Navigating the Cinematic Maze: A Data-Driven Approach to Optimal Release Dates for *The Maze Runner*

Sylvanne Braganza (sbraganz)

Yukta Butala (ybutala)

Clarissa Gunawan (cgunawan)

Kelly Hong (sewonh)

Cece Zhang (xiyuzhan)

Analytical Marketing

January 29, 2024

In helping the production company Gotham Group with the post-production of *The Maze Runner*, we have been tasked to consider seasonality and competition to decide on an optimal release date. Given a dataset of movie characteristics for all major releases in the US domestic market from 2006 through 2014, as well as their release dates, we can perform k-means clustering and topic modeling to analyze and understand the underlying themes or topics present in this collection of movies and consider the best windows for a release.

The 10 topics derived from our LDA analysis are atmospheric, action, animation, fantasy, sci-fi, superhero, revenge, survival, romance, and comedy. This can be interpreted from movie tags for the movies sorted into each topic. We can also consider genre and creative type. (Appendix 1.1) *The Maze Runner* belongs to the Survival topic. This topic's top labels are "post-apocalyptic" and "dystopia", overlapping with *The Maze Runner*'s description's top words on these as well as "survival." (Appendix 1.2)

When choosing a release date for our target movie, *The Maze Runner*, we must consider similarities to other movies that may be released around the same time. The studio wants to avoid releasing the movie in a week in which it is too "similar" to others to avoid competition at the box office. To quantitatively consider similarity, we came up with a method for calculating a "similarity score" between *The Maze Runner* and every other movie. We computed the "Euclidean distance" between the 10-dimensional topic score for *The Maze Runner* and every other movie. This informed us about how closely related *The Maze Runner* is to other movies based on the identified topics. A lower Euclidean distance indicates a higher similarity, suggesting that movies with similar thematic elements are closer in the 10-dimensional topic space. Based on this measure, the top ten most similar movies to *The Maze Runner* are *The Twilight Saga: New Moon*, *Daybreakers*, *28 Weeks Later*, *The Conjuring*, *Underworld: Evolution*, *1408*, *Insidious*, *The Hunger Games: Catching Fire*, *Doomsday*, and *Resident Evil: Extinction*. (Appendix 2.1) We can validate this model, confirming it makes sense, by considering metrics such as: "genre," "creative_type," "source," "rating," "production_method," and "budget." (Appendix 2.2) Since these movies were found to be similar, they should fall within a similar range of values and categories for these metrics. For example, *The Maze Runner* is a thriller/suspense genre; other movies were classified as horror, action, drama, and adventure, which can be considered to be within a similar scope in terms of genre. The Maze Runner is Science Fiction, and the other movies all align to a similar fictional creative type, including Science Fiction, Fantasy, and Contemporary Fiction. All of these movies share a Live Action production method. Based on these identifiers, the model does make sense, and so we can build on it to determine the optimal week for the release of our movie.

After conducting a similarity score analysis between *The Maze Runne*r and all other movies, we refined our focus to examine films released in 2014, aligning with the launch year of our target movie. We categorized all release dates into weeks, excluding those without any movie launches, resulting in a dataset comprising 50 weeks. When deciding on a metric for average similarity for each week, we weighed two options: averaging the similarity scores across all movies within a week or selecting the closest movie each week. We opted for the latter for the following reasons: (1) opting for the closest movie effectively minimizes the impact of films extremely dissimilar from *The Maze Runner*; (2) we assumed that people wouldn't attend movies of the same genre in a week.

To consider movie screening for more than one week, we assumed a two-week screening period in our analysis. Using the similarity scores obtained for each week, we lagged the data by one week to account for movies launched in the previous week (Appendix 3.1) The resulting dataset included a new column of lagged similarity scores. Subsequently, we plotted a line graph of both regular and lagged similarity scores on weeks. (Appendix 3.2) Using this graph, we were able to identify potential weeks for the launch of *The Maze Runner* by pinpointing instances where both regular and lagged similarity scores were relatively higher compared to other weeks. This approach ensures that in a selected week, new movie launches are not overly similar to our target movie, and movies still screening from the previous week also exhibit dissimilarity, minimizing competition.

