Data-Driven Audience Targeting to Maximize Movie Profitability

Tiffany Ding and Soryan Kumar

September 2020

1 Non-Technical Executive Summary

The movie industry is a competitive market teeming with movie producers looking to create successful and profitable movies. However, generating profit in the movie industry is rarely as simple as hiring talented actors and screenwriters under a large budget because the deciding factor between a hit movie and a box-office flop is audience reception. Since viewer preferences are constantly shifting, it can be difficult to assess aggregate trends in movie genre preferences. Additionally, film producers are faced with the tradeoff of catering to niche populations within a genre or to more general audiences across genres to optimize viewer reception of their movie. By tracking viewer preferences for multiple genres, we aim to answer the question: **How can movie producers identify the optimal target audience in order to maximize profit?**

We obtain user ratings from the MovieLens dataset and movie budget and profit information from the Movie Industry dataset. We generate *genre preference vectors* for each user using the frequency of their ratings for movie of each genre. Our analysis relies on the assumptions that users tend to watch more movies in the genre that they prefer and that the movies a user rates are an unbiased sample of the movies that they watch. We later test this assumption and find evidence in support of it. Our approach has the benefit of not relying on the numerical user ratings themselves, which are difficult to compare between users (i.e. a rating of 3 may communicate a bad movie to one person and a mediocre movie to another person).

These genre preference vectors are clustered via k-means to generate three viewer archetypes, which can be interpreted as *comedy lovers, drama lovers, and those with wide/eclectic tastes*. For each movie, we compute an *audience composition vector* containing the proportions of the movie's viewers that fall into each of the three mutually exclusive viewer preference clusters. For example, a movie could have an audience composition of 40% comedy lovers, 10% drama lovers, and 50% of viewers with eclectic tastes. Average movie profitability was computed across all possible audience compositions to determine the

most profitable audience composition: 20% comedy lovers, 20% drama lovers, and 60% viewers with wide/eclectic tastes.

Subsequent analyses were conducted to determine optimally profitable audience compositions across different genres. These analyses gave rise to the following genre-specific recommendations:

1. Drama movies should add humor to cater more to comedy lovers.
2. Comedy movies should focus on incorporating other genre elements to appeal to drama lovers and those with wide/eclectic interests.
3. Romance movies should strike a balance between drama and comedy to appeal to an optimal audience of 45% comedy lovers and 25% drama lovers.
4. Sci-fi movies should limit their comedic elements since movies with fewer comedy lovers tend to have higher profits on average.
5. Thriller, action, adventure, and crime movies should target the optimal audience composition presented above: 20% comedy lovers, 20% drama lovers, and 60% viewers with wide/eclectic tastes.

These recommendations can be used by movie executives to help determine movie content, advertise effectively to key audiences, and predict the profitability of a given movie.

2 Technical Exposition

2.1 Data

We utilize two datasets in our analysis: (1) MovieLens user rating data and (2) the Movie Industry dataset.

MovieLens is a movie recommendation system launched in 1997 that recommends movies to users based off their ratings of movies that they have previously watched. The MovieLens dataset consists of 58,098 movies dating back to 1995 and roughly 28 million user ratings across 283,228 distinct users, each with their own anonymized user ID.

The Movie Industry dataset details various movie attributes for 6,820 movies dating back to 1986. We use the Movie Industry dataset in conjunction with the MovieLens dataset to obtain additional information about each movie, including genre, gross revenue, and budget.

2.2 Methodology

A high-level overview of our workflow is as follows:

1. Define a vector embedding to represent each user's genre preferences.
2. Use this vector to cluster users and determine dominant user archetypes.
3. Identify relationships between a movie's audience composition (defined in terms of cluster membership) and its profitability. Verify the statistical significance of these relationships and provide concrete advice to movie executives to help profit.

We provide details about each of these steps below.

