BUSA3020: Advanced Analytics Techniques

ASSIGNMENT 3: CLUSTERING

45345473

Amritha Jeyarathan

TABLE OF CONTENTS

I. INTRODUCTION

I A. PURPOSE

To utilise *Orange Data Miner* to:

1. Reduce the dataset into fewer variables using *Principal Component Analysis* (PCA)
2. Identify clusters, or market segments, of individuals based on their music and movie preferences
3. Compare the cluster solutions produced by the different algorithms
4. Profile clusters on individuals' music and movie preferences and on their demographics

II. METHODOLOGY

II A. EXPERIMENTATION OF VARIABLES AND FACTORS

II A. STEP 1: LOOKING AT THE DATA

The data represents 1,010 individuals surveyed on 41 features, which can be categorised into a certain variable (*figure 1.1*), the features *Music* and *Movies* were ignored due to their redundancy.

Variable Name	Description	Data Type
Music	Music	Nil (variable ignored)
	Slow songs or fast songs	Numeric
	Dance	Numeric
	Folk	Numeric
	Country	Numeric
	Classical music	Numeric
	Musical	Numeric
	Pop	Numeric
	Rock	Numeric
	Metal or Hardrock	Numeric
	Punk	Numeric
	Hiphop, Rap	Numeric
	Reggae, Ska	Numeric
	Swing, Jazz	Numeric
	Rock n roll	Numeric
	Alternative	Numeric
	Latino	Numeric
	Techno, Trance	Numeric
	Opera	Numeric
Movies	Movies	Nil (variable ignored)
	Horror	Numeric
	Thriller	Numeric
	Comedy	Numeric
	Romantic	Numeric
	Sci-fi	Numeric
	War	Numeric
	Fantasy/Fairy tales	Numeric
	Animated	Numeric
	Documentary	Numeric
	Western	Numeric

	Action	Numeric
Demographic	Age	Numeric
	Height	Numeric
	Weight	Numeric
	Number of siblings	Numeric
	Spending on healthy eating	Numeric
	Heathy eating	Numeric
	Daily events	Numeric
	Finances	Numeric
	Shopping centres	Numeric
	Branded clothing	Numeric
	Entertainment spending	Numeric
	Spending on looks	Numeric
	Spending on gadgets	Numeric
	Smoking	Categorical
	Gender	Categorical
	Left - right handed	Categorical
	Education	Categorical
	Only child	Nil (variable ignored)
	Village - town	Categorical
	House - block of flats	Categorical

Figure 1.1 Each feature can be related to either variable, *Music*, *Movie* or *Demographic*. Individuals surveyed on whether they: (a) listen to music or (b) watch movies were ignored as they are contradictory with other features and thus were considered redundant. *Only child* was also ignored, as its responses are reflected in the *Number of siblings* variable.

II A. STEP 2: CONVERTING THE VARIABLES

Excel was employed to convert the categorical features in figure 1.2 into Boolean values to ensure consistency.

Former Variable	Dummy Variable	Description	Justification
Smoking - Never Smoked - Tried Smoking - Former Smoker - Current Smoker	Tried Smoking	Boolean value. 1 = yes 0 = no	Three dummy variables were created in place of the smoking variable. *Tried Smoking*, *Former Smoker* and *Current Smoker*. A dummy variable was not created for *Never Smoked*, as a value of 0 in all three dummy variables infers that the individual has never smoked.
	Former Smoker	Boolean value. 1 = yes 0 = no	
	Current Smoker	Boolean value. 1 = yes 0 = no	
Alcohol - Never - Social Drinker - Drink a lot	Social Drinker	Boolean value. 1 = yes 0 = no	Two dummy variables were created in place of the alcohol variable. *Social Drinker* and *Drink a lot*. A dummy variable was not created for *Never*, as a value of 0 in both dummy variables infers that the individual has never consumed alcohol.
	Drink a lot	Boolean value. 1 = yes 0 = no	
Gender - Male - Female	Male	Boolean value. 1 = yes 0 = no	A *Male* dummy variable was created. A dummy variable was not created for *Female*, as a value of 0 in *Male* infers that the individual is female.

Left - Right Handed - Right Handed - Left Handed	Right Handed	Boolean value. 1 = yes 0 = no	A *Right Handed* dummy variable was created. A dummy variable was not created for *Left Handed*, as a value of 0 in *Right Handed* infers that the individual is left handed.
Education - Currently a primary school pupil - Primary School - Secondary School - College/Bachelor degree - Masters Degree - Doctorate Degree	Primary Education	Boolean value. 1 = yes 0 = no	Three dummy variables were created in place of the education variable. *Primary Education*, *Secondary Education* and *Tertiary Education*. The response *currently a primary school pupil* was included in the *Primary Education* variable A dummy variable was not created for *Masters Degree* or *Doctorate Degree*, as a value of 0 in all three dummy variables infers that the individual has a post graduate degree.
	Secondary Education	Boolean value. 1 = yes 0 = no	
	Tertiary Education	Boolean value. 1 = yes 0 = no	
Village – Town - City - Village	Resides in City	Boolean value. 1 = yes 0 = no	A *Resides in City* dummy variable was created. A dummy variable was not created for *Village*, as a value of 0 in *Resides in City* infers that the individual resides in a village.
House - Block of Flats - House/Bungalow - Block of Flats	Resides in House/Bungalow	Boolean value. 1 = yes 0 = no	A *Resides in House/Bungalow* dummy variable was created. A dummy variable was not created for *Block of Flats*, as a value of 0 in *Resides in House/Bungalow* infers that the individual resides in a block of flats.

Figure 1.2. The above variables were transformed into dummy variables with Boolean values. This ensures that the data is consistent among all variables and allows for further simplification of the data prior to the PCA process. Since the data above relates to the individuals' demographic, this process is particularly helpful for step three, as the average or most frequent value will be taken as the default value.

II A. STEP 3: IMPUTING THE DATA

The *Impute* function was employed in *Orange* to remove the 0.4% missing values by setting those in all variables to their average or most frequent value. This is to ensure the dataset is complete, as missing values can alter the outcome of PCA and clustering.

II A. STEP 4: SELECTING THE DATA

The *Select Columns* function was then utilised to retain features relating to music and movies and to ignore those relating to demographics (figure 1.3).

Features Retained	Features Ignored
Slow songs or fast songs	Age
Dance	Height
Folk	Weight
Country	Spending on healthy eating
Classical music	Heathy eating
Musical	Daily events

Pop	Finances
Rock	Shopping centres
Metal or Hardrock	Branded clothing
Punk	Entertainment spending
Hiphop, Rap	Spending on looks
Reggae, Ska	Spending on gadgets
Swing, Jazz	Tried smoking
Rock n roll	Former smoker
Alternative	Current smoker
Latino	Social drinker
Techno, Trance	Drink a lot
Opera	Male
Horror	Right handed
Thriller	Primary education
Comedy	Secondary education
Romantic	Tertiary education
Sci-fi	Resides in city
War	Resides in House/Bungalow
Fantasy/Fairy tales	
Animated	
Documentary	
Western	
Action	

Figure 1.3. The above variables were considered on whether or not they should be retained and utilised in the Principal Component Analysis or Ignored. Variables relating to individuals' demographic was ignored as they are considered important in the reduction of data, but more so, in profiling the clusters formed.

II A. STEP 5: PERFORMING A PCA

The *PCA* function was employed to perform a principal component analysis to retain features relating to music and movies and to ignore those relating to demographics (figure 1.4, figure 1.5 and figure 1.6).

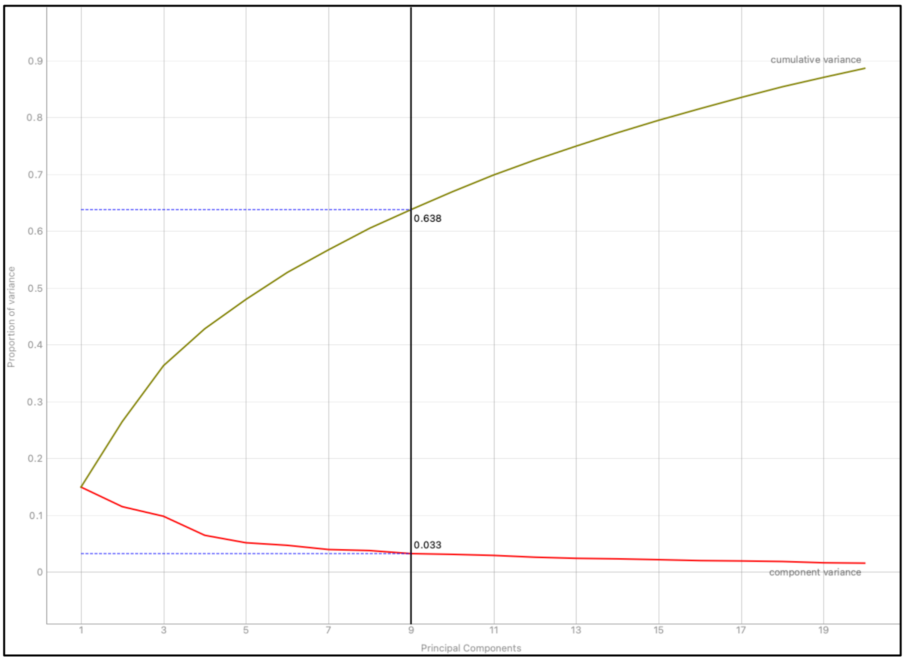


Figure 1.4. Scree plot obtained from the *Principal Component Analysis* performed on *Orange Data Miner*. Nine principal components provide a component variance of 3.3% and a cumulative variance of

Number of components	Component variance	Cumulative variance
1	15.0%	15.0%
2	11.5%	26.5%
3	9.8%	36.4%
4	6.5%	42.8%
5	5.2%	48.0%
6	4.7%	52.8%
7	4.0%	56.7%
8	3.8%	60.5%
9	3.3%	63.8%
10	3.1%	67.0%
11	3.0%	69.9%

Figure 1.5. The decision was made to use only nine principal components, as the component variance curve begins to flatten out at a faster rate and the cumulative variance increases at a decreasing rate, as shown in the scree plot in figure 1.4. Although nine components only reflect 63.8% of the original data, it appears to be sufficient in summarising the responses from music and movies variables. Additionally, the component variance decreases at a decreasing rate from 0.5% at nine variables, to increasing by only 0.2% at ten variables, 0.1% and eleven variables and so on.

