

The Looma Project: Exploration of Features and Story Types

Consultant: Qinghua Li

Client: Elaine McVey, The Looma Project

STOR 765: Statistical Consulting

Dr. Perry Haaland and Dr. Steve Marron

February 23, 2021

Abstract

The Looma Project (TLP) makes films for local small businesses and helps them with their product sales. In this study, the data contains 20 features and 8 story types measured on 70 films made by TLP. We examined the marginal conditional distributions of the features. One-way analysis of variance (ANOVA) was used to explore the association between features and story types. In addition, principal component analysis (PCA) and hierarchical clustering were used to examine possible clusters and their relation with story types. We identified 7 features that are significantly different among the story types using ANOVA. Using PCA analysis, we found that the first principal component (PC1) seems to capture the film quality, and the second PC (PC2) may distinguish highly thematic films versus those visually oriented. Also, 5 clusters are found using hierarchical clustering, and different story types dominate in different clusters.

Introduction

TLP¹ aims to help local small business owners and connect shoppers to the people and stories behind their products. To achieve the goal, TLP makes films that contain different story types for various products and places these films in local grocery stores such as Harris Teeter. In this way, TLP hopes to spread the notions of the goods and the stories behind the products.

There are 20 features describing the films from different aspects and 8 story types. In this study, we want to explore the association between features and story types, and try to find out some important features that can separate story types. In addition, PCA and hierarchical clustering will be used to explore possible clusters and their relation with story types.

Executive summary

The following shows the overview of major findings and indication of where in this report details can be found:

- Marginal conditional distributions show that features with similar means generally have similar distributions. Some features that have the highest means are slightly negative skewed (Results subsection 1).
- Using one-way ANOVA, we identified 7 features that differed significantly among the story types. These features are Ambassadorial Strength, Editorial Resonance, Authenticity of Delivery, Product Appeal, Human Centricity, Technical Quality and Sense of Place (Results subsection 2).
- Overall, the principal components did not align closely with story types. We found that PC1 seems to capture the film quality. The PC2 seems to distinguish between films that are highly thematic versus those that are visually oriented (Results subsection 3).
- Five clusters were identified using hierarchical clustering, and we used “sigclust” to determine which clusters are significantly different from each other (Results subsection 4.1).
- Clusters are not the same as the story types, but different story types dominate in different clusters (Results subsection 4.2).
- One of the differentiators among the clusters appears to be overall film quality (Results subsection 4.3).
- Two clusters are particularly interesting, both of which have high quality films. In particular, one cluster seems to contain films that are more visually oriented, and the other cluster contains films that are thematically oriented (Results subsection 4.3).
- Clusters with different quality films are the most significantly distinct (Results subsection 4.4).
- We used a heatmap that clustered both features and films, which revealed an interesting relationship between features and clusters (Results subsection 4.5).

Data

The data was provided by Elaine McVey, TLP, Durham, NC. The dataset contains 20 features measured on 70 films.

Features used in this analysis included:

- Emotional Strength
- Product-Story Integration
- Authenticity of Delivery
- Sense of Place
- Ambassadorial Strength
- Visual Memorability
- Message Simplicity
- Technical Quality
- Narrative Arc
- Protagonist Memorability
- Editorial Resonance
- Human Centricity
- Product Appeal
- Brand Appeal
- Cohesiveness
- Differentiator Strength
- Visual Narrative
- Product Recognition
- Protagonist Likability
- Music Effectiveness

In this analysis, we used these 20 features that were scored based on scale from -3 to 3. For 29 of the films, the scores are averages from 3 different film viewers. For the remaining 41 films, the features were scored by a single viewer, and the scores are integers from -3 to 3. Each film belongs to one of the 8 story types: Culture/Lifestyle (N = 6), Human/Personal (N = 13), Origin/History (N = 23), Place/Community (N = 4), Process (N = 7), Product (N = 11), Social Impact/Cause (N = 5) and Other (N = 1).

Results

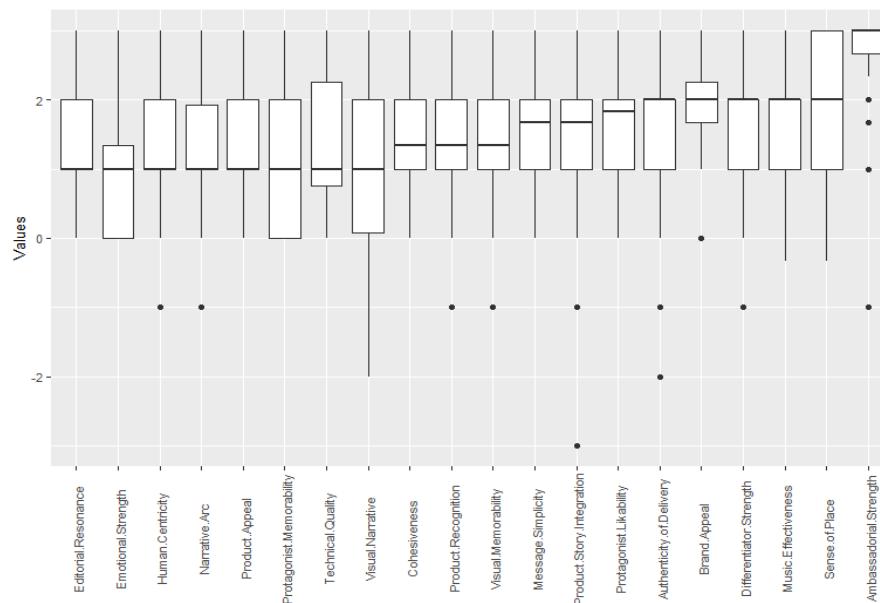
RStudio (1.3.1093) and R Markdown were used for the analysis. All files and outputs are stored on GitHub and are provided to our client for her further use.

