Final Project Marketing Analytics

Ryan O’Connor, Alexandra Garthe, Paolo Arante, Adwika Narain, Jasmin Hedvat, Nidhi Gurrala

Presentation Agenda

Introduction

Slides #3-4

EDA

Slides #5-11

Methodology

Slides #12-15, 17-19, 21-24, 26-27

Test Implications

Slides #16,20,25,28

Analysis Results

Slides #29

Managerial Implications

Slides #30

Introduction

- IMDb is an online database that provides detailed information about millions of films and television programs
- Our data set takes 1000 of the most popular movies from IMDb from 2006 to 2016

We seek to answer:

- **Where is the movie industry headed? How can the movie industry optimize revenue and ratings through its selection of genres and other factors in the future?**

Advantages and Limitations of Our Dataset

- Why our dataset?
  - Our dataset allows us to analyze multiple factors that may affect movie success including: genre, recency and run time, and predict how these factors affect future movie trends
  - Our data includes several measures of movie success: votes, rating, revenue, and metascore, allowing us to measure movie success more holistically

- Potential limitations:
  - The production time of movies:
    - Because movies are produced over a span of a few years, it takes a couple of years to see if a trend has taken place in the movie industry
    - Our data only spans 10 years, so it may be limited in finding some trends

IMDb Movie Data

Dataset from Kaggle.com consisted of 12 attributes/columns and 1000 observations.

Rank - the ranking of each movie (from 1 to 1000)
Title - the title of each movie
Genre - the label each movie can be classified under. Each movie can have more than one genre label. There are 20 Genres and 195 different combinations shown in the data set
Description - the biography of each movie that highlights what the movie is about.
Director - the name of the main director of the movie
Actors - the names of the leading actors in the movie.
Year - the year in which the movie was made (from 2006-2016)
Runtime - the movie runtime (in minutes)
Rating - the movie's rating determined by IMDB registered users (on a scale from [1,10]
Votes - the number of IMDB registered users that provide a rating for a movie on IMDB
Revenue - the gross revenue of each movie in millions of dollars.
Metascore - the movie's rating determined by professional movie critics (on a scale from [1, 100]).

Preparing our Data - Operative Results

The raw IMDB movie data consisted of 12 attributes/columns and 1000 observations. Of these observations, 13.3% of our Dependent variable Revenue had NULL values. From this, we deleted the records with NULL revenue values in our dataset to be able to analyze 867 observations. Ratings had 0.

1	Rank	Title	Genre	Description	Director	Actors	Year	Runtime	Rating	Votes	Revenue	Metascore
2	1	Guardians of	Action,Adver	A group of in	James Gunn	Chris Pratt, V	2014	121	8.1	757074	333.13	76.00
3	2	Prometheus	Adventure,M	Following clu	Ridley Scott	Noomi Rapa	2012	124	7	485820	126.46	65.00
4	3	Split	Horror,Thrille	Three girls a	M. Night Shy	James McAv	2016	117	7.3	157606	138.12	62.00
5	4	Sing	Animation,Co	In a city of hu	Christophe Lo	Matthew Mc	2016	108	7.2	60545	270.32	59.00
6	5	Suicide Squa	Action,Adver	A secret gov	David Ayer	Will Smith, J	2016	123	6.2	393727	325.02	40.00
7	6	The Great W	Action,Adver	European me	Yimou Zhang	Matt Damon	2016	103	6.1	56036	45.13	42.00
8	7	La La Land	Comedy,Drar	A jazz pianis	Damien Chaz	Ryan Gosling	2016	128	8.3	258682	151.06	93.00
9	9	The Lost City	Action,Adver	A true-life dr	James Gray	Charlie Hunn	2016	141	7.1	7188	8.01	78.00
10	10	Passengers	Adventure,Di	A spacecraft	Morten Tyldu	Jennifer Law	2016	116	7	192177	100.01	41.00

Preparing our Data: Adding Columns for Genre & Recency

The IMDB movie data consists of 20 different genres that movies can be labeled under. We used binary values to code each record of being labeled with that genre (1) or not (0)

Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Horror

Mystery
Romance
Sci-Fi
Thriller
Western
Sport
History
War
Music
Musical

N	O	P	Q	R	S	T	U
Adventure	Biography	Comedy	Crime	Drama	Family	Fantasy	Horror
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1
0	0	1	0	0	1	0	0
1	0	0	0	0	0	1	0
1	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0
1	1	0	0	0	0	0	0
1	0	0	0	1	0	0	0

The raw IMDB movie data consists of movies made from 2006 to 2016. We encoded each record with numerical values in number of years from 2021 that the movie was made.

*2021-year movie was made = recency.