2.2.1 Creating a vector embedding for genre preferences

For each user u, we compute a vector $\mathbf{g}^{(u)}$ that represents the genre preferences of user u. $\mathbf{g}^{(u)}$ is an n-dimensional vector, where $n = 20$ is the total number of movie genres. The movie genres we used were `Drama`, `Thriller`, `Comedy`, `Romance`, `Horror`, `Sci-Fi`, `Action`, `Adventure`, `Fantasy`, `Western`, `Crime`, `War`, `Musical`, `Children`, `Mystery`, `Animation`, `IMAX`, `Film-Noir`, `Documentary`, and `No genres listed`. Each index of $\mathbf{g}^{(u)}$ corresponds to a movie genre. For example, index 0 corresponds to `Drama`. The i^{th} element of $\mathbf{g}^{(u)}$ is computed as

$$\mathbf{g}_i^{(u)} = \frac{\text{\# of movies rated by user } u \text{ that are in genre } i}{\text{total \# of movies rated by user } u}$$

Note that some movies belong to more than one genre. In those cases, we consider the movie to contribute a uniform fraction to each of the genres it belongs to. For example, *Star Trek: The Motion Picture (1979)* is classified as both `Adventure` and `Sci-Fi`, so it contributes $\frac{1}{2}$ to the number of `Adventure` movies and $\frac{1}{2}$ to the number of `Sci-Fi` movies.

We make the assumption that the movies that a user decides to rate are an unbiased sample of the movies that the user has seen (e.g., a user who watches mostly action movies will mostly rate action movies). This assumption is consistent with the incentives of MovieLens users; since MovieLens is marketed as a movie recommendation system, users have an incentive to provide comprehensive information about their movie viewing history in order to receive movie recommendations that are more tailored to their personal preferences. Under this assumption,

$$\mathbf{g}_i^{(u)} \approx \frac{\text{\# of movies watched by user } u \text{ that are in genre } i}{\text{total \# of movies watched by user } u}$$

Thus, we can interpret $\mathbf{g}^{(u)}$ as a representation of the movie genres user u likes to watch.

2.2.2 Clustering users based on genre preferences

For each user u, genre preferences $\mathbf{g}^{(u)}$ are clustered via k-means clustering. Silhouette analysis is used to select the optimal number of clusters k based on

how close each user's genre preference is to their cluster and how far each user's genre preference is from the next nearest cluster. Clustering genre preferences serves to group users with similar tastes, which allows us to make broader generalizations about each preference group.

Uniform Manifold Approximation and Projection (UMAP) is applied to the genre preferences vectors ($\mathbf{g}^{(u)}$) for reduce the number of dimensions to allow for visualization of the clusters in two dimensions. Due to memory constraints, PCA is first used to reduce the number of genre dimensions to an intermediate value before reducing it down to two dimensions using UMAP.

In order to better interpret the meaning of each cluster, we compute the entropy of the genre preference vectors for the users in each cluster and plot the distribution in order to visualize how wide/niche the typical user's preferences are when it comes to movie genres. A larger value for entropy indicates that a user has a preference vector closer to the uniform distribution. This would imply that the user has wide genre interests since they are likely to watch movies in many genres. Average entropy is computed across all users in each cluster to determine if certain clusters have more general preferences than other clusters.

2.2.3 Identifying the optimal target audience

After clustering users based off their genre preferences, we investigate how the fraction of viewers in each cluster affects a movie's profitability. We compute each movie's profit as $profit = gross\ revenue - budget$. For each movie m, we also calculate the fraction of users who rated that movie that fall in each cluster. For $c \in \{\text{comedy lovers}, \text{drama lovers}, \text{wide/eclectic}\}$,

$$\mathbf{v}_c^{(m)} = \frac{\#\text{ of users in cluster } c \text{ who rated movie } m}{\text{total } \#\text{ of users who rated movie } m}$$

Relying on the assumption that MovieLens users who rate a given movie are an unbiased sample of people who watched the movie, we interpret $\mathbf{v}_c^{(m)}$ as the fraction of people who watched movie m that fall in cluster c. We use "audience composition" and "viewership breakdown" to refer to

$$\mathbf{v}^{(m)} = [\mathbf{v}_{\text{comedy}}^{(m)}, \mathbf{v}_{\text{drama}}^{(m)}, \mathbf{v}_{\text{wide/eclectic}}^{(m)}]$$

To determine the specific relationship between audience composition and profit, we put each movie into one of ten evenly-spaced bins depending on its $\mathbf{v}_{\text{comedy}}^{(m)}$. We then compute the average profit for all movies in each bin and plot the results. We repeat this process using $\mathbf{v}_{\text{drama}}^{(m)}$ and $\mathbf{v}_{\text{wide/eclectic}}^{(m)}$.

