Identifying Subcultures with Likert-Scale Data: Evidence from Applications of Clustering and Principal Component Analysis

Jordan Fox
University of Texas at Austin

5/13/21

Abstract

This paper seeks to demonstrate the use of two unsupervised learning techniques–hierarchical clustering and principal component analysis–by presenting their applications to cross-sectional, Likert scale survey data. In these applications, I identify several consumer subcultures that aren't explicitly labeled. These findings are presented in a way that would be relevant for marketing analytics, consumer science, or even sociological analysis.

1 Introduction: Subcultures and Marketing

A subculture is a sociocultural grouping that is distinguished from a dominant, or parent group. The boundaries that define them often fall along lines of nationality, age, vocation, religion, and lifestyle. They serve various social functions, for example, like giving their members the ability to find their identity, providing a means of retreat for alienated individuals, and connecting people with like-minded individuals or those with interests similar to their own. Notable examples of youth subcultures include hipsters, punks, nerds, and skaters; each of these is typically associated with different modes of recreation, as well as consuming different forms of media.

For this project, I focus my attention to consumer subcultures, sometimes described as market segments, with members that posses similar consumption habits, media preferences, and tastes. My intention is to see whether I can identify these in spite of the narrow range of responses that Likert scales (numeric responses that range from one to five) offer to their respondents. Ultimately, one's association with a particular subculture often implies a common preference for an activity or type of product.[1] The identification of such groups is thus relevant in the context of marketing as they allow marketers to focus their efforts on representative populations,[2] as opposed the voluminous number of people who make them up.

The organization of this paper is as follows. In the preliminaries, I describe the survey and the data collection process, as well as visualizing some of the labeled features of some of the respondents. Then, I use hierarchical cluster-ordered correlograms to see if it makes sense to compress respondents into clusters based on the data subset I'm inspecting. After identifying potential subcultures within the observations, I then make use of principal component analysis to compress the interests of a given cluster to test how much

[1] Solomon (2013)
[2] Schiffman, Kanuk, and others (2009)

of the variance is preserved under dimension reduction. Whether a cluster is indeed a subculture or not depends on the amount of this preserved variance.

2.1 Preliminaries: About The Survey & Its Respondents

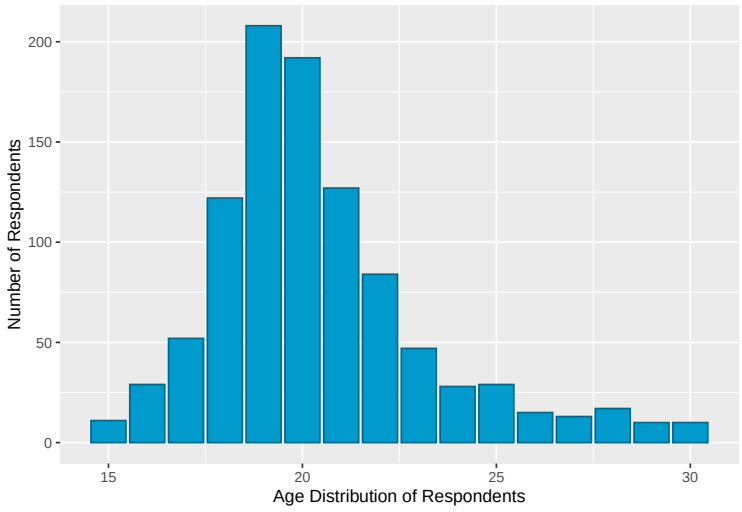
The survey was issued by the Social and Economic Sciences department at Comenius University in Bratislava, Slovakia. To collect the data, the instructor distributed the survey to a section of his students, and asked them to encourage their friends to fill it out. It consists of just over 1,000 responses to over 150 questions, which for the data set means I have 1,000 observations with 150 columns. Of these 150 columns, 139 are integers; the rest are categorical. Each column represents a different question, each of which belongs to a sub-topic, and is rated from one to five by the respondent. For each question, a "1" represents a response along the lines of "least interested in," "don't enjoy," or "not important," and a "5" being in the ballpark of "most interested in," "greatly enjoy," or "very important." The topics range from music and movie preferences, to hobbies, activities, and interests, phobias, health habits, personality traits, philosophy, views on life, spending habits, and demographics. Many of these observations have missing data points, which are dropped from the analysis.

