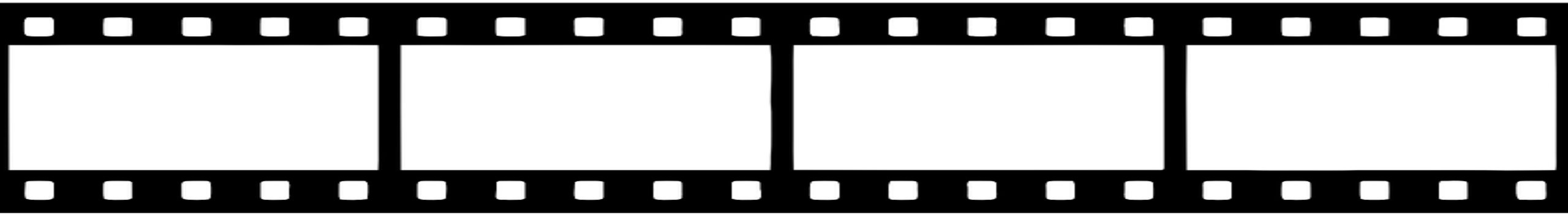
BUILD A DATA STORY

The Movies Dataset

Final Project



Jason Kanhai
Data Analyst for Shell Nanodegree
Sep-29-2021

Executive Summary

The key differentiators of the Top 3 genres of movies based on average popularity from the remaining bottom 17 genres are, in no order of significance:

1. The budget spent on the movie.
2. The month movie was released.
3. The revenue the movie generated.

Our recommendation for any upcoming movies in development that wish to generate revenue would be to:

- Create a movie that falls within the top 3 genres (Family, Adventure or Science Fiction)
- Ensure sufficient budget is allocated for the movie to ensure success at the box office and generate revenue.
- Ensure movies are released within the months of March to June.

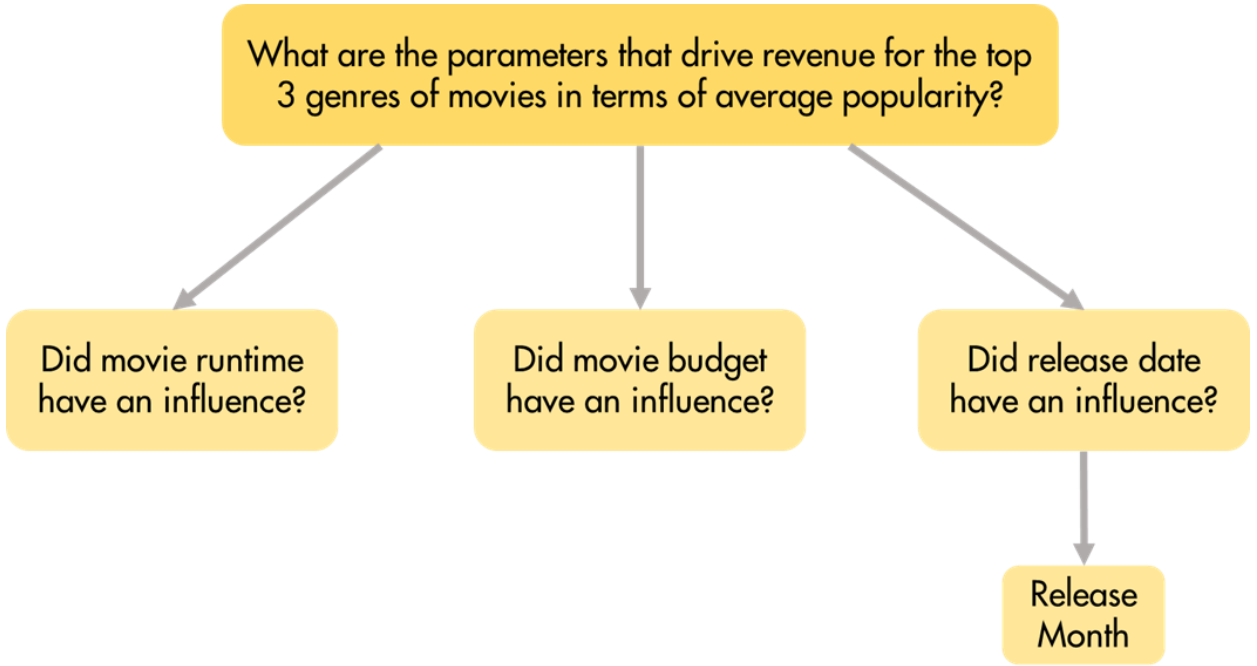
Problem Statement

What are the parameters that drive revenue for the top 3 genres of movies in terms of average popularity?

I focused my analysis on the following workstreams:

1. Identifying the top 3 genres in terms of average popularity
2. If runtime affected revenue for movies within these top 3 genres .
3. If budget affected revenue for these movies.
4. If the release month in affected revenue for these movies.