For a more detailed investigation, we perform a similar analysis except using a two-dimensional grid of bins, where one axis is $\mathbf{v}_{\text{comedy}}^{(m)}$ and the other axis is $\mathbf{v}_{\text{wide/eclectic}}^{(m)}$. We then estimate profit as a function of $\mathbf{v}_{\text{comedy}}^{(m)}$ and $\mathbf{v}_{\text{wide/eclectic}}^{(m)}$ by computing the average profit for all movies in each bin. Note that since

$\mathbf{v}^{(m)}_{\text{drama}} = 1 - \mathbf{v}^{(m)}_{\text{comedy}} + \mathbf{v}^{(m)}_{\text{wide/eclectic}}$, this approach is equivalent to estimating profit as a function of audience breakdown. After plotting the results, we are able to determine the audience breakdown that maximizes expected profit. We validate the statistical significance of our findings using Welch's t-test.

Finally, we investigate whether the optimal audience breakdown varies for different movie genres. For each genre g, we perform the 2-D binning approach from above but only using movies from genre g.

2.3 Results

2.3.1 User clusters

Figure 1 depicts the three main preference clusters among MovieLens users. k-means clustering was applied for $k \in [2, 20]$, and silhouette analysis was used to select the optimal cluster parameter $k = 3$.

PCA was used to first reduce the 20-dimensional genre preference vectors down to 10 dimensions. UMAP was subsequently used to reduce genre preference vectors down to 2 dimensions for visualization. The lack of spacing between clusters is hypothesized to result from information loss due to dimensionality reduction.

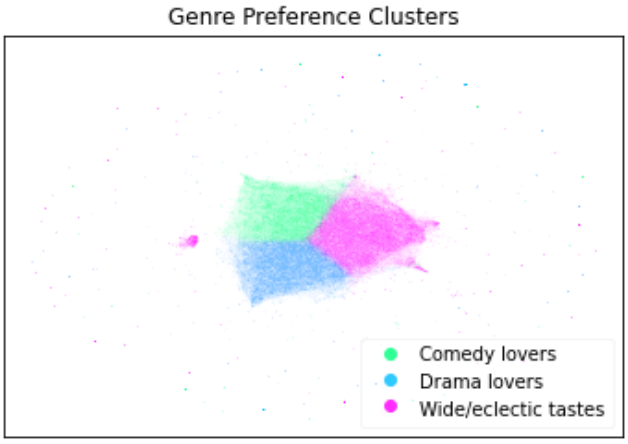

Figure 1: This figure depicts the three viewer clusters based on genre preferences: Comedy lovers, Drama lovers, and those with wide/eclectic tastes.

The centroids of each cluster are used to interpret the meaning of each cluster.

As shown in Figure 2 below, the gray line, which plots average genre preference across all users, has the highest peaks at 3 genres: drama, comedy, and action/adventure. The centroid of the green cluster depicts a strong preference for comedies with nearly 28% of all user reviews conducted on comedy films. Members of this cluster are assumed to have a strong affinity for comedies relative to the average user. The centroid of the blue cluster reveals a strong preference for dramas with roughly 33% of all user reviews conducted on drama films. These individuals are presumed to watch mostly dramas. The centroid of the magenta cluster doesn't reveal a strong preference for any given genre, though it does have a higher peak at action/adventure films than the average genre preference line. As such, the magenta cluster is considered to represent individuals of varied interests. We refer to viewers in the magenta cluster as viewers with wide/eclectic tastes.