All participants were of Slovakian nationality, and each had an age of between 15 and 30 at the time of the survey. As I explore in the next subsection, most of the respondents are from cities, have a high school education, and are female.

2.2 Preliminaries: Descriptive Analysis of Labeled Features

Before we delve into machine learning applications, it would be useful to first get a basic idea of some of the characteristics of our respondents, as well as the features of the data. To do so, I plot the distributions of some relevant variables in the following pages.

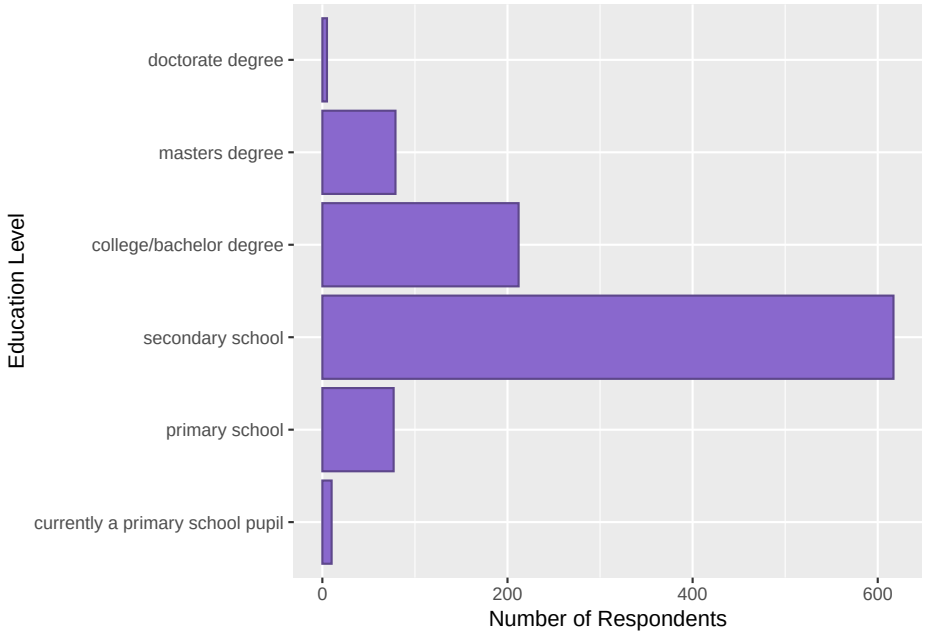
Figure 1: Distribution of Respondent Ages and Education Levels



The average age of respondents at the time of the survey was just over 20 years old (20.36, specifically). The distribution has a right skew, indicating that our observations tend to fall on the younger side of the

age scale. Considering how the survey was collected, the average age of respondent isn't too surprising. As in the US, in Slovakia the average age of a college student probably falls between 18 and 22.

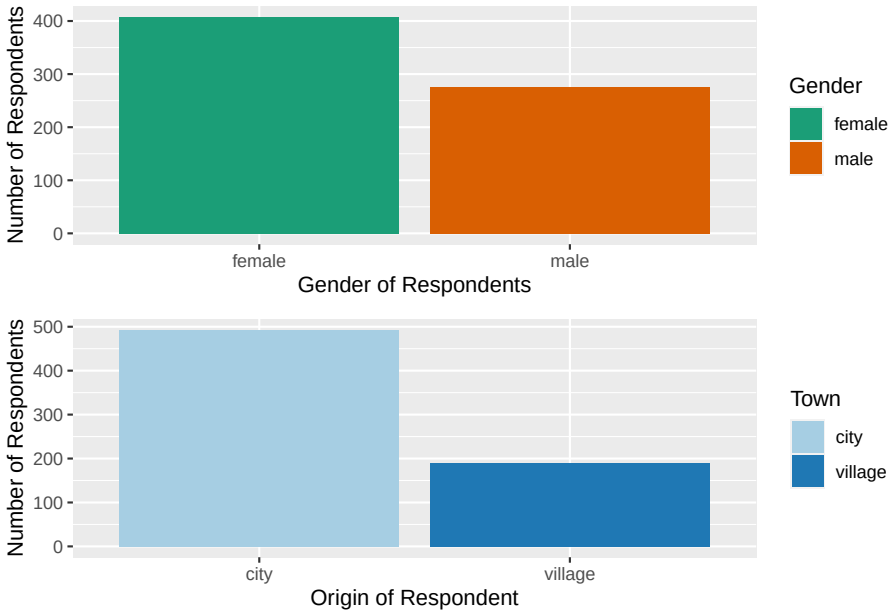
Figure 2: Distribution of Education Levels of Respondents