Issue Tree

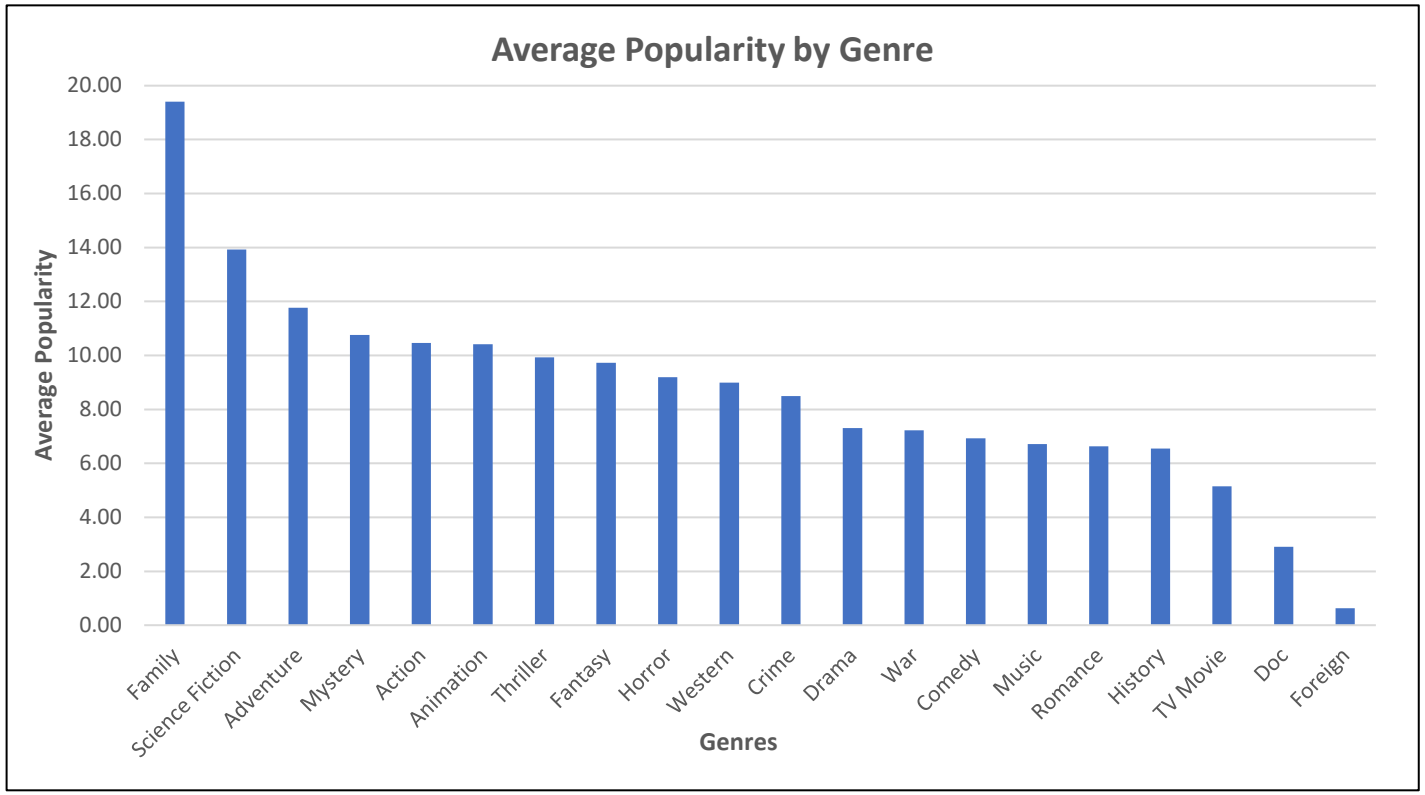


Problem Statement

Hypotheses

Sub-Issue

Top 3 Genres by Average Popularity



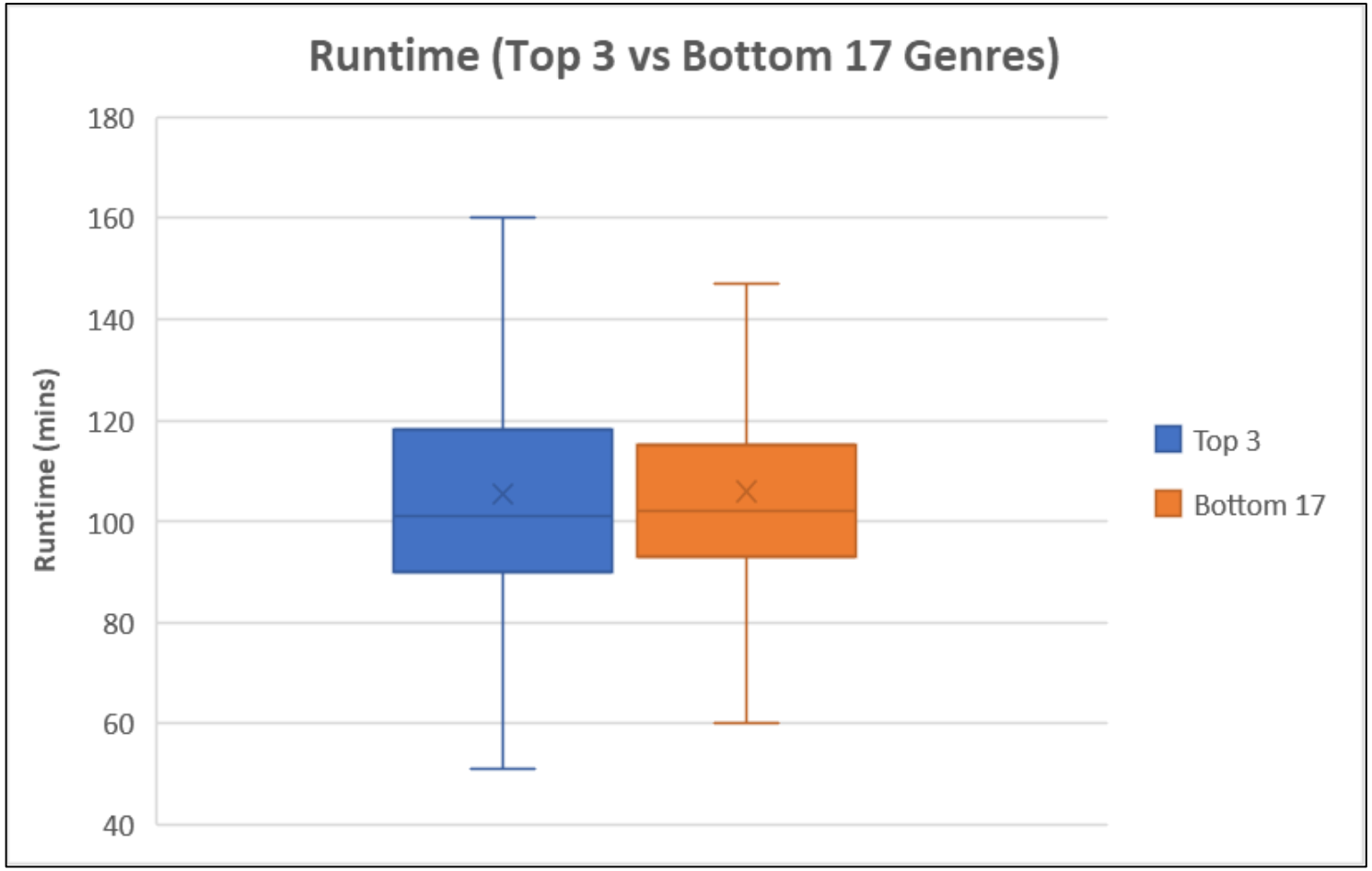
Takeaways

The Top 3 Genres of the entire dataset in terms of average popularity are:

- Family (19.41)
- Science Fiction (13.93)
- Adventure (11.76)

Genre	*Family*	*Science Fiction*	*Adventure*	*Mystery*	*Action*	*Animation*	*Thriller*	*Fantasy*	*Horror*	*Western*	*Crime*	*Drama*	*War*	*Comedy*	*Music*	*Romance*	*History*	*TV Movie*	*Doc*	*Foreign*
Min	0.00	0.30	0.14	0.22	0.00	0.03	0.08	0.19	0.10	0.18	0.20	0.00	0.32	0.00	0.11	0.01	0.06	5.15	0.00	0.24
Q1	5.46	6.86	6.38	5.74	5.04	6.81	5.93	5.83	6.07	5.80	5.14	3.22	2.00	3.62	2.47	2.42	3.43	5.15	0.51	0.36
Median	9.27	10.49	10.20	8.32	8.24	10.25	8.85	9.58	8.36	7.72	8.04	6.49	5.64	6.68	6.77	6.75	6.33	5.15	1.76	0.57
Mean	19.41	13.93	11.76	10.76	10.46	10.42	9.93	9.73	9.20	8.99	8.49	7.31	7.22	6.94	6.72	6.63	6.55	5.15	2.91	0.63
Q3	12.21	14.40	13.86	12.36	12.15	13.67	11.93	12.43	11.60	12.34	11.62	9.80	9.91	9.52	10.19	9.37	7.95	5.15	4.47	0.68
Max	547.49	147.10	213.85	154.80	294.34	34.85	140.95	41.05	72.88	23.50	30.21	146.16	36.71	48.31	18.83	34.46	31.60	5.15	15.30	1.41
IQR	6.76	7.53	7.48	6.62	7.11	6.86	6.00	6.60	5.53	6.53	6.47	6.58	7.91	5.90	7.72	6.95	4.52	0.00	3.95	0.33
Lower Limit	-4.68	-4.44	-4.83	-4.19	-5.63	-3.47	-3.06	-4.07	-2.21	-3.99	-4.57	-6.65	-9.87	-5.23	-9.10	-8.00	-3.34	5.15	-5.42	-0.14
Upper Limit	22.35	25.70	25.08	22.29	22.82	23.96	20.92	22.33	19.89	22.13	21.32	19.66	21.77	18.37	21.77	19.80	14.72	5.15	10.40	1.17

Runtime

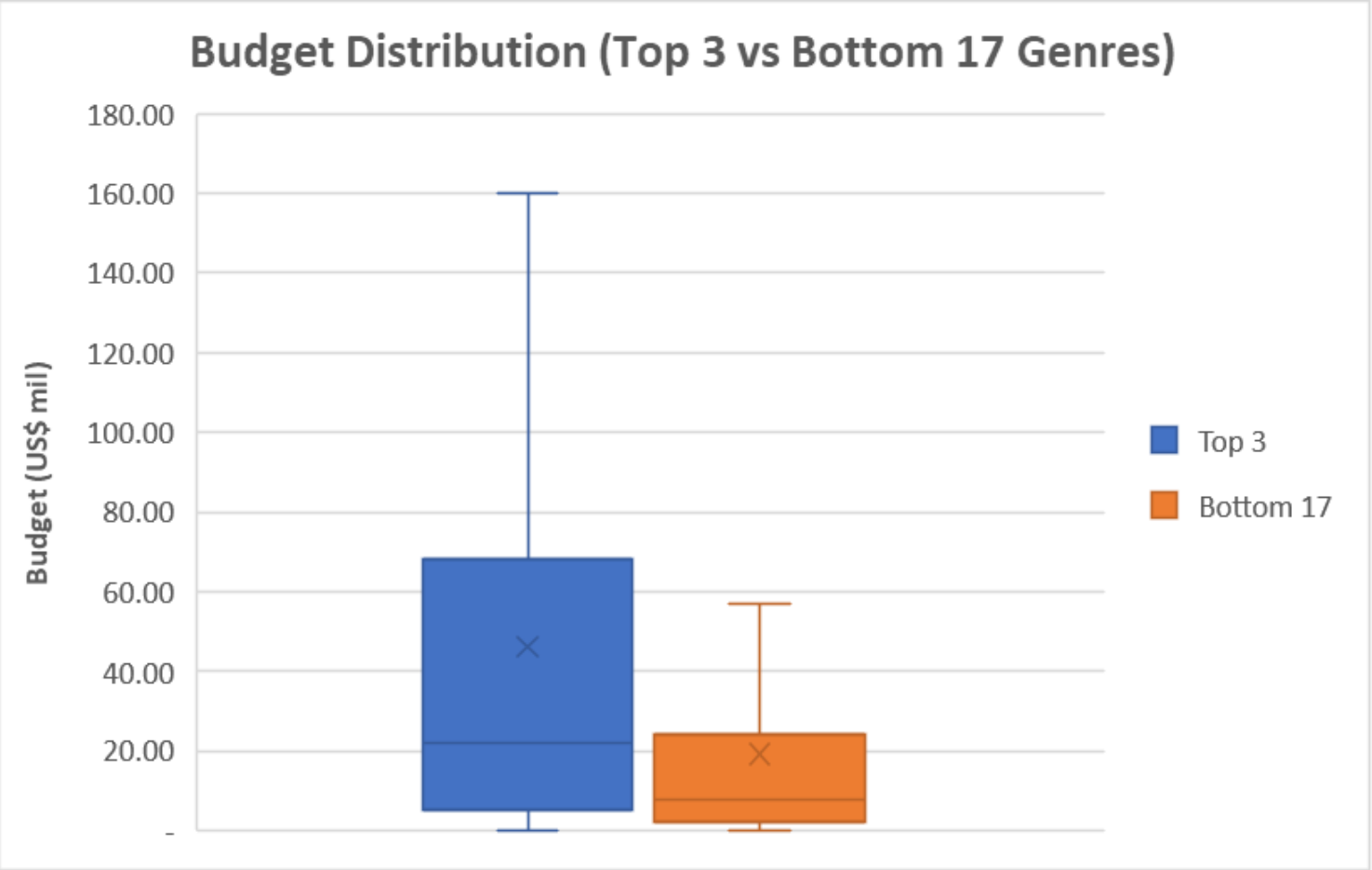


Takeaways

The runtime of movies that fall within the top 3 genres in terms of popularity, do not have differing average runtimes (105 minutes) from the bottom 17 genres of movies.

