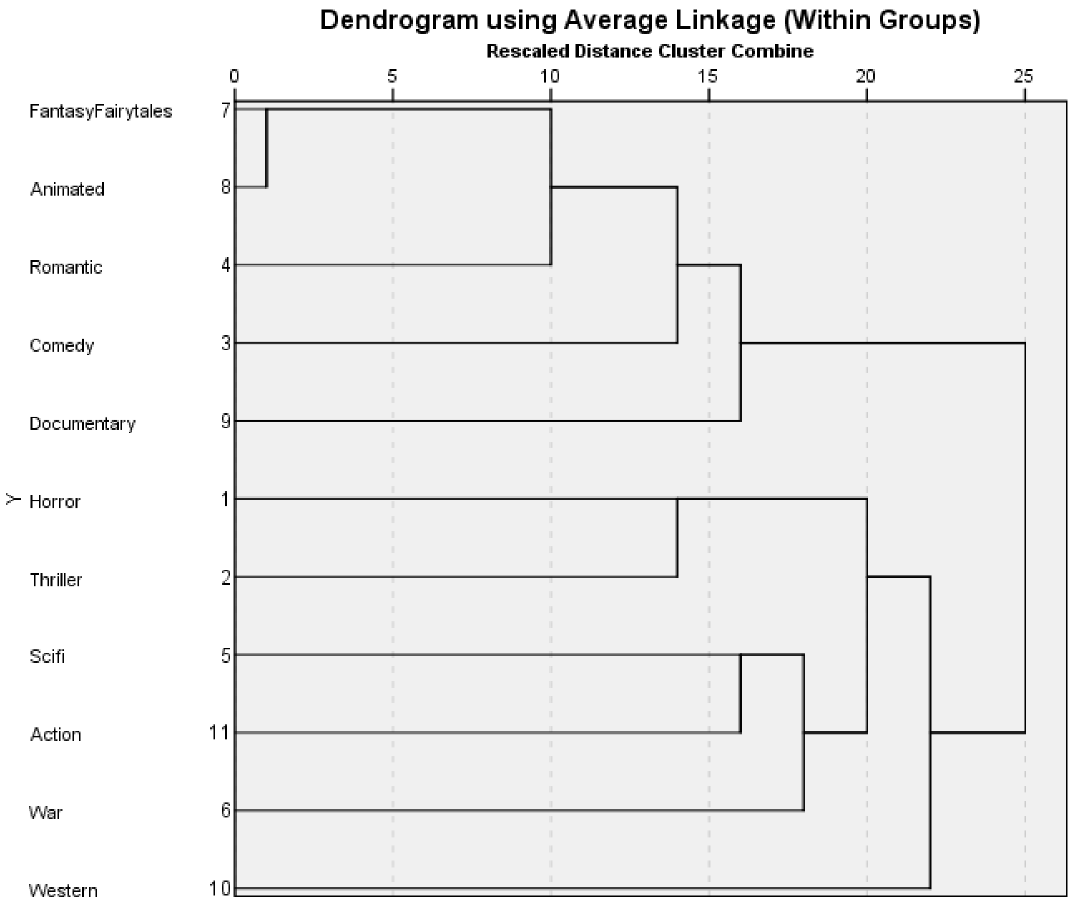


Figure 6. Dendrogram for Movie Preferences (Clusters saved by all Variables)

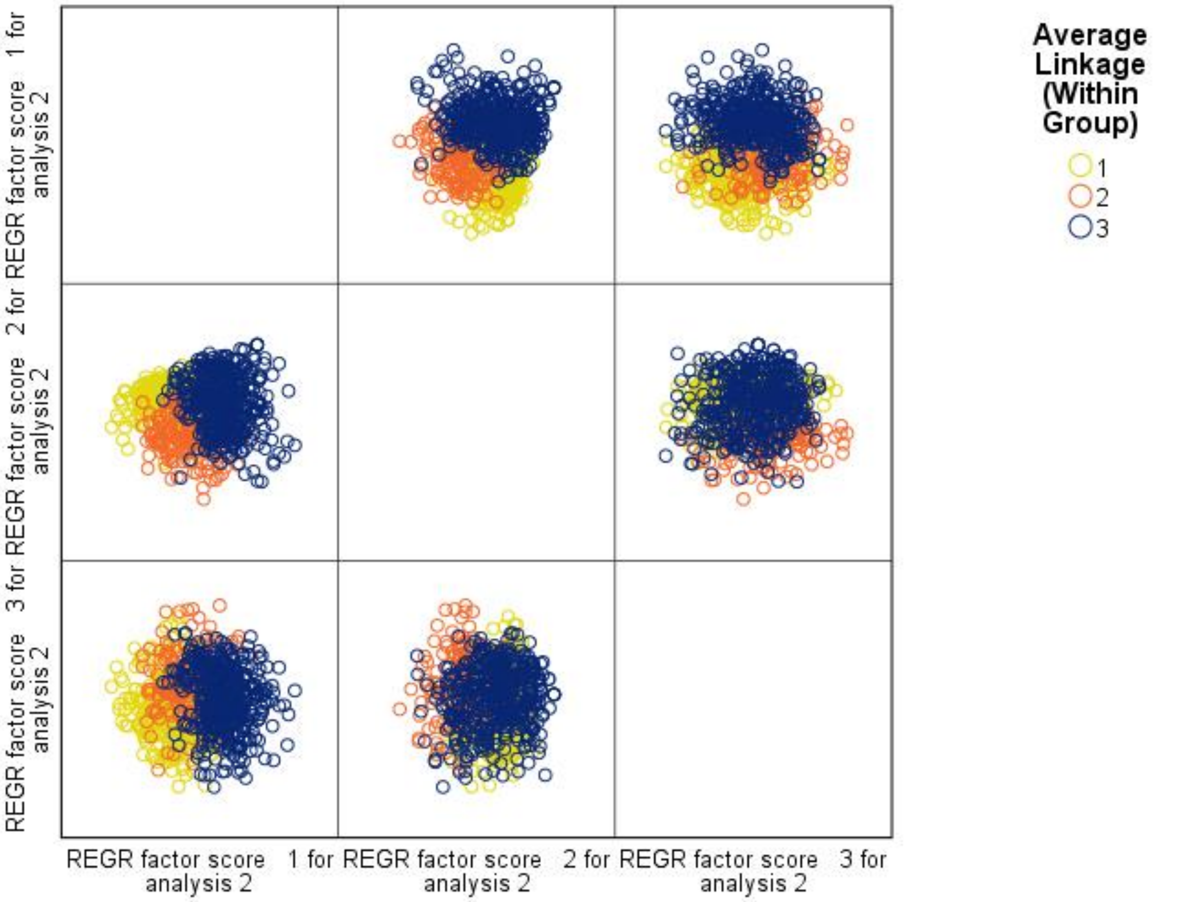


Figure 7. Hierarchical Movie Clustering Scatterplot Matrix of PC1, PC2, and PC3 (Cluster saved by all Cases)