Based on the Euclidean distance analysis, we recommend Weeks 14, 28, and 38 (Week ending April 4, July 11, and September 19, respectively) (Appendix 4.1). Week 14 is recommended as it has the highest minimum and average Euclidean distance compared to other weeks, indicating it has the most dissimilar movies in that week. The competing movies that week are *Sabotage* and *Captain America - The Winter Soldier,* which is an Action/Contemporary Fiction and Action/Superhero movie, respectively. Week 28 is also recommended considering that there are no movies released during that week. The competing movie that week would be from Week 27, which has a relatively high average Euclidean distance of 0.617. The competing movies would be *Earth to Echo*, *Tammy*, and *Deliver Us from Evil*, which are Adventure/Animated/Kids Fiction, Comedy/Contemporary Fiction, and Horror/Fantasy movies, respectively. Lastly, Week 38 is recommended considering the high relative minimum and weekly average Euclidean Distance, and also making it eligible for the “Awards Season”. The competing movies for the week would be *No Good Deed*, *Dolphin Tale 2*, *A Walk Among the Tombstone*s, and *This is Where I Leave You*, which is Thriller/Contemporary Fiction, Drama/Kids Fiction, Action/Historical Fiction, and Comedy/Contemporary Fiction, respectively. Analyzing the competing movies, it makes sense to release *The Maze Runner* in any of the 3

weeks due to the stark contrast between the competition and the movie being Thriller/Science-Fiction. It is also worth noting that these weeks do not fall within January and February, the first and second week of August, early September, and the week after Thanksgiving, as these tend to be slower months at the box office.

Our recommendation is based on the LDA method, so we want to assess the robustness of this method. When the goal is to group movies based on their latent topics and allow for the possibility that a movie can belong to multiple topics, as is the case here, then LDA may be the most suitable method. LDA provides a nuanced representation of movies in terms of topic distribution. However, if the focus is on creating distinct clusters where each movie is assigned to a single group, k-means or hierarchical clustering might be more appropriate. K-means is computationally efficient and may work well when clusters are well-separated, so we chose to compare the topic model result with this method for our assessment. We conducted K-means clustering using the movie genres in the LDA analysis for the same task. Our findings revealed 2 key reasons that support the appropriateness of LDA.

For K-means clustering, we set the number of clusters (k) to 10. Although the within-cluster sum of squares (wss) indicated k=9 as optimal, we assumed that the manual labeling used in LDA was performed by a movie professional. This implies that the number of topics was carefully chosen according to industry standards, leading us to set our k to 10. Subsequently, we associated each cluster with its respective genres (Appendix 5.1). However, the resulting clusters exhibited overlaps in movie types; clusters 2 and 7 were treated as separate clusters, but both predominantly consisted of horror movies. This observation indicates that the clusters do not effectively represent the diverse topics present in movies. Also, upon conducting a cross-tabulation between the results of K-means clustering and the genres in the dataset (Appendix 5.2), we observed that cluster 4 has diverse genres. This implies that it may not adequately represent distinct movies in the cluster.

Upon examining the distribution of movies across topics in LDA with 10 topics and K-means with 10 clusters (Appendix 5.3), the LDA shows a more balanced distribution of movies, evenly allocating them across the identified topics. Conversely, K-means clusters displayed a lumpier distribution, concentrating movies within specific clusters and resulting in a less uniform representation. The imbalance observed in K-means clusters reinforces our preference for the LDA model. The even distribution of movies across LDA topics highlighted its capacity to capture a diverse range of characteristics, contrasting with the concentrated grouping observed in K-means. This finding underscores LDA's nuanced and comprehensive representation, solidifying our bias toward favoring the LDA model for its ability to provide a more balanced representation of movie characteristics.

Appendices

Appendix 1.1

Movie tags for each of the 10 movies in each of the 10 topics can help us interpret each topic space:

	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
1	#N/A	action	animation	based on a book	sci-fi	superhero	revenge	dystopia	based on a true	comedy
2	visually appealin	espionage	pixar	fantasy	aliens	comic book	johnny depp	post-apocalyptic	true story	funny
3	atmospheric	stupid	funny	magic	time travel	marvel	quentin tarantin	zombies	romance	drugs
4	alternate reality	assassin	disney	remake	action	action	brad pitt	horror	drama	dark comedy
5	leonardo dicapri	james bond	talking animals	adventure	space	robert downey jr	violence	vampires	multiple storylin	emma stone
6	surreal	unrealistic	adventure	police	social commenta	stylized	bruce willis	predictable	denzel washingt	satire
7	cinematography	conspiracy	friendship	fairy tale	robots	based on a comi	violent	survival	russell crowe	high school
8	christian bale	robert downey jr	computer anima	franchise	special effects	scarlett johanss	world war ii	bad acting	ben affleck	seth rogen
9	thought-provokir	martial arts	computer anima	franchise	future	will ferrell	tim burton	religion	chick flick	nudity (topless)
10	dark	murder	cute	adapted from:bc	adventure	visually appealin	gore	cliche	sports	hilarious

Considering genre (top) and creative type (bottom) for each of the 10 movies in each of the 10 topics can help us qualitatively consider and further interpret each topic space:

	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
1	Thriller/Suspense	Action	Adventure	Adventure	Action	Adventure	Action	Horror	Drama	Comedy
2	Thriller/Suspense	Action	Thriller/Suspense	Adventure	Action	Action	Western	Drama	Thriller/Suspense	Comedy
3	Action	Action	Adventure	Adventure	Thriller/Suspense	Adventure	Musical	Action	Drama	Comedy
4	Thriller/Suspense	Adventure	Adventure	Adventure	Action	Action	Adventure	Thriller/Suspense	Drama	Comedy
5	Action	Thriller/Suspense	Thriller/Suspense	#N/A	Action	Action	Action	Adventure	Drama	Thriller/Suspense
6	Action	Thriller/Suspense	Musical	Adventure	Adventure	Comedy	Adventure	Action	Action	Romantic Comedy
7	Drama	Drama	Adventure	Comedy	Adventure	Action	Western	Drama	Thriller/Suspense	Comedy
8	Drama	Action	Drama	Adventure	Action	Action	Black Comedy	Horror	Drama	Drama
9	Thriller/Suspense	Action	Adventure	Adventure	Action	Drama	Musical	Adventure	Romantic Comedy	Comedy
10	Drama	#N/A	Thriller/Suspense	Adventure	Thriller/Suspense	Action	Thriller/Suspense	Horror	Drama	Comedy

1	Science Fiction	Contemporary Fiction	Fantasy	Fantasy	Science Fiction	Super Hero	Historical Fiction	Science Fiction	Dramatization	Contemporary Fiction
2	Historical Fiction	Contemporary Fiction	Kids Fiction	Fantasy	Science Fiction	Contemporary Fiction	Historical Fiction	Science Fiction	Historical Fiction	Contemporary Fiction
3	Contemporary Fiction	Contemporary Fiction	Kids Fiction	Fantasy	Science Fiction	Super Hero	Historical Fiction	Science Fiction	Dramatization	Contemporary Fiction
4	Historical Fiction	Historical Fiction	Kids Fiction	Fantasy	Science Fiction	Fantasy	Fantasy	Science Fiction	Dramatization	Contemporary Fiction
5	Super Hero	Contemporary Fiction	Kids Fiction	Fantasy	Science Fiction	Super Hero	Contemporary Fiction	Science Fiction	Dramatization	Fantasy
6	Super Hero	Contemporary Fiction	Fantasy	Fantasy	Science Fiction	Contemporary Fiction	Historical Fiction	Fantasy	Historical Fiction	Contemporary Fiction
7	Science Fiction	Contemporary Fiction	Kids Fiction	Contemporary Fiction	Science Fiction	Super Hero	Historical Fiction	Fantasy	Contemporary Fiction	Contemporary Fiction
8	Contemporary Fiction	Contemporary Fiction	Fantasy	Fantasy	Science Fiction	Contemporary Fiction	Contemporary Fiction	Science Fiction	Dramatization	Dramatization
9	Contemporary Fiction	Contemporary Fiction	Kids Fiction	Science Fiction	Science Fiction	Super Hero	Fantasy	Fantasy	Contemporary Fiction	Contemporary Fiction
10	Historical Fiction	#N/A	Contemporary Fiction	Fantasy	Science Fiction	Contemporary Fiction	Contemporary Fiction	Science Fiction	Dramatization	Contemporary Fiction

For example, Topic 1, "Atmospheric" represents movies that emphasize mood and setting, encompassing genres such as Suspense, Drama, and Action as well as Historical Fiction, Contemporary Fiction, and Super Hero creative types. Topic 3, "Animation" represents movies that emphasize family-friendly stories and techniques, encompassing genres such as Adventure and Musical/Drama as well as Kids Fiction or Fantasy creative types. Topic 8, "Survival" represents movies that emphasize extreme situations, encompassing genres such as Horror, Drama, Action, Suspense, and Adventure and Science Fiction and Fantasy creative types.