	Top 3	Bottom 17
Min	23.00	6.00
Q1	90.00	93.00
Median	101.00	102.00
Q3	118.25	115.00
Max	216.00	287.00
IQR	28.25	22.00
Lower Limit	47.63	60.00
Upper Limit	160.63	148.00
Mean	105.41	105.90

Budget

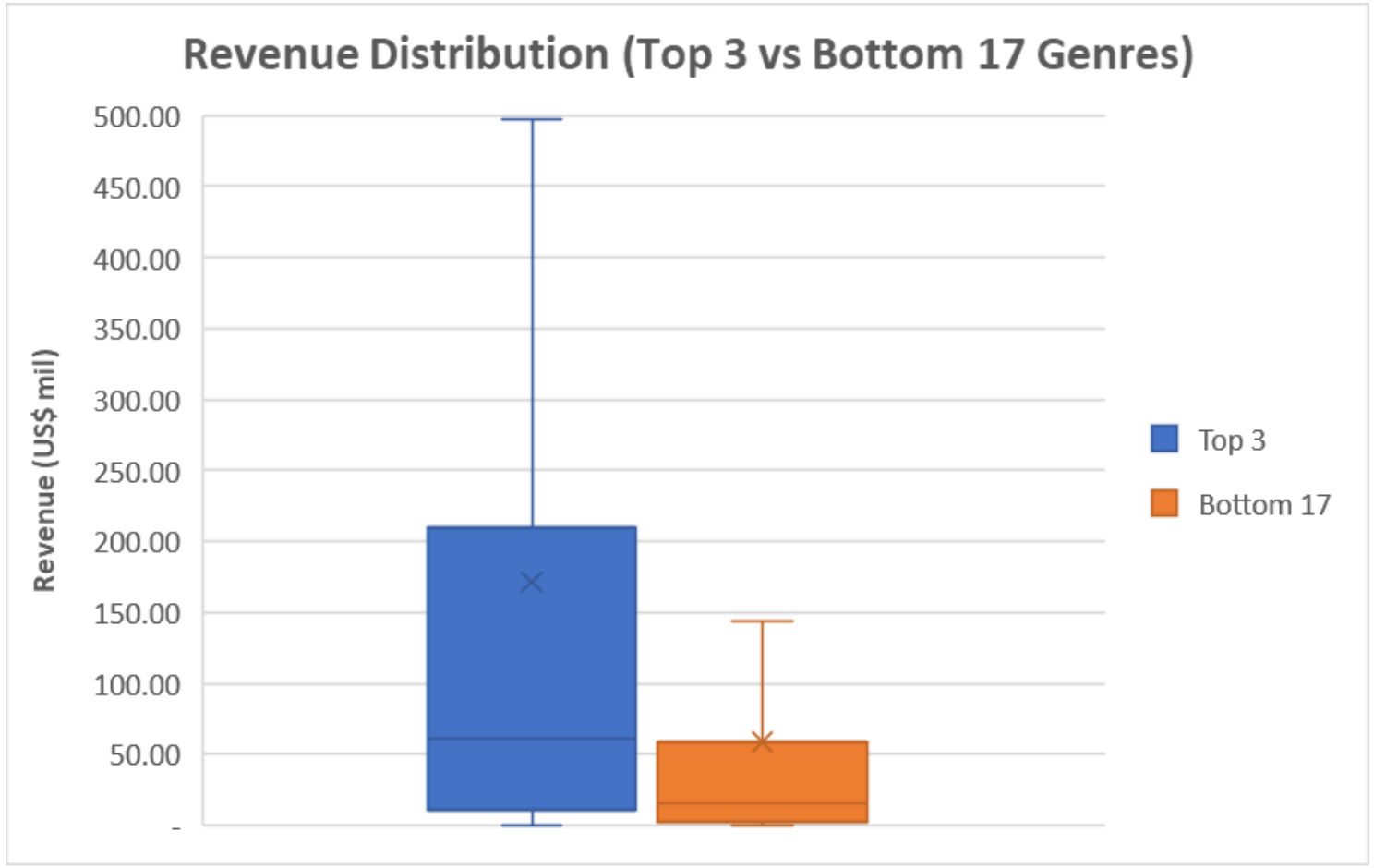


Takeaways

The mean budget allocated to the top 3 genres of movies (US $46.11 million) is on average greater than the average budget of the bottom 17 genres (US $19.16 million). The top 3 genres also cover a winder range of values for budget.

	Top 3	Bottom 17
Min	0.00	0.00
Q1	5.10	2.00
Median	22.00	7.70
Q3	68.00	24.00
Max	380.00	280.00
IQR	62.90	22.00
Lower Limit	-89.24	-31.00
Upper Limit	162.34	57.00
Mean	46.11	19.16