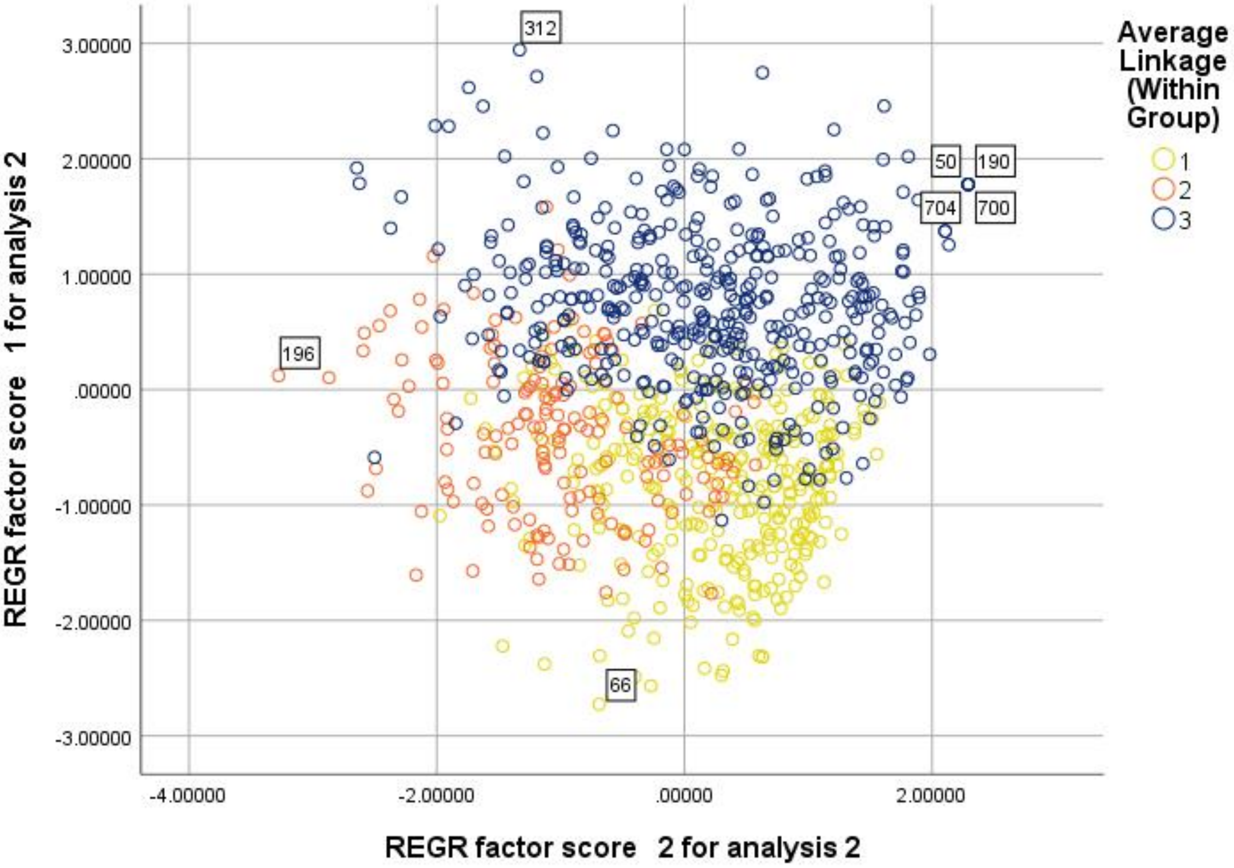
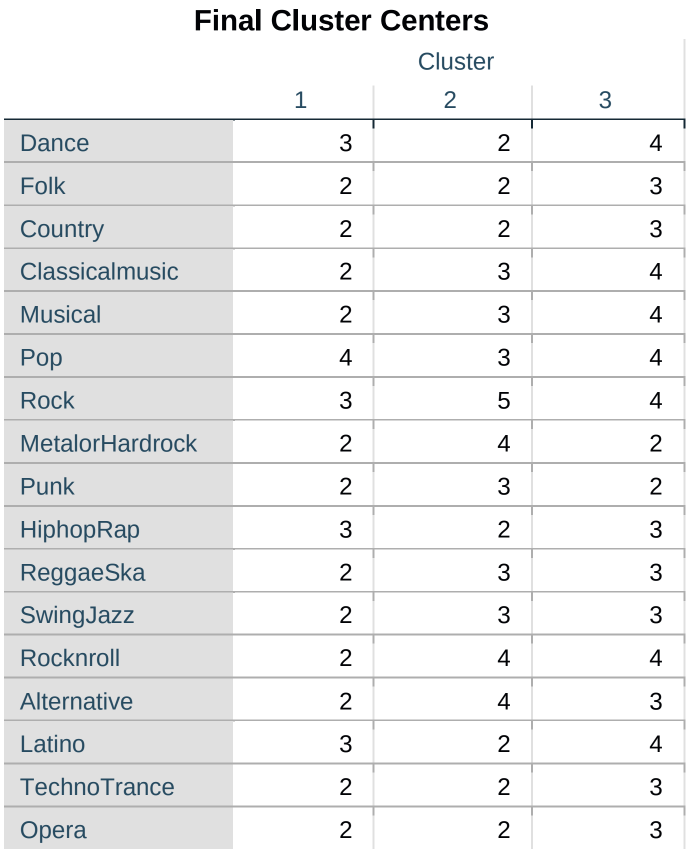


Figure 8. Hierarchical Movie Clustering Scatterplot of PC1 and PC2

2.2 K-means Clustering

Music Clustering



Final Cluster Centers

	Cluster		
	1	2	3
Dance	3	2	4
Folk	2	2	3
Country	2	2	3
Classicalmusic	2	3	4
Musical	2	3	4
Pop	4	3	4
Rock	3	5	4
MetalorHardrock	2	4	2
Punk	2	3	2
HiphopRap	3	2	3
ReggaeSka	2	3	3
SwingJazz	2	3	3
Rocknroll	2	4	4
Alternative	2	4	3
Latino	3	2	4
TechnoTrance	2	2	3
Opera	2	2	3