The Maze Runner's genre and creative type:

The Maze Runner	Thriller/Suspense	Science Fiction

Appendix 1.2

The Maze Runner description has the following words:

Action	6
Amnesia	7
based on a book	7
Dystopia	17
plot holes	17
Post-apocalyptic	10
Survival	13
Teen	6

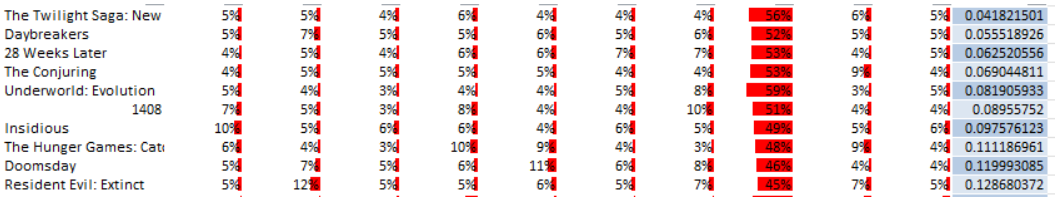
Topic 8, or the "Survival" topic, has the following top labels:

apocalypse	2%
bad acting	2%
cliche	2%
dystopia	8%
horror	2%
plot holes	2%
post-apocalyptic	8%
predictable	2%
religion	2%
survival	2%
vampires	2%
will smith	2%
zombies	5%

Appendix 2.1

Calculated Euclideon Distance:

Movie	Euclidean Distance
The Twilight Saga: New	0.041822
Daybreakers	0.055519
28 Weeks Later	0.062521
The Conjuring	0.069045
Underworld: Evolution	0.081906
1408	0.089558
Insidious	0.097576
The Hunger Games: Catc	0.111187
Doomsday	0.119993
Resident Evil: Extinct	0.128680



The Twilight Saga: New	5%	5%	4%	6%	4%	4%	4%	56%	6%	5%	0.041821501
Daybreakers	5%	7%	5%	5%	6%	5%	6%	52%	5%	5%	0.055518926
28 Weeks Later	4%	5%	4%	6%	6%	7%	7%	53%	4%	5%	0.062520556
The Conjuring	4%	5%	5%	5%	5%	4%	4%	53%	9%	4%	0.069044811
Underworld: Evolution	5%	4%	3%	4%	4%	5%	8%	59%	3%	5%	0.081905933
1408	7%	5%	3%	8%	4%	4%	10%	51%	4%	4%	0.08955752
Insidious	10%	5%	6%	6%	4%	6%	5%	49%	5%	6%	0.097576123
The Hunger Games: Cat	6%	4%	3%	10%	9%	4%	3%	48%	9%	4%	0.111186961
Doomsday	5%	7%	5%	6%	11%	6%	8%	46%	4%	4%	0.119993085
Resident Evil: Extinct	5%	12%	5%	5%	6%	5%	7%	45%	7%	5%	0.128680372

Appendix 2.2

Title	genre	creative type	source	rating	production_method
The Twilight Saga: New I	Drama	Fantasy	Based on Fiction Book/Short Story	PG-13	Live Action
Daybreakers	Horror	Science Fiction	Original Screenplay	R	Live Action
28 Weeks Later	Horror	Science Fiction	Original Screenplay	R	Live Action
The Conjuring	Horror	Fantasy	Original Screenplay	R	Live Action
Underworld: Evolution	Action	Fantasy	Original Screenplay	R	Live Action
1408	Horror	Contemporary Fictic	Based on Fiction Book/Short Story	R	Live Action
Insidious	Horror	Fantasy	Original Screenplay	R	Live Action
The Hunger Games: Catc	Adventure	Science Fiction	Based on Fiction Book/Short Story	PG-13	Live Action
Doomsday	Action	Science Fiction	Original Screenplay	R	Live Action
Resident Evil: Extinction	Action	Science Fiction	Based on Game	R	Live Action
The Maze Runner	Thriller/Susp	Science Fiction	Based on Fiction Book/Short Story	PG-13	Live Action

Genre: *The Maze Runner* falls under Thriller/Suspense, aligning it with other "similar" movies categorized as Horror, Action, Drama, and Adventure. These are all within the same scope.