1. Marginal conditional distributions

Before we started the analysis, we first examined the properties of the features. Boxplots were used to show the minimum, 1st quartile, median, 3rd quartile and maximum of the features. A boxplot is a good method for graphically depicting the distribution of data based on the quartiles. The 1st quartile is the median of the lower half of the data, and the 3rd quartile is the median of the upper half of the data. InterQuartile Range (IQR) is the distance between the 1st and 3rd quartiles, and it measures variability in a spirit similar to the median. In boxplots, the box is drawn from the 1st and 3rd quartiles. A longer box indicates a larger IQR, so that the middle half of the data is more spread. The horizontal line in the box denotes the median of the data. The minimum and maximum are calculated based on the 1.5 IQR rule, which are the ends of two whiskers (vertical lines outside the box) in the boxplot. In our data, X axis represents the feature names, and Y axis is the feature values.

As shown in Figure 1, features are ordered by their overall median values, and most features have values ranging from -2 to 3. Editorial Resonance has the smallest median and Ambassadorial Strength the biggest median. There are a few outliers, but they do not dominate. The figure shows the data are skewed, for example, look at Ambassadorial Strength. In this figure, medians for Editorial Resonance, Human Centricity, Narrative Arc and Product Appeal are same as their 1st quartile, and medians for Authenticity of Delivery, Differentiator Strength, Music Effectiveness and Ambassadorial Strength are at their 3rd quartile. This seems to be caused by the fact that the data are discrete since there are 41 out of 70 films scored by a single film viewer.

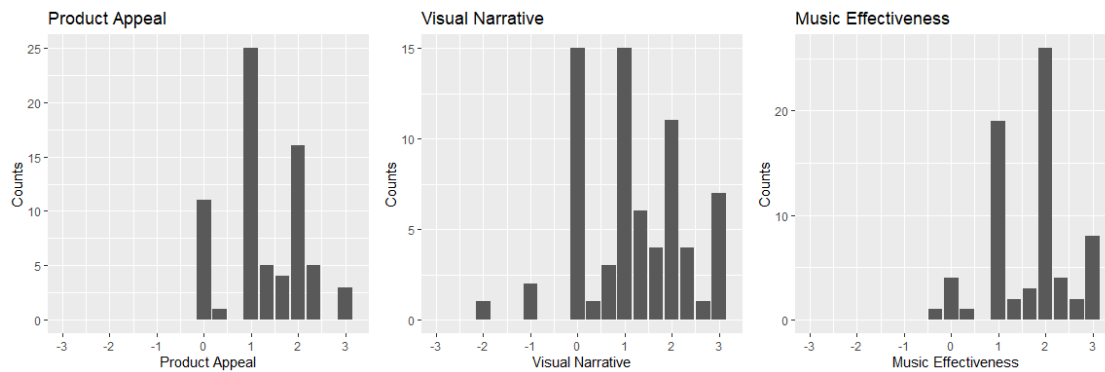
Figure 1: Boxplot of the 20 Numeric Features Ordered by Feature Median



Boxplot for the features. There are a few outliers, but they do not dominate. Some features have their medians at the 1st quartile, and some features have their medians at the 3rd quartile. This may suggest that the data are discrete.

To further explore the distributions, we made bar plots for 3 different features: Product Appeal with median at the 1st quartile, Visual Narrative with median located between the quartiles, and Music Effectiveness with median at the 3rd quartile. As shown in Figure 2, most of the values are integers and the ratings that are averages are equally spaced between the integers. Product Appeal (left most panel) has most values at 1, and Music Effectiveness (right most panel) has most values at 2. This explains why their medians are coincide with the 1st and 3rd quartiles, respectively.

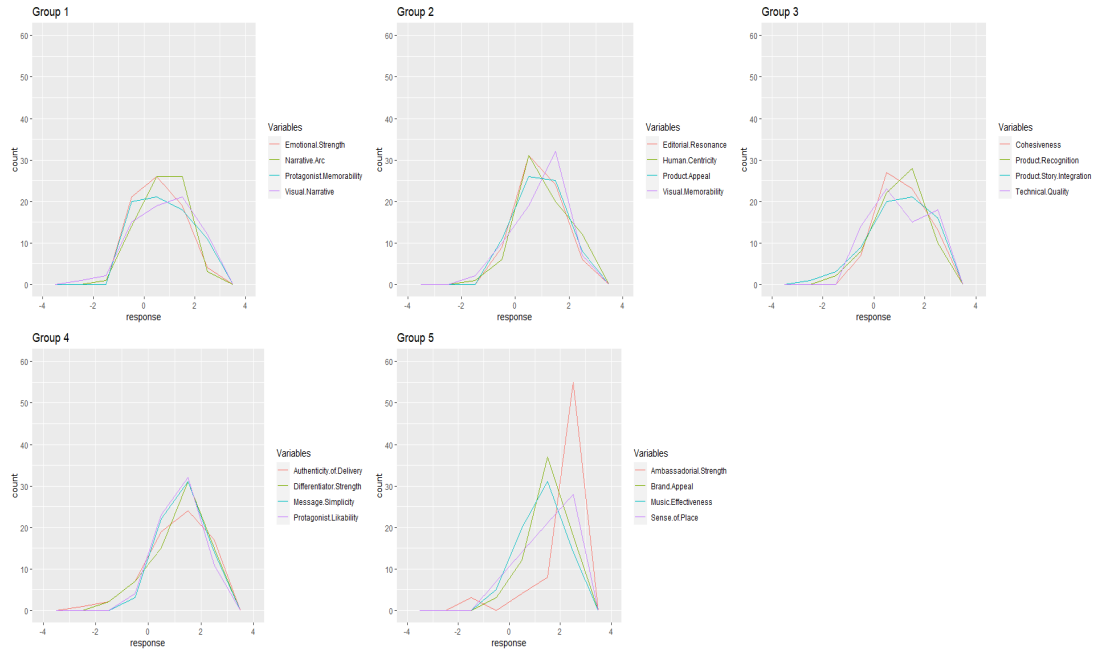
Figure 2: Bar Plots for 3 Different Features



We show 3 features that are representative of the type of skewness of the data: Product Appeal (median = 1st quartile), Visual Narrative (median is between the quartiles) and Music Effectiveness (median = 3rd quartile). Most of the values are integers and the ratings that are averages are equally spaced between the integers.