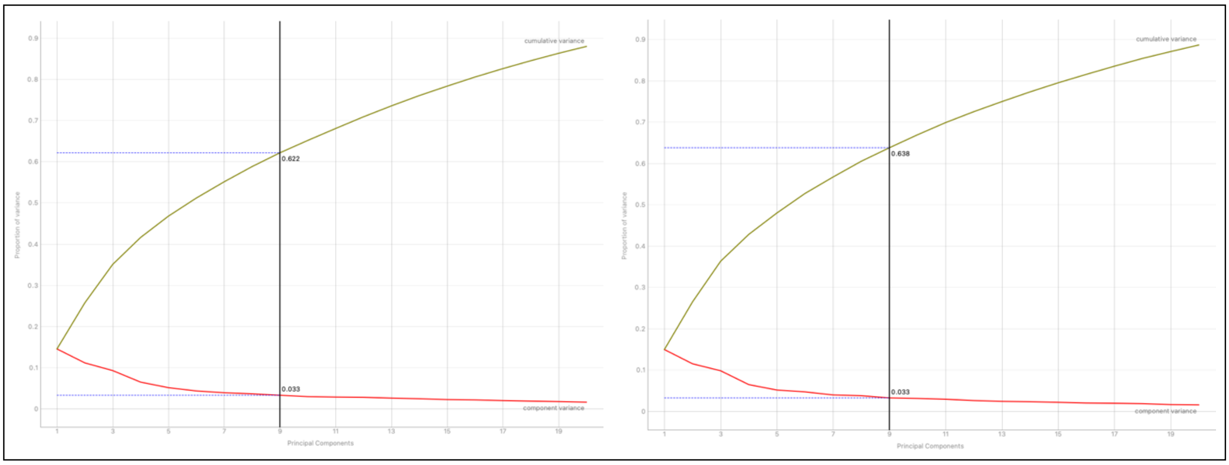
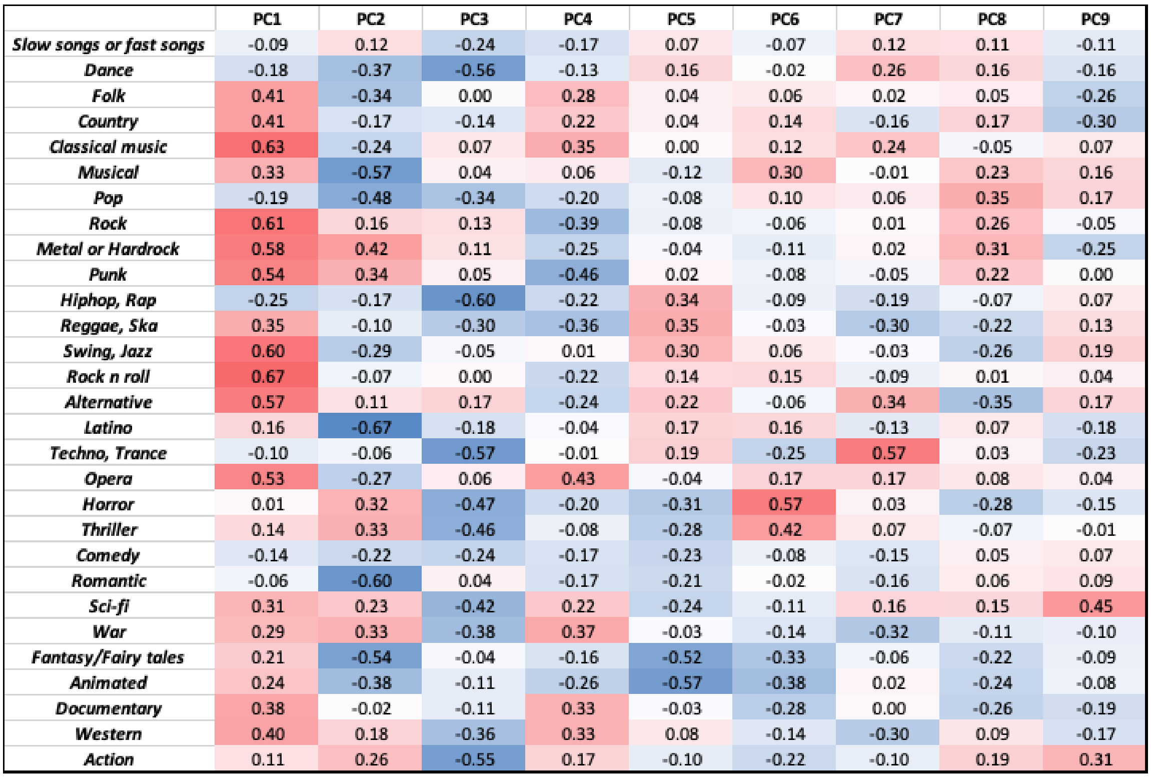


Figure 1.6. (left to right) a scree plot of the nine variables when it is normalised and a scree plot of the variables when it is not normalised. It was decided to not normalise the variables and to instead, leave them unstandardised. This is because all of the variables being considered for PCA lie on the same scale of measurement i.e., a rating of 1-5. Additionally, leaving the variable unstandardised allows for a higher cumulative variance of 63.8%, compared to 62.2% when normalised.

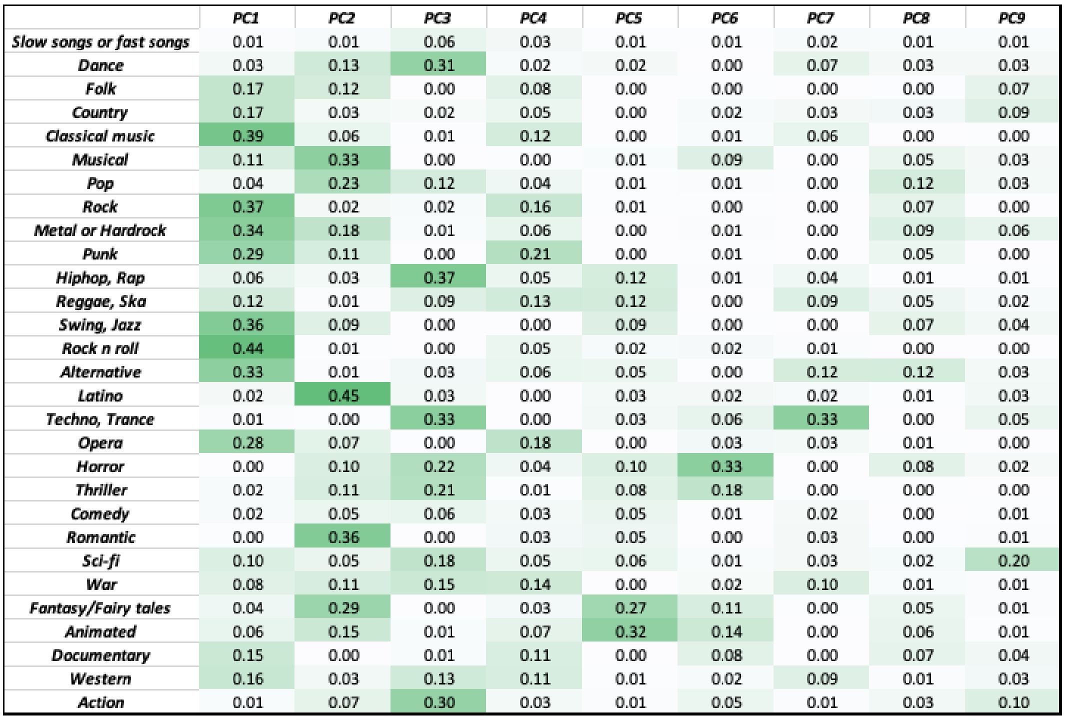
II A. STEP 6: ASSESSING CORRELATION AND SHARED VARIANCE

Based on the data obtained following the Principal Component Analysis, correlation and shared variance matrices were created in excel (figure 1.7 and figure 1.8).



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Slow songs or fast songs	-0.09	0.12	-0.24	-0.17	0.07	-0.07	0.12	0.11	-0.11
Dance	-0.18	-0.37	-0.56	-0.13	0.16	-0.02	0.26	0.16	-0.16
Folk	0.41	-0.34	0.00	0.28	0.04	0.06	0.02	0.05	-0.26
Country	0.41	-0.17	-0.14	0.22	0.04	0.14	-0.16	0.17	-0.30
Classical music	0.63	-0.24	0.07	0.35	0.00	0.12	0.24	-0.05	0.07
Musical	0.33	-0.57	0.04	0.06	-0.12	0.30	-0.01	0.23	0.16
Pop	-0.19	-0.48	-0.34	-0.20	-0.08	0.10	0.06	0.35	0.17
Rock	0.61	0.16	0.13	-0.39	-0.08	-0.06	0.01	0.26	-0.05
Metal or Hardrock	0.58	0.42	0.11	-0.25	-0.04	-0.11	0.02	0.31	-0.25
Punk	0.54	0.34	0.05	-0.46	0.02	-0.08	-0.05	0.22	0.00
Hiphop, Rap	-0.25	-0.17	-0.60	-0.22	0.34	-0.09	-0.19	-0.07	0.07
Reggae, Ska	0.35	-0.10	-0.30	-0.36	0.35	-0.03	-0.30	-0.22	0.13
Swing, Jazz	0.60	-0.29	-0.05	0.01	0.30	0.06	-0.03	-0.26	0.19
Rock n roll	0.67	-0.07	0.00	-0.22	0.14	0.15	-0.09	0.01	0.04
Alternative	0.57	0.11	0.17	-0.24	0.22	-0.06	0.34	-0.35	0.17
Latino	0.16	-0.67	-0.18	-0.04	0.17	0.16	-0.13	0.07	-0.18
Techno, Trance	-0.10	-0.06	-0.57	-0.01	0.19	-0.25	0.57	0.03	-0.23
Opera	0.53	-0.27	0.06	0.43	-0.04	0.17	0.17	0.08	0.04
Horror	0.01	0.32	-0.47	-0.20	-0.31	0.57	0.03	-0.28	-0.15
Thriller	0.14	0.33	-0.46	-0.08	-0.28	0.42	0.07	-0.07	-0.01
Comedy	-0.14	-0.22	-0.24	-0.17	-0.23	-0.08	-0.15	0.05	0.07
Romantic	-0.06	-0.60	0.04	-0.17	-0.21	-0.02	-0.16	0.06	0.09
Sci-fi	0.31	0.23	-0.42	0.22	-0.24	-0.11	0.16	0.15	0.45
War	0.29	0.33	-0.38	0.37	-0.03	-0.14	-0.32	-0.11	-0.10
Fantasy/Fairy tales	0.21	-0.54	-0.04	-0.16	-0.52	-0.33	-0.06	-0.22	-0.09
Animated	0.24	-0.38	-0.11	-0.26	-0.57	-0.38	0.02	-0.24	-0.08
Documentary	0.38	-0.02	-0.11	0.33	-0.03	-0.28	0.00	-0.26	-0.19
Western	0.40	0.18	-0.36	0.33	0.08	-0.14	-0.30	0.09	-0.17
Action	0.11	0.26	-0.55	0.17	-0.10	-0.22	-0.10	0.19	0.31

Figure 1.7. A correlation matrix was formed between the principal components and the features these components are comprised of using excel. Those opaque and in red represent a higher positive correlation while those in blue indicate a negative relationship. E.g., PC1 appears to have above-average positive correlations with *rock n roll*, *classical music*, *rock* and *swing, jazz*. While PC2 has higher negative correlations with *Latino* and *romantic*. Interesting to note, is that there none of the principal components have a very high correlation coefficient i.e., above 0.7 with any of the features of music and movies.