Creative type: *The Maze Runner* belongs to the Science Fiction creative type, alongside other movies that share the same classification of Sci-Fi, or similar fictional creative types such as Fantasy and Contemporary Fiction.

Source: *The Maze Runner* is based on a Fiction Book/Short Story, aligning it with the other “similar movies” that were mostly either also based on Fiction Books/Short Stories or Original Screenplay, a similar concept.

Rating: *The Maze Runner* is PG-13; all of the “similar” movies are rated either PG-13 or higher.

Production method: All of the movies, including *The Maze Runner*, are Live Action.

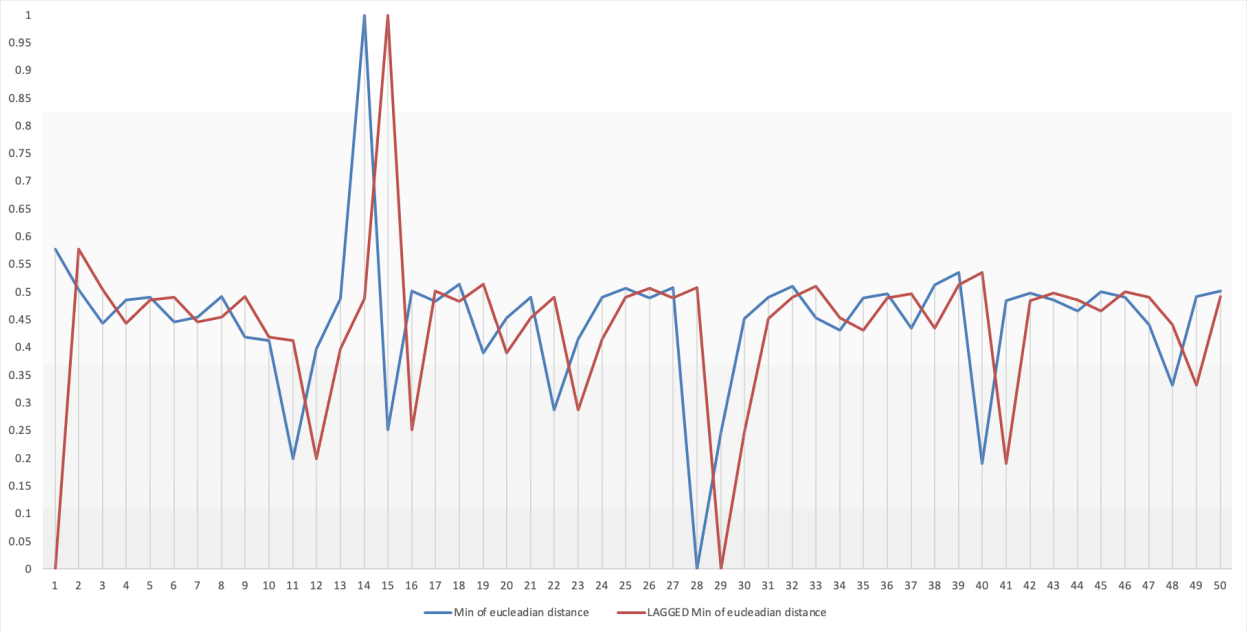
Appendix 3.1

Calculated Euclidean Distance between *The Maze Runner* and the most similar movie released each week as well as the lagged similarity score to take into account the previous week's releases:

Week number	Min of eucleadian distance	LAGGED Min of eucleadian distance
1	0.576818401	n/a
2	0.504657713	0.576818401
3	0.443097021	0.504657713
4	0.484930704	0.443097021
5	0.490371525	0.484930704
6	0.445680667	0.490371525
7	0.454631665	0.445680667
8	0.491971823	0.454631665
9	0.418152147	0.491971823
10	0.412057266	0.418152147
11	0.199043921	0.412057266
12	0.397479068	0.199043921
13	0.487977125	0.397479068
14	0.998463369	0.487977125
15	0.250885431	0.998463369
16	0.501062931	0.250885431
17	0.483681121	0.501062931
18	0.514420501	0.483681121
19	0.39041022	0.514420501
20	0.453059677	0.39041022
21	0.490846318	0.453059677
22	0.287502298	0.490846318
23	0.414636924	0.287502298
24	0.490608979	0.414636924
25	0.506335415	0.490608979
26	0.488708035	0.506335415
27	0.507330556	0.488708035
28	n/a	0.507330556
29	0.247163483	n/a
30	0.451504285	0.247163483
31	0.490371525	0.451504285
32	0.510809305	0.490371525
33	0.453359859	0.510809305
34	0.431039265	0.453359859
35	0.489845093	0.431039265
36	0.496641165	0.489845093
37	0.434098078	0.496641165
38	0.512327292	0.434098078
39	0.535367064	0.512327292
40	0.190415565	0.535367064
41	0.483931295	0.190415565
42	0.498487357	0.483931295
43	0.484930704	0.498487357
44	0.465709098	0.484930704
45	0.500182419	0.465709098
46	0.490846318	0.500182419
47	0.44143356	0.490846318
48	0.331191703	0.44143356
49	0.492294731	0.331191703
50	0.501329659	0.492294731

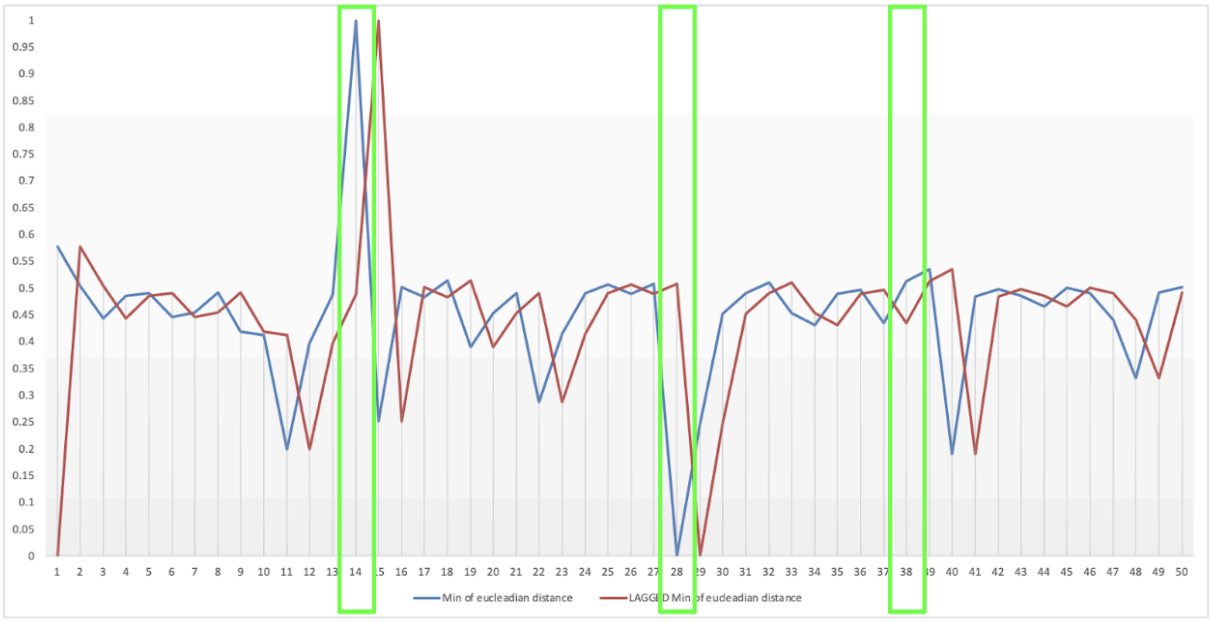
Appendix 3.2

Visual representation of Euclidean Distance between *The Maze Runner* and the most similar movie released that week, taking into account movies released the previous week, for easier weekly comparisons:



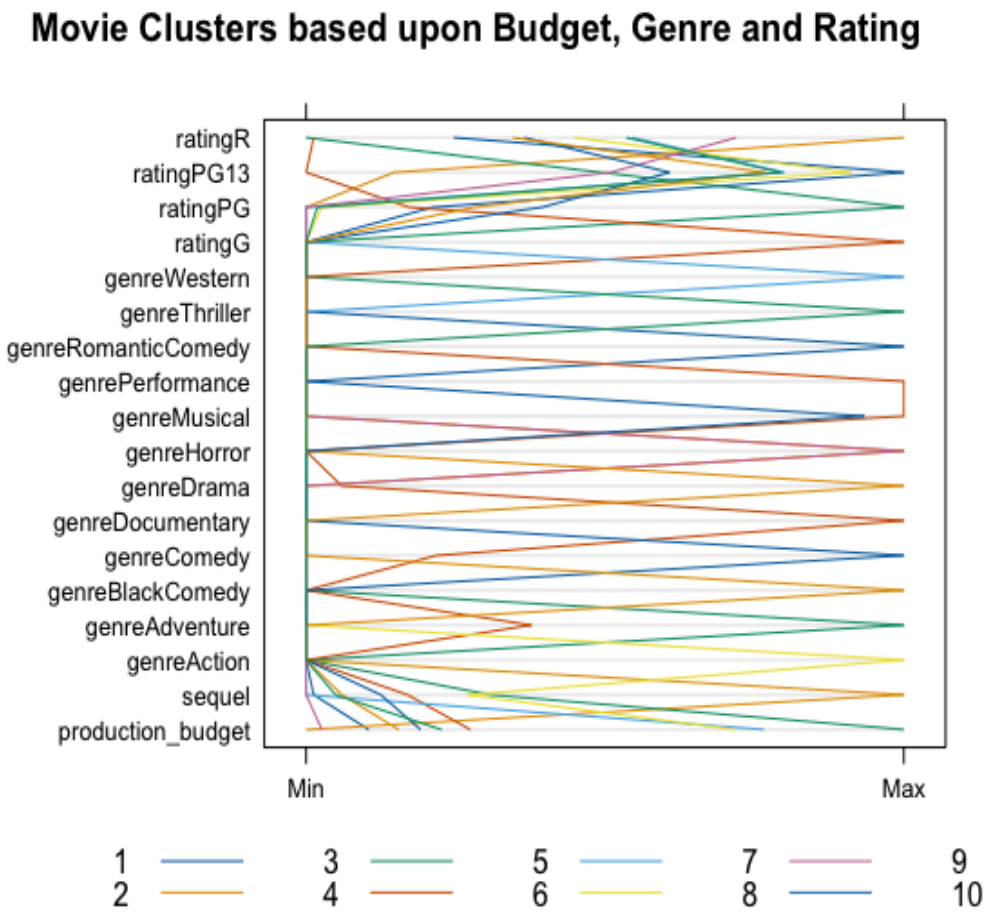
Appendix 4.1

Visual representation of Euclidean Distance between *The Maze Runner* and the most similar movie released that week, taking into account movies released the previous week, for easier weekly comparisons with our recommended weeks marked in green:



Appendix 5.1

Parallel graph of the K–means Clustering Results:



We defined the K-means clustering results as:

Cluster 1 Romantic comedy
Cluster 2 Horror
Cluster 3 Adventure
Cluster 4 Animation
Cluster 5 Western, Historical Fiction
Cluster 6 Action
Cluster 7 Horror
Cluster 8 Comedy
Cluster 9 Drama
Cluster 10 Thriller