Revenue

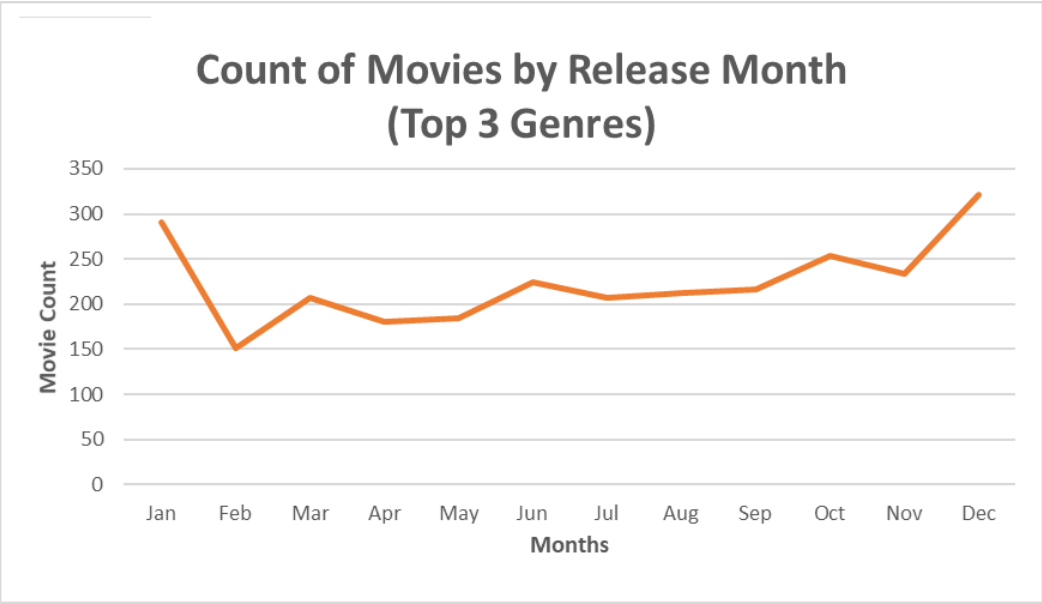


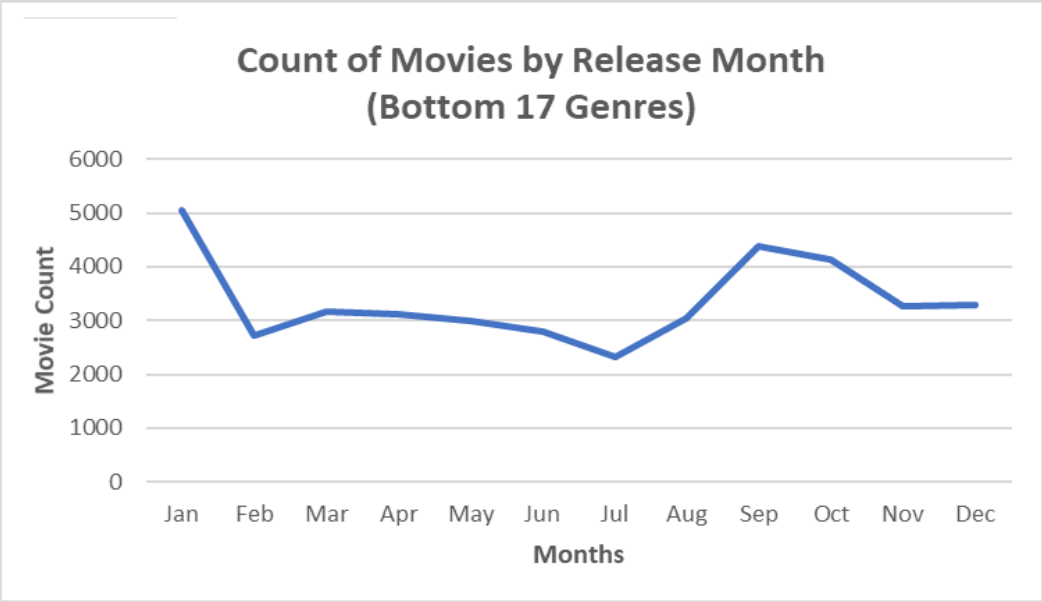
Takeaways

For the top 3 genres of movies, the average revenue (US $172.94 million) is greater than that of the bottom 17 genres (US $58.72 million).

	Top 3	Bottom 17
Min	0.00	0.00
Q1	11.81	2.09
Median	63.25	15.13
Q3	211.71	58.92
Max	1519.56	2787.97
IQR	199.90	56.83
Lower Lim	-288.05	-83.15
Upper Lim	511.57	144.16
Mean	172.94	58.72

