Is there a consistent correlation between preference of certain movie genres?

Vinicius Rech Verza

***Abstract* — The purpose of this report is to analyze if there are significant correlations between the preference of certain movie genres considering sex. The approaches were graphical visualizations and Principal Component Analysis performed in a data-set from The Faculty of Social and Economic Sciences at Comenius University in Bratislava containing results from a survey. This paper contains information on how the data was handled, cleaned, tested, methodologies and conclusion.**

***Index Terms* — Survey, Movie Genres, Factorial Analysis, PCA, Principal Component Analysis.**

I. Introduction

Movies, a type of entertainment made possible after the invention of the motion camera machine in 1884 by Auguste and Louis Lumière [1]. This creation allowed the brothers to shot a scene from the back of a moving train in 1896. From this historical mark onwards the movie industry was born and, with that, several genres appeared in the films scene, from drama and horror to sci-fi and documentaries, bringing a world of possibilities for the spectator [2]. This paper has the goal of analyzing if there are correlations with certain movie genres and if this is somehow connected to preference by sex. The analysis was made possible by a data-set found at Kaggle website named 'Young People Survey' [3]. This study is about a survey made by students from The Faculty of Social and Economic Sciences at Comenius University in Bratislava in which people who participated in the survey answered their preferences in various themes, such as music, movies, hobbies and interests, phobias, health habits, personality traits, views on life and spending habits. The study also collected some demographic data such as age, height, weight, siblings, gender, education level and more. The data-set consisted of 150 attributes and 1010 observations. In order to make this analysis possible, the tools used were Microsoft Excel, SPSS, JASP and Tableau. Following to the next step now you will find detailed information about the preparation of the data-set as well as first impressions, review, cleaning and tiding.

Manuscript received May 04, 2020, to be revised. The review of this paper is coordinated by Prof. Theo Mendonça.

Vinicius Rech Verza is with the National College of Ireland, Department of Computing Science, Dublin, Ireland (email: viniverza@gmail.com).

II. Background

The techniques used on this report are brief EDA - Exploratory Data Analysis and PCA – Principal Component Analysis.

Exploratory Data Analysis (EDA) is a methodology of analyzing and summarizing main points and characteristics of a data-set usually supported by visual methods such as graphics or plots and summary statistics [4]. It is a process to help the analyst spot problems and anomalies in the data, discover patters, create and verify assumptions and even determine if a hypothesis can be rejected or not [5]. For this project, only a few aspects of EDA were relevant for this data-set. Only initial investigations were performed in order to spot outliers and anomalies and graphical visualizations were created using Tableau and SPSS in order to have a 'visual' idea on what to expect from this data-set.

Principal Component Analysis (PCA) is an arm of Factor Analysis, which is basically a technique for data reduction which takes a large set of variables and mathematically finds a way to reduce/summarize them by components. It is performed by looking for strong intercorrelations into groups within a set of variables [6]. According to Statistics Solutions website, PCA is the most common method of Factor Analysis used by researchers. It starts creating the first factor by extracting the maximum variance. Following that, it removed the first factor's variance and extracts the maximum variance from the second factor and the process goes on like that until the last factor [7].

In order to perform a PCA analysis, a series of events must take place before the actual analysis, which is exploring if the data is suitable for this type of analysis by checking Bartlett's Test of Sphericity and obtained significance of Kaiser-Meyer-Olkin (KMO). Considering the data is suitable for PCA, the next step would be determining how many components the data will have by checking a scree plot and eigenvalue, lastly interpret results, understanding if the data is assumed to be correlated or uncorrelated and deciding if the data-set has to be rotated or not. If so, understanding what types of rotation are compatible for better results [8].

III. Data Preparation

The data-set was already very organized and tidied already but some actions were still needed in order to have it ready for the report. As the survey collected much more information that are not needed for this analysis, all the attributes that are not related to movie preferences have been removed. They were all the music related columns (19), all the hobbies and interests columns (32), all the phobias columns (10), all the health habits columns (3), all the personality traits, views on life and opinions columns (57), all the spending habits columns (7), some of the demographics columns (age, height, weight, siblings number, education level, left of right handed, 'I am the only child', 'I spent most of my childhood in a' and 'I lived most of my childhood in a'). Lastly, the column 'I really enjoy watching movies' from the movies category was also removed as it was not really relevant for this analysis.