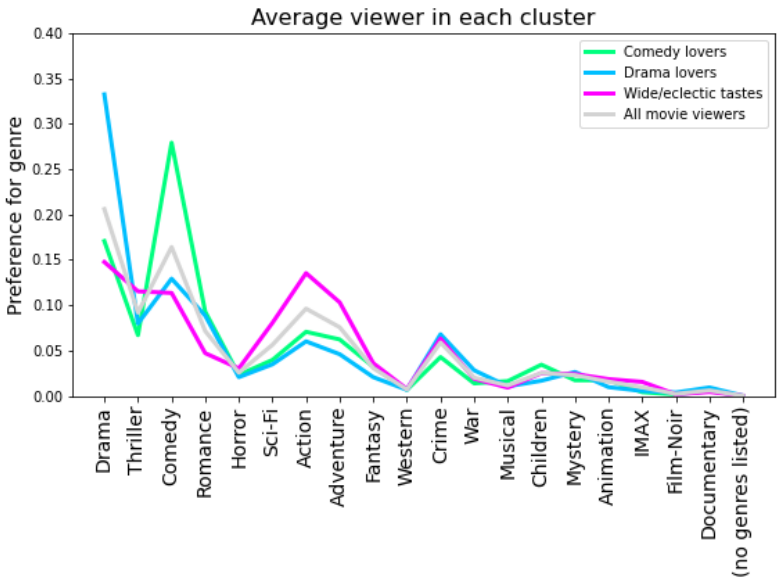

Figure 2: The genre preferences at the centroid of each cluster, as well as the average genre preferences for all MovieLens users.

We also consider the entropy of the genre preference vectors in each cluster to support our characterizations above. As shown in Figure 3, the entropy distributions for each of the clusters are generally skewed towards users with narrow genre preferences (low entropy). Drama and Comedy clusters have a very similar distribution since they are predominantly composed of users with relatively niche interests. The Wide preference cluster has a much higher proportion of individuals with higher entropy, which corresponds to more varied

interests.

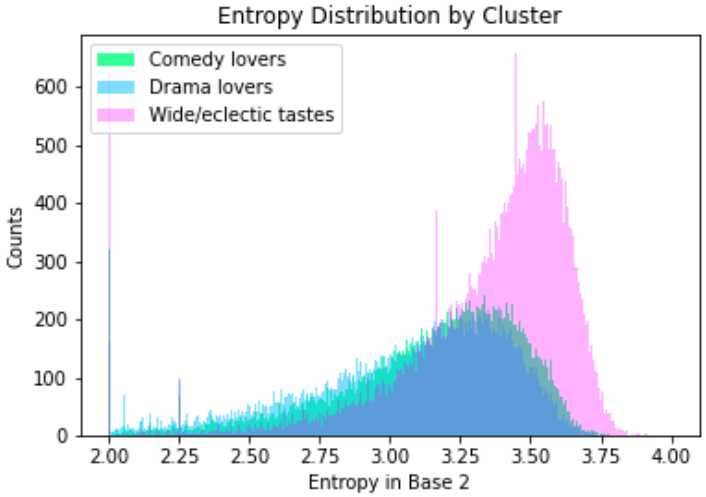

Figure 3: This figure depicts the entropy distribution across the three preference clusters. Note that the blue portion of the histogram is where all three histograms intersect. The Drama cluster histogram is actually behind the Comedy histogram and has a very similar shape.

2.3.2 What is the ideal target audience?

Figure 4 illustrates the relationship between a movie's audience breakdown and its profit. In the previous section, we identified three primary types of movie viewers: comedy lovers, drama lovers, and viewers with wide/eclectic tastes. Figure 4 helps us see how a movie's profitability is related to the fraction of its viewers that fall into each of these three archetypes.

Focusing first on the green line, we see that the fraction of viewers who are comedy lovers does not have a strong relation to a movie's profitability, although intermediate values (0.1-0.5) seems to result in higher profits on average.

Turning our attention to the blue line, we see that movies tend to do best if drama lovers comprise approximately 15% of their audience. Appealing to too many drama lovers can be detrimental to a movie's profitability.

Finally, when we consider the magenta line, we see that appealing to viewers who have wide or eclectic movie taste tends to increase profits. However, catering too much to those with wide/eclectic tastes can also be risky, as we see that movies with a wide/eclectic fraction 0.7 or greater have confidence intervals

that encompass a wide range of values.