AG
Recency
7
9
5
5
5
5
5
5
5

Exploratory Data Analysis - Rows and Variables

```
> glimpse(data)
Rows: 867
Columns: 33
$ Rank        <int> 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 29, 30, 31, 32, 33, 34,…
$ Title       <chr> "Guardians of the Galaxy", "Prometheus", "Split", "Sing", "Suicide Squad", "The Great Wall", "La La Land", …
$ Genre       <chr> "Action,Adventure,Sci-Fi", "Adventure,Mystery,Sci-Fi", "Horror,Thriller", "Animation,Comedy,Family", "Actio…
$ Description <chr> "A group of intergalactic criminals are forced to work together to stop a fanatical warrior from taking con…
$ Director    <chr> "James Gunn", "Ridley Scott", "M. Night Shyamalan", "Christophe Lourdelet", "David Ayer", "Yimou Zhang", "D…
$ Actors      <chr> "Chris Pratt, Vin Diesel, Bradley Cooper, Zoe Saldana", "Noomi Rapace, Logan Marshall-Green, Michael Fassbe…
$ Year        <int> 2014, 2012, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016,…
$ Runtime     <int> 121, 124, 117, 108, 123, 103, 128, 141, 116, 133, 127, 133, 107, 109, 87, 139, 123, 118, 116, 120, 137, 92,…
$ Rating      <dbl> 8.1, 7.0, 7.3, 7.2, 6.2, 6.1, 8.3, 7.1, 7.0, 7.5, 7.8, 7.9, 7.7, 6.4, 6.6, 8.2, 6.7, 8.1, 8.0, 6.7, 7.9, 6.…
$ Votes       <int> 757074, 485820, 157606, 60545, 393727, 56036, 258682, 7188, 192177, 232072, 93103, 323118, 118151, 8612, 12…
$ Revenue     <dbl> 333.13, 126.46, 138.12, 270.32, 325.02, 45.13, 151.06, 8.01, 100.01, 234.02, 169.27, 532.17, 248.75, 2.87, …
$ Metascore   <int> 76, 65, 62, 59, 40, 42, 93, 78, 41, 66, 74, 65, 81, 70, 61, 71, 58, 69, 81, 49, 96, 56, 32, 60, 36, 39, 67,…
$ Action      <int> 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0,…
$ Adventure   <int> 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0,…
$ Biography   <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Comedy      <int> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,…
$ Crime       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Drama       <int> 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1,…
$ Family      <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Fantasy     <int> 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,…
$ Horror      <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,…
$ Mystery     <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Romance     <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Sci.Fi      <int> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0,…
$ Thriller    <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Western     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,…
$ Sport       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ History     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ War         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Animation   <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,…
$ Music       <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Musical     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Recency     <int> 7, 9, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 7, 5, 5, 5, 5,…
```

There are 867 observations

After preparing our data there are 33 columns

Exploratory Data Analysis - Data Types, Zeros, NULL

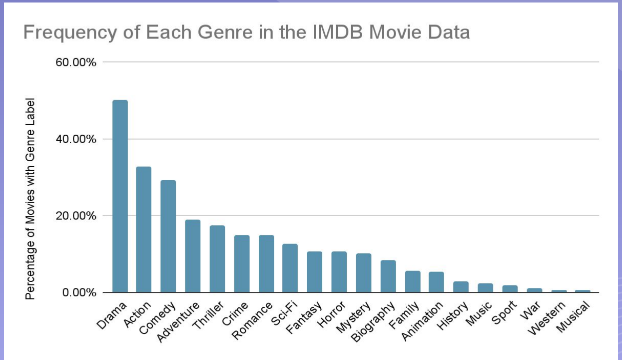
```
> df_status(data)
```

	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
1	Rank	0	0.00	0	0	0	0	integer	867
2	Title	0	0.00	0	0	0	0	character	866
3	Genre	0	0.00	0	0	0	0	character	195
4	Description	0	0.00	0	0	0	0	character	867
5	Director	0	0.00	0	0	0	0	character	535
6	Actors	0	0.00	0	0	0	0	character	863
7	Year	0	0.00	0	0	0	0	integer	11
8	Runtime	0	0.00	0	0	0	0	integer	91
9	Rating	0	0.00	0	0	0	0	numeric	50
10	Votes	0	0.00	0	0	0	0	integer	866
11	Revenue	0	0.00	0	0	0	0	numeric	810
12	Metascore	0	0.00	0	0	0	0	integer	82
13	Action	583	67.24	0	0	0	0	integer	2
14	Adventure	616	71.05	0	0	0	0	integer	2
15	Biography	794	91.58	0	0	0	0	integer	2
16	Comedy	612	70.59	0	0	0	0	integer	2
17	Crime	738	85.12	0	0	0	0	integer	2
18	Drama	431	49.71	0	0	0	0	integer	2
19	Family	818	94.35	0	0	0	0	integer	2
20	Fantasy	774	89.27	0	0	0	0	integer	2
21	Horror	775	89.39	0	0	0	0	integer	2
22	Mystery	778	89.73	0	0	0	0	integer	2
23	Romance	738	85.12	0	0	0	0	integer	2
24	Sci.Fi	757	87.31	0	0	0	0	integer	2
25	Thriller	716	82.58	0	0	0	0	integer	2
26	Western	862	99.42	0	0	0	0	integer	2
27	Sport	851	98.15	0	0	0	0	integer	2
28	History	842	97.12	0	0	0	0	integer	2
29	War	857	98.85	0	0	0	0	integer	2
30	Animation	820	94.58	0	0	0	0	integer	2
31	Music	847	97.69	0	0	0	0	integer	2
32	Musical	862	99.42	0	0	0	0	integer	2
33	Recency	0	0.00	0	0	0	0	integer	11

```
>
```

How many movies are labeled with each genre?

Action = 32.76%
Adventure = 18.95%
Animation = 5.42%
Biography = 8.42%
Comedy = 29.41%
Crime = 14.88%
Drama = 50.29%
Family = 5.65%
Fantasy = 10.73%
Horror = 10.61%
Mystery = 10.27%
Romance = 14.88%
Sci-Fi = 12.69%
Thriller = 17.42%
Western = .58%
Sport = 1.85%
History = 2.88%
War = 1.15%
Music = 2.31%
Musical = .58%



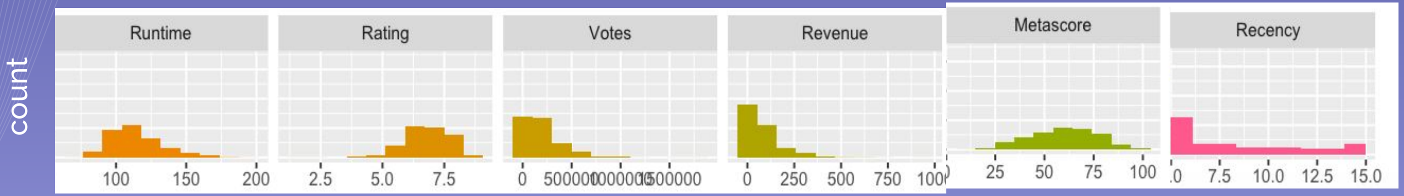
Exploratory Data Analysis - Frequency

freq(data)