Release Month



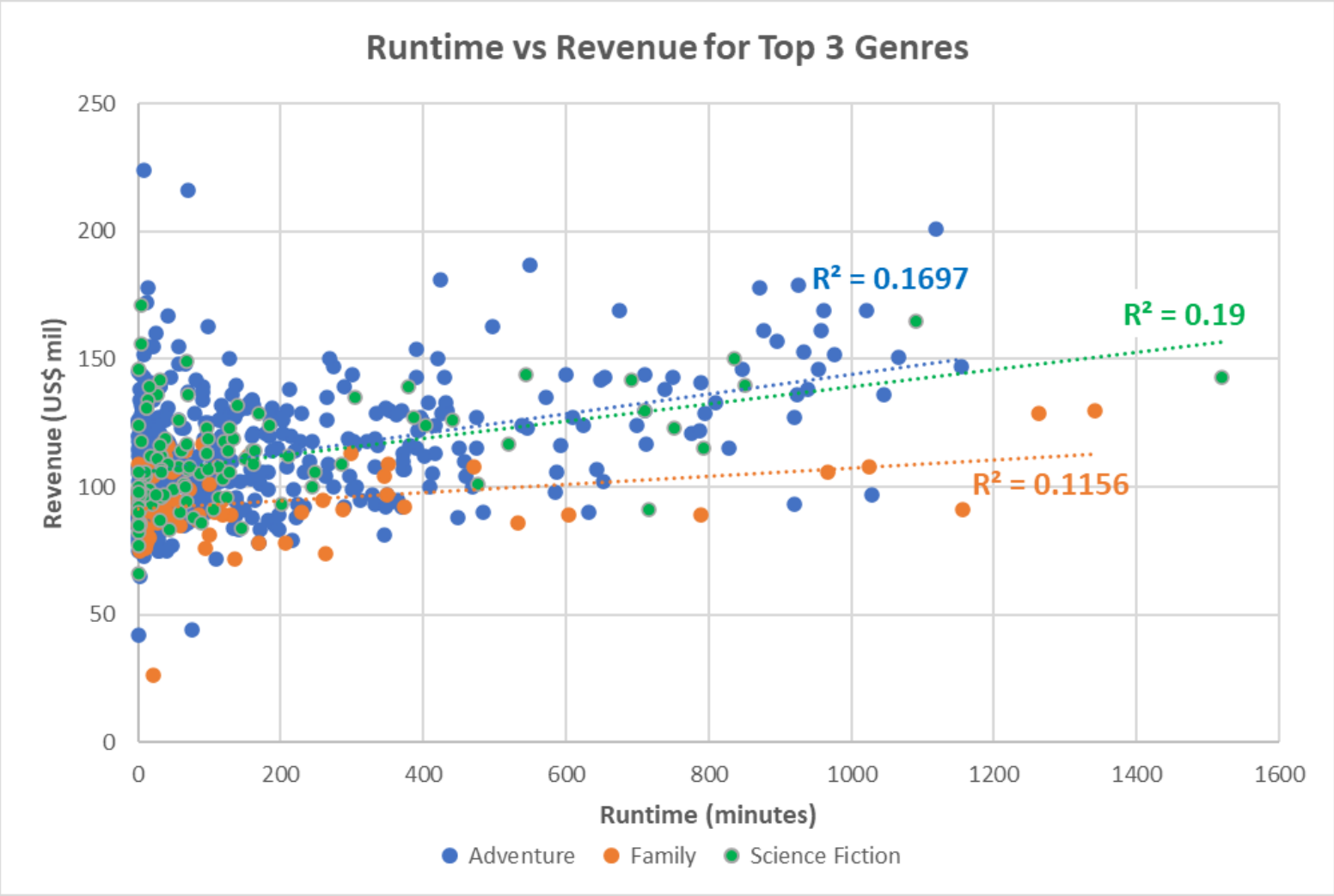


Takeaways

Over time, movies that fall within the top 3 genres have a peak release around Christmas time (Dec & January).

Movies that fall within the bottom 17 genres have a peak release in January but also a noticeable peace around September.

Runtime vs Revenue

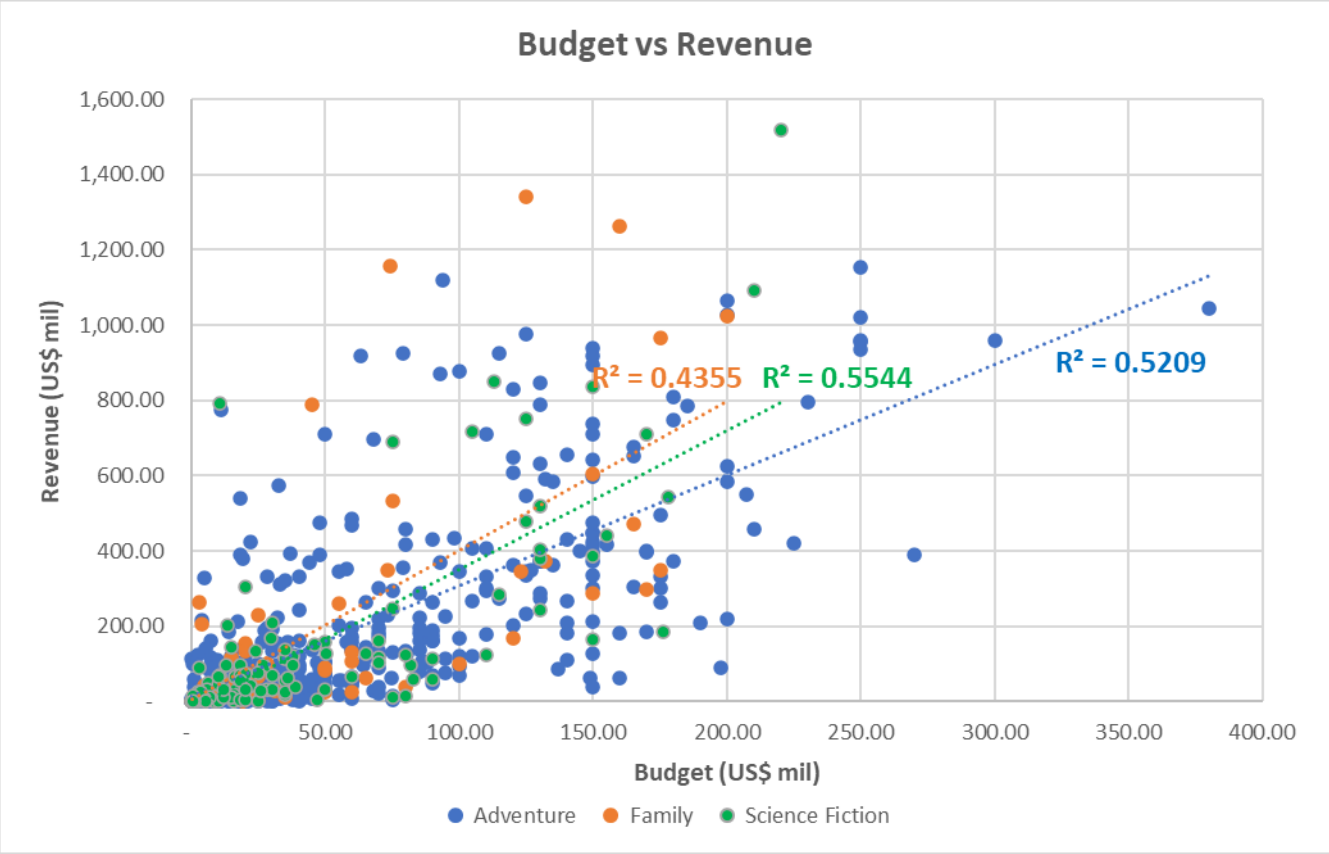


Takeaways

A weak although positive relationship exists between movie runtime and the revenue it generates as demonstrated by the R^2 values (0.11 – 0.19) depending on the genre.

This weak relationship suggest that runtime isn't the strongest determining factor to a movie's revenue. Any potential movies in development should not worry about runtime.