After all these steps, the data-set was left with 12 attributes (Gender, Horror, Thriller, Comedy, Romantic, Sci-fi, War, Fantasy/Fairy Tales, Animated, Documentary, Western and Action). Following that, two attributes had been renamed for better handling and interpreting the data. The attribute 'Sci-fi' has been renamed as 'SciFi' and the attribute 'Fantasy/Fairy Tales' has been renamed as 'Fantasy' only.

Lastly, a visual analysis was performed through the whole data-set using filters in Microsoft Excel in order to spot any missing values from the file. A few observations had some attributes missing and were removed in order to avoid interfering with the final results. There were 6 missing values for Gender, 2 for Horror, 1 for Thriller, 3 for Comedy, 3 for Romantic, 2 for SciFi, 2 for War, 1 for Fantasy, 2 for Animated, 8 for Documentary, 3 for Western and 2 for Action, totalizing 35 missing attributes.

After the tiding, the final data-set had 975 observations and 12 attributes.

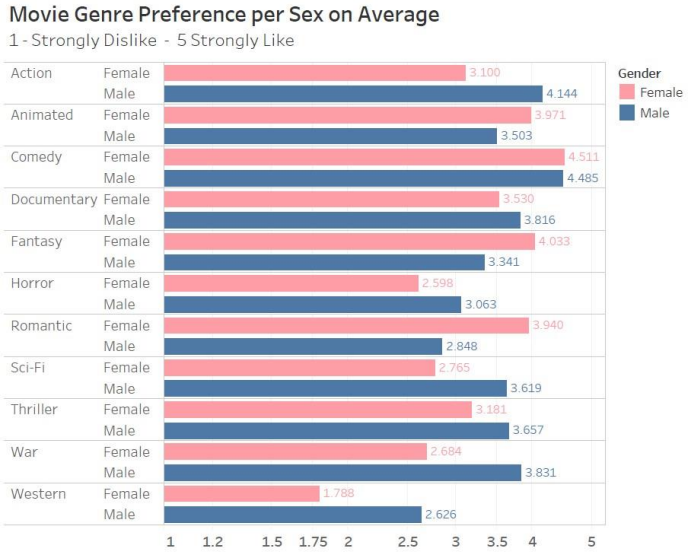
IV. Methodology and Calculations

The first step in this analysis was to observe the data and create a simple visual representation of the results in the data to have a better understanding of what the results from PCA might be. Following that, analyze the data and interpret if the Principal Component Analysis was actually compatible by observing the KMO (Kaiser-Meyer-Olkin) level of significance as well as Bartlett's Test of Sphericity results. Following that, determining the number of components by checking the scree plot, results of eigenvalue to visually compare this data-set to another randomly generated data-set with the same parameters in order to have another evidence on how many components this analysis must have.

Following that, generating a Path Diagram to better understand the relations between the components and a 3D scatterplot of the components with gender classification. Lastly, interpreting the matrixes generated by SPSS such as KMO and Bartlett's Tests, Correlation Matrix, Communalities, Total Variance Explained, Component Matrix, Pattern Matrix, Structure Matrix and Component Correlation Matrix.

V. Analysis and Results

The first graphic, generated in Tableau, was created to have some initial impressions on what to expect from the results, so we can visually compare with the findings from the PCA later on.



Higher resolution image can be found on Appendix A – Movie Genre Preference per Sex on Average.

This graphic shows us the mean results of preferences of movie genres by sex. It is useful to understand easily and compare to what PCA obtained as a method of 'double-checking' for errors or misinterpretations. As we can see, Comedy has almost no preference difference between genres while Action, Documentary, Horror, Sci-fi, Thriller, War and Western are more appealing to men and Animated, Fantasy and Romantic are more appealing to women.