	Genre	frequency	percentage	cumulative_perc
1	Action,Adventure,Sci-Fi	50	5.77	5.77
2	Comedy,Drama,Romance	33	3.81	9.58
3	Drama	31	3.58	13.16
4	Drama,Romance	30	3.46	16.62
5	Action,Adventure,Fantasy	26	3.00	19.62
6	Animation,Adventure,Comedy	26	3.00	22.62
7	Comedy	26	3.00	25.62
8	Comedy,Drama	24	2.77	28.39
9	Comedy,Romance	23	2.65	31.04
10	Crime,Drama,Mystery	18	2.08	33.12
11	Action,Adventure,Drama	17	1.96	35.08
12	Crime,Drama,Thriller	17	1.96	37.04
13	Action,Crime,Drama	16	1.85	38.89
14	Action,Adventure,Comedy	14	1.61	40.50
15	Adventure,Family,Fantasy	14	1.61	42.11
16	Biography,Drama	14	1.61	43.72
17	Action,Comedy,Crime	12	1.38	45.10
18	Biography,Drama,History	12	1.38	46.48
19	Drama,Thriller	12	1.38	47.86
20	Action,Adventure,Thriller	11	1.27	49.13
21	Action,Crime,Thriller	11	1.27	50.40
22	Horror,Thriller	11	1.27	51.67
23	Animation,Action,Adventure	9	1.04	52.71
24	Action,Thriller	8	0.92	53.63
25	Adventure,Comedy,Drama	8	0.92	54.55
26	Biography,Crime,Drama	8	0.92	55.47
27	Biography,Drama,Sport	8	0.92	56.39
28	Crime,Drama	8	0.92	57.31
29	Comedy,Crime,Drama	7	0.81	58.12
30	Horror	7	0.81	58.93
31	Horror,Mystery,Thriller	7	0.81	59.74
32	Action,Adventure,Crime	6	0.69	60.43
33	Action,Biography,Drama	6	0.69	61.12
34	Action,Drama,Thriller	6	0.69	61.81
35	Action,Sci-Fi,Thriller	6	0.69	62.50
36	Biography,Comedy,Drama	6	0.69	63.19
37	Drama,Mystery,Romance	6	0.69	63.88
38	Drama,Mystery,Thriller	6	0.69	64.57
39	Horror,Mystery	6	0.69	65.26
40	Mystery,Thriller	6	0.69	65.95
41	Action,Adventure,Mystery	5	0.58	66.53
42	Action,Comedy	5	0.58	67.11
43	Action,Horror,Sci-Fi	5	0.58	67.69
44	Adventure,Drama,Fantasy	5	0.58	68.27
45	Animation,Comedy,Family	5	0.58	68.85
46	Comedy,Crime	5	0.58	69.43

Overall, there are 195 different genre combinations between the 20 categories that we have (Action, Adventure, and Sci-Fi) are the most frequent combination next to (Comedy, Drama, and Romance.)

Exploratory Data Analysis - Histograms and Numerical and Categorical Representations




```
Runtime
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     867        0       91        1    114.7    20.43     90.0     94.0    101.0    112.0    124.0    140.0    150.7

lowest :  66  80  81  83  84, highest: 170 172 180 187 191
---------------------------------------------------------------------------------------------------
Rating
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     867        0       50    0.999    6.814   0.9722     5.33     5.70     6.30     6.90     7.40     7.90     8.10

lowest : 1.9 2.7 3.9 4.0 4.1, highest: 8.4 8.5 8.6 8.8 9.0
---------------------------------------------------------------------------------------------------
Votes
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     867        0      866        1   191845   188386     8270    21040    61079   136323   268077   435560   547381

lowest :     178     277     279     291     391, highest: 1045588 1047747 1222645 1583625 1791916
---------------------------------------------------------------------------------------------------
Revenue
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     867        0      810        1    83.35    97.48    0.236    1.588   13.650   48.040  114.665  210.122  294.380

lowest :   0.01   0.02   0.03   0.04   0.05, highest: 533.32 623.28 652.18 760.51 936.63
---------------------------------------------------------------------------------------------------
Metascore
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     867        0       82        1    59.55    19.39     32.0     36.0     47.0     60.0     72.0     81.0     85.7

lowest :  11  15  18  19  22, highest:  95  96  98  99 100
---------------------------------------------------------------------------------------------------
Recency
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     867        0       11    0.981    8.539    3.571        5        5        6        8       11       14       14

lowest :  5  6  7  8  9, highest: 11 12 13 14 15

Value          5     6     7     8     9    10    11    12    13    14    15
Frequency    202   112    93    88    64    62    57    47    51    49    42
```