Budget vs Revenue

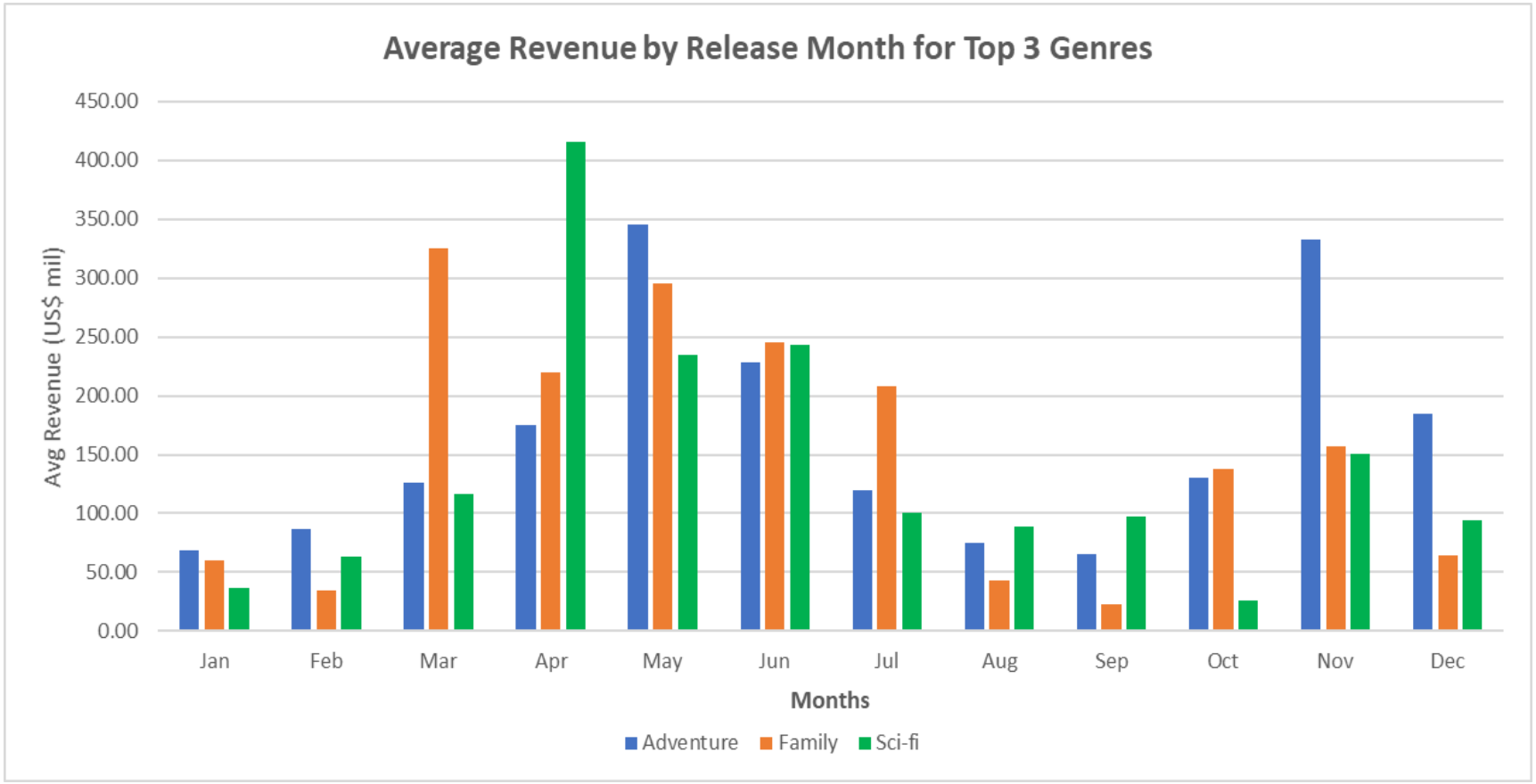


Takeaways

A significant positive relationship exists between movie budget and the revenue it has generated.

This is demonstrated by the R^2 values (0.44- 0.55). This suggests that for a movie to make a decent enough revenue, a proportional budget must be allocated.

Release Month vs Revenue



Takeaways

For movie that fall within the Top 3 Genres, it appears that average revenue is higher between the months of March to June.

An additional peak appears around November although smaller than the former.

If movies that fall within these genres seek to do well at the box office, they should aim to release within these months.

Release Month vs Revenue

Adventure

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Min	0.03	0.44	1.10	0.01	0.48	0.27	0.00	0.00	0.00	0.88	0.25	0.00
Q1	6.13	6.21	7.79	26.24	57.02	10.85	16.63	9.81	2.73	16.42	31.66	10.57
Median	30.82	27.30	38.79	70.64	292.83	101.87	57.88	31.76	24.38	44.24	331.93	89.77
Mean	68.61	86.96	126.56	174.56	345.05	228.64	119.45	74.68	65.00	130.78	332.71	184.77
Q3	97.92	123.21	179.25	214.68	528.15	357.04	148.44	102.50	124.88	146.52	515.51	241.02
Max	355.69	469.16	712.17	1153.30	1045.71	1065.66	933.96	415.69	328.20	954.31	1021.10	1118.89
IQR	91.78	116.99	171.46	188.44	471.14	346.19	131.81	92.69	122.16	130.10	483.84	230.44
Lower Limit	-131.54	-169.28	-249.40	-256.42	-649.69	-508.44	-181.09	-129.22	-180.51	-178.72	-694.10	-335.09
Upper Limit	235.60	298.70	436.43	497.34	1234.86	876.33	346.15	241.54	308.11	341.67	1241.28	586.68