Here, we can see that the average level of schooling for a respondent was equivalent to a high-school education in the US. The second-most common education level for the respondents was a bachelor's degree, followed by what appears to be a tie between respondents who have completed a master's degree and those who have completed elementary/middle school. A small number of respondents had either a doctorate or were currently enrolled as a primary school student.

The fact that high-school graduates make up over 50% of the respondents is not at all surprising; given how the survey was administered, one would expect the average respondent to be someone that is either in college or is of the age of a college student. The existence of respondents that were in primary school is a bit strange given that the age for primary school in the US is well below 15, as is the presence of PhD-holding respondents. However, the latter is less surprising considering that many doctorates are awarded to people under the age of 30.

Figure 3: Male/Female and City/Village-Origin Distributions



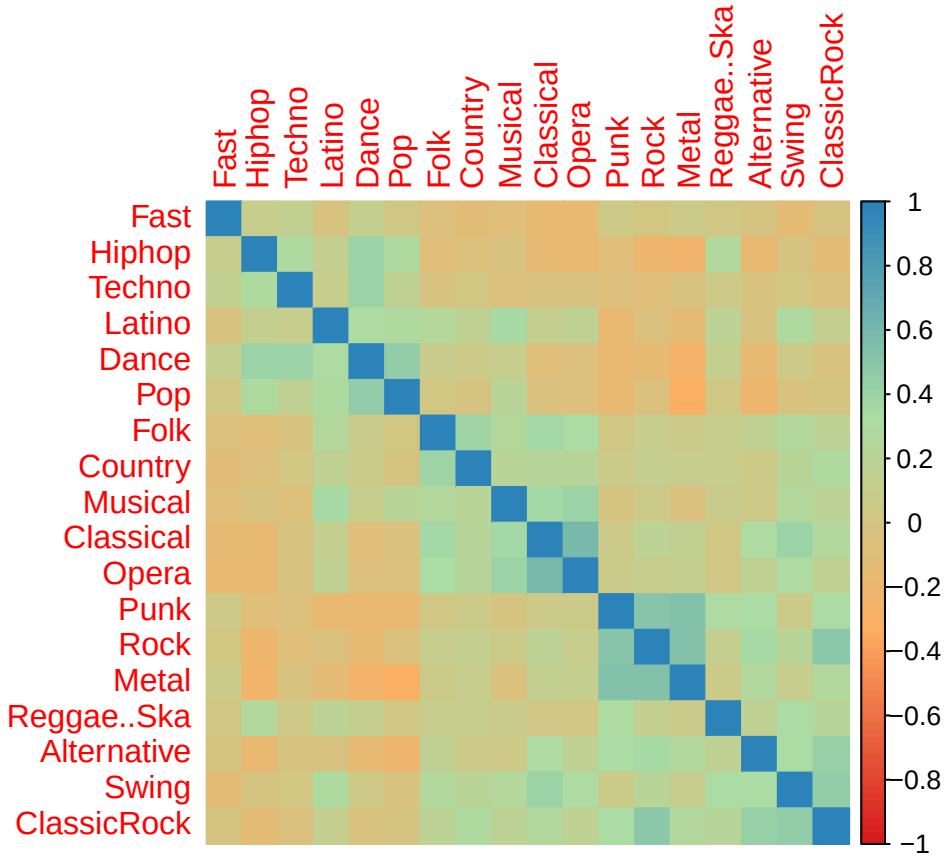
It appears that women are slightly over-represented in this data set, making up about 60% of the survey respondents. While there are various explanations, the most obvious one is that women tend to be more likely to attain higher education[3]. Likewise, it could be more reflective of the population of the university, or the makeup of the students majoring in social sciences. Meanwhile, respondents from villages appear to be dramatically under-represented, making up just ~30% of the respondents.