Methodology- Agenda of Tests

#01.
K-means clustering

#02.
Regression analysis

#03.
Logistic regression analysis

#04.
Genre perception map

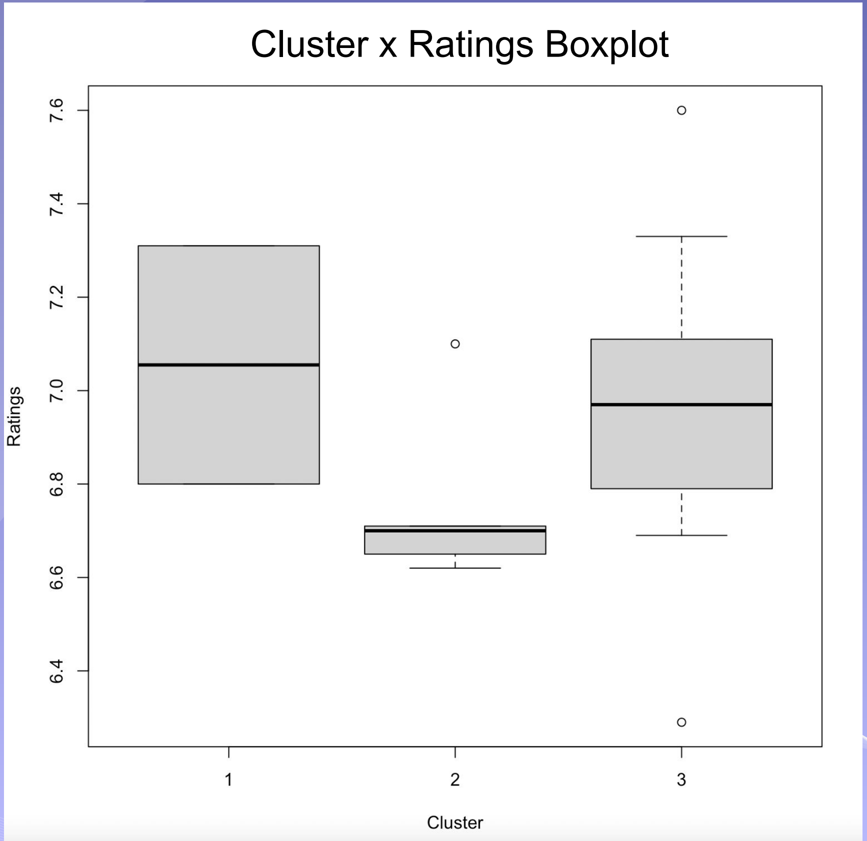
#01. K-Means Clustering

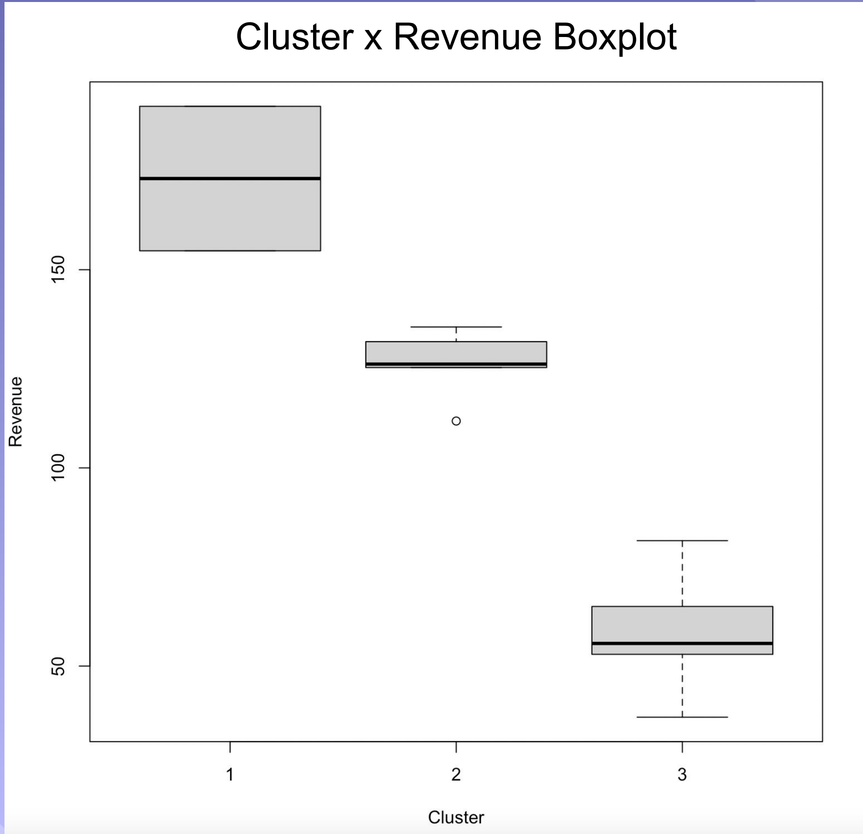
1. For our clustering data, we took the average revenue and average rating for each genre (20 total)
2. We selected 3 clusters for our analysis
3. In conducting this analysis, we discovered the following 3 clusters
   a. Cluster 1: 2 genres
   b. Cluster 2: 5 genres
   c. Cluster 3: 13 genres

Cluster Genres

Cluster Features	Most Grossing; Popular	Grossing; Least Popular	Least Grossing; Popular
Cluster Number	1	2	3
Genres	*Adventure*, Animation	*Action*, Family, Fantasy, *Sci-Fi*, Western	*Biography*, *Comedy*, *Crime*, *Drama*, *Horror*, *Mystery*, *Romance*, *Thriller*, Sport, History, War, Music, Musical
Average Revenue (millions) [Avg Overall = 83.35]	173	126.14	58.24
Average Rating (out of 10) [Avg Overall = 6.814]	7.06	6.76	6.98

Cluster Boxplots





Test Implications

1. Movies in the adventure & animation genres have the highest average revenues & ratings
2. Cluster 3 holds a significant amount of movie genres that lag behind the rest in terms of amount of gross revenue
3. The difference between average revenue of each cluster is significantly large, especially clusters 2 and 3
4. The difference between average rating of each cluster is significantly small, as they are each saturated with similar ratings