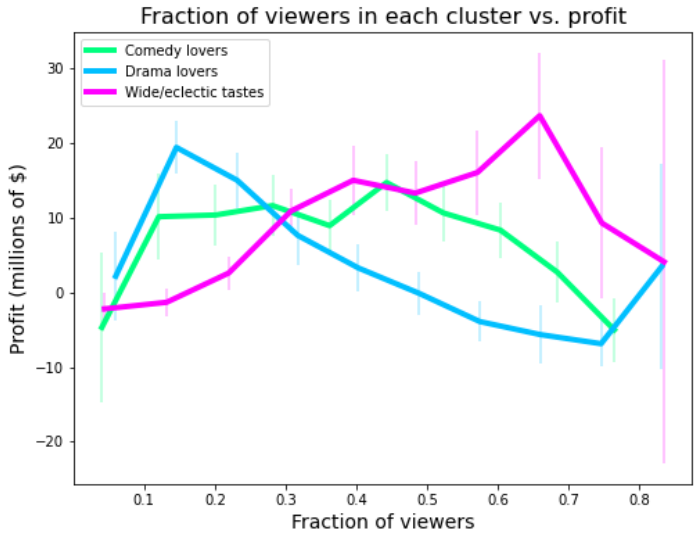

Figure 4: This figure illustrates how a movie's viewership relates to its profitability. The error bars represent a 95% confidence interval.

Figure 5 provides further insight into the relationship between the composition of a movie's viewership and its profitability. Whereas each line in Figure 4 is computed by binning movies based on the fraction of viewers in a single cluster, Figure 5 first bins movies based on the fraction of viewers in the wide/eclectic cluster and then further separates each of those bins by the fraction of viewers who are drama lovers. Note that since the sum of the fraction of viewers in each of the three clusters must equal 1, knowing the fraction of viewers in the wide/eclectic cluster and the fraction of viewers who are drama lovers fully specifies a movie's viewership composition. The fraction of viewers who are comedy lovers can be computed as 1 minus the sum of the other two fractions.

Figure 5 tells us that movies with **an audience comprised of 60% viewers with wide/eclectic tastes, 20% drama lovers, and 20% comedy lovers result in the highest profit on average.** This result is statistically significant ($p < 0.001$) using Welch's t-test between movies close [1] to the optimal audience breakdown and all other movies.

Does the ideal target audience depend on the genre of the movie being made?

[1] We define 'close' to mean within 5% of the optimal fraction for each cluster.

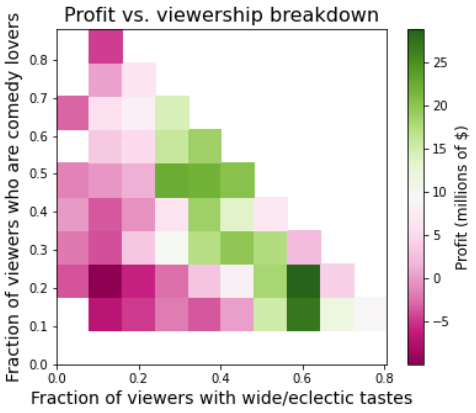

Figure 5: Average profit as a function of a movie's viewership breakdown. Bins with fewer than 20 movies are excluded.

Figure 6 tells us that the answer is yes. The audience composition that results in the highest profit shifts depending on the genre of the movie. Based on the plots in Figure 6, we make the following recommendations:

1. Producers of *drama* movies should consider **adding humor** to their movie in order to appeal more to comedy lovers, as our analysis shows that drama movies with an audience breakdown of 35% wide/eclectic, 45% comedy lovers, and 20% drama lovers yield the highest profit. Drama movies that cater primarily to drama lovers (corresponding to the bottom left corner of the plot), actually yield negative profit on average.

2. Producers of *comedy* movies should **focus on incorporating elements from other genres**. We find that comedy movies with audiences containing a lower fraction of comedy lovers tend to yield higher profits. Comedy movies that appeal only to comedy lovers (likely because these movies focus solely on humor at the expense of a compelling plot and other fulfilling movie elements) often tank at the box office. We find that the optimal audience breakdown for comedy movies is 50% wide/eclectic, 25% comedy lovers, and 25% drama lovers.