3.1 Identifying Clusters Using Music, Movie, and Activity/Interest Preferences

In this section, I use an application of clustering–specifically hierarchical cluster-ordered correlograms–to identify some subcultures that are present in the university population's student body, but unapparent in the survey data. My intention is to first focus my analysis on patterns within *the observations* before turning to the columns of the data set. The correlograms are presented in the next three figures.

[3]Becker, Hubbard, and Murphy (2010)

Figure 4: Hierarchical Cluster-Ordered Correlogram of Music Preferences



The above correlogram's rows and columns are ordered based on hierarchical clustering, which reveals two parent clusters, which I call the "digital" and "analog" clusters. The borders of these can be seen at the pop/folk boundary. The digital cluster is made up of fast music, hip-hop, techno, latino, dance, and pop music; the analog cluster is made up of folk, country, punk, rock, reggae, and swing. A closer inspection of the analog and digital clusters reveals several branches, indicating subgroups within these clusters. For the digital group, the most visible being the dance/pop correlation. There also appears to be a mild association between techno and hip-hop as well.

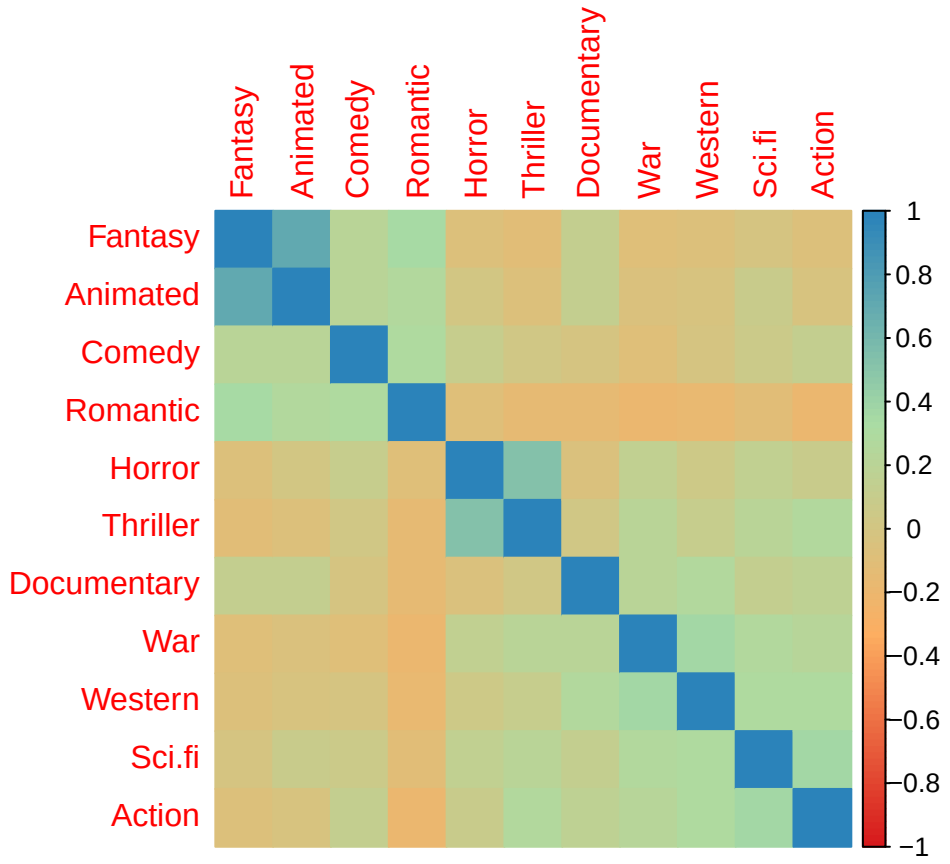
For the analog cluster, the most visible grouping is the metal/rock/punk cluster. Interestingly, these three genres of music are more correlated with each other than any of the other genres; they also appear to be negatively correlated with just about every other genre of music, indicating that people who listen to them tend to be insulated from other kinds of music. This is consistent with the literature on metal music and metal subcultures, which finds that metal listeners are more insolated and socially alienated[4] than listeners of other kinds of music. Other groups seen in the analog cluster are those for the folk/country and classical/musical/opera listeners.