Appendix 5.2

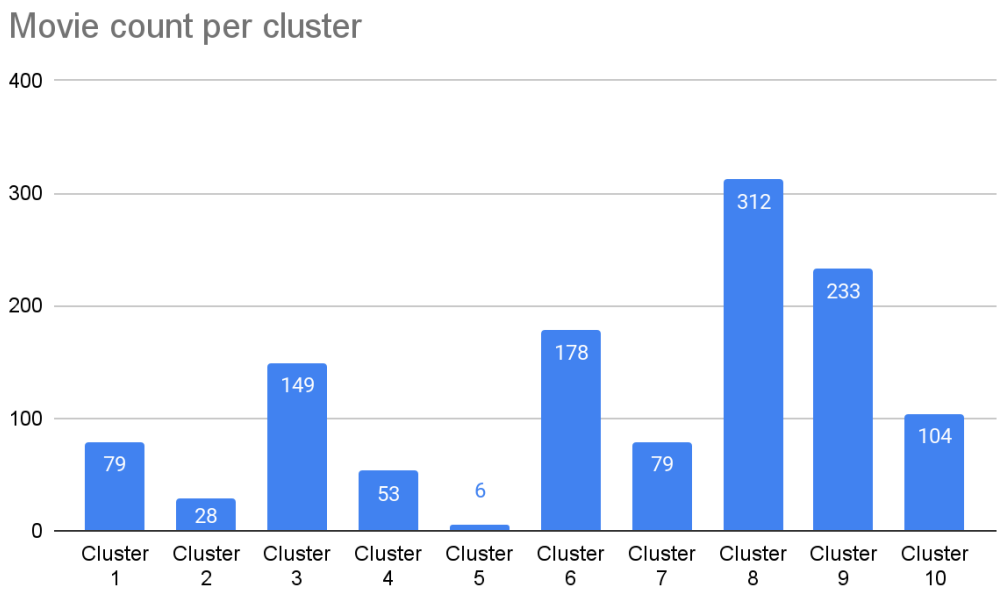
Cross Tabulation between the genres provided in the data and the clusters from K-means clustering:

	Cluster									
genre	1	2	3	4	5	6	7	8	9	10
	0	0	0	0	0	0	0	3	0	0
Action	0	0	0	0	0	178	0	0	0	0
Adventure	0	0	149	20	0	0	0	0	0	0
BlackComedy	0	0	0	0	0	0	0	0	11	0
Comedy	0	0	0	11	0	0	0	298	0	0
Documentary	0	0	0	9	0	0	0	0	0	0
Drama	0	0	0	3	0	0	0	0	222	0
Horror	0	28	0	0	0	0	80	0	0	0
Musical	0	0	0	2	0	0	0	11	0	0
Performance	0	0	0	8	0	0	0	0	0	0
RomanticComedy	79	0	0	0	0	0	0	0	0	0
Thriller	0	0	0	0	0	0	0	0	0	163
Western	0	0	0	0	6	0	0	0	0	0

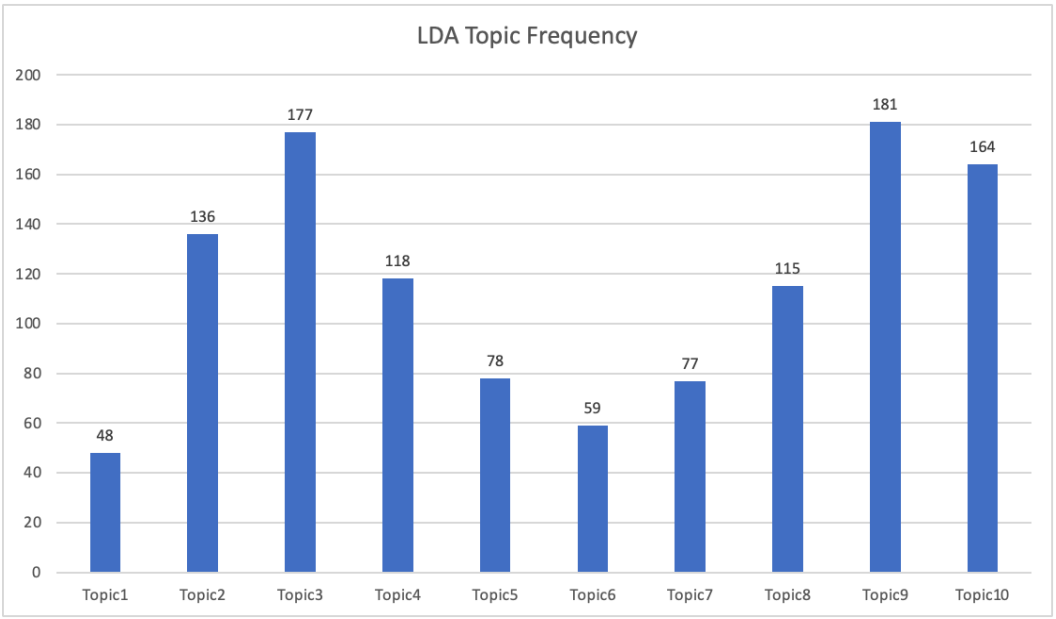
Appendix 5.3

Frequency charts for k-means and LDA analysis:

K - Means



LDA



Movies are categorized into topics by identifying the topic with the highest probability for each film.