Table 9. Final Cluster Centers for K-means Music Clustering

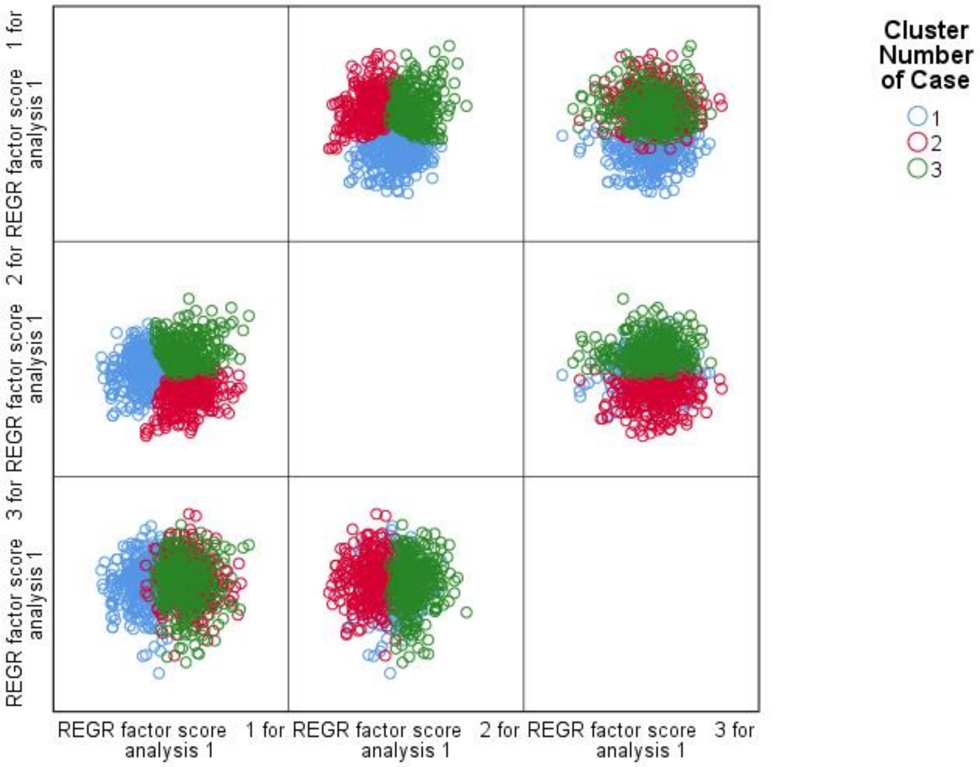


Figure 9. K-means Music Clustering Scatterplot Matrix of PC1, PC2, and PC3

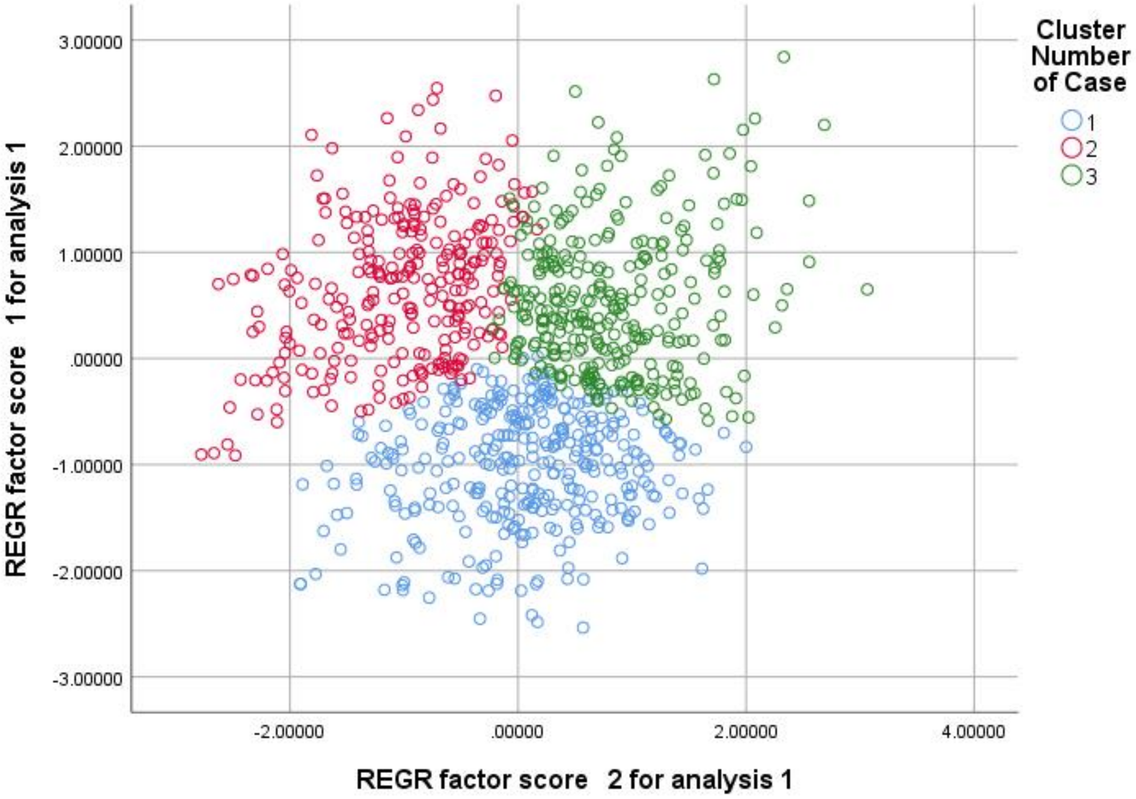
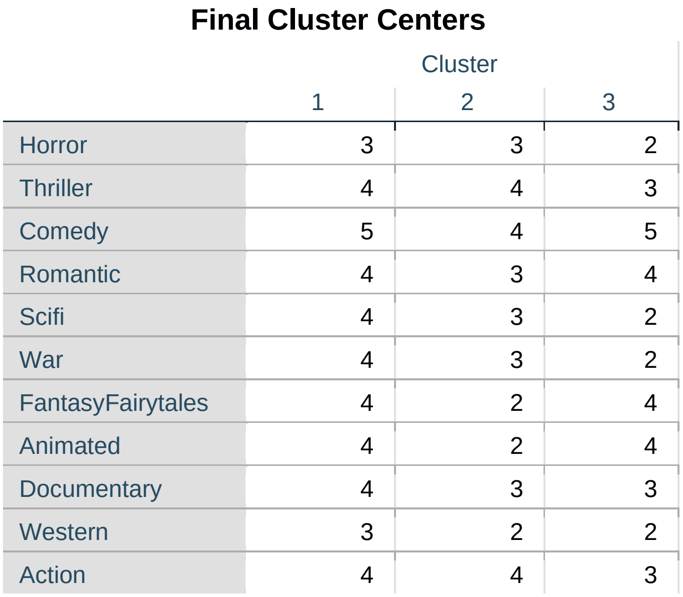


Figure 10. K-means Music Clustering Scatterplot of PC1 and PC2

Movies Clustering



Final Cluster Centers

	Cluster		
	1	2	3
Horror	3	3	2
Thriller	4	4	3
Comedy	5	4	5
Romantic	4	3	4
Scifi	4	3	2
War	4	3	2
FantasyFairytales	4	2	4
Animated	4	2	4
Documentary	4	3	3
Western	3	2	2
Action	4	4	3