Third, marginal conditional distributions were explored using frequency polygons as shown in Figure 3. In this analysis, features are divided into 5 groups based on their overall means from the lowest to highest, namely the features in the first group have the smallest feature mean and the ones in the fifth group have the biggest feature mean. Y axis was set to 0-60 for easy comparison. Figure 2 shows that there are a lot of similarities among the features within each group except group 5. Features in group 5 are slightly negative skewed, especially for Ambassadorial Strength. Based on examination of Figure 1 - 3, we decided not to use transformations, but we decided to standardize the data for PCA and hierarchical clustering.

Figure 3: Frequency Polygons for 20 Numeric Features Ordered by Feature Mean



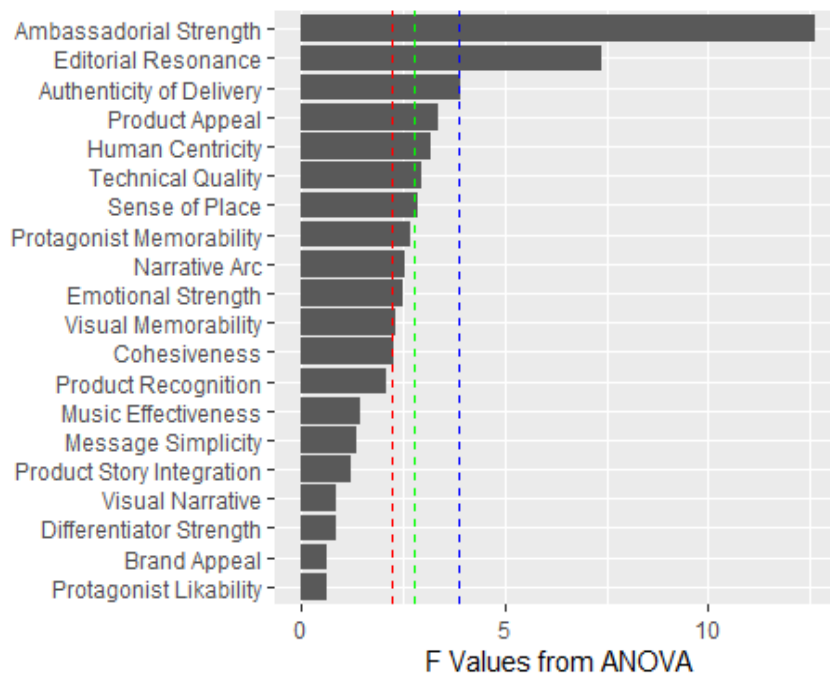
Y axis represents counts and X axis represents feature values. Features have been grouped based on their mean values from lowest (group 1) to highest (group 5). Features within each group have similar distributions except group 5.

2. One-way ANOVA: Features by Story Types

We used One-way ANOVA to explore how well each feature separates the story types. Since there is only one observation in story type “Other”, we did not include it in the ANOVA.

First, a Pareto plot was made as shown in Figure 4. In this figure, features were ordered by significance level. There are 12 features with p value < 0.05 , 7 using BH adjusted p value, and 3 using Bonferroni adjusted p value. We can see that too many important features are identified using p value without any adjustment, and too few using Bonferroni adjusted p value. Therefore, we recommended use of the BH adjusted p value significant features for further analysis. These features include Ambassadorial Strength, Editorial Resonance, Authenticity of Delivery, Product Appeal, Human Centricity, Technical Quality and Sense of Place.

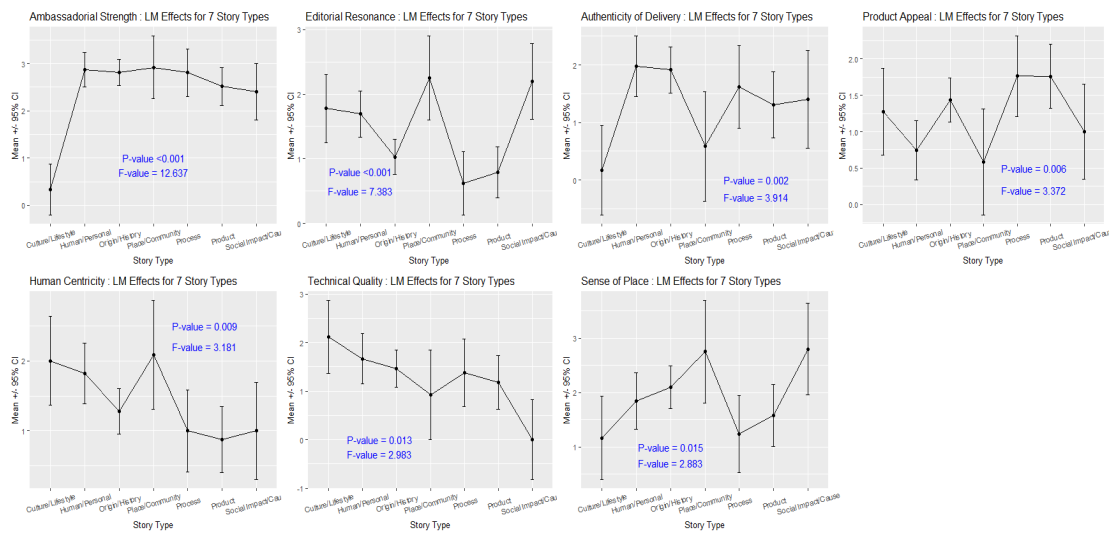
Figure 4: Ranking of the Features by Significance from ANOVA



The red and vertical dashed line represents the p value of 0.05; the green is the BH adjusted p value; and the blue represents Bonferroni adjusted p value. Seven important features are recommended as significant using BH adjusted p value.