While practically no one listens to only a single type of music, I argue that it's reasonable to characterize the smallest of these clusters as subcultures. Most of the digital cluster could be characterized as variants of club-goers, although this may be a bit of a generalization given the number of interests that make it up.

[4]Bryson (1996), Stack, Gundlach, and Reeves (1994)

This also applies (to a greater extent) to the analog cluster; it certainly doesn't make sense to lump country music aficionados in with listeners of metal, but when we break this group down into its branches, sensible distinctions begin to appear. These are: (1)*Rockers*, or people who prefer metal and punk, (2)*Theater-Goers*, who tend to listen to classical music, opera, or musical theater, and (3)*Traditionals*, who enjoy country and folk music. Given that preferences for these types of music are negatively correlated with virtually every other genre, characterizing these are distinct subcultures makes sense.

Figure 5: Hierarchical Cluster-Ordered Correlogram of Movie Preferences



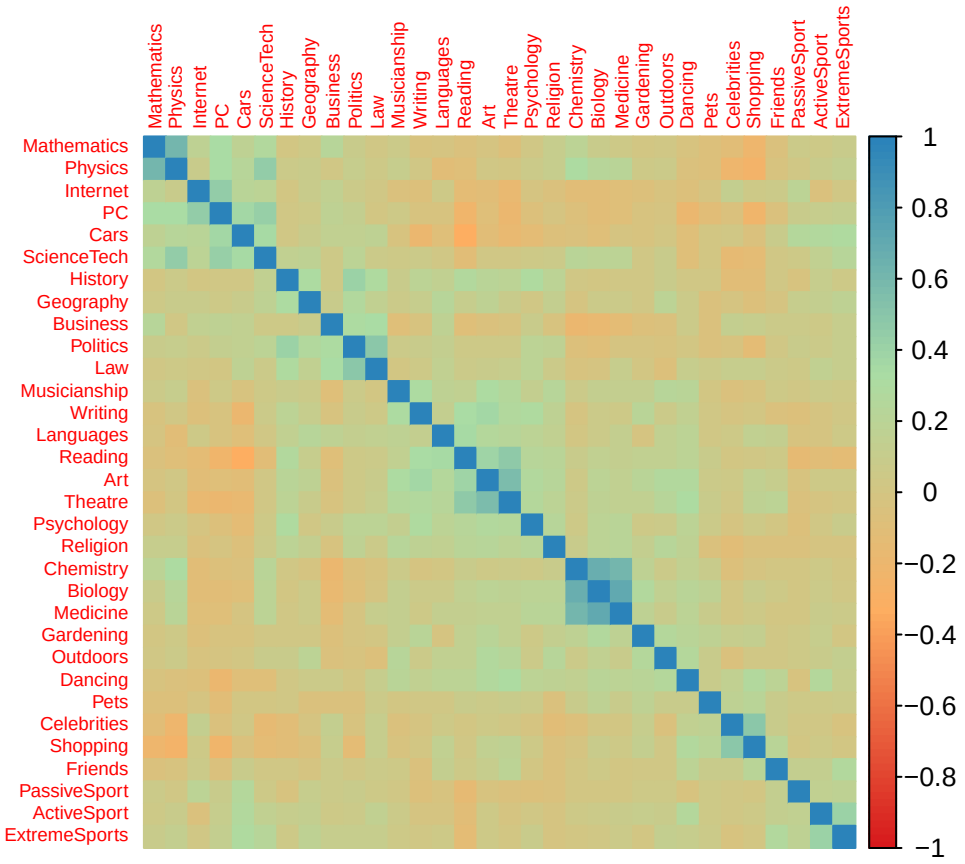
The above correlogram, this time focusing on movie preferences, also shows two large parent clusters. Here, it's a bit difficult to characterize the differences between these. One might be tempted to characterize the cluster encompassing fantasy, animated, comedy, and romantic as escapist, but then we see that sci-fi, action, and thriller are included in the opposite parent cluster, making this distinction less sensible. Instead, I decide to call the large and small parent clusters (A) and (B), respectively.