Table 10. Final Cluster Centers for K-means Movie Clustering

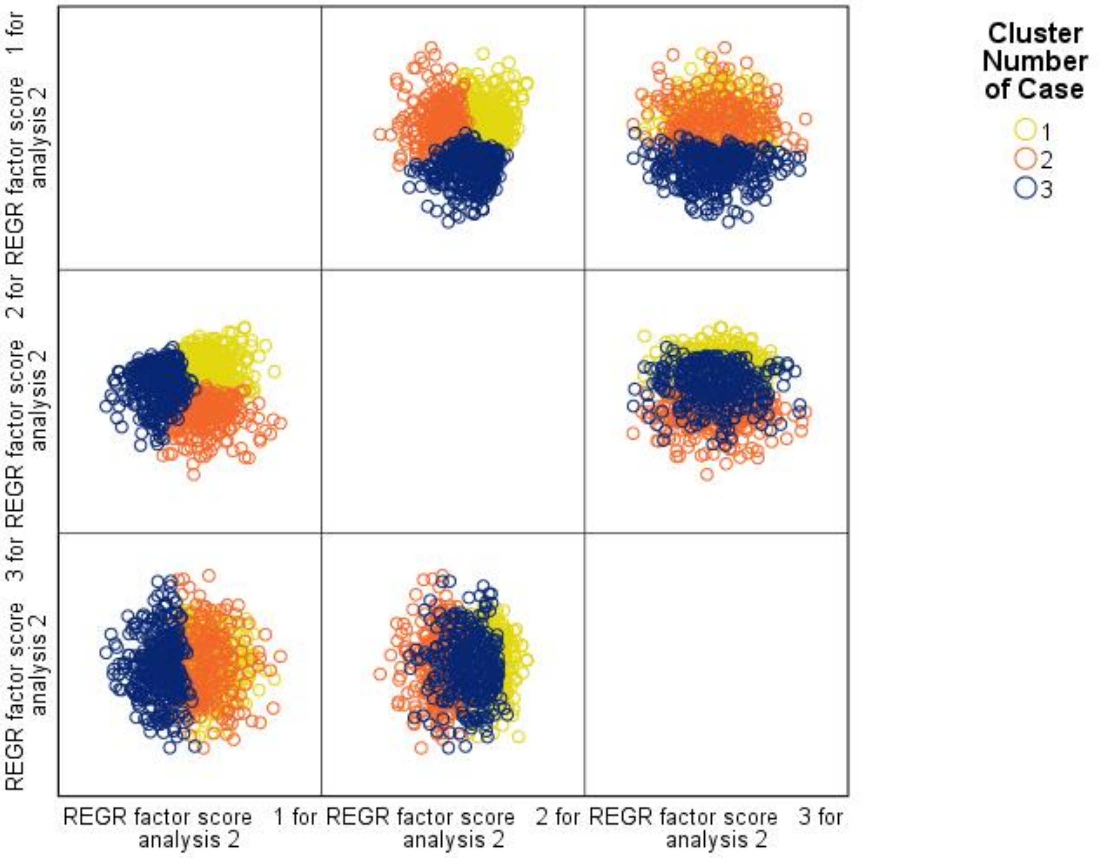


Figure 11. K-means Movie Clustering Scatterplot Matrix of PC1, PC2, and PC3

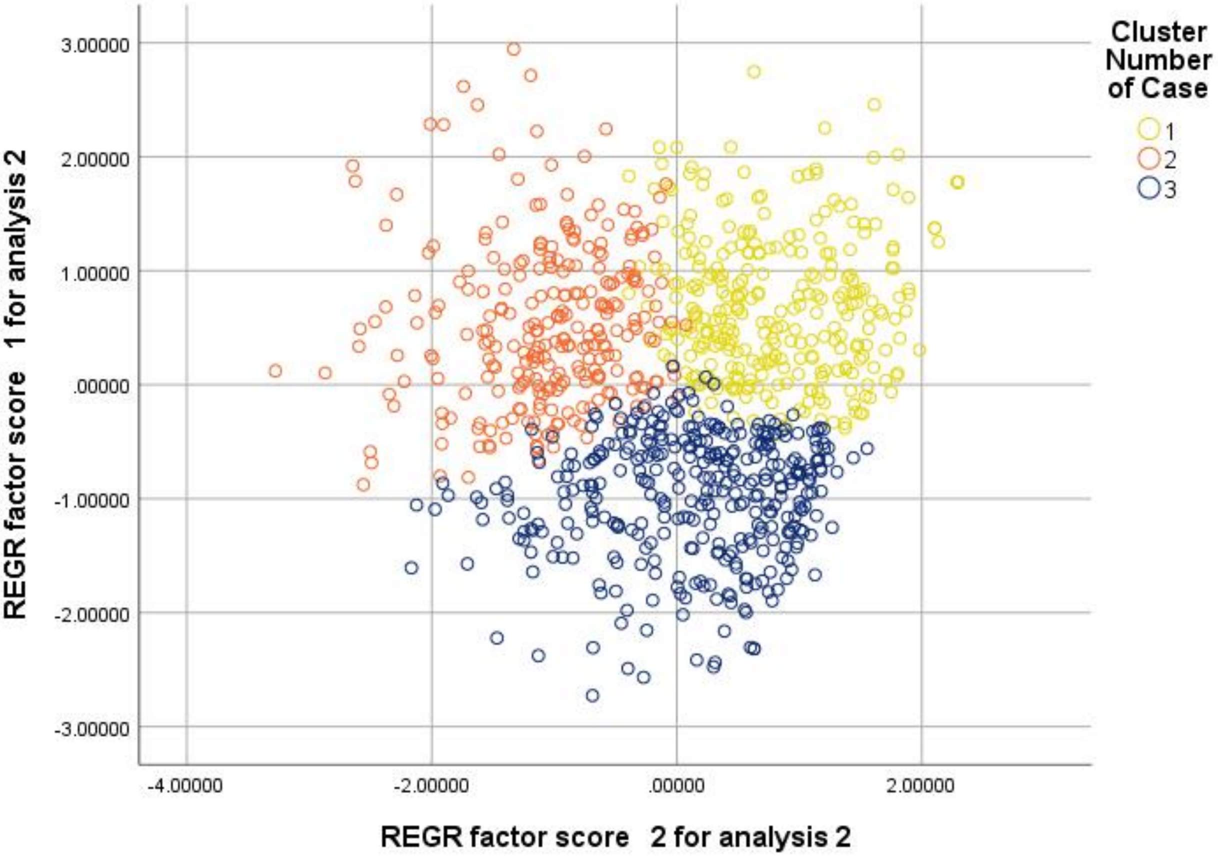
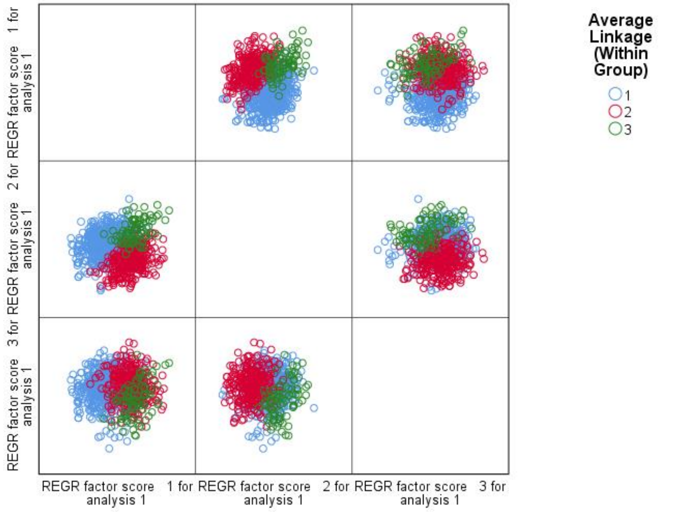