	PC1	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*	*PC7*	*PC8*	*PC9*
Slow songs or fast songs	0.01	0.01	0.06	0.03	0.01	0.01	0.02	0.01	0.01
Dance	0.03	0.13	0.31	0.02	0.02	0.00	0.07	0.03	0.03
Folk	0.17	0.12	0.00	0.08	0.00	0.00	0.00	0.00	0.07
Country	0.17	0.03	0.02	0.05	0.00	0.02	0.03	0.03	0.09
Classical music	0.39	0.06	0.01	0.12	0.00	0.01	0.06	0.00	0.00
Musical	0.11	0.33	0.00	0.00	0.01	0.09	0.00	0.05	0.03
Pop	0.04	0.23	0.12	0.04	0.01	0.01	0.00	0.12	0.03
Rock	0.37	0.02	0.02	0.16	0.01	0.00	0.00	0.07	0.00
Metal or Hardrock	0.34	0.18	0.01	0.06	0.00	0.01	0.00	0.09	0.06
Punk	0.29	0.11	0.00	0.21	0.00	0.01	0.00	0.05	0.00
Hiphop, Rap	0.06	0.03	0.37	0.05	0.12	0.01	0.04	0.01	0.01
Reggae, Ska	0.12	0.01	0.09	0.13	0.12	0.00	0.09	0.05	0.02
Swing, Jazz	0.36	0.09	0.00	0.00	0.09	0.00	0.00	0.07	0.04
Rock n roll	0.44	0.01	0.00	0.05	0.02	0.02	0.01	0.00	0.00
Alternative	0.33	0.01	0.03	0.06	0.05	0.00	0.12	0.12	0.03
Latino	0.02	0.45	0.03	0.00	0.03	0.02	0.02	0.01	0.03
Techno, Trance	0.01	0.00	0.33	0.00	0.03	0.06	0.33	0.00	0.05
Opera	0.28	0.07	0.00	0.18	0.00	0.03	0.03	0.01	0.00
Horror	0.00	0.10	0.22	0.04	0.10	0.33	0.00	0.08	0.02
Thriller	0.02	0.11	0.21	0.01	0.08	0.18	0.00	0.00	0.00
Comedy	0.02	0.05	0.06	0.03	0.05	0.01	0.02	0.00	0.01
Romantic	0.00	0.36	0.00	0.03	0.05	0.00	0.03	0.00	0.01
Sci-fi	0.10	0.05	0.18	0.05	0.06	0.01	0.03	0.02	0.20
War	0.08	0.11	0.15	0.14	0.00	0.02	0.10	0.01	0.01
Fantasy/Fairy tales	0.04	0.29	0.00	0.03	0.27	0.11	0.00	0.05	0.01
Animated	0.06	0.15	0.01	0.07	0.32	0.14	0.00	0.06	0.01
Documentary	0.15	0.00	0.01	0.11	0.00	0.08	0.00	0.07	0.04
Western	0.16	0.03	0.13	0.11	0.01	0.02	0.09	0.01	0.03
Action	0.01	0.07	0.30	0.03	0.01	0.05	0.01	0.03	0.10

Figure 1.8. A shared variance matrix was also formed by taking the squared correlation coefficients between the principal components and the features these components are comprised of using excel. Shared variance refers to the variance shared among a set of variables; variables that are highly correlated (whether positively or negatively) will share a lot of variance. Values that are opaquer in green suggest a higher shared variance.

Principal Component	Degree of shared variance	Feature	Shared variance	Interpretation
PC1	Moderate degree: 40% - 60%	Rock n Roll	44%	It appears that due to its higher shared variance, PC1 attempts to capture the following features the most, in comparison to other principal components: rock n roll, classical music, rock, swing, jazz, metal or hard rock, alternative, punk, opera, folk, country, western and documentary.
	Less than moderate degree: 20% - 40%	Classical music	39%	
		Rock	37%	
		Swing, Jazz	36%	
		Metal or Hardrock	34%	
		Alternative	33%	
		Punk	29%	
		Opera	28%	
	Low degree: 10% - 20%	Folk	17%	
		Country	17%	
		Western	16%	
		Documentary	15%	
PC2	Moderate degree: 40% - 60%	Latino	45%	PC2 attempts to capture the following features the most, in comparison to other principal components: latino, romantic, musical, fantasy/fairy tales and pop.
	Less than moderate degree: 20% - 40%	Romantic	36%	
		Musical	33%	
		Fantasy/Fairy tales	29%	
		Pop	23%	
PC3	Less than moderate degree: 20% - 40%	Hiphop, Rap	37%	PC3 attempts to capture the following features the most, in comparison to other principal components: hiphop, rap, techno, trance (tied with PC7), dance, action, thriller, war, comedy, slow songs or fast songs.
		Techno, Trance	33%	
		Dance	31%	
		Action	30%	
		Thriller	21%	
	Low degree: 10% - 20%	War	15%	
	Very low degree: 0% - 10%	Comedy	6%	
		Slow songs or fast songs	6%	
PC4	Low degree: 10% - 20%	Raggae, ska	13%	PC4 attempts to capture the following feature the most, in comparison to other principal components: raggae, ska
PC5	Less than moderate degree: 20% - 40%	Animated	32%	PC5 attempts to capture the following feature the most, in comparison to other principal components: animated
PC6	Less than moderate degree: 20% - 40%	Horror	33%	PC6 attempts to capture the following feature the most, in comparison to other principal components: horror
PC7	Less than moderate degree: 20% - 40%	Techno, trance	33%	PC6 attempts to capture the following feature the most, in comparison to other principal components: techno, trance (tied with PC3)
PC8	N/A	N/A	N/A	PC7 does not seem to capture a particular feature any more than other principal components, however it does share some variance with certain features (refer to figure 1.8)
PC9	Less than moderate degree: 20% - 40%	Sci-fi	20%	PC6 attempts to capture the following feature the most, in comparison to other principal components: sci-fi

Figure 1.9. An analysis of the shared variance matrix in figure 1.8. The table above identifies the features each principal component captures the most i.e., the principal component of which each feature has the highest shared variance with. Shared variances have also been categorised as *very low degree, low degree*, *low to moderate degree* and *moderate degree*. Note that despite having the highest shared variance, this may fall under very low degree. All 29 features are listed above under a principal component, except for the feature *Techno, trance* as it is equally captured the most by PC3 and PC7. Note that PC8 does not particularly capture any feature 'the most'.

II B. EXPERIMENTATION OF ALTERNATIVE CLUSTERING METHODS

II B. k-MEANS CLUSTERING

The *k-Means Clustering* function was employed on the dataset produced from the PCA (figure 2.1 and 2.2)

	Silhouette scores	
Number of clusters	Variables normalised	Variables not normalised
2	0.081	0.143
3	0.082	0.125
4	0.083	0.124
5	0.083	0.119
6	0.083	0.116
7	0.087	0.111
8	0.086	0.105

Figure 2.1. Silhouette scores were compared among different numbers of clusters in the case of variables being normalised and variables not being normalised. Interestingly, the optimum number of clusters (i.e. the highest silhouette score) for *k-Means* if variables are normalised is seven. This is a stark difference from the optimum number of clusters of two when variables are not normalised. It was decided not to normalise the variables, to ensure that the results from the PCA were being fully reflected in the clustering process. A number of two clusters were chosen, since the silhouette score significantly drops thereafter. Below is a comparison between two and three clusters used.

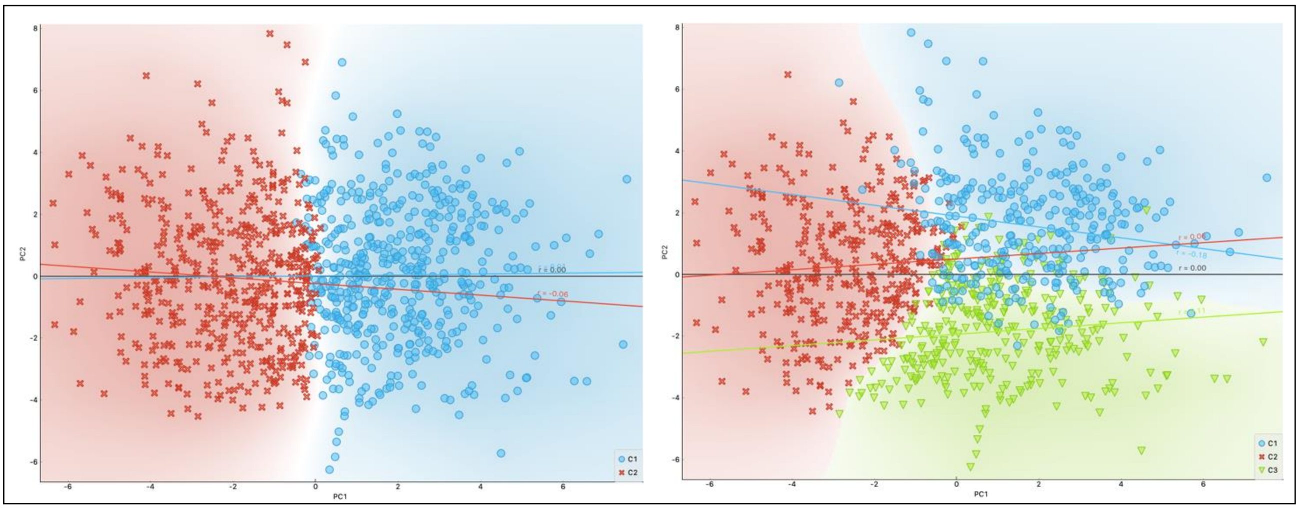


Figure 2.2. (left to right) A scatter plot from *k-Means* against PC1 and PC2 using two clusters and three clusters. From observation, both options provide an equally clear divide between clusters, however, the scatter plot displaying two clusters appears to be tidier to the viewer.

II B. HIERARCHICAL CLUSTERING

The *Hierarchical Clustering* function was utilised on the dataset produced from the PCA (figure 2.3 and 2.4).

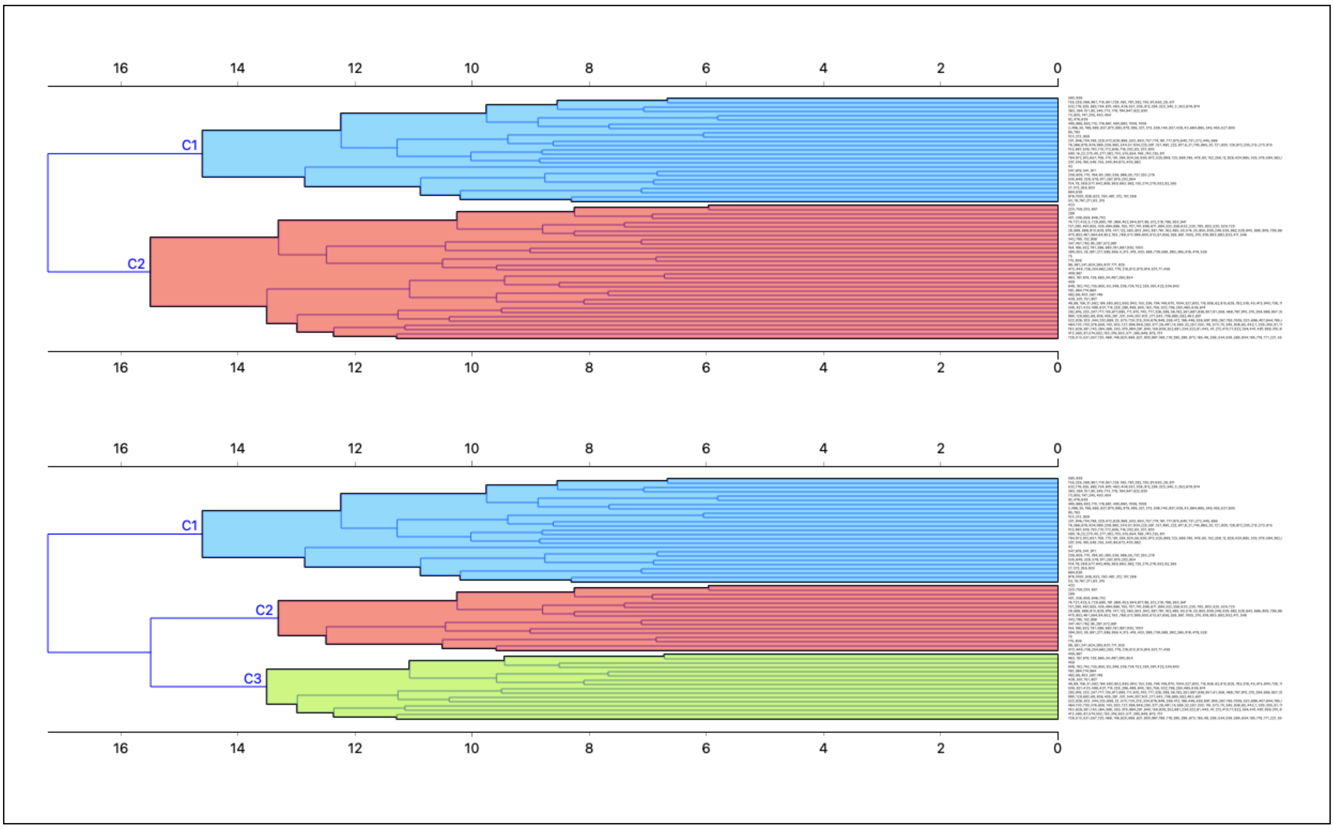


Figure 2.3. (top to bottom) Output from *Hierarchical Clustering* with two clusters and three clusters. Both options show clusters of almost equal quantity and as such, the decision to use two clusters or three clusters are reasonable given the data.