When starting out the Principal Component Analysis, the first step taken was to check if the data-set was suitable for factorial analysis by analyzing the significance level of the Kaiser-Meyer-Olkin Measure of Sampling Adequacy and Bartlett's Test of Sphericity.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.659
Bartlett's Test of Sphericity	Approx. Chi-Square	2040.305
	df	55
	Sig.	0.000

Higher resolution image can be found on Appendix B – KMO and Bartlett's Test Results.

As we can observe above in the test results, the KMO value of 0.659 > 0.600, so that tells us that the level of adequacy of this data-set is relevant for PCA. The level of significance of Bartlett's Test of Sphericity of 0.000 < 0.05 also suggesting that factorial analysis is appropriate for this data.

The next step was to determine how many components are appropriate for the analysis. The PCA run could tell us some preliminary information on that. It was performed in SPSS with Extract of Eigenvalue = 1 so we could have a Total Variance Explained table in order to determine the number of

components and their percentage of cumulative variance.

Total Variance Explained

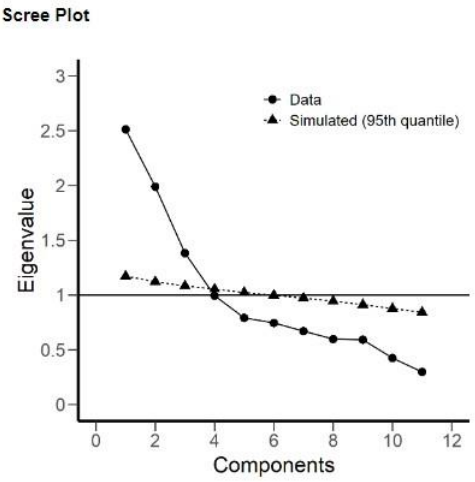
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.513	22.844	22.844	2.513	22.844	22.844	2.216
2	1.990	18.086	40.930	1.990	18.086	40.930	2.065
3	1.383	12.576	53.506	1.383	12.576	53.506	1.758
4	0.994	9.038	62.544				
5	0.791	7.192	69.736				
6	0.746	6.779	76.515				
7	0.670	6.093	82.607				
8	0.598	5.438	88.045				
9	0.592	5.382	93.427				
10	0.425	3.864	97.291				
11	0.298	2.709	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Higher resolution image can be found on Appendix C – Total Variance Explained Table.

As we can observe on the results above, SPSS estimated that this analysis should consider 3 components based on the Eigenvalue greater than 1. According to the table, only the first three components have the Eigenvalue > 1, which the first component explains 22.84% of the data, the second component explains 18.08% and the third explains 12.57%, accounting for 53.50% of total variance explained. This is the first indication that possibly the best number of components was 3.



Higher resolution image can be found on Appendix D – Scree Plot.

After that, a scree plot was generated in JASP, which is an open-source software for statistical analysis [9], for an easier way to literally visualize the eigenvalues compared and to understand the best number of components for the data. As we can observe in the plot, the elbow (which is the drastic change in the line) is situated in the component 4, which tells us that the first 3 components explain the variance better than the remaining. Again, a second indication that three components is the best choice for this analysis.

With the number of components defined as 3, the next step was to analyze the Correlation Matrix to see if there are significant correlation values within some variables to have a better idea on what to expect from the components and its attributes.