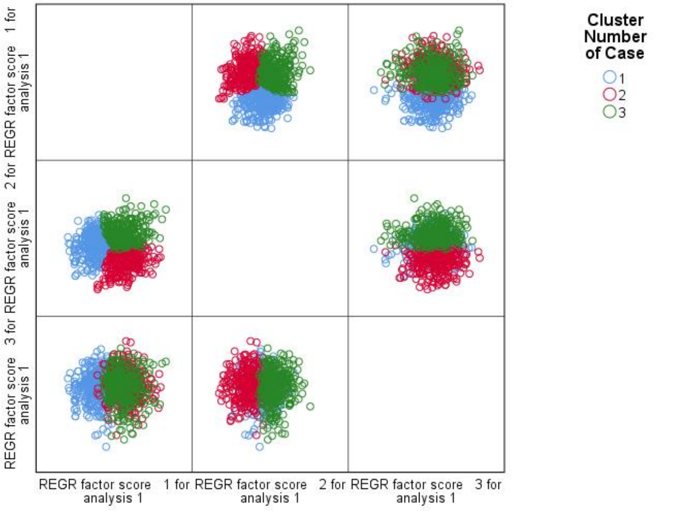
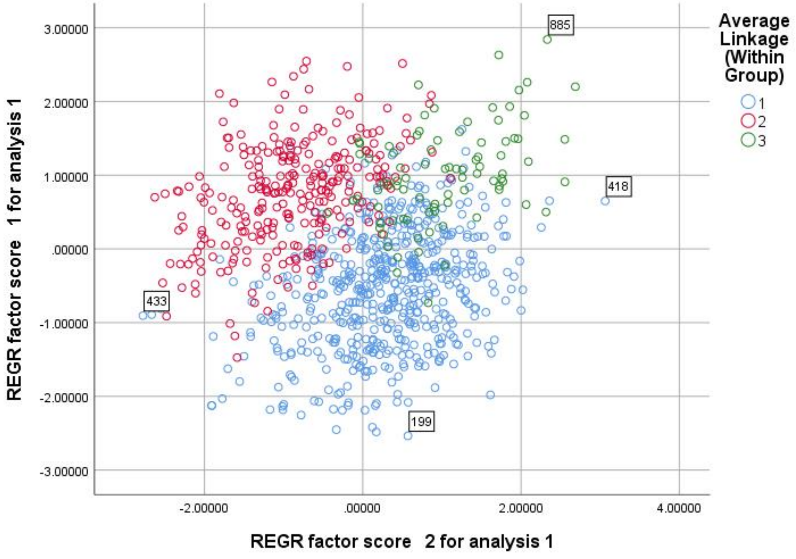
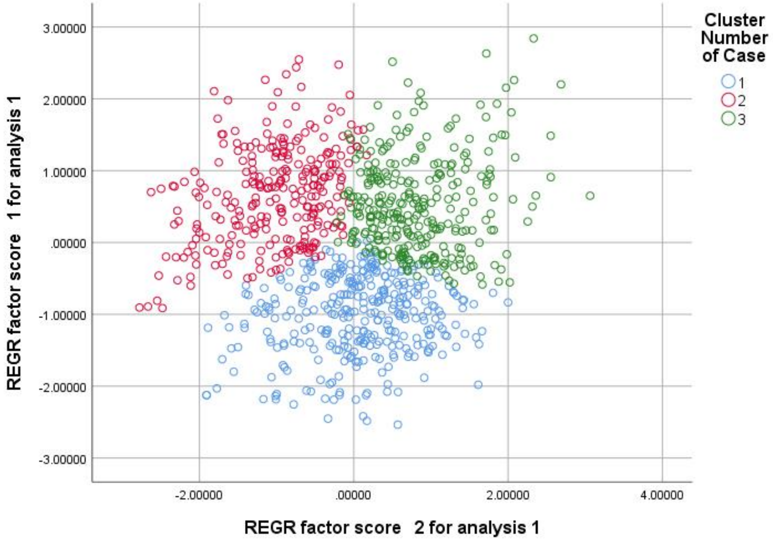
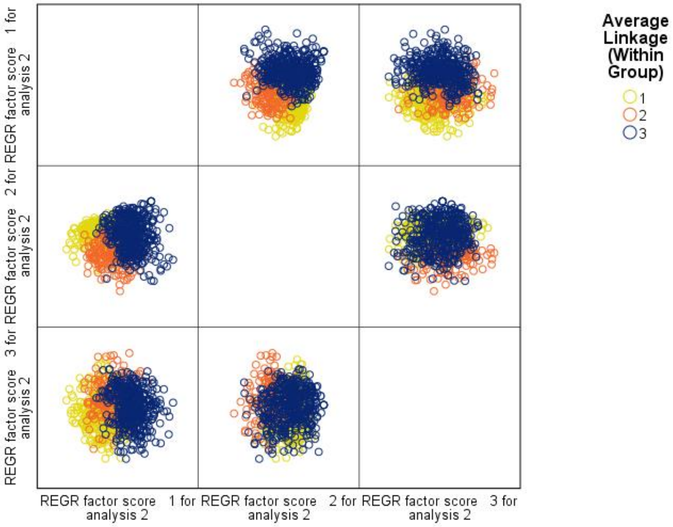
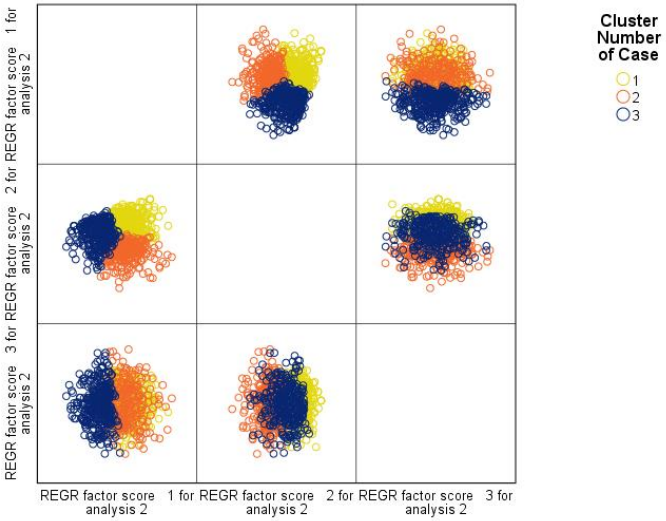
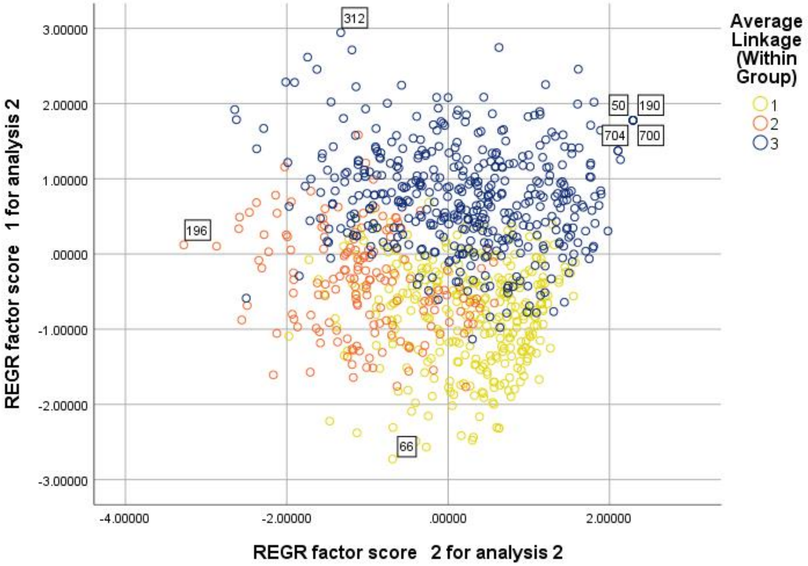
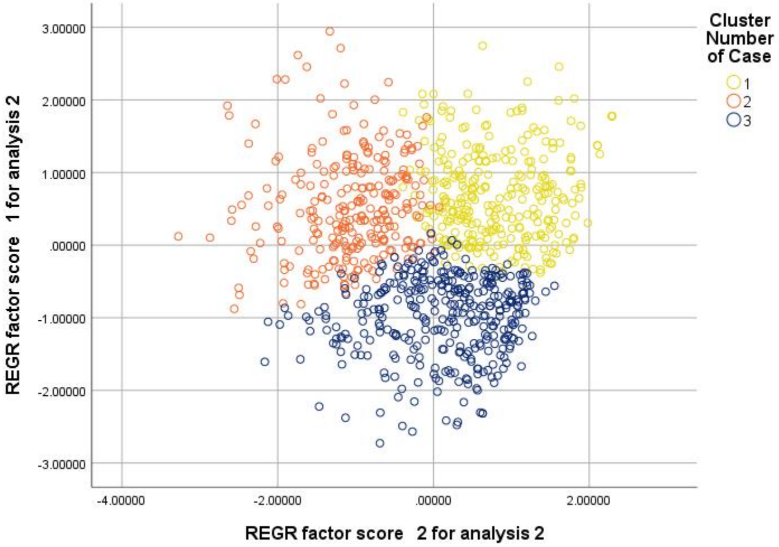