For (A), the association that is most obvious at first glance is the horror/thriller cluster. Additionally, there appear to be slight associations that make up war/western and sci-fi/action clusters, but these are not strong enough to warrant a particular subculture. Thus, the first consumer subculture that I identify from this figure is what I call the *Thrill-Seekers*; these are people who tend to watch movies for suspense and excitement.

Looking at (B), the association that stands out the most is the one between fantasy and animated films. This is the strongest association between any two genre of film, and constitutes the second consumer

subculture identified in this subset of the data: *Fantasy Fans*. These might be people who are into fantasy franchises like Lord of The Rings and Dungeons & Dragons, or animated studios like Pixar/Dreamworks or even Studio Ghibli. There is also a slight association between war and western films, again within (B), which might be worth characterizing as a subculture if it were stronger.

Figure 6: Hierarchical Cluster-Ordered Correlogram of Interest and Activity Preferences



Although less defined than the parent clusters in the previous section, we can see that three emerge in the ordered correlogram of activities and interests. The first is made up of technical interests, such as mathematics, physics, science and technology, and computers. I characterize this cluster as the "tech cluster" , and it is likely made up of individuals majoring STEM fields and with backgrounds in computing and technology. The second cluster, which I call the "humanities cluster" consists of more academic-oriented interests, ranging from law and history to theater, psychology, and art. The third is an assemblage of social (sports, fun with friends, dancing) and outdoor (gardening, outdoors) activities, which I characterize as the "wellness cluster." This also includes the well-defined association between chemistry, biology, and medicine, which I explore in the coming paragraphs.

Within each of these, there are several distinct clusters of interests that could be indicative of subcultures. Within the tech cluster, the strongest correlation is between mathematics and physics, followed by PC and internet; I characterize these as the *Mathematicians* and *Onlines* subcultures, respectively. Within the humanities cluster, we have the politics/law/business cluster, which I characterize as the *Policy Wonks*. These might be people who majored in interdisciplinary humanities like politics, economics, law, or history,

and may even be newly-professional college grads. Next, also in the humanities cluster, we have the *Fine Artists*; this cluster is made up of people who prefer reading, art, and theater.

Finally in the last parent cluster, there are at least two associations that are easily identifiable. The first is at the intersection of celebrities and shopping, which I call the *Socialites*. Finally, a subculture that one might expect to find lumped in with the tech cluster is the one defined by common interests in chemistry, biology, and medicine. I characterize this association as the *Scientists* subculture. These are likely individuals that are majoring in these disciplines, or on track to work in a field which utilizes all three, such bio-tech or medicine.

3.2 Verifying Subcultures With Principal Component Analysis

Now that some potential subcultures have been identified, I move my focus away from the observations and toward the columns which make up the data set. To confirm that these associations make up distinct groups of people with disparate consumption habits and interests, I'll compress these columns into a single principal component with the goal of identifying how much variance is preserved. In the event that a principal component of two or more interests maintains a significant amount of the variance (that is, we do not lose much information by collapsing the columns of data), then I argue that these interests can plausibly constitute a microculture or subculture. The reasoning is this: if a single group, or component, makes up enough of the variation in a particular set of interests, then it is likely that this component is comprised of individuals with like-minded hobbies and attitudes, and as such makes up a subculture.

First, I'll review the potential subcultures I've identified across music, movie, and interest preferences. These were:

1. *Rockers*: **Metal, Rock, and Punk Music**
2. *Theater-Goers*: **Musicals, Opera, and Classical Music**
3. *Traditionals*: **Folk and Country Music**
4. *Thrill-Seekers*: **Horror and Thriller Movies**
5. *Fantasy Fans*: **Fantasy and Animated Movies**
6. *Mathematicians*: **Mathematics and Physics**
7. *Policy Wonks*: **Law, History, and Politics**
8. *Fine Artists*: **Reading, Art, and Theater**
9. *Onlines*: **PC and Internet**
10. *Socialites*: **Shopping and Celebrities**
11. *Scientists*: **Chemistry, Biology, and Medicine**

Beyond the compression into a single component, I also perform a rotation of each to confirm that the individual scores for the factor loadings are weighted in the same direction. The results from many of these

are reported and commented on in the following paragraphs, but the output from each is left to the appendix for aesthetic reasons. However, the cumulative variance is reported for each of the potential subcultures.