3. Producers of *romance* movies should **strike a balance between comedy and drama** in order to appeal to the optimal audience of 30% viewers with wide/eclectic tastes, 45% comedy and 25% drama lovers. Comedic romance movies ("rom-coms") have a wide following, but romance movies that appeal too heavily to comedy lovers are at risk of hurting their profits.

4. Producers of *sci-fi* movies should **err on the serious side** and limit the comedic elements, as the sci-fi movies that appeal to fewer comedy lovers

have higher profits on average. The optimal audience for sci-fi movies is 60% wide/eclectic, 10% comedy lovers, and 30% drama lovers.

5. Producers of *thriller, action, adventure, crime* movies should **target** audiences similar to the audience we identified as the optimal audience breakdown we identified in our genre-agnostic analysis from above (**60% viewers with wide/eclectic tastes, 20% drama lovers, and 20% comedy lovers**)

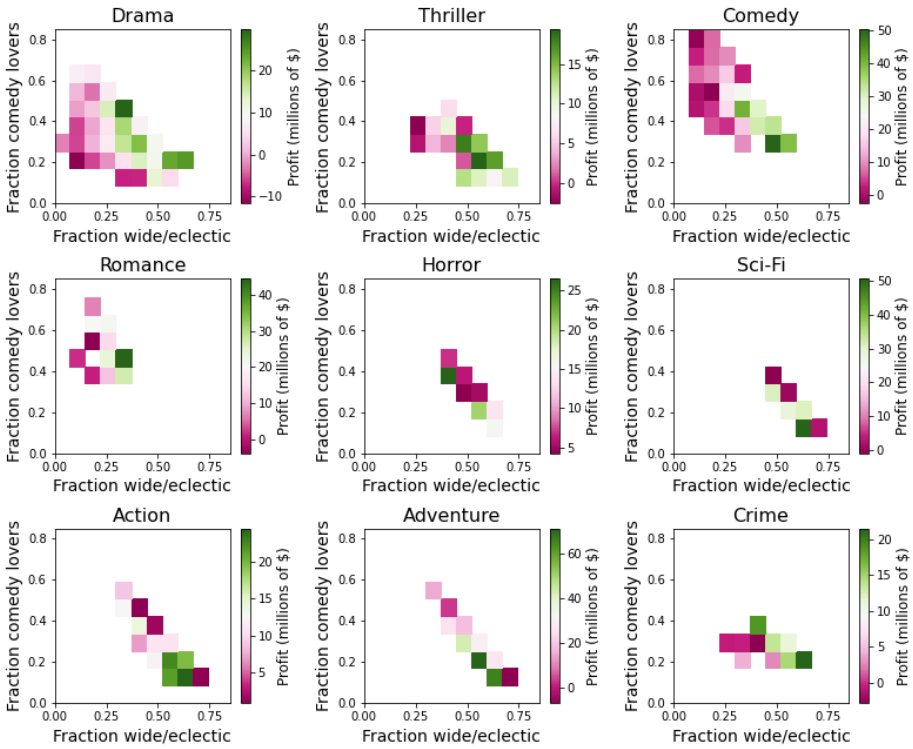

Figure 6: Average profit as a function of a movie's viewership breakdown for nine common movie genres. Bins with fewer than 20 movies are excluded.

2.4 Discussion

Our data analysis approach has several key advantages:

1. We do not rely on user ratings, which have numerous problems. The first problem is the way in which they are collected. To assess data quality, we created a MovieLens account so we could understand how the data

was collected. We found that before a user enters in their own rating of a movie, they are shown the average ratings given by other users and also a prediction of what MovieLens thinks the user will rate the movie. We believe that seeing these ratings injects an anchoring bias[2] into the rating that the user subsequently provides, causing it to be an inaccurate reflection of how much the user enjoyed the movie. The second problem is that each user has their own calibration of the rating scale (e.g., Person A may only give a rating of `3` to movies they think were truly terrible, whereas Person B may give a rating of `3` to movies that they think are mediocre.) Rather than deal with the problems associated with the rating numbers themselves, we use a more robust approach by only using the *existence* of ratings to compute our metrics. This approach relies on the assumptions that users tend to watch more movies in the genre that they prefer, and user ratings are an unbiased sample of the movies they watch. These assumptions are further tested in 4.1.