To further explore how well these 7 features differ among the story types, we made effect plots for each feature using “emmeans” package. As shown in Figure 5, these features are ordered by significance level. The first is Ambassadorial Strength which has the smallest BH adjusted p value, and the last one is Sense of Place which has the least significant p value among the 7 features. The mean score and corresponding 95% confidence intervals (CIs) are shown on the Y axis. Story types are shown on the X axis. Each feature has at least 1 story type that has different mean score and corresponding 95% CI. Take Ambassadorial Strength in the top left panel for example, Culture/Lifestyle has a lower mean and non-overlapping 95% CI compared with other story types, indicating that Ambassadorial Strength is a strong feature that can separate Culture/Lifestyle from the other story types. This makes sense because Ambassadorial Strength is about how closely the protagonist are connected to the brand or product, and it has little connection with Culture/Lifestyle.

Figure 5: Effect Plots of Important Features for Different Story Types



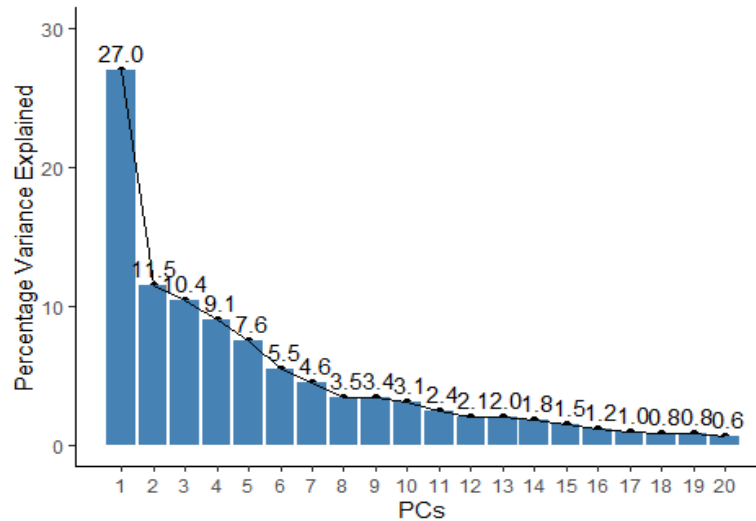
Y axis is the mean score and corresponding 95%. X axis represents story types. Features are ordered by significance level. Each feature has at least 1 story type with different mean score and corresponding 95%

3. PCA

We tried PCA using both standardization and non-standardization for the data. Data standardization is to make columns with mean 0 and standardized deviation 1. In this way, we put features on the same scale. This is useful when the data are measured in different units. In the TLP films study, our data have a common -3 to 3 scale $[-3, 3]$ indicated by marginal distribution plots in Figure 1 and Figure 3. Although the results for PCA and hierarchical clustering are similar with or without data standardization, as we discussed above, all results shown here were calculated using the standardized data.

We first made a variance plot as shown in Figure 6. We can see that PC1 explains 27.0% of the variance, PC2 explains 11.5% of the variance, and the percentage of variance decreases as PC increases.

Figure 6: Percentage of Variance Explained by Different PCs.



PC variance bar plot. Variance explained by PC decreases as PC increases.

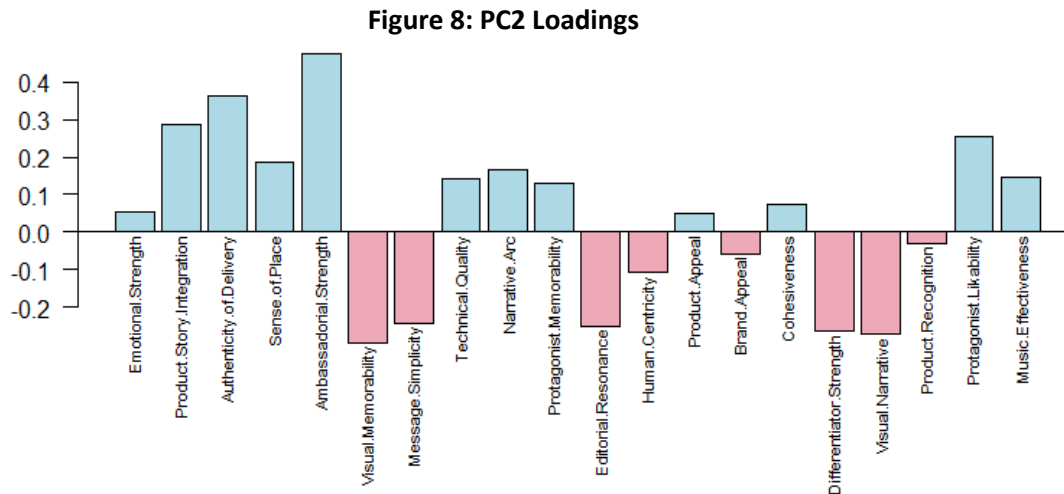
To find out what PC1 and PC2 represent, a biplot was made as shown in Figure 7. In this biplot, the numbers represent the sequential order of the 70 films in the data. The red arrows are the graphical display of the loadings for PC1 and PC2. The loadings in biplot are unit vectors. For example, the arrow for Ambassadorial Strength, which is at the top of the graph, has its X coordinate from the PC1 loading and its Y coordinate from the PC2 loading. We can see that almost all features point in one direction for PC1, suggesting that PC1 may represent overall quality for the films. Since PC1 seems to capture the overall quality for the films and all the loadings are negative, suggesting that lower values on PC1 may indicate higher scores on the features relative to the overall mean. For example, the point labeled 47 corresponds to film id 29, *When Inspiration Strikes*. It is on the far right side and has the lowest quality. The point labeled 33 corresponds to film id 46, *Brewed with Gratitude*. It is on the far left side and has the highest quality. PC2, on the other hand, contrasts the features going upwards and the ones going downwards. The upwards features include Ambassadorial Strength, Authenticity of Delivery, Product Story Integration and Protagonist Likability. The features going upwards are mostly about protagonists in the films.