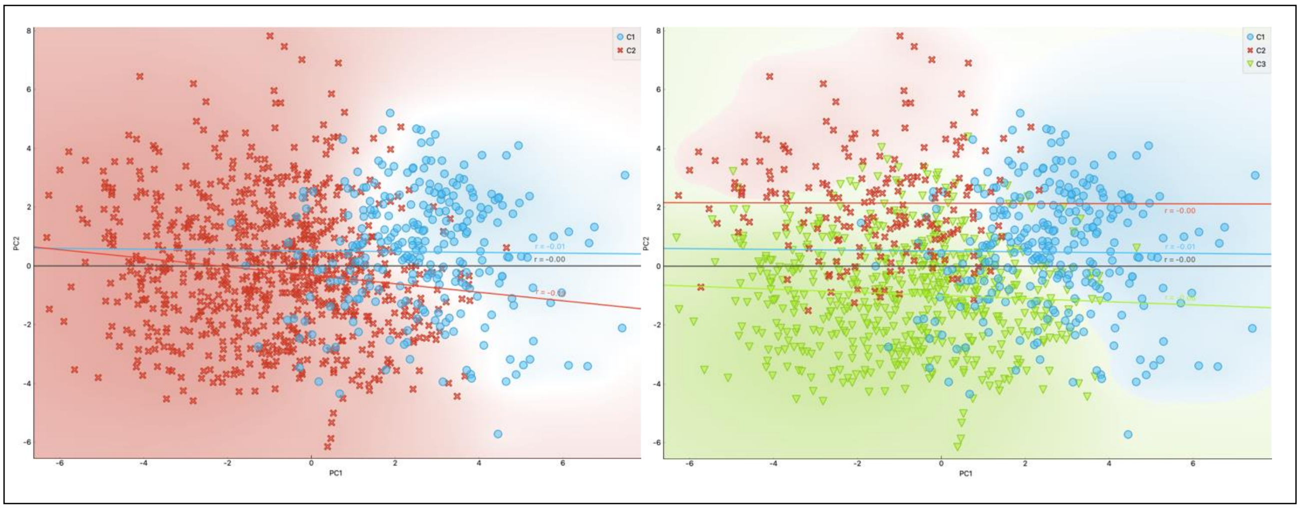


Figure 2.4. (left to right) A scatter plot from *Hierarchical Clustering* against PC1 and PC2 using two clusters and three clusters. The Euclidean distance metric was chosen, particularly due to the data being fully numerical. It was also decided not to normalise the distances, to ensure that the PCA results were fully reflected and to remain consistent with the decision to not normalise in k-Means for effective comparison. Due to the sizeable nature of the dataset, it was decided to set the number of clusters to two, as any greater number of clusters only resulted in its visualisation being cluttered and disarrayed.

III. RESULTS

III A. COMPARING ALTERNATIVE CLUSTERING METHODS

The solutions from *k-MEANS Clustering* and *Hierarchical Clustering* were compared keeping the number of variables constant at two (figure 2.5 and 2.6).

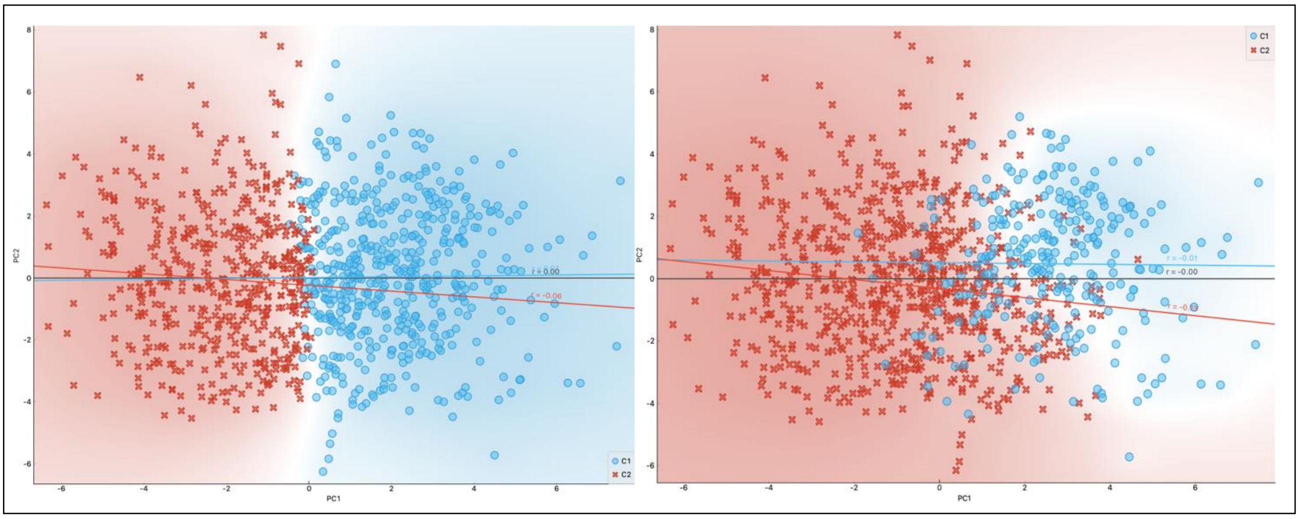
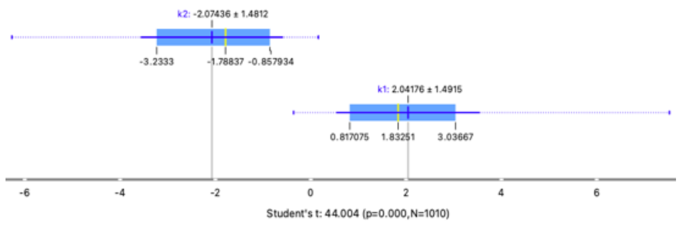
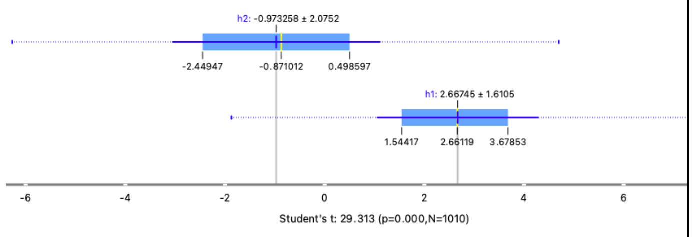
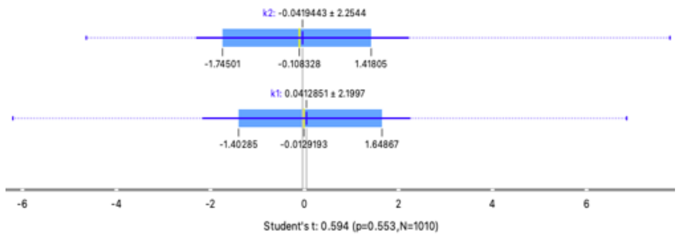
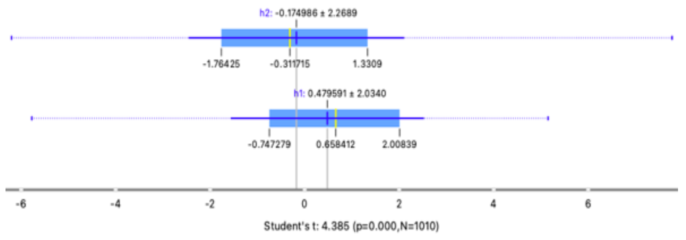
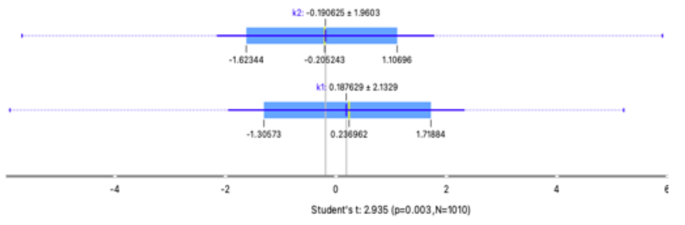
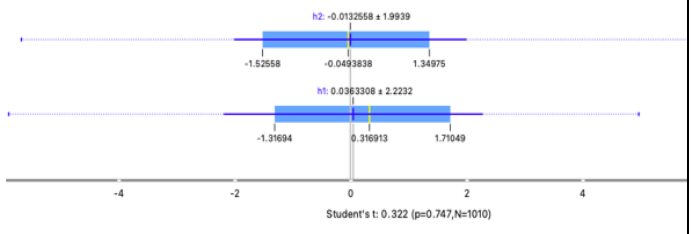
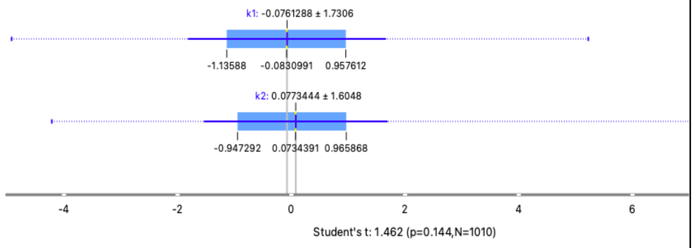
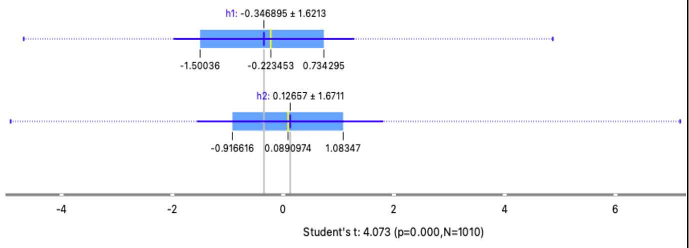
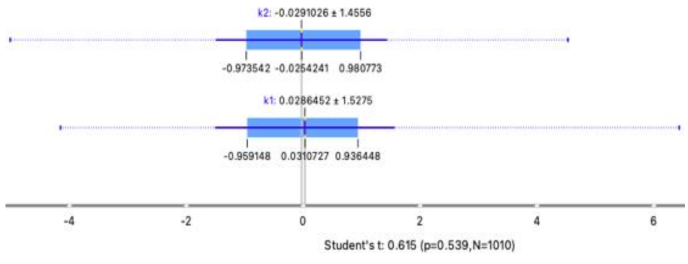
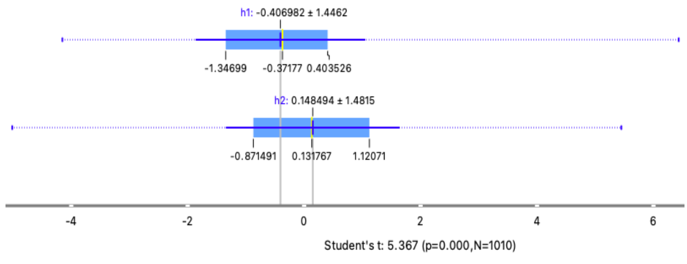


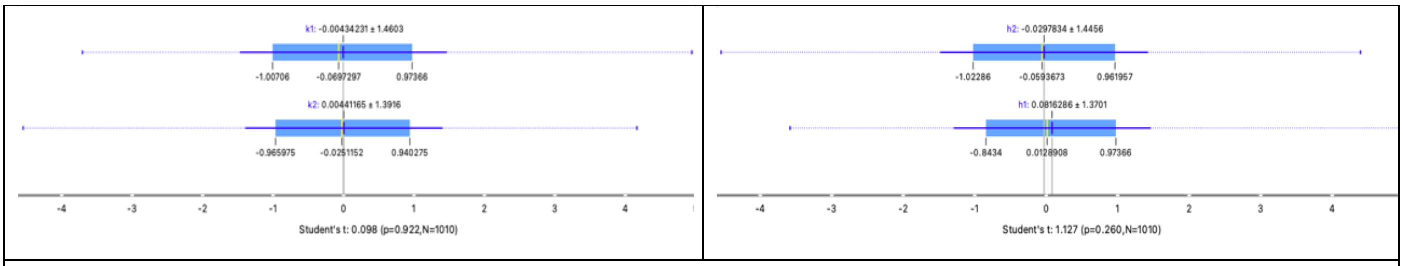
Figure 2.5. (left to right) a comparison of the *k-Means* scatter plot against the *Hierarchical Clustering* scatter plot using two clusters. From the figure, it is obvious that the *k-Means* method appears to do a better job at clustering, as there is a clear distinction between the two clusters with not much overlap in data. The method also reveals a gradient of 0.01 and -0.06 for cluster one and cluster two respectively. *Hierarchical Clustering* on the other hand has a lot of overlapping points and it is far more difficult to discern the clusters. It can also be seen that the data is not evenly divided between clusters The method reveals a gradient of -0.01 and -0.13 for cluster one and cluster two respectively, differing from those in *k-Means*.