Correlation Matrix

		Horror	Thriller	Comedy	Romantic	SciFi	War	Fantasy	Animated	Documentary	Western	Action
Correlation	Horror	1.000	0.503	0.103	-0.127	0.171	0.142	-0.083	0.010	-0.069	0.087	0.132
	Thriller	0.503	1.000	0.001	-0.164	0.238	0.221	-0.086	-0.024	0.044	0.132	0.286
	Comedy	0.103	0.001	1.000	0.283	0.042	-0.073	0.213	0.181	-0.017	-0.032	0.116
	Romantic	-0.127	-0.164	0.283	1.000	-0.102	-0.203	0.355	0.235	-0.097	-0.141	-0.187
	SciFi	0.171	0.238	0.042	-0.102	1.000	0.278	-0.013	0.062	0.134	0.277	0.363
	War	0.142	0.221	-0.073	-0.203	0.278	1.000	-0.067	-0.028	0.234	0.395	0.295
	Fantasy	-0.083	-0.086	0.213	0.355	-0.013	-0.067	1.000	0.682	0.138	-0.029	-0.057
	Animated	0.010	-0.024	0.181	0.235	0.062	-0.028	0.682	1.000	0.144	-0.011	0.015
	Documentary	-0.069	0.044	-0.017	-0.097	0.134	0.234	0.138	0.144	1.000	0.261	0.129
	Western	0.087	0.132	-0.032	-0.141	0.277	0.395	-0.029	-0.011	0.261	1.000	0.320
	Action	0.132	0.286	0.116	-0.187	0.363	0.295	-0.057	0.015	0.129	0.320	1.000

Higher resolution image can be found on Appendix E – Correlation Matrix.

As we can observe on the matrix above, there are some reasonably strong correlations amongst some of the variables, like Horror-Thriller being the highest (0.503), followed by Sci-Fi-Action (0.363) and also some negative correlations, such as War-Romantic (-0.203). This basically means that generally people who like Horror tend to like Thriller, or vice-versa, for example. On the other hand, people who like War tend to dislike Romantic movies, or vice-versa. With this table it is possible to have an idea of relations on which genres are going to be clustered together in the components.

It is also important to highlight that due to the type of results we are interpreting on this analysis (which is finding correlations), the rotation defined was Direct Oblimin, which is one of the few compatible with correlation checks that provided the best results.

Component Correlation Matrix

Component	1	2	3
1	1.000	-0.059	0.138
2	-0.059	1.000	-0.033
3	0.138	-0.033	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Higher resolution image can be found on Appendix F – Component Correlation Matrix.

This shows us the strength of the relationship between our factors. As the correlations are low within all the components, we can expect similar results for either types of rotations (Oblimin or Varimax). If we had values above 0.3, that would indicate that we would have different results within the rotations.

Communalities		
	Initial	ion
Horror	1.000	0.668
Thriller	1.000	0.629
Comedy	1.000	0.366
Romantic	1.000	0.458
SciFi	1.000	0.400
War	1.000	0.500
Fantasy	1.000	0.752
Animated	1.000	0.681
Documentary	1.000	0.476
Western	1.000	0.514
Action	1.000	0.442
Extraction Method: Principal Component Analysis.		

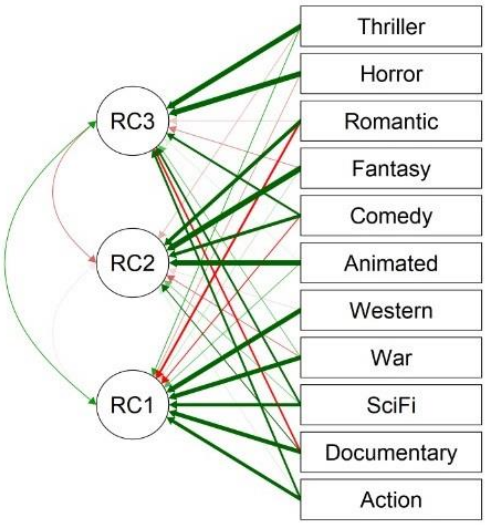
Higher resolution image can be found on Appendix G – Communalities Table.

This table indicates the variance by the factors within the given variables. For achieving best results, the values should be at least 0.4 or greater. In this case there is only one variable that has a lower value, which is 0.366 for Comedy.

Component Matrix[a]			
	Component		
	1	2	3
War	0.643		-0.258
Action	0.623	0.232	
Western	0.586	0.209	-0.355
Thriller	0.586		0.527
SciFi	0.568	0.278	
Romantic	-0.506	0.421	0.157
Fantasy	-0.329	0.801	
Animated	-0.207	0.798	
Comedy		0.463	0.371
Horror	0.446		0.681
Documentary	0.284	0.334	-0.532
Extraction Method: Principal Component Analysis.			
a. 3 components extracted.			