#02. Regression (Runtime, Genre, Recency on Revenue)

```
lm(formula = Revenue ~ Runtime + Recency + Action + Adventure +
    Biography + Comedy + Crime + Drama + Family + Fantasy + Horror +
    Mystery + Romance + SciFi + Thriller + Western + Sport +
    History + War + Animation + Music + Musical, data = Movie)

Residuals:
    Min      1Q  Median      3Q     Max
-196.45  -43.31   -8.83   29.18  744.63

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -141.6268    23.8594  -5.936 4.26e-09 ***
Runtime        1.7237     0.1731   9.959  < 2e-16 ***
Recency        2.5803     0.9084   2.840  0.00461 **
Action        17.5673     8.0995   2.169  0.03037 *
Adventure     39.0748     8.6288   4.528 6.80e-06 ***
Biography      0.3827    11.9170   0.032  0.97439
Comedy         3.2748     8.9218   0.367  0.71367
Crime         -5.3842     9.1784  -0.587  0.55762
Drama        -38.6647     8.2387  -4.693 3.14e-06 ***
Family        11.7811    13.4215   0.878  0.38031
Fantasy       27.0761    10.6869   2.534  0.01147 *
Horror       -23.4960    11.0051  -2.135  0.03305 *
Mystery      -16.8083    10.3834  -1.619  0.10587
Romance      -12.0237     9.5594  -1.258  0.20882
SciFi1        33.2466    10.1570   3.273  0.00111 **
Thriller       3.6806     9.0477   0.407  0.68426
Western      -12.4619    37.8205  -0.330  0.74186
Sport          3.5707    22.0086   0.162  0.87115
History      -18.6364    18.2702  -1.020  0.30800
War          -13.2715    27.0873  -0.490  0.62429
Animation    103.9028    14.3659   7.233 1.06e-12 ***
Music          2.8530    22.2079   0.128  0.89781
Musical       -1.0256    43.1852  -0.024  0.98106
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 83.08 on 845 degrees of freedom
Multiple R-squared:  0.3706,	Adjusted R-squared:  0.3542
F-statistic: 22.61 on 22 and 845 DF,  p-value: < 2.2e-16
```

1. Genres are dummy coded (0=no, 1=yes)
2. Animation, Adventure, Sci-Fi, Fantasy & Action genres all have significantly positive effects on revenue
3. Drama & Horror genres have significantly negative effects on revenue
4. Recency & Runtime have significantly positive effects on revenue
5. Low R-Squared value

Regression (Runtime, Genre, Recency on Ratings)

```
lm(formula = Rating ~ Runtime + Recency + Action + Adventure +
    Biography + Comedy + Crime + Drama + Family + Fantasy + Horror +
    Mystery + Romance + SciFi + Thriller + Western + Sport +
    History + War + Animation + Music + Musical, data = Movie)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4588 -0.3921  0.0631  0.4831  1.8623

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.345545   0.212560  20.444  < 2e-16 ***
Runtime      0.016987   0.001542  11.016  < 2e-16 ***
Recency      0.035524   0.008093   4.389 1.28e-05 ***
Action      -0.150750   0.072157  -2.089  0.03699 *
Adventure   -0.023570   0.076873  -0.307  0.75922
Biography    0.348277   0.106167   3.280  0.00108 **
Comedy       0.073522   0.079483   0.925  0.35523
Crime        0.015582   0.081770   0.191  0.84892
Drama        0.332838   0.073398   4.535 6.60e-06 ***
Family      -0.065177   0.119570  -0.545  0.58583
Fantasy     -0.102566   0.095208  -1.077  0.28166
Horror      -0.288337   0.098043  -2.941  0.00336 **
Mystery      0.208074   0.092505   2.249  0.02475 *
Romance     -0.176018   0.085163  -2.067  0.03905 *
SciFi1       0.150163   0.090487   1.659  0.09739 .
Thriller     0.085410   0.080605   1.060  0.28963
Western      0.145587   0.336938   0.432  0.66579
Sport       -0.027582   0.196072  -0.141  0.88816
History     -0.056118   0.162767  -0.345  0.73035
War          0.679248   0.241317   2.815  0.00499 **
Animation    1.011826   0.127984   7.906 8.29e-15 ***
Music        0.355444   0.197847   1.797  0.07276 .
Musical     -0.517002   0.384731  -1.344  0.17937
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7402 on 845 degrees of freedom
Multiple R-squared:  0.3048,    Adjusted R-squared:  0.2867
F-statistic: 16.84 on 22 and 845 DF,  p-value: < 2.2e-16
```

1. Genres are dummy coded (0=no, 1=yes)
2. Animation, War, Mystery, Drama, & Biography genres all have significantly positive effects on rating
3. Action, Horror, & Romance genres have significantly negative effects on rating
4. Runtime & Recency have significantly positive effects on rating
5. Low R-Squared value

Regression (Runtime & Recency on Revenue & on Rating)

```
Call:
lm(formula = Revenue ~ Runtime + Recency, data = Movie)

Residuals:
    Min      1Q  Median      3Q     Max 
-187.60  -58.01  -27.24   26.39  829.99 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -112.1557    22.1732  -5.058 5.17e-07 ***
Runtime        1.4703     0.1825   8.057 2.59e-15 ***
Recency        3.1399     1.0621   2.956   0.0032 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.05 on 865 degrees of freedom
Multiple R-squared:  0.0842,	Adjusted R-squared:  0.08208 
F-statistic: 39.76 on 2 and 865 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Rating ~ Runtime + Recency, data = Movie)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5888 -0.4812  0.0471  0.5427  2.0488 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.599788   0.181078  25.402  < 2e-16 ***
Runtime     0.016886   0.001490  11.330  < 2e-16 ***
Recency     0.032307   0.008674   3.725 0.000208 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8089 on 865 degrees of freedom
Multiple R-squared:  0.1501,	Adjusted R-squared:  0.1482 
F-statistic: 76.39 on 2 and 865 DF,  p-value: < 2.2e-16
```

1. Genres were left out of analysis
2. Runtime & Recency have significantly positive effects on rating and revenue
3. Even lower R-Squared values without genre

Models at a Glance

	Model #01	Model #02	Model #03	Model #04
Independent Variables	Runtime, Recency, Genre	Runtime, Recency, Genre	Runtime, Recency	Runtime, Recency
Dependent Variable	Revenue	Ratings	Revenue	Ratings
Significant Positive	Runtime & Recency Animation, Adventure, Sci-Fi, Fantasy & Action	Runtime & Recency Animation, War, Mystery, Drama, & Biography	Runtime & Recency	Runtime & Recency
Significant Negative	Drama & Horror	Action, Horror, & Romance	N/A	N/A
R-Squared Value	0.3706	0.3048	0.0842	0.1501