Family

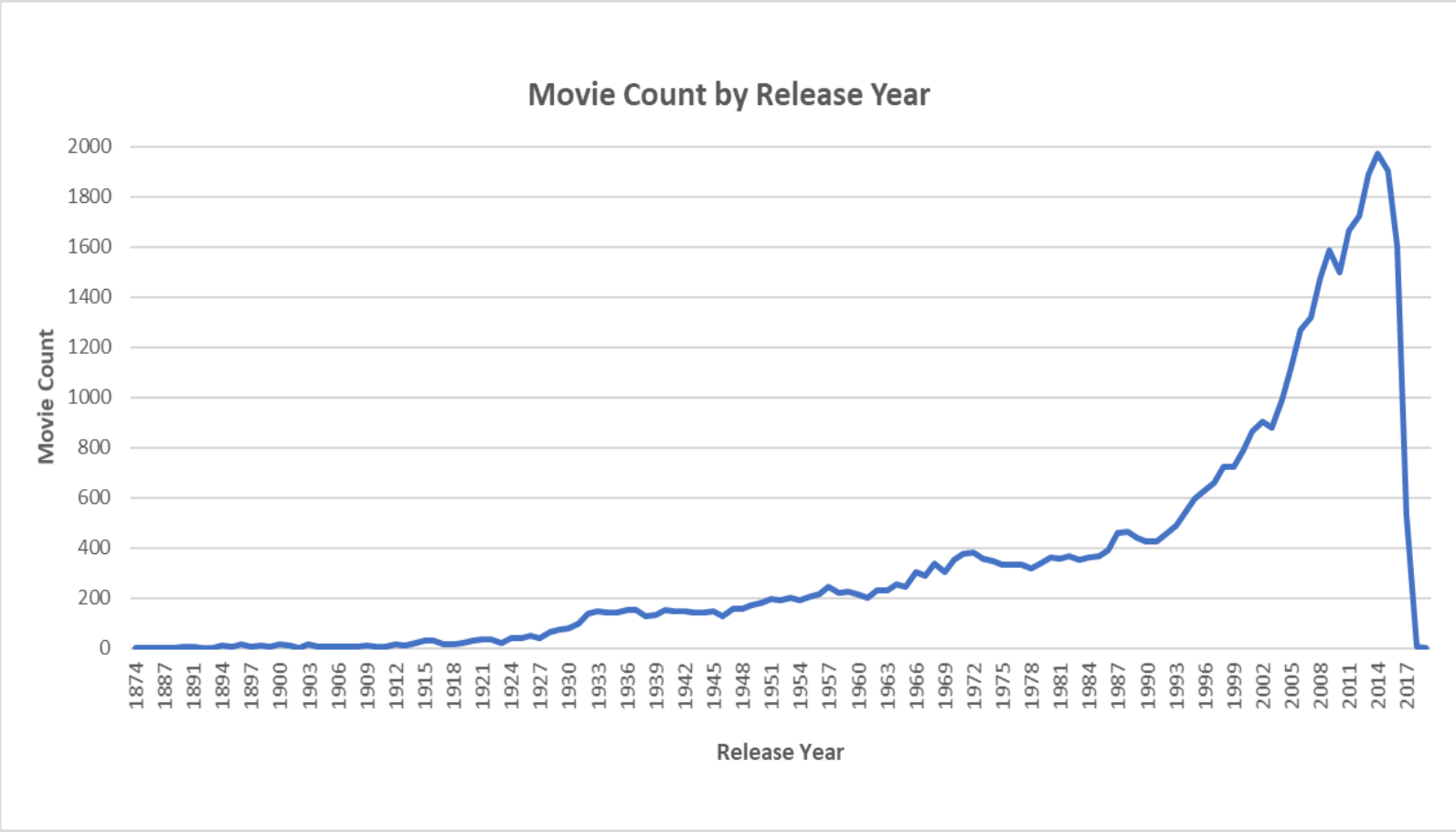
Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Min	9.41	1.59	3.30	0.04	0.65	4.00	0.30	2.61	0.00	0.00	0.40	11.35
Q1	34.84	6.82	20.66	27.39	224.69	24.33	26.29	5.94	0.64	20.24	8.99	38.07
Median	60.27	15.00	72.42	47.09	324.57	83.76	42.37	12.51	20.08	61.90	66.12	64.80
Mean	60.27	34.62	324.66	219.32	295.62	245.05	208.52	42.75	23.15	137.63	157.26	64.80
Q3	85.70	38.20	263.59	55.53	395.50	233.72	153.70	85.56	42.58	179.55	302.17	91.52
Max	111.13	135.68	1262.89	966.55	532.68	1156.73	1342.00	107.14	52.42	603.90	471.22	118.24
IQR	50.86	31.38	242.93	28.15	170.82	209.39	127.41	79.63	41.94	159.31	293.19	53.45
Lower Limit	-41.45	-40.25	-343.74	-14.83	-31.54	-289.76	-164.83	-113.51	-62.26	-218.72	-430.79	-42.10
Upper Limit	161.99	85.27	627.99	97.75	651.73	547.81	344.81	205.01	105.49	418.51	741.96	171.69

Science Fiction

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Min	3.00	0.05	0.36	19.17	0.03	0.10	0.34	0.02	0.55	0.07	0.00	11.21
Q1	6.64	19.14	31.07	38.88	54.60	6.83	2.78	1.42	3.55	0.42	57.11	57.13
Median	32.25	33.40	64.57	64.08	124.55	70.69	30.52	9.20	10.22	2.02	104.95	107.45
Mean	36.48	63.29	116.60	415.31	234.25	242.97	100.30	89.04	97.26	26.11	150.18	94.59
Q3	44.46	93.66	123.73	611.90	384.01	302.16	73.19	81.86	27.11	2.50	156.00	131.43
Max	96.06	183.99	691.21	1519.56	850.00	1091.41	519.31	477.20	716.39	125.54	752.10	168.84
IQR	37.82	74.52	92.66	573.02	329.41	295.33	70.41	80.44	23.56	2.08	98.89	74.30
Lower Limit	-50.10	-92.64	-107.92	-820.65	-439.51	-436.16	-102.84	-119.24	-31.79	-2.69	-91.23	-54.31
Upper Limit	101.20	205.44	262.72	1471.43	878.13	745.15	178.80	202.52	62.45	5.61	304.34	242.87

Descriptive Statistics of Revenue by Release month for Top 3 Genres

Limitations & Biases (Data Collection)



Biases in Coverage: It should be noted that while this dataset covers movies between 1874 and 2020, as can be seen from the line plot on the right, a significant portion of the data is dominated by recent movies, especially those released after the year 2000.

Given this, all variables with be dominated by data associated with movies that were released by the year 2000 onwards.

Likewise, the budgets and revenues for movies in the past have not been adjusted for inflation and represent their value relative to their release date.

Limitations & Biases (Data Processing & Insight)

Data Processing

It should be noted that 0 and blanks values existed in the following variables used in the analysis:

- Budget (MAR) – many values were missing for older movies
- Revenue (MAR) – likewise, many values were missing for older movies.
- Genre (MCAR)
- Release Date (MCAR)

In analyses using these variables, records were removed if there was a zero or blank value, that would have inadvertently skewed the analaysis

Data Insight

While from the provided analyses we can see that revenue has a somewhat strong correlation with budget and release date, it is possible that revenue for the top 3 genres of movies are also controlled by other variables not analysed (eg. Production company, actor) and variables not existing in the provided dataset (marketing, etc).

Next Steps

The proposed next steps for this analysis suggests that:

1. For movies to succeed in popularity they should fall within the top 3 Genres (Adventure, Family, Science Fiction).

2. Once within these genres, to generate revenue, a proportional budget should be allocated to the movie

3. For financial success at the box office, movies within these genres should be released between the months of March to June.

4. Runtime does not have a significant effect on the revenue a movie within these genres is able to generate.