Higher resolution image can be found on Appendix H – Component Matrix.

The Component Matrix above shows which items interrelate with the components, positively or negatively. It contains the estimates of correlations between the components and the items. According to the matrix, the first component has the items War, Action, Western, Thriller, Sci-Fi, Horror and Documentary positively correlated while Romantic, Fantasy and Animated negatively correlated. The second component has Action, Western, Sci-Fi, Romantic, Fantasy, Animated, Comedy and Documentary all positively correlated and the third component has War, Western and Documentary negative correlated while Thriller, Romantic, Comedy and Horror positively correlated. There was no Rotated Component Matrix generated in SPSS even though Oblimin rotation was selected for this analysis.



Higher resolution image can be found on Appendix I – Path Diagram.

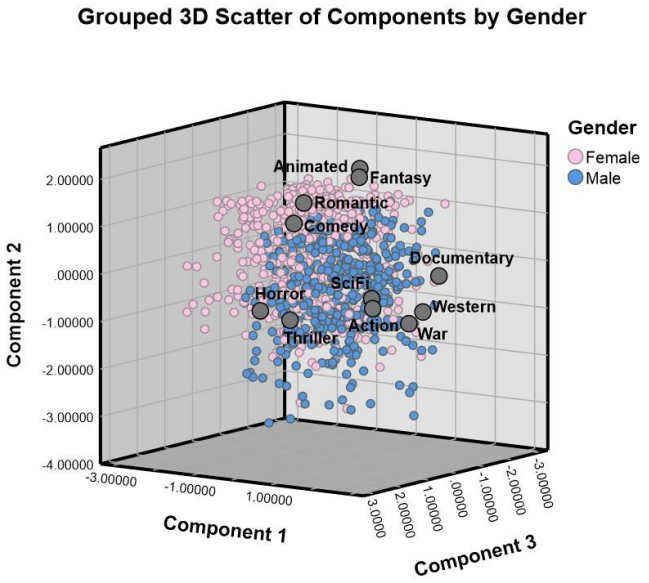
This Path Diagram exemplifies better the findings of the items and their impact on the components. The green lines show the positive correlations while the red lines show the negative correlations. The thickness of the lines represents the impact power of their correlations, the thicker the line, the stronger the correlation within the item and the component. As we can see, Component 1 has strong positive correlations with Western, War, Sci-Fi, Documentary and Action. Component 2, Romantic, Fantasy, Comedy and Animated and Component 3 has Thriller, Horror and Comedy positively correlated. There are some considerably strong negative correlations as well, which are the case of Documentary for PC3, Romantic for PC1 and Comedy for PC1.

Pattern Matrix[a]			
	Component		
	1	2	3
Western	0.715		
War	0.674		
Documentary	0.646		-0.322
Action	0.544		0.314
SciFi	0.523		0.293
Fantasy		0.853	
Animated	0.182	0.814	
Romantic	-0.286	0.588	
Comedy		0.520	0.313
Horror			0.821
Thriller	0.167		0.748
Extraction Method: Principal Component Analysis. Rotation Method: Oblimin with Kaiser Normalization.			
a. Rotation converged in 10 iterations.			

Higher resolution image can be found on Appendix J – Pattern Matrix.

This table shows 'regression-like' coefficients of each variable on the factors. It shows us which items are loading in each component that are above 0.15, which was the set value in SPSS when running the PCA. If we consider a the most significant values within the items (0.30 and above), that would replicate the results on the path diagram previously commented

on this report. Component 1 would be Western, War, Documentary, Action and Sci-Fi. Component 2 Fantasy, Animated, Romantic and Comedy and Component 3 Horror and Thriller.



Higher resolution image can be found on Appendix K – 3D Scatter Plot.

Lastly, this plot visually demonstrates the three components scattered in 3D as well as all the observations found, divided by gender. It clearly shows us that there are some differences in the preferences when you consider their sex. Apparently, females tend to like more Animated, Fantasy Comedy and Romantic movies (PC2), while males tend to relate more to PC1 items, that are Western, War, Sci-Fi, Documentary and Action. The PC3 item preferences are also more predominant in males, which are Horror and Thriller.