Test Implications

1. Runtime & recency both have significantly positive effects on revenue & ratings, as indicated in all 4 models
2. Genre can more accurately predict revenue and ratings, as indicated by the R-Squared values in each model
3. Being classified in the animation genre has a significantly positive effect on movie revenue and rating
4. Being classified in the horror genre has a significantly negative effect on movie revenue and rating

#03. Logistic Regression

1. We used logistic regression to learn how genre is dependent on rating, revenue, and metascore
2. We ran regression on each specific genre as the dependent variable
3. Criteria:
   a. Regression analysis for each genre must include at least 2 or more statistically significant variables
   b. Model AIC must be less than 540

Logistic Regression
(Revenue, Rating, and Metascore based on Biography Genre)

```
Call:
glm(formula = Biography ~ Revenue + Rating + Metascore, family = binomial("logit"),
    data = Movie)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9971  -0.4660  -0.3210  -0.2252   3.2403

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.998603   1.371513  -6.561 5.34e-11 ***
Revenue     -0.005678   0.001735  -3.274 0.001062 **
Rating       0.845000   0.229964   3.674 0.000238 ***
Metascore    0.016181   0.010217   1.584 0.113267
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 501.14  on 867  degrees of freedom
Residual deviance: 451.37  on 864  degrees of freedom
AIC: 459.37

Number of Fisher Scoring iterations: 6
```

1. Revenue and rating are statistically significant
   a. Revenue negatively associated with biography genre
   b. Rating positively associated with biography genre
2. Metascore is not statistically significant

Logistic Regression
(Revenue, Rating, and Metascore based on Animation Genre

```
Call:
glm(formula = Animation ~ Revenue + Rating + Metascore, family = binomial("logit"),
    data = Movie)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1462  -0.3401  -0.2461  -0.1808   2.9752

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.34026    1.53715  -4.125 3.71e-05 ***
Revenue      0.00537    0.00109   4.926 8.39e-07 ***
Rating       0.04780    0.28150   0.170  0.86516
Metascore    0.03853    0.01438   2.679  0.00738 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 365.52  on 867  degrees of freedom
Residual deviance: 316.20  on 864  degrees of freedom
AIC: 324.2

Number of Fisher Scoring iterations: 6
```

1. Revenue and Metascore have a significantly positive association with the animation genre
2. Rating is not statistically significant

Logistic Regression
(Revenue, Rating, and Metascore based on Horror Genre)

```
Call:
glm(formula = Horror ~ Revenue + Rating + Metascore, family = binomial("logit"),
    data = Movie)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8081  -0.5151  -0.3887  -0.2310   2.9813

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.868102   0.838193   3.422 0.000622 ***
Revenue     -0.009587   0.002408  -3.982 6.84e-05 ***
Rating      -0.826390   0.164334  -5.029 4.94e-07 ***
Metascore    0.017039   0.008663   1.967 0.049185 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 586.86  on 867  degrees of freedom
Residual deviance: 523.90  on 864  degrees of freedom
AIC: 531.9

Number of Fisher Scoring iterations: 6
```

1. Revenue, rating, and metascore are statistically significant
2. Revenue and rating have a negative association with the horror genre
    a. Rating has a large association
3. Metascore has a positive association with Horror genre

Models at a Glance

	Model #01	Model #02	Model #03
Independent Variables	Revenue, Rating, Metascore	Revenue, Rating, Metascore	Revenue, Rating, Metascore
Dependent Variable	Biography	Animation	Horror
Revenue	Statistically Significant Negatively Associated	Statistically Significant Positively Associated	Statistically Significant Negatively Associated
Rating	Statistically Significant Positively Associated	Not Statistically Significant Positively Associated	Statistically Significant Largely Negatively Associated
Metascore	Not Statistically Significant Positively Associated	Statistically Significant Positively Associated	Statistically Significant Positively Associated
AIC	459.37	324.2	531.9

Test Implications

1. Only 3 genres were statistically relevant according our criteria
    a. Biography
    b. Animation
    c. Horror
2. Rating has a significantly large positive impact on the Biography genre
3. Both metascore and revenue have a significantly large positive impact on the Animation genre
4. Both rating and revenue have a significantly large negative impact on the Horror genre