Count		**Hierarchical Clustering**		
		Cluster 1	Cluster 2	**Total**
k-Means Clustering	Cluster 1	257	252	**509** (50.40%)
	Cluster 2	13	488	**501** (49.60%)
	Total	**270** (26.73%)	**740** (73.27%)	**1010** (100%)

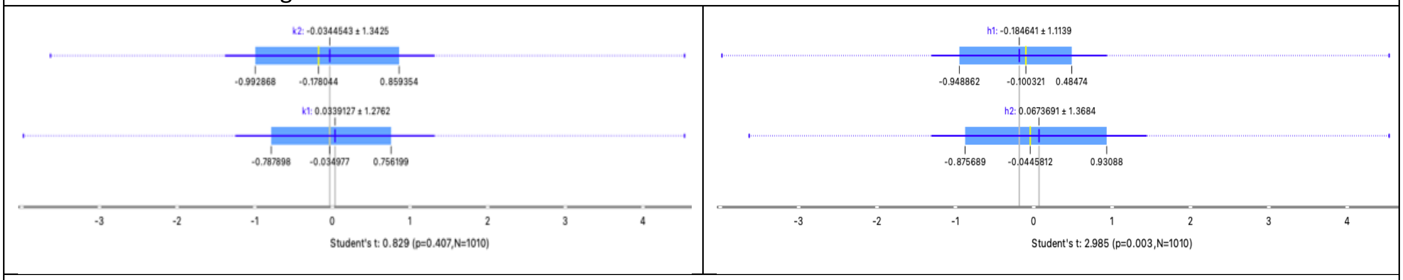
Figure 2.6. A pivot table exhibiting how much of the data is captured and shared in each cluster for both *k-Means* and *Hierarchical Clustering*. The *k-Means* method shows an even divide with cluster 1 representing 50.40% of the overall data and cluster 2, representing 49.60%. *Hierarchical Clustering,* however, appears to be uneven with cluster 1 only comprising of 26.73% of the overall data and cluster 2 comprising of 73.27%. Despite this, we can see that for cluster 1, both *k-Means* and *Hierarchical* share 257 individuals and for cluster 2, share 488 individuals.

Box plots were assessed among *k-Means* and *Hierarchical Clustering* for each principal component, to compare the means, median and interquartile range.

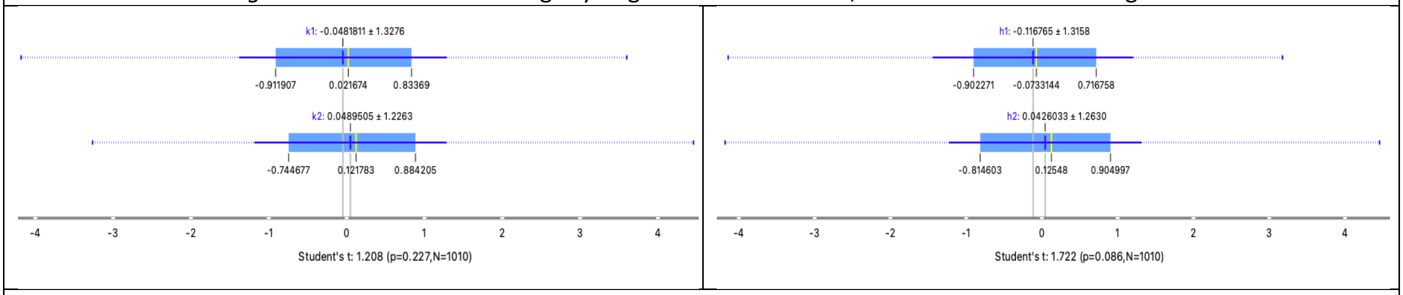
<table>
<tr><th>k-Means Clustering</th><th>Hierarchical Clustering</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2"><b>Box Plot PC1</b>: both <i>k-Means</i> and <i>Hierarchical Clustering</i> result in means that are far apart for PC1, which is reflected in the statistically significant test statistic. However, <i>Hierarchical Clustering</i> has a lower test statistic and overestimates the means slightly (2.66 and -0.87 for cluster one and two respectively, compared to 1.83 and -1.79). It also reports a wider range for cluster two compared to <i>k-Means</i>.</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2"><b>Box Plot PC2</b>: <i>k-Means Clustering</i> results in a mean among clusters almost identical to each other, while <i>Hierarchical Clustering</i> results in means that are slightly far apart for PC2, which is reflected in a higher test statistic. <i>Hierarchical Clustering</i> also reports a slighter higher mean for cluster one compared to <i>k-Means</i> (0.66 compared to -0.01).</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2"><b>Box Plot PC3</b>: <i>Hierarchical Clustering</i> results in a mean among clusters almost identical to each other at 0.32 for cluster one and -0.05 for cluster two. While <i>k-Means Clustering</i> has a slighter larger variation between means 0.24 for cluster one and -0.21 for cluster two, which is reflected in a higher test statistic and lower p-value.</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2"><b>Box Plot PC4</b>: <i>k-Means Clustering</i> results in a mean among clusters almost identical to each other (about 0), while <i>Hierarchical Clustering</i> results in means that are slightly more far apart for PC4, which is reflected in a higher test statistic.</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2"><b>Box Plot PC5</b>: <i>k-Means Clustering</i> results in a mean among clusters almost identical to each other (about 0), while <i>Hierarchical Clustering</i> has means that have a slightly larger variation for PC5, which is reflected in a higher test statistic.</td></tr>
</table>



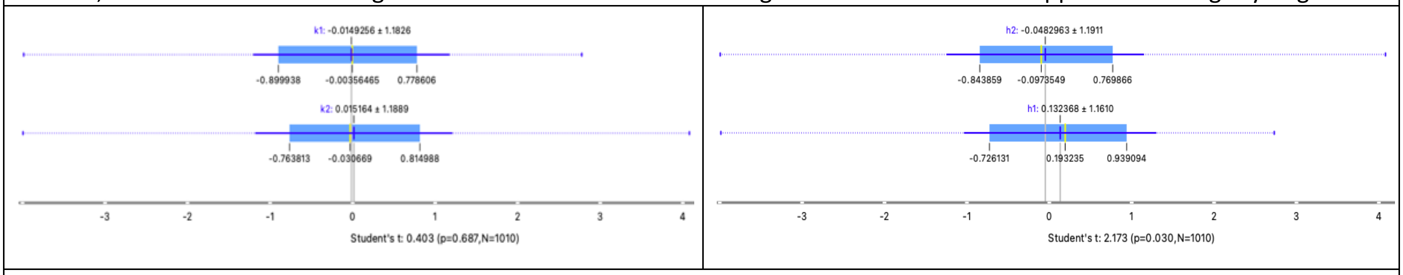
Box Plot PC6: Both methods yield identical results. However, *k-Means Clustering* results in a mean among clusters almost identical to each other (about 0), while *Hierarchical Clustering* has means that have a ever so slightly larger variation for PC6, which is reflected in a higher test statistic.



Box Plot PC7: *k-Means Clustering* results in a mean among clusters almost identical to each other (about 0), while *Hierarchical Clustering* has means that have a slightly larger variation for PC7, which is reflected in a higher test statistic.

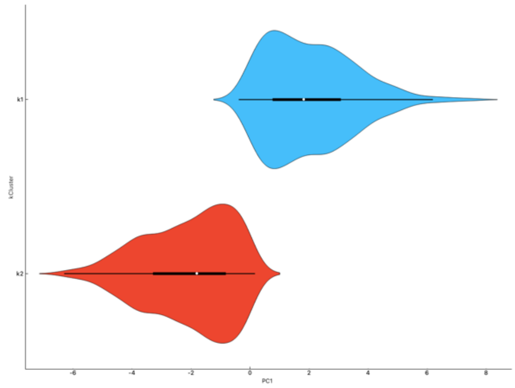
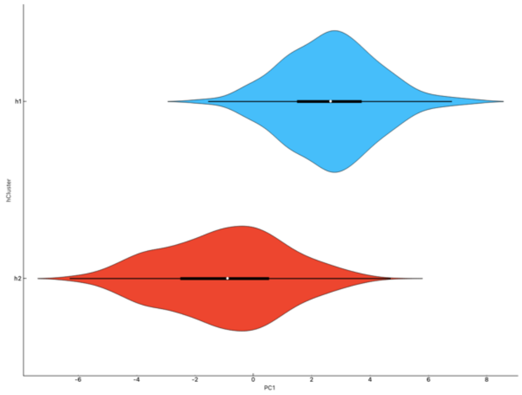
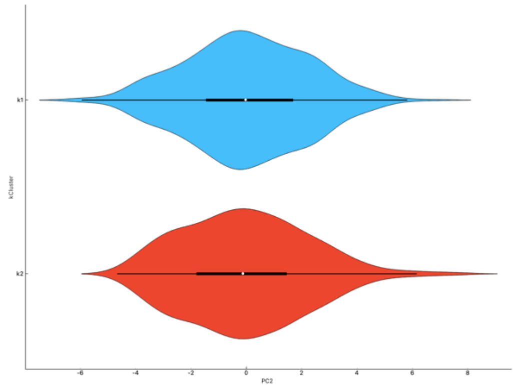
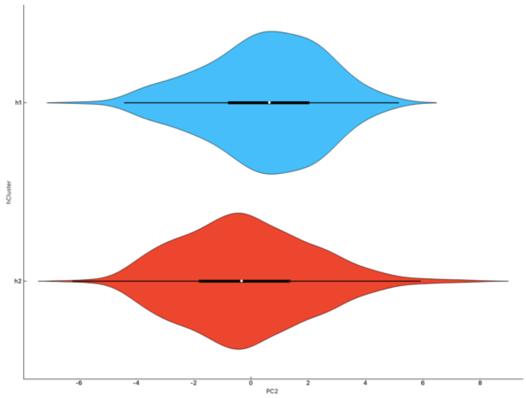
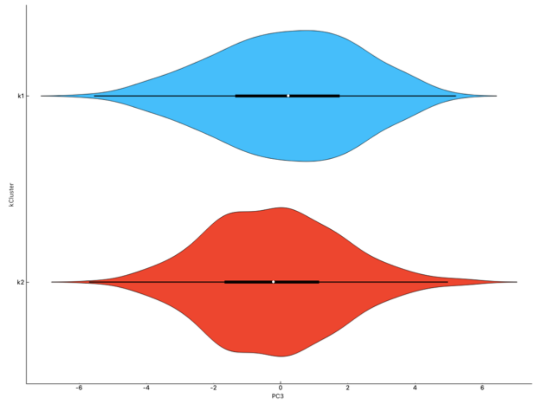
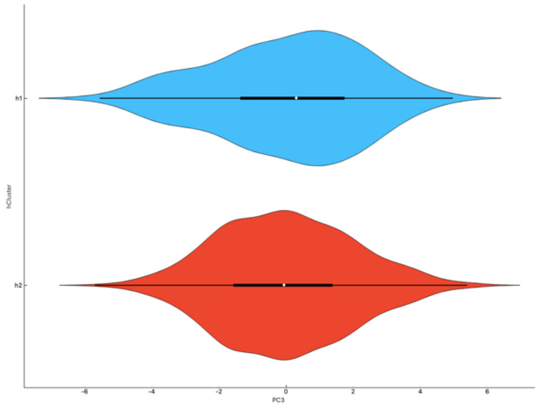