Figure 7: Biplot for Features and Films



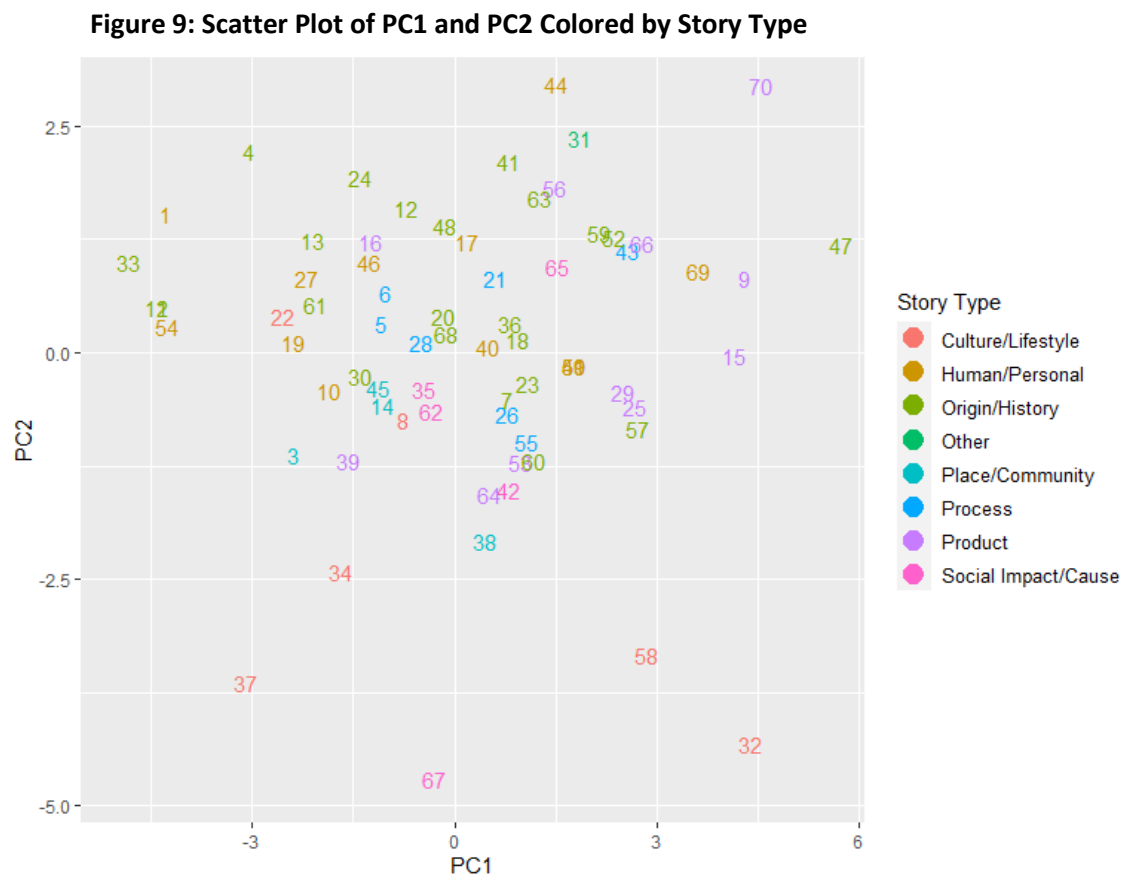
This biplot shows first two principal components, PC1 reflects overall quality measure, and PC2 suggests features that inversely related with each other.

It is hard to tell the features going downwards in Figure 6 because they are all clustered together. Therefore, we made a PC2 loadings bar plot. As shown in Figure 8, the features pointing downwards include Visual Memorability, Visual Narrative, Message Simplicity, Editorial Resonance and Differentiator Strength. These features are mostly about visual impact of the films. As discussed above in Figure 7, features going upwards on PC2 may relate more with protagonists in the films. Thus, we conclude that PC2 may distinguish between films that are highly thematic versus those that are visually oriented.



Y axis represents loading values and X axis represents features. PC2 distinguishes features with positive loading values and negative ones.

Furthermore, we wanted to examine how PCs align with story types. In this analysis, scatter plots were made using the different combination of PCs. The story types were not really distinguishable as clusters on any of these plots. PC1 and PC2 are the most important 2 PCs. So, for simplicity, we show only the scatter plot using PC1 and PC2 in Figure 9 where the numbers represent the sequential order of the 70 films in the data. We also made scatter plot using film names. Because it is hard to read the text when the figure is small, we put it in the 04_PCA PDF output so that our client can access it in the repository. There are a few films in the story type Culture/Lifestyle and Social Impact/Cause with high values on PC2, but they do not form clearly distinct clusters. In other words, the PCs only weakly align with story types.