The criteria for being a subculture or microculture here is as follows. For groups that are made up of two interests, the principal component they are compressed into must preserve at least 70% of the variance of the two vectors of interests; for groups that contain three interests, the resulting principal component must preserve at least 65%. These correspond to information losses of 35% and 40% respectively. Losing some information is to be expected, as most people tend to diversify their music consumption. However, a significant reduction in variance below these thresholds implies that additional components are needed to explain a majority of the information, and would cast doubt on a single group driving the preference for those interests. Clusters which fail the criteria are marked with an [X].

Total Information Preserved Under A Single Principal Component

1. *Rockers*: **Metal, Rock, and Punk Music – 68.36%**

2. *Theater-Goers*: **Musicals, Opera, and Classical Music – 64.29% [X]**

3. *Traditionals*: **Folk and Country Music – 69.32% [X]**

4. *Thrill-Seekers*: **Horror and Thriller Movies – 77.01%**

5. *Fantasy Fans*: **Fantasy and Animated Movies – 80.57%**

6. *Mathematicians*: **Mathematics and Physics – 80.7%**

7. *Policy Wonks*: **Law, History, and Politics – 58.24% [X]**

8. *Fine Artists*: **Reading, Art, and Theater – 65.17%**

9. *Onlines*: **PC and Internet – 72.93%**

10. *Socialites*: **Shopping and Celebrities – 74.46%**

11. *Scientists*: **Chemistry, Biology, and Medicine – 77.78%**

We can see the interests of each of the *Rockers*, *Thrill-Seekers*, *Fantasy Fans*, *Mathematicians*, *Fine Artists*, *Socialites*, and *Scientists* subcultures pass my threshold for variance preservation. A closer inspection of the factor loadings reveals that the weights for most of the interests within the components are negative. This is not necessarily surprising, because most of the interests that make up these clusters are typically positively correlated with one another, but are negatively correlated with the rest of the interests in the subset of the data we're analyzing (as was the case for metal music in Figure 4).

Meanwhile, the clusters that did not meet the variance preservation criteria had interests that were generally more positively correlated with other interests. I argue that this bolsters the case that those groupings which passed the criteria are in fact groups of people with tastes that are separate from the general university population, and thus represent different subcultures. For example, classical music and opera are culturally accepted genres, and respondents in general might have neutral or slightly positive attitudes toward them. If it were the case that a general acceptance or neutrality toward them was driving their average scores more so than a dedicated subgroup of fans or adherents, then it would be expected

that less of the variation would be preserved if they were to be compressed, as the interests of the general population tends to be more dispersed and less focused than the members of certain subcultures.

Not surprisingly, this is the case for two of the three clusters which failed to meet the threshold criteria. With the exception of folk and country, which make up the interests for the *Traditionals*, the interests of the rest of the non-subcultures tend to be more positively rated than those that make up, say, the *Mathematicians* or *Scientists* subcultures. One exception to this rule is the *Socialites*, whose interests of shopping and celebrities are generally correlated with other interests, yet still stands out as a subculture after analyzing the principal component resulting from the compression of these interests. A table of these average ratings is reported in the appendix of code chunks at the end of this document.

Conclusion

Despite the limited range of responses that respondents were able to give (and the poor definition of parent clusters when analyzing one subset of data), I was able to identify eight distinct groups of respondents that were latent in the survey data. Most of these groups were defined by interests that were seen as less favorable to the general public, possibly indicating that the associations between their respective interests were being driven by dedicated fandoms or subsets of respondents, which I argue is representative of subcultures in the respondent pool. Clusters that were suspected of being subcultures but did not pass the specified threshold had interests that were seen as favorable across respondents.