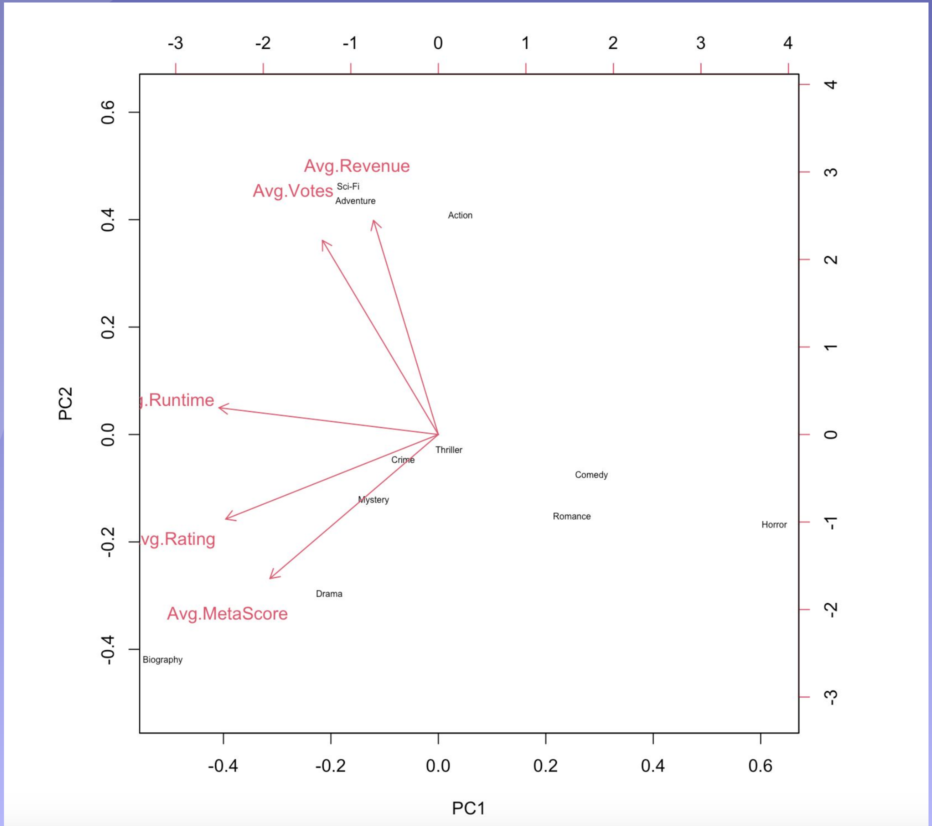
#04. Genre Perception Map

- Sorted data - using only genres that have more than 50 movies (12)
- All variables have calculated averages to create a brand perception map

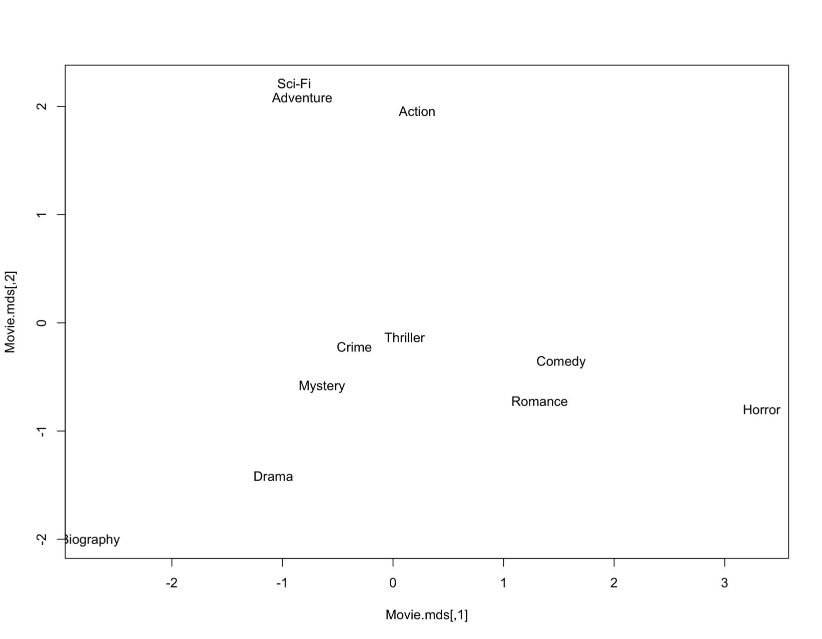
A1 | fx | Genre

Genre	No of Movies	Runtime	Rating	Votes	Revenue	MetaScore	Avg Runtime	Avg Rating	Avg Votes	Avg Revenue	Avg MetaScore
Biography	73	8932	535.1	12108173	4121.9	5032	122.36	7.33	165865.38	56.46	68.93
Mystery	89	10345	620.6	18529807	4861.86	5448	116.24	6.97	208200.08	54.63	61.21
Horror	92	9582	578.4	9993740	3413.58	5144	104.15	6.29	108627.61	37.1	55.91
Sci-Fi	110	12922	744	34473598	14910.78	6413	117.47	6.76	313396.35	135.55	58.3
Romance	129	14592	863.1	17185704	5482.89	7202	113.12	6.69	133222.51	42.5	55.83
Crime	130	15249	897.5	26123301	8034.62	7652	117.3	6.9	200948.47	61.8	58.86
Thriller	152	17392	1031.5	30153358	10645.31	9065	114.42	6.79	198377.36	70.03	59.64
Adventure	251	29635	1706.3	66597741	38846.11	14488	118.07	6.8	265329.65	154.77	57.72
Comedy	255	27195	1708.7	39006151	19316.45	14848	106.65	6.7	152965.3	75.75	58.23
Action	284	33280	1887.7	70231996	35587.77	15341	117.18	6.65	247295.76	125.31	54.02
Drama	437	51619	3072.4	72565761	21861.73	28126	118.12	7.03	166054.37	50.03	64.36

Genre Perception Map



Genre Distance Map



Test Implications

1. The Sci-fi and Adventure, and to an extent Action, genres are leaders in generating revenue and votes for all top genres sampled
2. The Comedy and Romance, and to an even larger extent Horror, genres are not representative of the any of the attributes for all top genres sampled
3. The Biography and Drama genre gets the highest ratings out of all top genres sampled
4. Similar consumer perception of genres:
    a. Comedy and Romance
    b. Crime, Mystery and Thriller
    c. Sci-fi, Adventure and Action

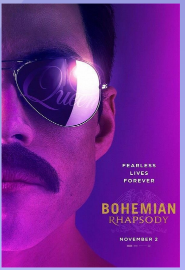
Results/Interpretations

Where is the movie industry headed? How can the movie industry optimize revenue and ratings through its selection of genres and other factors in the future?