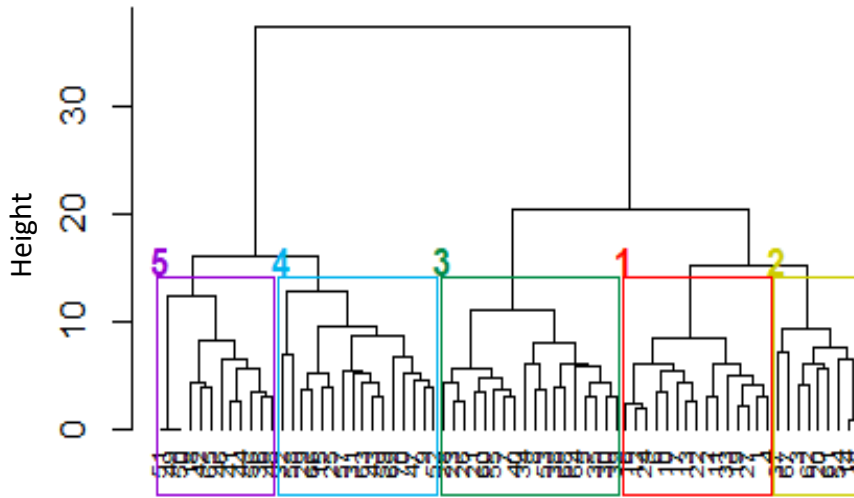
The first two principal components are shown. Different story types do not form clearly distinct clusters.

4. Hierarchical clustering

Since PCA shows story types do not form clearly distinct clusters on PCs, we wanted to explore features using hierarchical clustering and tried to find potential clusters. In this section, standardized data was again used for analysis. Euclidean distance was used to calculate the dissimilarities between sets of observations, and Ward was used as the linkage criterion.

Five clusters are identified as shown in Figure 10. There are 15 films in cluster 1, 9 in cluster 2, 18 in cluster 3, 16 in cluster 4 and 12 in cluster 5. The order of the clusters is 5, 4, 3, 1 and 2 so that they will have labels shown in Figure 11 (see below).

Figure 10: Cluster Dendrogram

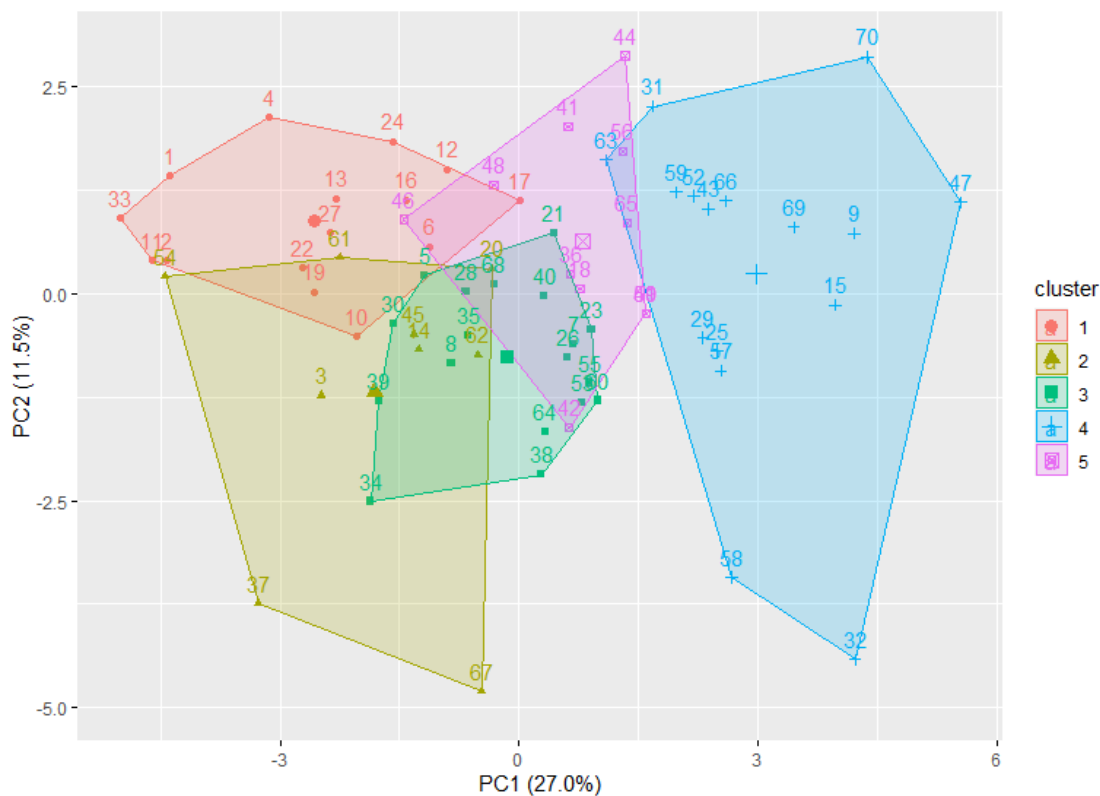


Five clusters are identified using Euclidean distance and Ward linkage criterion.

To better visualize the clusters, a cluster plot using PC1 and PC2 was made as shown in Figure 11. We can see that cluster 4 is really different from the other clusters. There is little or no overlapping between cluster 1 vs 3, cluster 1 vs 4, cluster 1 vs 5, cluster 2 vs 4 and cluster 3 vs 4. We show the statistical significances of the differences among the clusters in Table 1 (see below).

Cluster 4 has the highest values on PC1, and clusters 1 and 2 have the lowest values on PC1. As discussed above in Figure 7, higher values on PC1 represent lower scores on the features relative to their means, so, lower values on PC1 suggest better films. Consequently, clusters 1 and 2 have the higher rated films and cluster 4 has the lower rated films. Although, clusters 1 and 2 have similar values on PC1, suggesting they are similar in film quality. Cluster 1 has higher values than cluster 2 on PC2. Based on the interpretation of PC2 that we proposed above in Figure 7 & 8, this suggests that films in cluster 1 tend to be more thematic and less visual than films in cluster 2.

Figure 11: Cluster Plot using PC1 and PC2



Cluster 4 is really different from other clusters.

Since cluster 4 is really different from the other clusters in Figure 12, it would be interesting to examine the significance level between every two clusters. Statistical Significance of Clustering (SigClust) was conducted to assess whether there is statistically significance between every two clusters, and we used “sigclust” package for the test. As shown in Table 1, there are statistically significant differences (as shown by bold type in Table 1) for cluster 1 vs 3, cluster 1 vs 4, cluster 1 vs 5, cluster 2 vs 4, and cluster 3 vs 4.