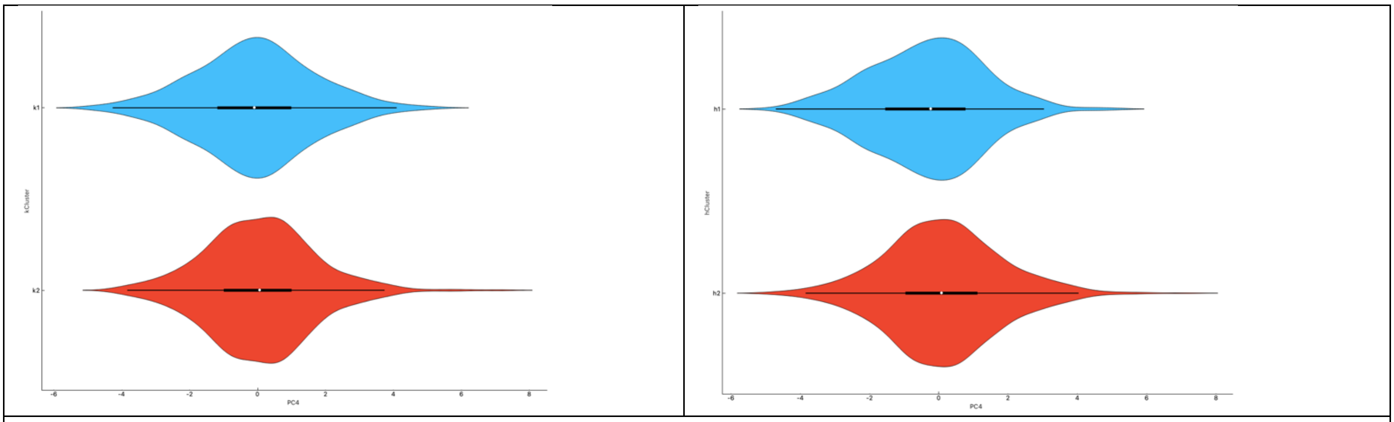
Box Plot PC8: Both methods yield almost identical results. However, *k-Means Clustering* results in a mean among clusters almost identical to each other (about 0), while *Hierarchical Clustering* has means that have a ever so slightly larger variation for PC8, which is reflected in a higher test statistic. The *k-Means* range for cluster one in also appears to be slightly larger.



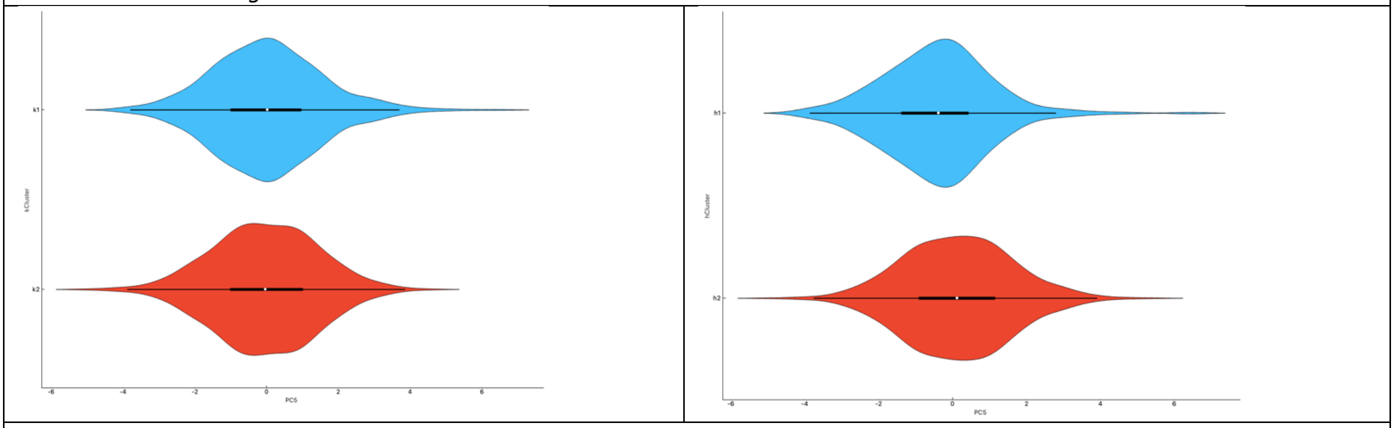
Box Plot PC9: Both methods yield almost identical results. However, *k-Means Clustering* results in a mean among clusters almost identical to each other (about 0), while *Hierarchical Clustering* has means that have a ever so slightly larger variation for PC8, which is reflected in a higher test statistic.

Violin plots were also considered among *k-Means* and *Hierarchical Clustering* for each principal component to compare the distribution of the clusters.

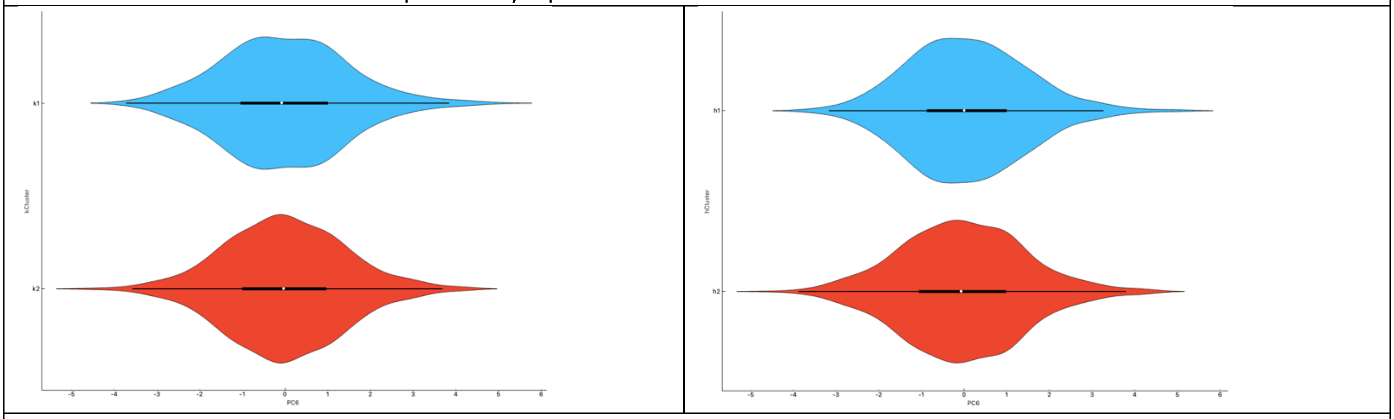
<table>
<tr><th>k-Means Clustering</th><th>Hierarchical Clustering</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2"><b>Violin Plot PC1</b>: Both <i>k-Means</i> and <i>Hierarchical Clustering</i> provide vastly different results in their violin plot of principal component one. For cluster one, <i>k-Means</i> shows that there is a very high probability that the individuals take on a value of approximately 1, while <i>Hierarchical</i> underestimates this probability and exhibits that there is a higher probability that members take on a value of approximately 3. Similarly for cluster two, <i>k-Means</i> identifies a shorter probability distribution, mostly centred at approximately -1. While <i>Hierarchical Clustering</i> is also mostly centred at -1, this distribution is longer.</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2"><b>Violin Plot PC2</b>: Both methods are almost identical in plotting clusters in principal component two. However, <i>k-Means</i> seems to be slightly more specific in showing multipal peaks. For instance, in cluster one the data is heavily centred just under 0 and there is another peak at approximately 2. In cluster two, the data is also mostly concentrated around 0, however there is also another peak of a high probability of individuals taking on the value of -3. While this is also represented in <i>Hierarchical Clustering</i>, <i>k-Means</i> seems to be more effective in exhibiting clearer peaks and troughs in its plot.</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2"><b>Violin Plot PC3</b>: Again, for principal component three, <i>k-Means</i> and <i>Hierarchical Clustering</i> show very similar results. However, <i>k-Means</i> appears to be ‘smoother’ than its <i>Hierarchical</i> counterpart. For intance, <i>k-Means</i> shows only one peak in cluster one and two peaks in cluster two. While this is also reflected in <i>Hierarchical Clustering</i>, smaller peaks can be seen more clearly. As such, Hierarchical clustering appears to be more effective in plotting principal component three.</td></tr>
</table>



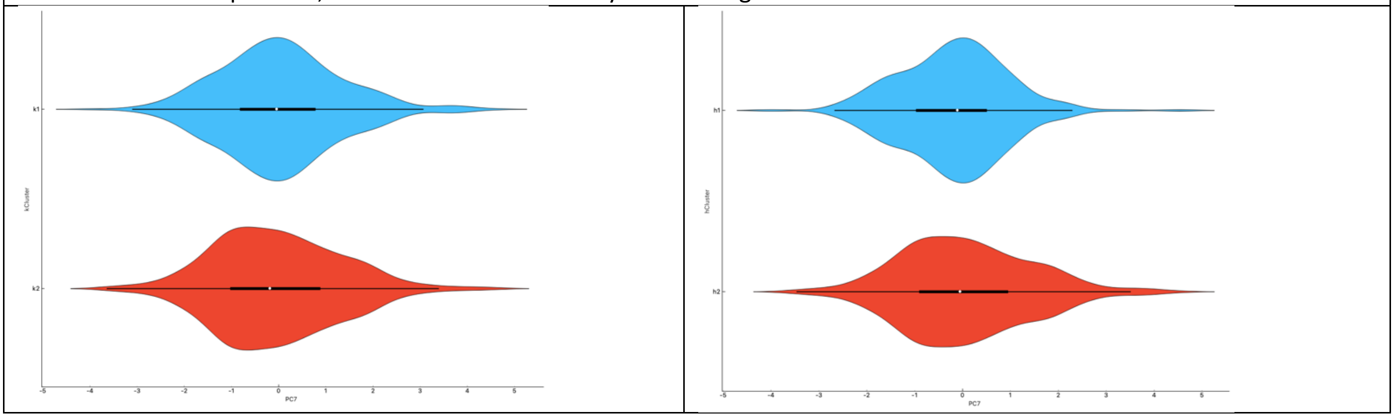
Violin Plot PC4: Both methods are almost identical in plotting clusters in principal component four. *k-Means* seems to be slightly more specific in showing multiple peaks for cluster two particularly in showing troughs. *Hierarchical* however appears more effective for cluster one. It is also interesting to note that the peak shown for cluster one appears slightly fatter in *Hierarchical Clustering*.