VI. Final Considerations

It was fairly clear after the Principal Component Analysis that some genres of movies tend to correlate in the preferences of people. People who like more 'easy-going' movies like Fantasy, for example, tend to like Romantic and Animated movies more and have the tendency to dislike violent movies, which would indicate as a negative correlation. This pattern follows the other way around: People who like violent movies like Action and War tend to not give much importance to Romantic of Fantasy movies. Even though during the PCA, Horror and Thriller movies were combined separately from PC1, the results are similar and the negative correlations between them and the PC2 also exist.

Furthermore, when considering the sex of participants who answered the survey, it is noticeable that women relate more to PC2 while men relate more to PC1 and PC3. This is also backed by the visual analysis done with the plot created with Tableau, in the beginning of this report.

References

[1] Auguste and Louis Lumière [Internet]. Wikipedia. [Cited 2020May04]. Available from: https://en.wikipedia.org/wiki/Auguste_and_Louis_Lumi%C3%A8re

[2] History of Film [Internet]. Wikipedia. [Cited 2020May04]. Available from: https://en.wikipedia.org/wiki/History_of_film

[3] Young People Survey [Internet]. Kaggle. [Cited 2020May04]. Available from: https://www.kaggle.com/miroslavsabo/young-people-survey#responses.csv

[4] Exploratory Data Analysis [Internet]. Wikipedia. [Cited 2020May04]. Available from: https://en.wikipedia.org/wiki/Exploratory_data_analysis

[5] What is Exploratory Data Analysis? [Internet]. Towards Data Science [Cited 2020May04]. Available from: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[6] Factor Analysis [Internet]. Statistics Solutions. [Cited 2020May04]. Available from: https://www.statisticssolutions.com/factor-analysis-sem-factor-analysis/

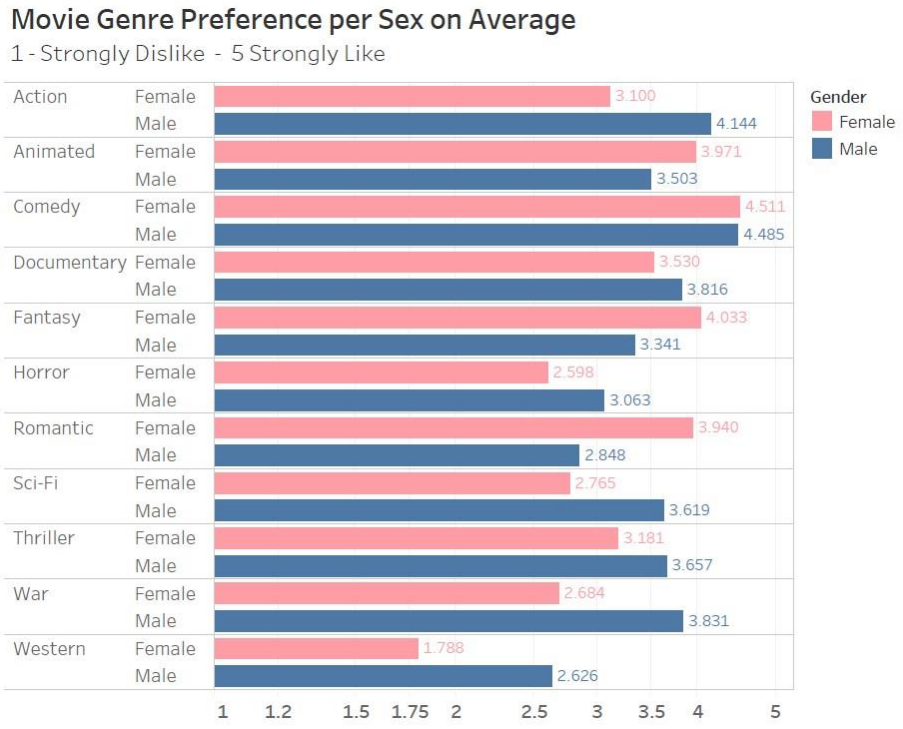
[7] A Practical Introduction to Factor Analysis: Exploratory Factor Analysis [Internet]. UCLA. [Cited 2020May04]. Available from: https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/