Table 1: P values between different clusters.

Cluster	P-value
1 vs 2	0.308
1 vs 3	0.036
1 vs 4	< 0.001
1 vs 5	0.001
2 vs 3	0.222
2 vs 4	0.001
2 vs 5	0.148
3 vs 4	0.001
3 vs 5	0.249
4 vs 5	0.101

Shows significant difference in cluster 1 vs 3, 1 vs 4, 1 vs 5, 2 vs 4 and 3 vs 4.

To examine the relation between story types and clusters, we first made a contingency table as shown in Table 2. We can see that each story type has different numbers of films, and different story types dominate in different clusters.

Table 2: Contingency table for story types and clusters.

Clusters Story Types	1	2	3	4	5	Sum
Culture/Lifestyle	1	1	2	2	0	6
Human/Personal	5	1	1	1	5	13
Origin/History	7	2	5	5	4	23
Place/Community	0	3	1	0	0	4
Process	1	0	5	1	0	7
Product	1	0	3	6	1	11
Social Impact/Cause	0	2	1	0	2	5
Other	0	0	0	1	0	1
Sum	15	9	18	16	12	70

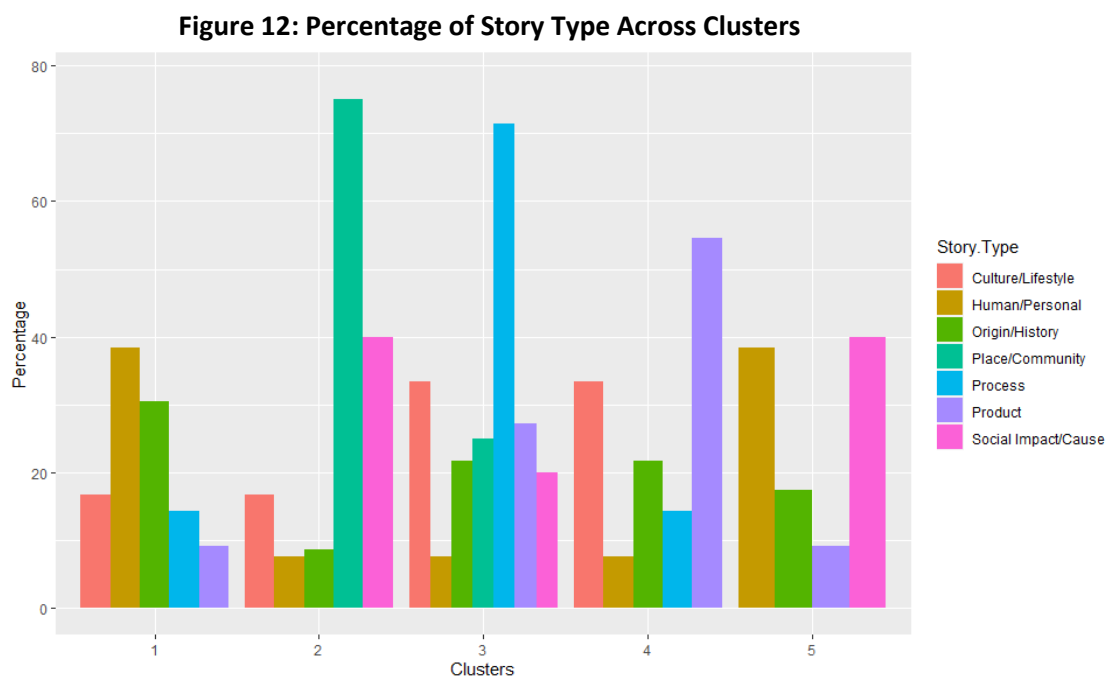
Shows different story types dominate in different clusters.

Because there are large differences in the number of films of each story type as shown in Table 2, we found it useful to normalize for this by computing the percentage within each story type as shown in Figure 12. In this Figure, the story type “Other” was omitted because there is only 1 observation in this group.

As discussed above in Figure 11, cluster 4 has lower rated films. Product and Culture/Lifestyle have a higher percentage in cluster 4, and we conclude that these story types have a higher proportion of lower rated films.

On the other hand, clusters 1 and 2 have higher rated films. Human/Personal and Origin/History have a higher percentage in cluster 1, indicating that these story types have a higher proportion of higher rated films. Place/Community and Social Impact/Cause have a higher percentage in cluster 2, indicating that these story types also have a higher proportion of higher rated films.

High values in PC2 seem to match with films more thematic and low values are more visual films. Figure 12 shows dominant story types in cluster 1 are Human/Personal and Culture/Lifestyle. We think this makes sense because these films have specific story lines or themes. In contrast the story types in cluster 2 are Place/Community and Social Impact/Caus. We think these story types may tend to more visually oriented because they intend to present a feeling or impression.