Appendix of Principal Component Analysis Results

Rockers

```
##              PC1
## Rock  -0.5768606
## Metal -0.5806621
## Punk  -0.5745114
```

```
## Importance of first k=1 (out of 3) components:
##                             PC1
## Standard deviation     1.4321
## Proportion of Variance 0.6836
## Cumulative Proportion  0.6836
```

Threatre-Goers

```
##                 PC1
## Classical 0.5991391
## Musical   0.5115384
## Opera     0.6159228
```

```
## Importance of first k=1 (out of 3) components:
##                             PC1
## Standard deviation     1.3888
## Proportion of Variance 0.6429
## Cumulative Proportion  0.6429
```

Traditionals

```
##               PC1
## Folk    0.7463511
## Country 0.6655524
```

```
## Importance of first k=1 (out of 2) components:
##                             PC1
## Standard deviation     1.2862
## Proportion of Variance 0.6932
## Cumulative Proportion  0.6932
```

Thrill-Seekers

```
##                 PC1
## Horror   -0.8025295
## Thriller -0.5966124
```

```
## Importance of first k=1 (out of 2) components:
##                             PC1
```

```
## Standard deviation     1.6013
## Proportion of Variance 0.7701
## Cumulative Proportion  0.7701
```

Fantasy Fans

```
##                 PC1
## Fantasy  -0.7071068
## Animated -0.7071068
```

```
## Importance of first k=1 (out of 2) components:
##                           PC1
## Standard deviation     1.3082
## Proportion of Variance 0.8557
## Cumulative Proportion  0.8557
```

Mathematicians

```
##                    PC1
## Mathematics -0.7071068
## Physics     -0.7071068
```

```
## Importance of first k=1 (out of 2) components:
##                           PC1
## Standard deviation     1.270
## Proportion of Variance 0.807
## Cumulative Proportion  0.807
```

Policy Wonks

```
##                PC1
## Politics 0.6026186
## Law      0.6113157
## Business 0.5129756
```

```
## Importance of first k=1 (out of 3) components:
##                           PC1
## Standard deviation     1.3218
## Proportion of Variance 0.5824
## Cumulative Proportion  0.5824
```

Fine Artists

```
##               PC1
## Reading 0.5398844
## Art     0.5819416
## Theatre 0.6081684
```

```
## Importance of first k=1 (out of 3) components:
##                             PC1
## Standard deviation     1.3982
## Proportion of Variance 0.6517
## Cumulative Proportion  0.6517
```

Onlines

```
##                 PC1
## Internet -0.7071068
## PC       -0.7071068
```

```
## Importance of first k=1 (out of 2) components:
##                             PC1
## Standard deviation     1.2077
## Proportion of Variance 0.7293
## Cumulative Proportion  0.7293
```

Socialites

```
##                   PC1
## Celebrities 0.7071068
## Shopping    0.7071068
```

```
## Importance of first k=1 (out of 2) components:
##                             PC1
## Standard deviation     1.2203
## Proportion of Variance 0.7446
## Cumulative Proportion  0.7446
```

Scientists

```
##                  PC1
## Biology   -0.5949273
## Chemistry -0.5622796
## Medicine  -0.5743720
```

```
## Importance of first k=1 (out of 3) components:
##                             PC1
## Standard deviation     1.5275
## Proportion of Variance 0.7778
## Cumulative Proportion  0.7778
```

References

Becker, Gary S, William HJ Hubbard, and Kevin M Murphy. 2010. "Explaining the Worldwide Boom in Higher Education of Women." *Journal of Human Capital* 4 (3): 203–41.

Bryson, Bethany. 1996. "" Anything but Heavy Metal": Symbolic Exclusion and Musical Dislikes." *American Sociological Review*, 884–99.

Schiffman, Leon G, Leslie Lazar Kanuk, and others. 2009. "Consumer Behavior." *Harlow, England: Prentice Hall*.

Solomon, Michael R. 2013. *Consumer Behaviour: Buying, Having, and Being, 10th Global Edition*. Prentice-Hall, New Jersey.

Stack, Steven, Jim Gundlach, and Jimmie L Reeves. 1994. "The Heavy Metal Subculture and Suicide." *Suicide and Life-Threatening Behavior* 24 (1): 15–23.