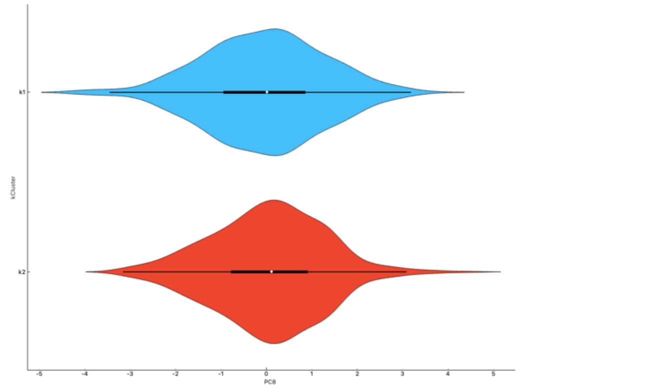
Violin Plot PC5: Although both methods are quite similar in its distribution for principal component five, *k-Means* again appears to be slightly more specific in showing multiple peaks. For instance, in cluster two the data is heavily centred just under 0 and just above 0, there is a clear trough inbetween. However, in *Hierarchical Clustering*, this is 'smoothed out' and it is difficult to ascertain whether the probability dips inbetween those two values.

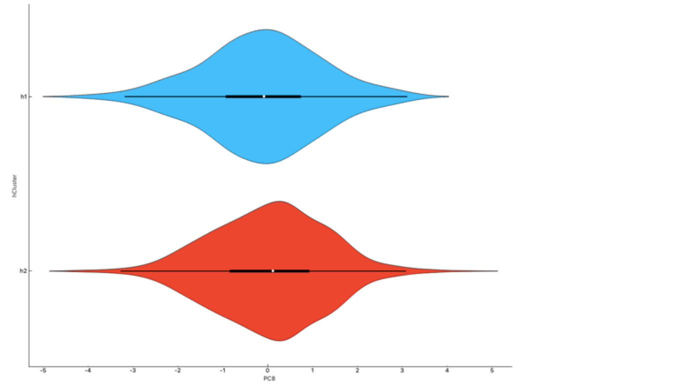


Violin Plot PC6: Both *k-Means* and *Hierarchical Clustering* are almost identical in plotting clusters in principal component six. *k-Means* appears to be slightly more specific in showing multiple peaks for cluster while *Hierarchical* seems more effective for cluster two. Despite this, both methods show a very similar range in distribution.

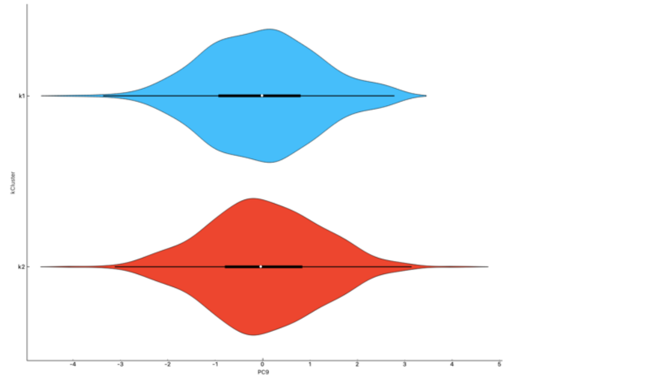


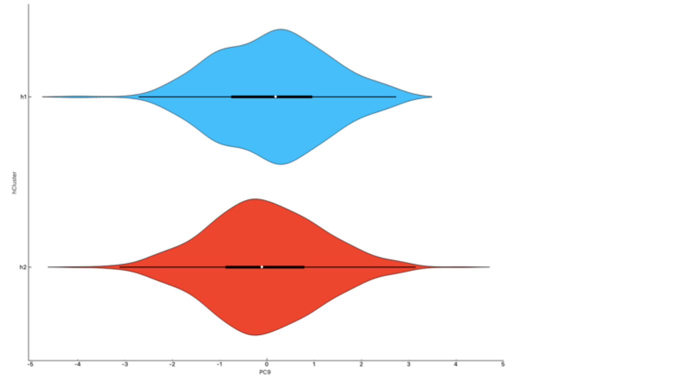
Violin Plot PC7: Here, *Hierarchical Clustering* appears to be more specific in terms of exhibiting certain peaks in probability of individuals assuming a given value. This is particularly evident in values below 0 in cluster one and values above 1 for cluster two, when compared against *k-Means*. *Hierarchical Clustering* also makes it clear that there is a slight dip in distribution for cluster two at the value 3, after which the probability assumed slightly increases. This is something that is not observed in *k-Means*.





Violin Plot PC8: In terms of distribution, both methods provide almost identical results for cluster two, but different results for cluster one. In *k-Means Clustering*, it can be seen that the probability that an individual takes on a value just above 0 is highest; the probability of an individual assuming a value of -1 is also high. This is not particularly distinguishable in *Hierarchical Clustering*, which shows only one distinct peak.





Violin Plot PC9: Both *k-Means* and *Hierarchical Clustering* are almost indistinguishable for cluster two. However, *Hierarchical Clustering* appears to be more specific in displaying peaks in the distribution for principal component nine, i.e. one just above 0, and another at approximately -1. While this is also shown in *k-Means*, it is more apparent in *Hierarchical Clustering*.

III B. PROFILING OF CLUSTERS

The below figures exhibit the specific features chosen in profiling the clusters formed by the *k-Means* method.

III B (i) AGE

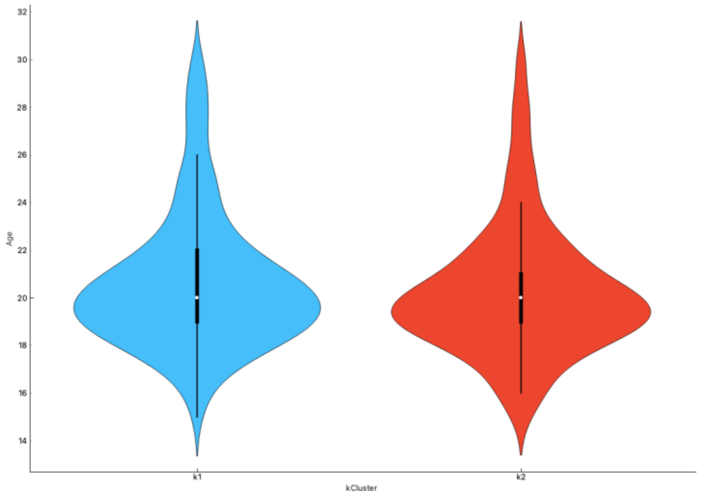


Figure 2.6. (left) The violin plot graphs age against each cluster in *k-means*. From the plot, it appears that both clusters tend to comprise mainly of individuals in their late teens and early twenties. However, individuals in cluster 1 have a higher likelihood of also being in their mid to late twenties. This is signified by the longer upper range and the ‘fatter’ tail in cluster one, showing a higher probability of individuals having the ages of 23 to 31 in comparison to cluster 2.

III B. (ii) HEIGHT

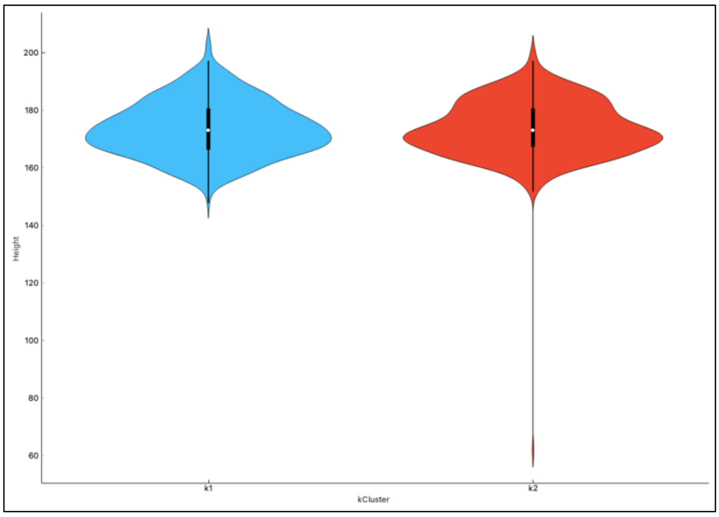


Figure 2.6. (left) The violin plot graphs height against each cluster in *k-means*. From the plot, it seems that generally, both clusters comprise of individuals that are of similar stature. However, cluster 2 tends to also include shorter individuals with a height of approximately 60 cm, as can be seen. While there is a high likelihood that this could be an error or outlier, it is still important to note. There is also a possibility that taller individuals may be represented in cluster 1, as shown by the slightly fatter and longer upper tail at the height of approximately 200m. This is consistent with the previous figure showing cluster 1 including much older individuals. Hence, it would not be surprising that these very individuals could also be taller in height as well.

III B. (iii) WEIGHT

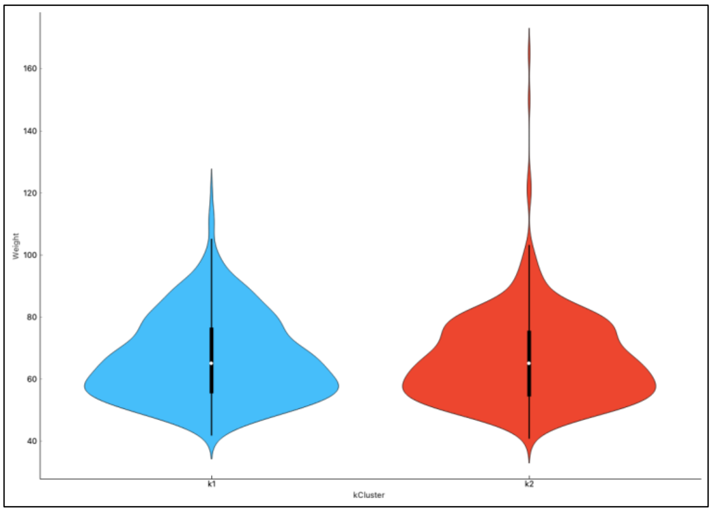


Figure 2.6. (left) The violin plot graphs weight against each cluster in *k-means*. The plot shows both clusters comprising of individuals with similar weights. However, cluster 1 tends to have a higher probability of individuals weighing 80 – 100 kg, as indicated by its fatter tail. Similarly, cluster 2 contains a higher probability of its individuals have a higher weight, particularly from 140 – 170 kg. As such, it could be concluded that cluster 2 may contain individuals of higher weights in comparison to cluster 1.

III B. (iv) NUMBER OF SIBLINGS

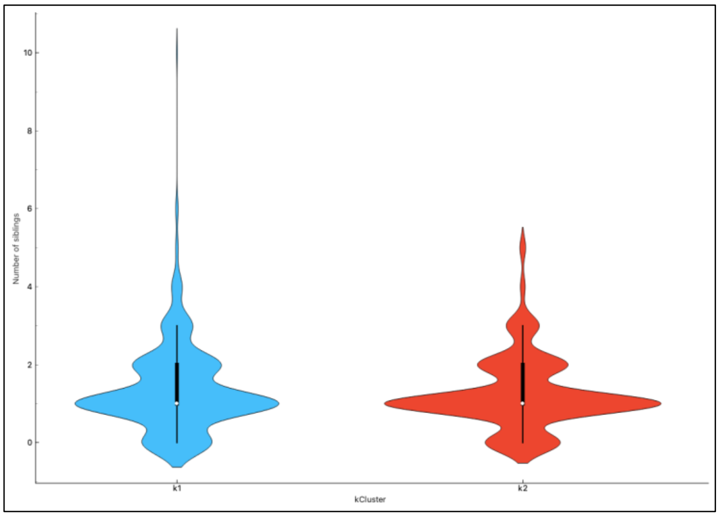


Figure 2.6. (left) The violin plot graphs the number of siblings an individual has against each cluster in *k-means*. The plot shows both clusters tend to comprise of individuals with a similar number of siblings, shown by the shared mean of one and a similar range of data. However, cluster 1 has a higher probability of individuals having at least six or more siblings compared to cluster 2, as shown by the longer upper tail which extends to approximately eleven. Cluster two also shows a wider body at one compared to cluster 1, indicating that cluster two may mainly comprise of individuals having only one sibling.