[8] Choosing The Right Type of Rotation in PCA and EFA [Internet]. James Dean Brown. [Cited 2020May04]. Available from: http://hosted.jalt.org/test/PDF/Brown31.pdf

[9] JASP Statistical Software [Internet]. JASP. [Cited 2020May04]. Available from: https://jasp-stats.org/download/

APPENDIX

Appendix A – Movie Genre Preference per Sex on Average



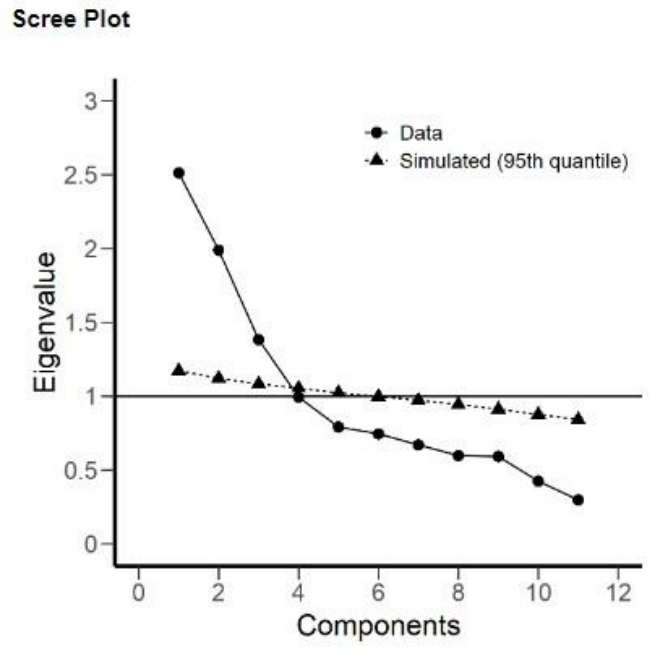
Appendix B – KMO and Bartlett's Test Results

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.659
Bartlett's Test of Sphericity	Approx. Chi-Square	2040.305
	df	55
	Sig.	0.000

Appendix C – Total Variance Explained Table

Total Variance Explained							
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.513	22.844	22.844	2.513	22.844	22.844	2.216
2	1.990	18.086	40.930	1.990	18.086	40.930	2.065
3	1.383	12.576	53.506	1.383	12.576	53.506	1.758
4	0.994	9.038	62.544				
5	0.791	7.192	69.736				
6	0.746	6.779	76.515				
7	0.670	6.093	82.607				
8	0.598	5.438	88.045				
9	0.592	5.382	93.427				
10	0.425	3.864	97.291				
11	0.298	2.709	100.000				
Extraction Method: Principal Component Analysis.							
a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.							

Appendix D – Scree Plot



Appendix E – Correlation Matrix

Correlation Matrix												
		Horror	Thriller	Comedy	Romantic	SciFi	War	Fantasy	Animated	Documentary	Western	Action
Correlation	Horror	1.000	0.503	0.103	-0.127	0.171	0.142	-0.083	0.010	-0.069	0.087	0.132
	Thriller	0.503	1.000	0.001	-0.164	0.238	0.221	-0.086	-0.024	0.044	0.132	0.286
	Comedy	0.103	0.001	1.000	0.283	0.042	-0.073	0.213	0.181	-0.017	-0.032	0.116
	Romantic	-0.127	-0.164	0.283	1.000	-0.102	-0.203	0.355	0.235	-0.097	-0.141	-0.187
	SciFi	0.171	0.238	0.042	-0.102	1.000	0.278	-0.013	0.062	0.134	0.277	0.363
	War	0.142	0.221	-0.073	-0.203	0.278	1.000	-0.067	-0.028	0.234	0.395	0.295
	Fantasy	-0.083	-0.086	0.213	0.355	-0.013	-0.067	1.000	0.682	0.138	-0.029	-0.057
	Animated	0.010	-0.024	0.181	0.235	0.062	-0.028	0.682	1.000	0.144	-0.011	0.015
	Documentary	-0.069	0.044	-0.017	-0.097	0.134	0.234	0.138	0.144	1.000	0.261	0.129
	Western	0.087	0.132	-0.032	-0.141	0.277	0.395	-0.029	-0.011	0.261	1.000	0.320
	Action	0.132	0.286	0.116	-0.187	0.363	0.295	-0.057	0.015	0.129	0.320	1.000