Figure 12. K-means Movie Clustering Scatterplot of PC1 and PC2

Summary & Comparison

	Hierarchical Clustering	**K-means Clustering**
Music Preferences		
		

	Hierarchical Clustering	K-means Clustering
Movie Preferences		
		

Both cluster analysis produces two different solutions, with the K-means Clustering outperforming the Hierarchical Clustering.

3. Cluster Profiles

Based on the clustering analysis, K-means is a better method to compare profile clusters on their music, movie preferences and on demographics.

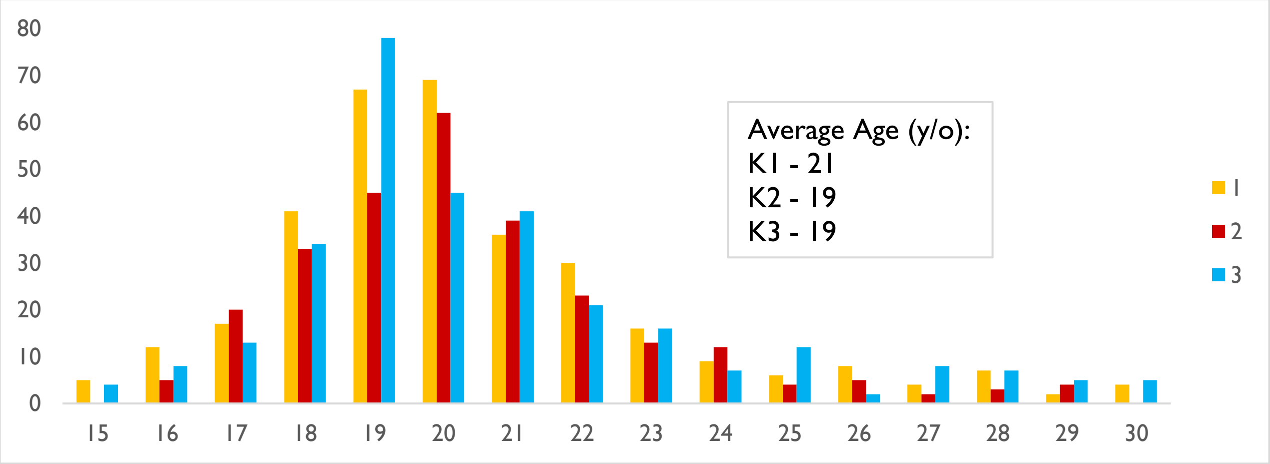
Music Clustering

	K-means Cluster			
Villagetown	**1**	**2**	**3**	**Grand Total**
city	27.38%	20.92%	22.12%	70.43%
village	9.42%	8.76%	11.39%	29.57%
Grand Total	**36.80%**	**29.68%**	**33.52%**	**100.00%**

Table 11. K-means Music Clusters by Location (Music Preference)

	K-means Cluster			
Gender	**1**	**2**	**3**	**Grand Total**
female	21.27%	15.46%	23.36%	60.09%
male	15.57%	14.25%	10.09%	39.91%
Grand Total	**36.84%**	**29.71%**	**33.44%**	**100.00%**

Table 12. K-means Music Clusters by Gender (Music Preference)



Average Age (y/o):
K1 - 21
K2 - 19
K3 - 19

Figure 13. K-means Music Clustering Bar Chart (By Age)

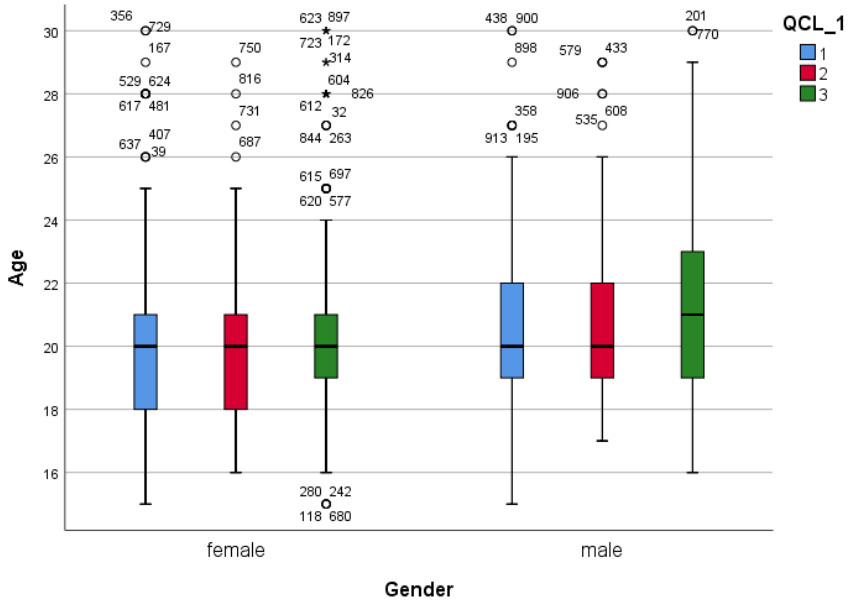


Figure 14. K-means Music Clustering Boxplot (By Age & Gender)

Movie Clustering

	K-means Cluster			
Villagetown	1	2	3	Grand Total
city	21.69%	21.69%	27.05%	70.43%
village	10.51%	7.56%	11.50%	29.57%
Grand Total	**32.20%**	**29.24%**	**38.55%**	**100.00%**

Table 13. K-means Music Clusters by Location (Movie Preference)

	K-means Cluster			
Gender	1	2	3	Grand Total
female	12.94%	11.95%	35.20%	60.09%
male	19.19%	17.43%	3.29%	39.91%
Grand Total	**32.13%**	**29.39%**	**38.49%**	**100.00%**