2. Our approach allows us to discover cross-genre clusters. A naive approach would be to group users by their favorite genre, but we recognized that people's movie preferences are more complex than that. Some genres are very similar to each other (e.g. Action and Adventure) whereas others are very different. By considering user preferences over *all* genres, we are able to identify the primary axes of variability in genre preferences.

3. We offer specific advice to movie executives for producing more profitable movies given their genre choices.

4. Our analysis offers something that MovieLens doesn't already do. Rather than try to predict movies that a user will enjoy, which is what MovieLens is designed to do, we repurpose the MovieLens dataset in order to provide useful insights to movie executives about ideal target audiences.

Directions for future work include:

1. Validating the above results using a different embedding space for clustering users. Clustering users based off their genre preference vectors results in well-defined and interpretable clusters, though our analysis could be strengthened by incorporating a more rating-specific approach that adequately controls for user cognitive biases.

2. Obtaining other user-specific data including age and other demographic characteristics could be useful for generating more meaningful clusters that control for specific user characteristics. This would help solidify recommendations for targeting age-specific audiences.

[2]For more information, see the work of Amos Tversky and Daniel Kahneman

3 Conclusion

In this analysis, we clustered user preferences to form three user preference groups: Comedy lovers, Drama lovers, and those with wide/eclectic tastes. We then used profitability as a metric for determining optimal audience composition as a percentage of each cluster. We concluded our analysis with genre-specific guidelines for maximizing profit by targeting the appropriate audience.

We envision our guidelines being used by movie executives at three points in the production process. First, these metrics can be used to assess the profitability of a given movie idea and whether it would be in the executive's best interest to pursue it. Second, when determining movie content, executives should assemble focus groups with a composition that matches the ideal target audience we identified above. Third, when advertising the movie, executives should split their budget so that comedy lovers, drama lovers, and movie viewers with wide tastes are reached in the same proportion as the ideal target audience.

In conclusion, we believe that our analysis will be invaluable to those in the movie industry seeking to take a more data-driven approach to audience targeting.

4 Appendix

4.1 Robustness check #1

To test our assumption that $\mathbf{g}^{(u)}$ reflects user u's genre preferences, we do the following: First, we compute a vector $\mathbf{r}^{(u)}$ where the i^{th} entry ($\mathbf{r}_i^{(u)}$) is the average rating given by user u to movies in genre i. Then we perform a linear regression of $\mathbf{r}_i^{(u)}$ on $\mathbf{g}_i^{(u)}$ for all genres i and all users u. The resulting coefficient is $\beta = 10.44 \pm 0.01$ (p-value = 0.0). This strong positive correlation between the fraction of movies that a user watches from a given genre and the average rating that the user gives to movies in that genre supports our use of $\mathbf{g}^{(u)}$ as a measure of what movie genres user u enjoys watching.

4.2 Robustness check #2

Figure 4 shows a positive relationship between the fraction of viewership in the wide/eclectic tastes cluster and a movie's profit. A skeptical reader may question if movie quality is a confounding factor (i.e., viewers with wide/eclectic tastes are more willing to watch movies if they are higher quality and higher quality movies tend to make more money). However, there are two arguments against this. First, if movie quality was a confounding factor, we wouldn't expect the trend to turn downward when the cluster fraction exceeds 0.65. Second, we performed additional analysis and found that there is no statistically significant relationship between the fraction of viewership in the wide/eclectic tastes cluster

and the IMDb rating of the movie, which we view as a proxy for movie quality.

4.3 Robustness check #3

A skeptical reader may also question whether trends in movie profitability are driven by budget and genre rather than by audience composition. Using a linear regression t-test, we have determined that there is a significant association ($p < 0.01$) between audience composition and movie profitability given budget and genre as control factors. This provides the base for our recommendations to movie producers across different genres.