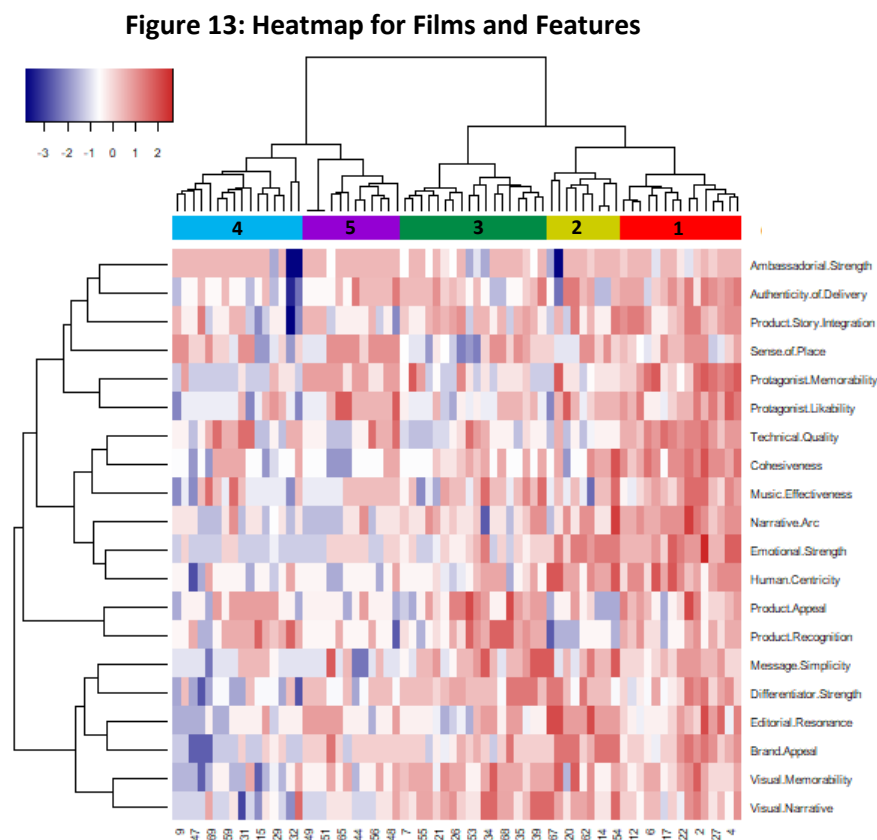


Percentage of story types across clusters. Percentage of different story types is distinct across clusters.

Furthermore, we are also interested in the relationship between clusters and the 20 features. Therefore, a heatmap was plotted with film dendrogram on the top and feature dendrogram on the left. Standardized data was again used for this analysis. Euclidean distance was used to calculate the dissimilarities and Ward was used as the linkage criterion.

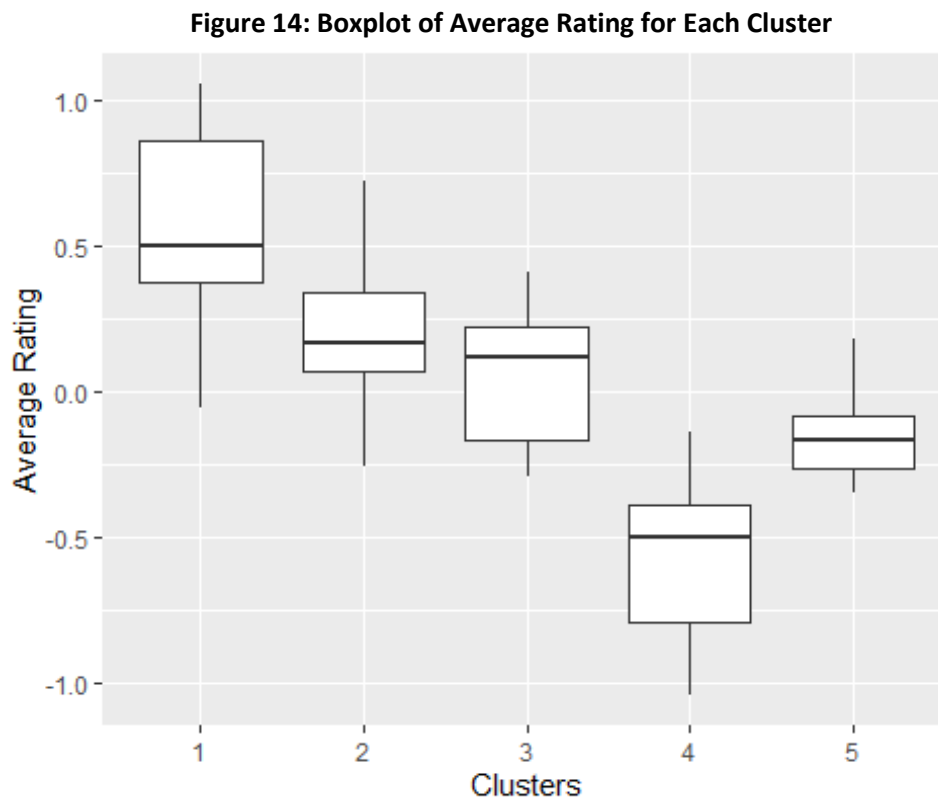
As shown in Figure 13, we used same colors and cluster numbers for the 5 clusters as those in Figure 10 & 11, where blue color represents cluster 4, purple color cluster 5, green color cluster 3, yellow-green cluster 2, and red color cluster 1. Also, blue color in the heatmap means lower values, and red color means higher values. In general, features have lower values in clusters 4 and 5 and higher values in clusters 1 and 2. Features have mixed values for cluster 3. These findings also indicate that clusters 1 and 2 have higher rated films, and clusters 4 and 5 lower rated films.

As we go from left to right, films are generally higher rated, which we discussed before is captured by PC1. The features at the top are generally more thematic, and the features at the bottom are generally more visual, which we said before is captured by PC2. So, the heatmap interestingly represents information from both the PCA and cluster analyses. There are many interesting patterns in the heatmap, which suggests that different combinations of the features can be used to distinguish different aspects among the films (Fig. 13).



Features have lower values in clusters 4 and 5, and higher values in clusters 1 and 2.

As discussed above, clusters 1 and 2 have higher rated films and cluster 4 has lower rated films. To further confirm these, we calculated average rating using the standardized data again. Average ratings refer to the row means for the clusters. Boxplot was made using the average ratings as shown in Figure 14. The figure shows cluster 4 has the lowest average ratings, and clusters 1 and 2 the highest average ratings. Therefore, we conclude that cluster 4 has the lowest rated films and clusters 1 and 2 have the highest rated films.



Average ratings are the row means of standardized data. Cluster 4 has the lowest average ratings and cluster 1 the highest average ratings.

References

1. <https://theloomaproject.com/>