Table 14. K-means Music Clusters by Gender (Movie Preference)

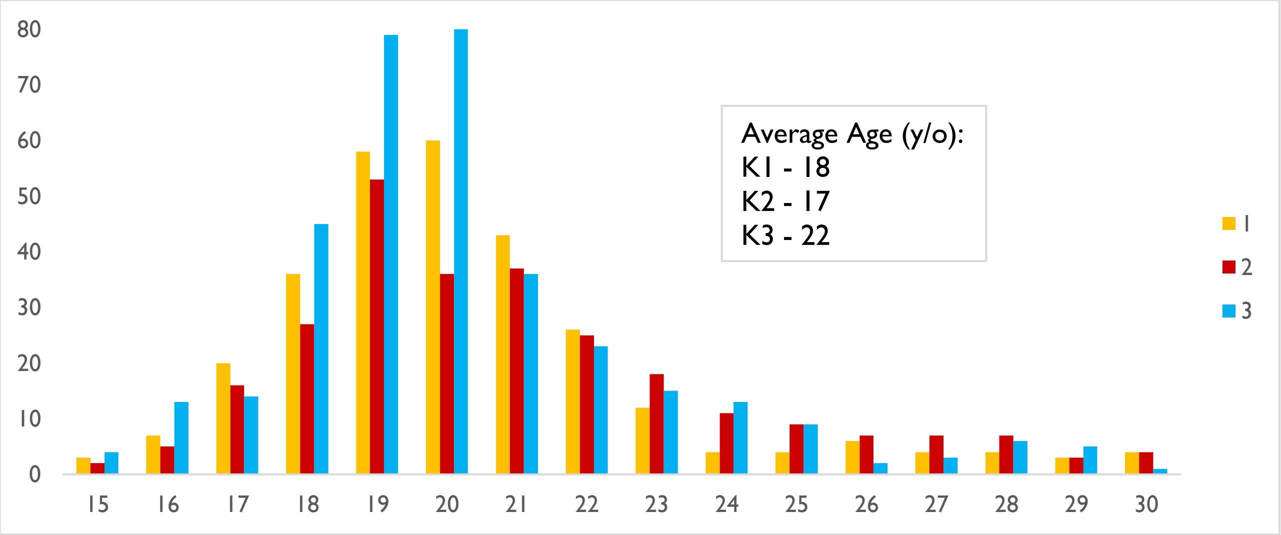


Figure 15. K-means Movie Clustering Bar Chart (By Age)

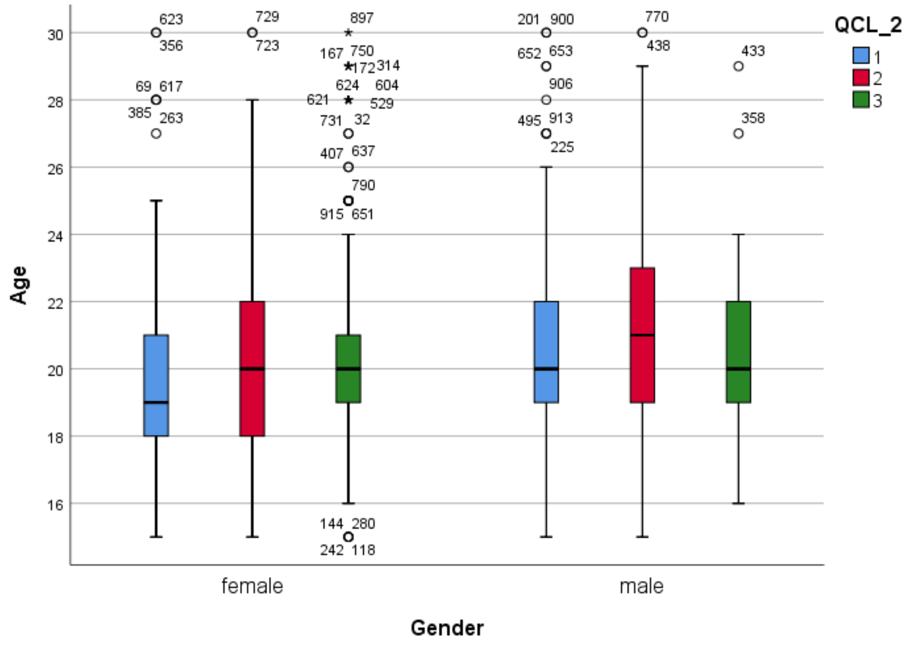
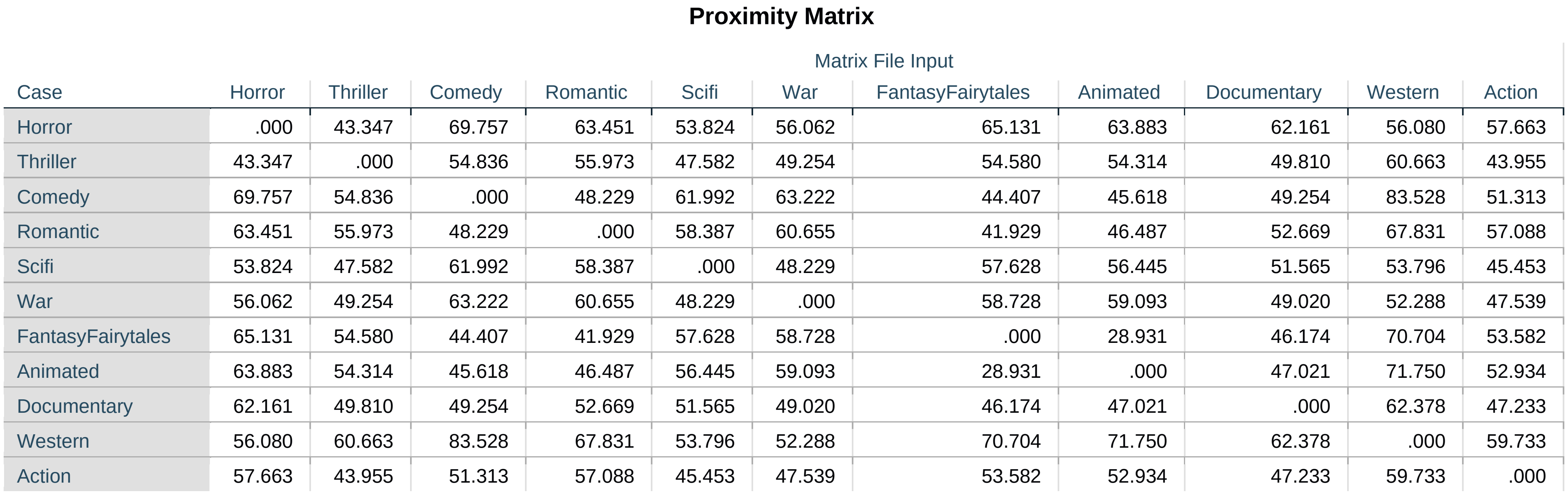


Figure 16. K-means Movie Clustering Boxplot (By Age & Gender)