III B. (v) GENDER

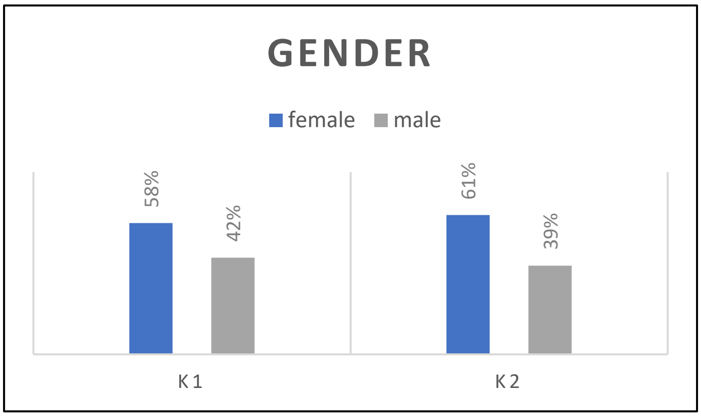


Figure 2.6. (left) The column graph plots an individual's gender against each cluster in *k-means*. Both clusters mainly comprise of females with 58% of cluster 1 being female and 61% in cluster 2 being female. This is likely due to the original data having bias, with the majority of those surveyed being female (599 females in a total of 1,010). Despite this, it can be argued that cluster 1 represents a slighter higher male population compared to cluster 2 (42% compared to 39%). Similarly, it can also be said that cluster 2 represents a slightly higher female population in comparison to cluster 1 (61% compared to 58%).

III B. (vi) HANDEDNESS

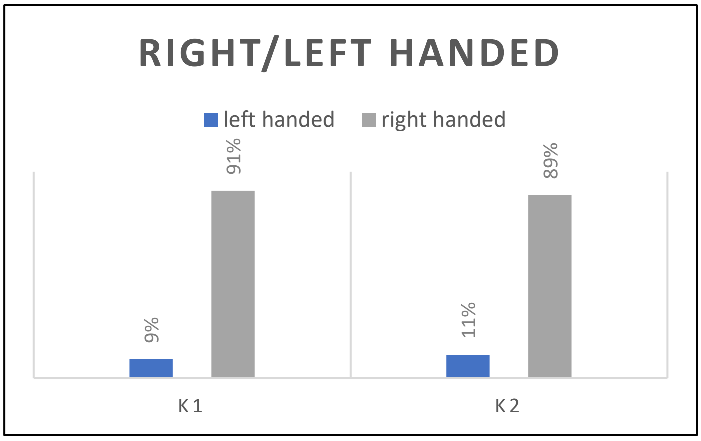


Figure 2.6. (left) The column graph plots whether an individual is right handed or left handed against each cluster in *k-means*. Both clusters mainly comprise of individuals that are right handed, with 91% of cluster 1 and 89% in cluster 2 being right handed. This is primarily due to the original data being skewed towards individuals being right handed, with the majority of those surveyed being female (906 right-handed individuals in a total of 1,010). Despite this, it can be argued that cluster 1 represents a slighter higher population of right-handed individuals compared to cluster 2 (91% compared to 89%). Similarly, it can also be said that cluster 2 represents a slightly higher population of left handed individuals in comparison to cluster 1 (11% compared to 9%). Regardless, it can also be concluded that an individual being either right handed or left handed have very little bearing on their music and movie preferences.

III B. (vii) EDUCATION

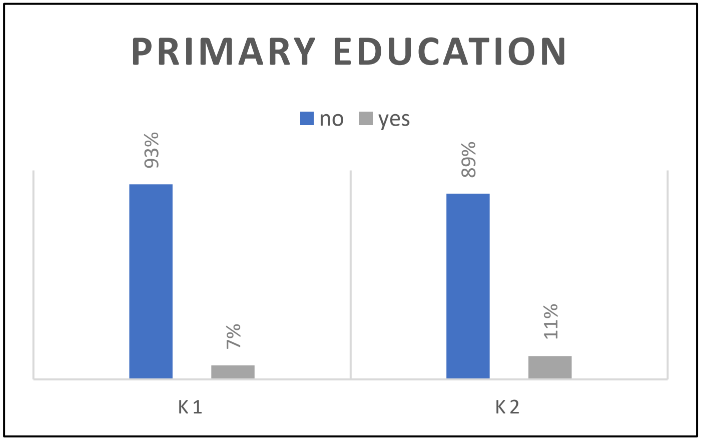


Figure 2.6. (left) The column graph plots whether the highest level of education an individual has received is primary-level education against each cluster in *k-means*. Both clusters mainly comprise of individuals not having primary education listed as their highest level of education, likely due to the overall data having fewer individuals having primary education as their highest level (90 individuals in a total of 1,010). It can be argued that cluster 2 comprises of a higher population having achieved primary education as their highest level of education compared to cluster 1 (11% compared to 7%).

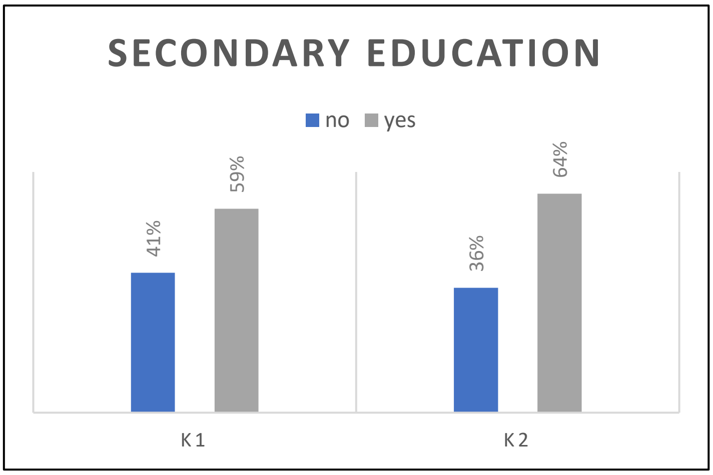


Figure 2.6. (left) The column graph plots whether the highest level of education an individual has received is secondary-level education against each cluster in *k-means*. Both clusters mainly comprise of individuals having secondary education being their highest level of education with cluster 1 having 59% of these individuals and cluster 2, 64%. This is consistent with other figures, as it is known that most of the individuals surveyed are in their early twenties. Thus, these very individuals may be in the process of obtaining a tertiary-level education. Despite this, it may be argued that cluster 1 represents a greater population of those that do not have secondary education as their highest level of education (41% compared to 36%).

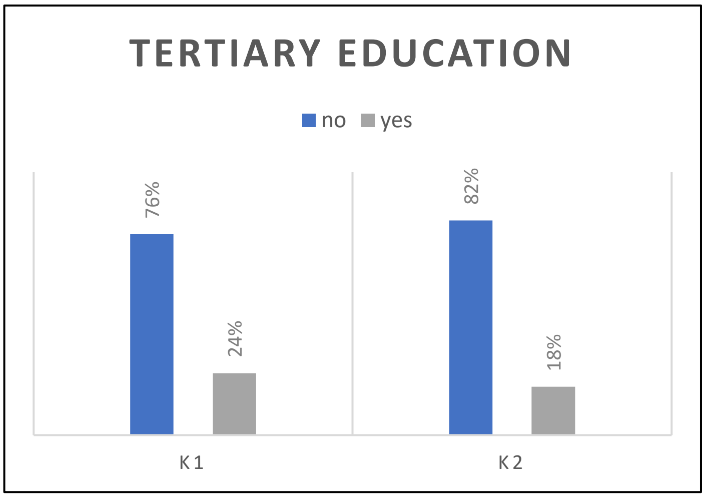


Figure 2.6. (left) The column graph plots whether the highest level of education an individual has received is tertiary-level education against each cluster in *k-means*. Both clusters mainly comprise of individuals having tertiary education not being their highest level of education with cluster 1 having 76% of these individuals and cluster 2, 82%. It could be concluded that cluster 1 represents a greater population of those that have tertiary education as their highest level of education in comparison to cluster 2 (24% compared to 18%).

III B. (viii) PLACE OF RESIDENCE

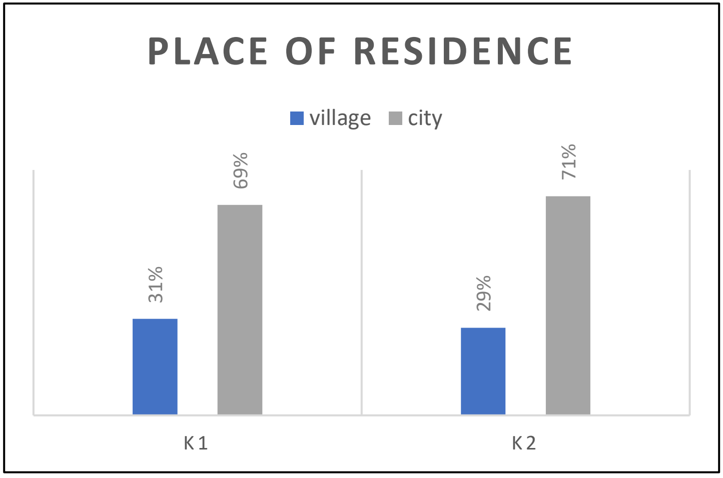


Figure 2.6. (left) The column graph plots whether an individual lives in a city or village against each cluster in *k-means*. Both clusters mainly comprise of individuals residing in a city, with cluster 1 having 69% of these individuals and cluster 2, 71%. This is likely owing to the original data being biased towards those living in a city, as the majority of those surveyed came from cities (707 individuals in a total of 1,010). It could hence be argued that cluster 1 represents a greater population of those residing in a village in comparison to cluster 2 (31% compared to 29%). While cluster 1 represents a greater population of individuals living in a city.

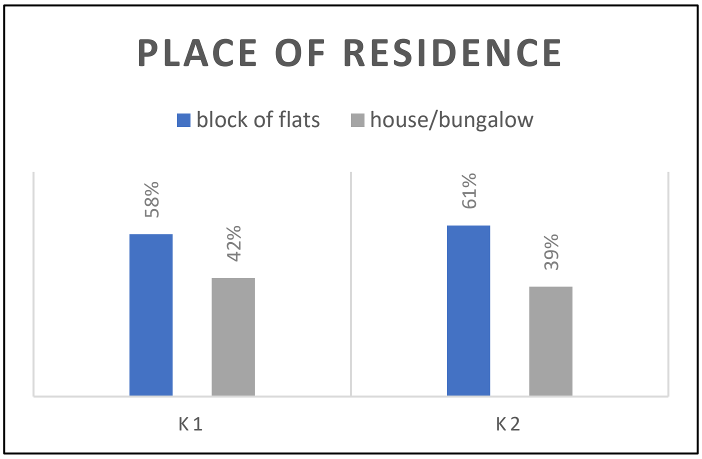


Figure 2.6. (left) The column graph plots whether an individual lives in a house/bungalow or in a block of flats against each cluster in *k-means*. Both clusters mainly comprise of individuals residing in a block of flats, with cluster 1 having 58% of these individuals and cluster 2, 61%. This is likely owing to the original data being biased towards those living in a block of flats, as the majority of those surveyed reside in a flat (599 individuals in a total of 1,010). It could hence be argued that cluster 1 represents a greater population of those residing in a house/bungalow in comparison to cluster 2 (42% compared to 39%). While cluster 2 represents a greater population of individuals living in a city.