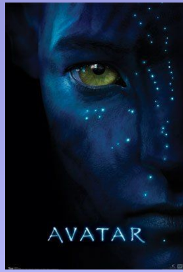
1. Genre on revenue
   a. Adventure (18.95%) and Animation (5.42%) lead the movie industry with the highest impact on revenues with Sci-Fi (12.69%), Fantasy (10.73%), Action (32.76%) following behind
   b. Drama (50.29%) & Horror (10.61%) have a significantly negative impact on revenue
2. Genre on ratings
   a. Animation and War (1.15%) lead the movie industry with the highest impact on ratings with Biography (8.42%), Drama, and Mystery (10.27%) following behind
   b. Horror, Romance (14.88%), and Action have a significantly negative impact on ratings
3. Other factors
   a. Movies with increased runtimes and older years tend to do better

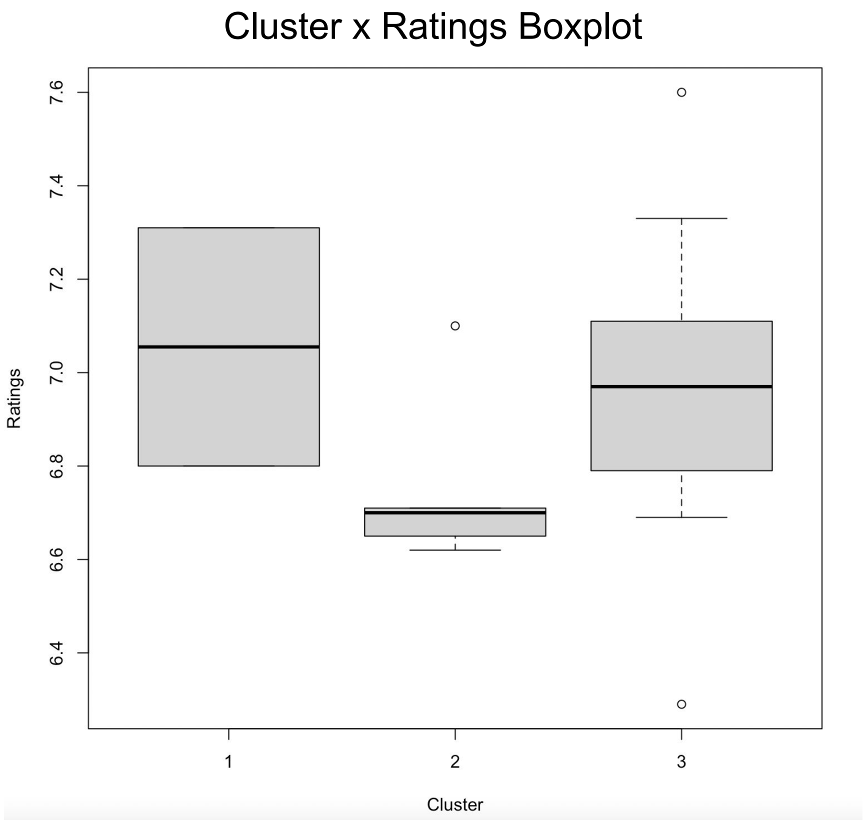
Managerial Implications

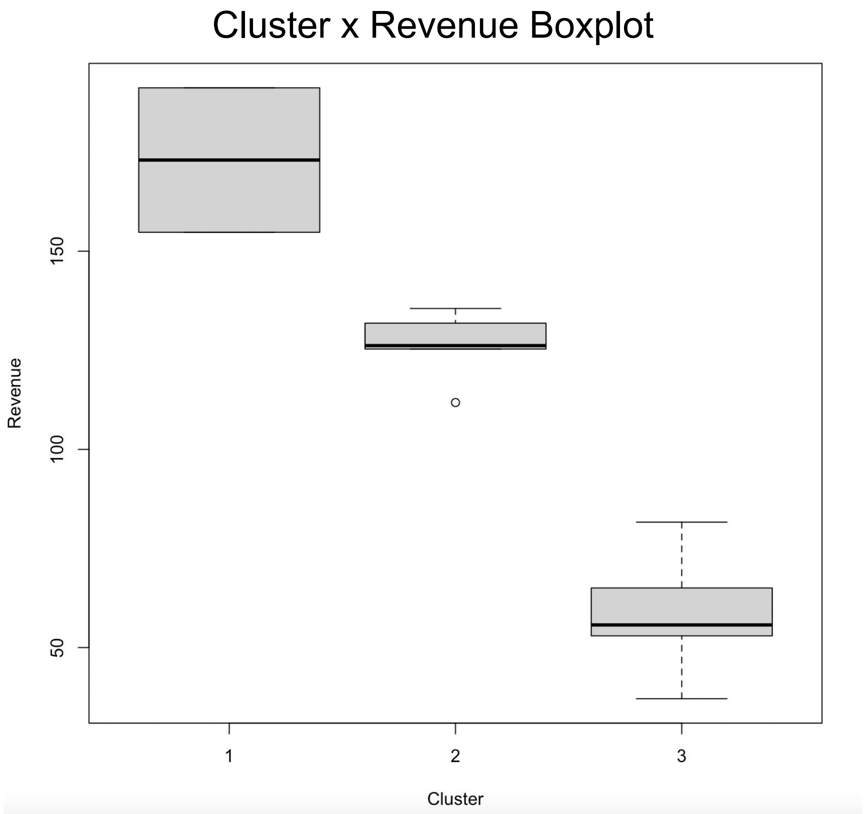
- Producers will need to curate their content based on whether they want to be part of the market trends or differentiate from what has been created
  - The top four genre combinations are: “Action, Adventure, and Sci-Fi” (50), “Comedy, Drama, Romance” (33), “Drama” (31), “Drama, Romance” (30)
  - Based our genre distance map. Other perceived combinations that could find success are: “Comedy and Romance” and “Crime, Mystery, and Thriller”
- Regardless of choice, the success of a movie will increase over time as it can reach viewers outside of the cinema and streaming services
  - In addition, an increased runtime may make viewers feel like they are getting more for their money



	Mainstream	Niche
Positive Revenues	Sci-Fi, Adventure	Animation, Fantasy
Positive Ratings	Biography, *Drama	War, Biography
Avoid	Romance, Comedy, Horror	Generally other genres




Cluster x Ratings Boxplot
Ratings
7.6
7.4
7.2
7.0
6.8
6.6
6.4
1
2
3
Cluster


Cluster x Revenue Boxplot
Revenue
150
100
50
1
2
3
Cluster