Appendix F – Component Correlation Matrix

Component Correlation Matrix			
Component	1	2	3
1	1.000	-0.059	0.138
2	-0.059	1.000	-0.033
3	0.138	-0.033	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Appendix G – Communalities Table

Communalities

	Initial	ion
Horror	1.000	0.668
Thriller	1.000	0.629
Comedy	1.000	0.366
Romantic	1.000	0.458
SciFi	1.000	0.400
War	1.000	0.500
Fantasy	1.000	0.752
Animated	1.000	0.681
Documentary	1.000	0.476
Western	1.000	0.514
Action	1.000	0.442

Extraction Method: Principal Component Analysis.

Appendix H – Component Matrix

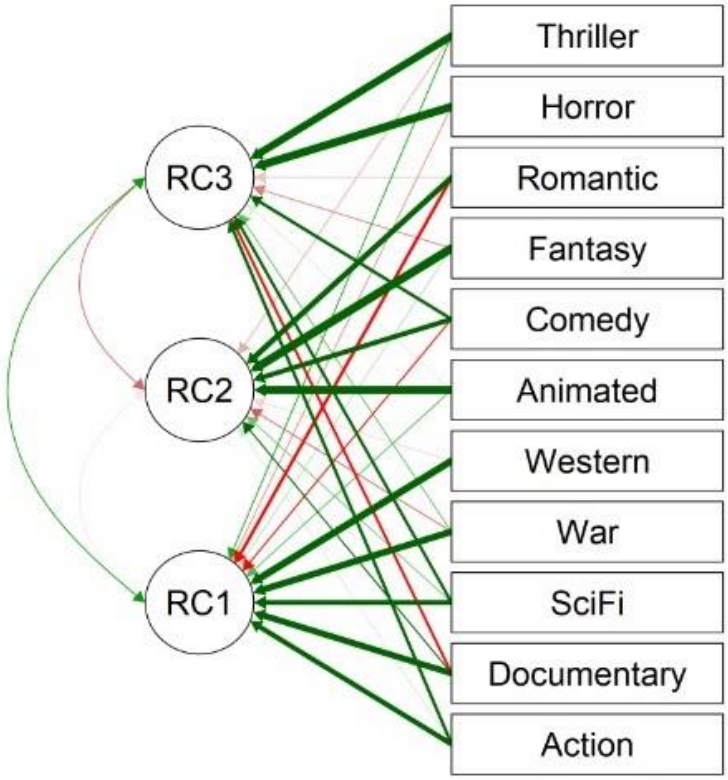
Component Matrix[a]

	Component		
	1	2	3
War	0.643		-0.258
Action	0.623	0.232	
Western	0.586	0.209	-0.355
Thriller	0.586		0.527
SciFi	0.568	0.278	
Romantic	-0.506	0.421	0.157
Fantasy	-0.329	0.801	
Animated	-0.207	0.798	
Comedy		0.463	0.371
Horror	0.446		0.681
Documentary	0.284	0.334	-0.532

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Appendix I – Path Diagram



Appendix J – Pattern Matrix

Pattern Matrix[a]

	Component		
	1	2	3
Western	0.715		
War	0.674		
Documentary	0.646		-0.322
Action	0.544		0.314
SciFi	0.523		0.293
Fantasy		0.853	
Animated	0.182	0.814	
Romantic	-0.286	0.588	
Comedy		0.520	0.313
Horror			0.821
Thriller	0.167		0.748

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 10 iterations.

Appendix K – 3D Scatter Plot