4. Appendices

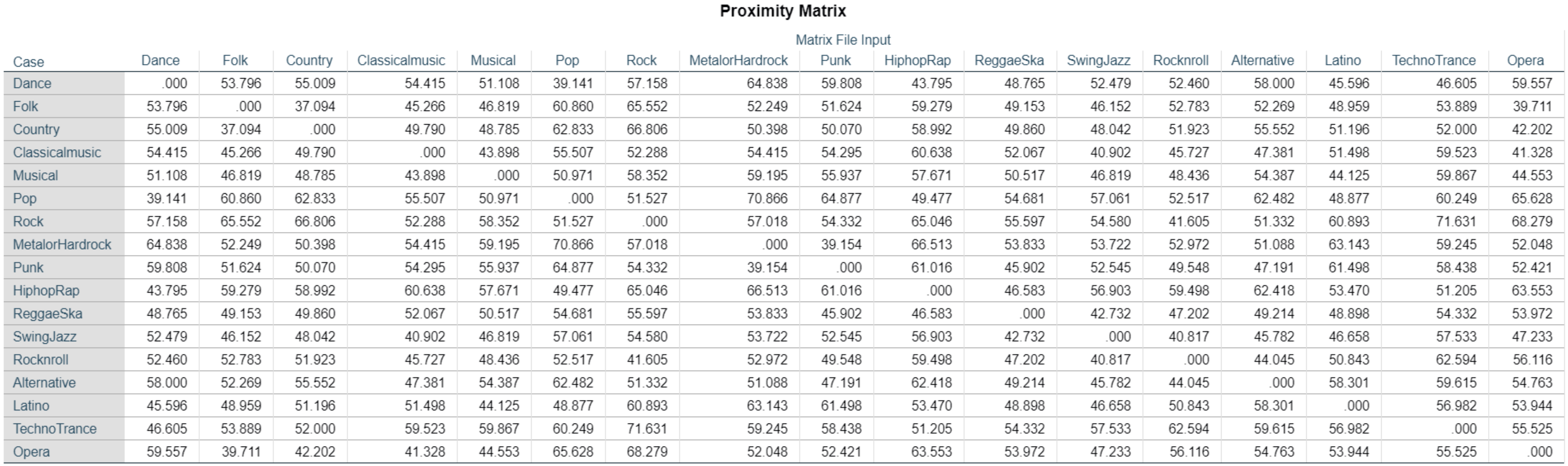
Appendix 1. Proximity Matrix of Movie Preferences (Within Groups)



Proximity Matrix

	Matrix File Input										
Case	Horror	Thriller	Comedy	Romantic	Scifi	War	FantasyFairytales	Animated	Documentary	Western	Action
Horror	.000	43.347	69.757	63.451	53.824	56.062	65.131	63.883	62.161	56.080	57.663
Thriller	43.347	.000	54.836	55.973	47.582	49.254	54.580	54.314	49.810	60.663	43.955
Comedy	69.757	54.836	.000	48.229	61.992	63.222	44.407	45.618	49.254	83.528	51.313
Romantic	63.451	55.973	48.229	.000	58.387	60.655	41.929	46.487	52.669	67.831	57.088
Scifi	53.824	47.582	61.992	58.387	.000	48.229	57.628	56.445	51.565	53.796	45.453
War	56.062	49.254	63.222	60.655	48.229	.000	58.728	59.093	49.020	52.288	47.539
FantasyFairytales	65.131	54.580	44.407	41.929	57.628	58.728	.000	28.931	46.174	70.704	53.582
Animated	63.883	54.314	45.618	46.487	56.445	59.093	28.931	.000	47.021	71.750	52.934
Documentary	62.161	49.810	49.254	52.669	51.565	49.020	46.174	47.021	.000	62.378	47.233
Western	56.080	60.663	83.528	67.831	53.796	52.288	70.704	71.750	62.378	.000	59.733
Action	57.663	43.955	51.313	57.088	45.453	47.539	53.582	52.934	47.233	59.733	.000

Appendix 2. Proximity Matrix of Music Preferences (Within Groups)



Proximity Matrix

	Matrix File Input																
Case	Dance	Folk	Country	Classicalmusic	Musical	Pop	Rock	MetalorHardrock	Punk	HiphopRap	ReggaeSka	SwingJazz	Rocknroll	Alternative	Latino	TechnoTrance	Opera
Dance	.000	53.796	55.009	54.415	51.108	39.141	57.158	64.838	59.808	43.795	48.765	52.479	52.460	58.000	45.596	46.605	59.557
Folk	53.796	.000	37.094	45.266	46.819	60.860	65.552	52.249	51.624	59.279	49.153	46.152	52.783	52.269	48.959	53.889	39.711
Country	55.009	37.094	.000	49.790	48.785	62.833	66.806	50.398	50.070	58.992	49.860	48.042	51.923	55.552	51.196	52.000	42.202
Classicalmusic	54.415	45.266	49.790	.000	43.898	55.507	52.288	54.415	54.295	60.638	52.067	40.902	45.727	47.381	51.498	59.523	41.328
Musical	51.108	46.819	48.785	43.898	.000	50.971	58.352	59.195	55.937	57.671	50.517	46.819	48.436	54.387	44.125	59.867	44.553
Pop	39.141	60.860	62.833	55.507	50.971	.000	51.527	70.866	64.877	49.477	54.681	57.061	52.517	62.482	48.877	60.249	65.628
Rock	57.158	65.552	66.806	52.288	58.352	51.527	.000	57.018	54.332	65.046	55.597	54.580	41.605	51.332	60.893	71.631	68.279
MetalorHardrock	64.838	52.249	50.398	54.415	59.195	70.866	57.018	.000	39.154	66.513	53.833	53.722	52.972	51.088	63.143	59.245	52.048
Punk	59.808	51.624	50.070	54.295	55.937	64.877	54.332	39.154	.000	61.016	45.902	52.545	49.548	47.191	61.498	58.438	52.421
HiphopRap	43.795	59.279	58.992	60.638	57.671	49.477	65.046	66.513	61.016	.000	46.583	56.903	59.498	62.418	53.470	51.205	63.553
ReggaeSka	48.765	49.153	49.860	52.067	50.517	54.681	55.597	53.833	45.902	46.583	.000	42.732	47.202	49.214	48.898	54.332	53.972
SwingJazz	52.479	46.152	48.042	40.902	46.819	57.061	54.580	53.722	52.545	56.903	42.732	.000	40.817	45.782	46.658	57.533	47.233
Rocknroll	52.460	52.783	51.923	45.727	48.436	52.517	41.605	52.972	49.548	59.498	47.202	40.817	.000	44.045	50.843	62.594	56.116
Alternative	58.000	52.269	55.552	47.381	54.387	62.482	51.332	51.088	47.191	62.418	49.214	45.782	44.045	.000	58.301	59.615	54.763
Latino	45.596	48.959	51.196	51.498	44.125	48.877	60.893	63.143	61.498	53.470	48.898	46.658	50.843	58.301	.000	56.982	53.944
TechnoTrance	46.605	53.889	52.000	59.523	59.867	60.249	71.631	59.245	58.438	51.205	54.332	57.533	62.594	59.615	56.982	.000	55.525
Opera	59.557	39.711	42.202	41.328	44.553	65.628	68.279	52.048	52.421	63.553	53.972	47.233	56.116	54.763	53.944	55.525	.000