Project 1: Movie Performance Analysis

By: Sanureet Bhullar, Julia Brunett, Shante Snell, Vasu Manikarnika, and Dana Walker

Overarching Question: How does a movie's budget predict its performance?

How do we define performance?

- Performance = Revenue
  - How much a movie made.
- Performance = Ratings
  - How popular a movie is.

Data Retrieval

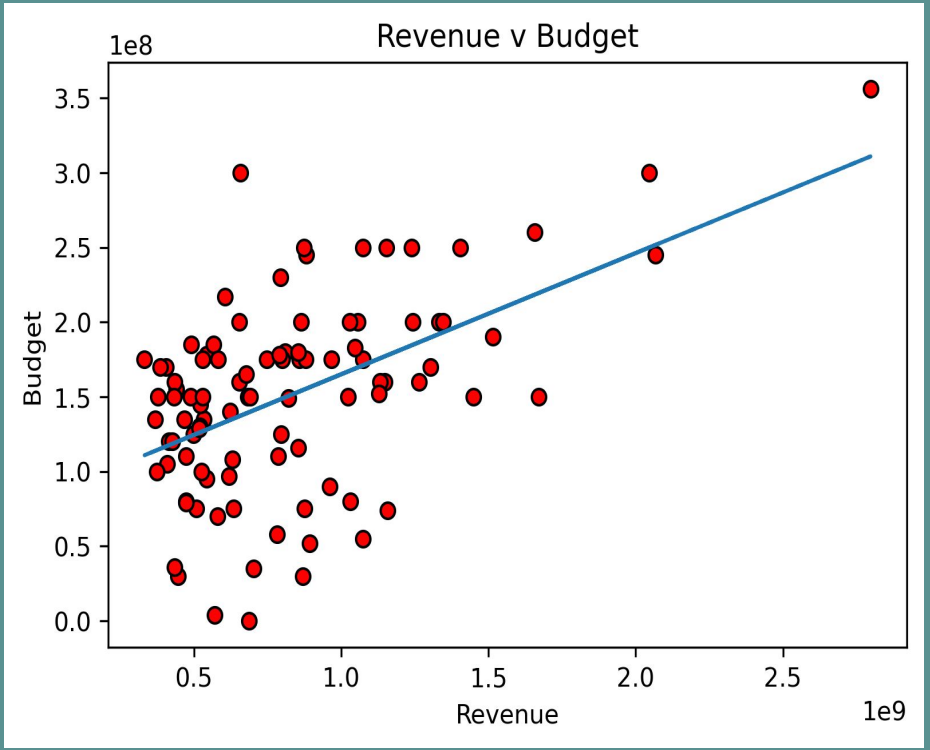
- **Retrieved from:**
  - **TMDb**: The Movie Database API
  - **OMDb**: Open Movie Database API
- Created initial DataFrames from API calls
- Fixed data type formatting of some columns

Data Cleanup & Exploration

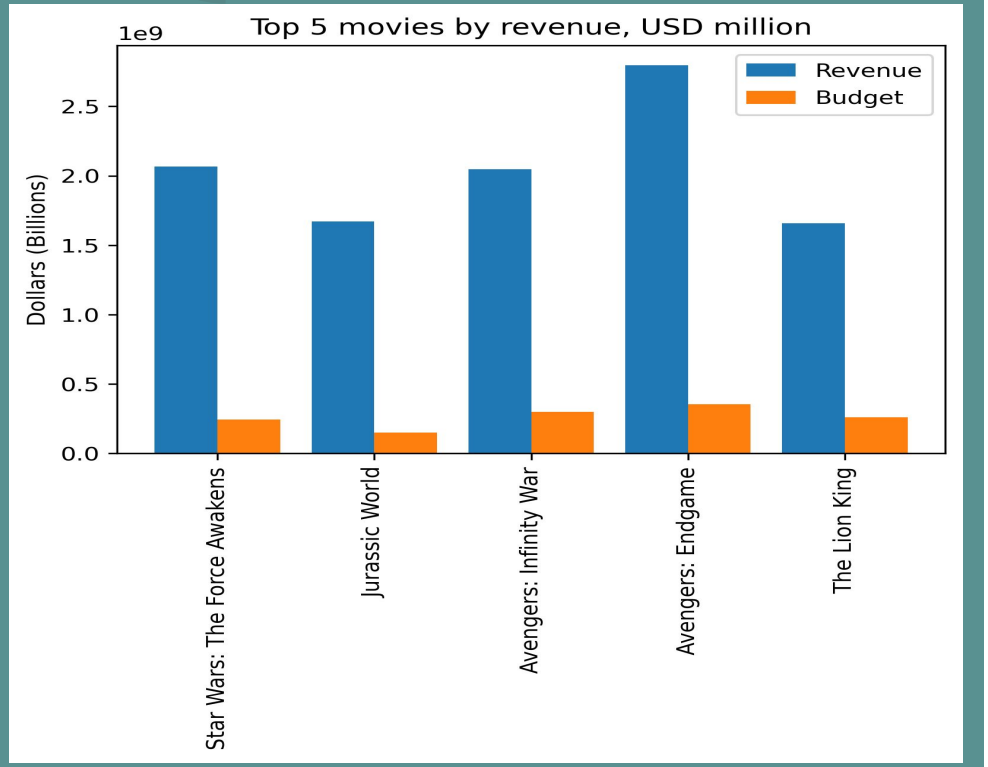
- Used Pandas to merge the files to create 3 individual csv files for analysis
- Renamed some columns to be more readable and only kept columns needed for each of the DataFrame csv files

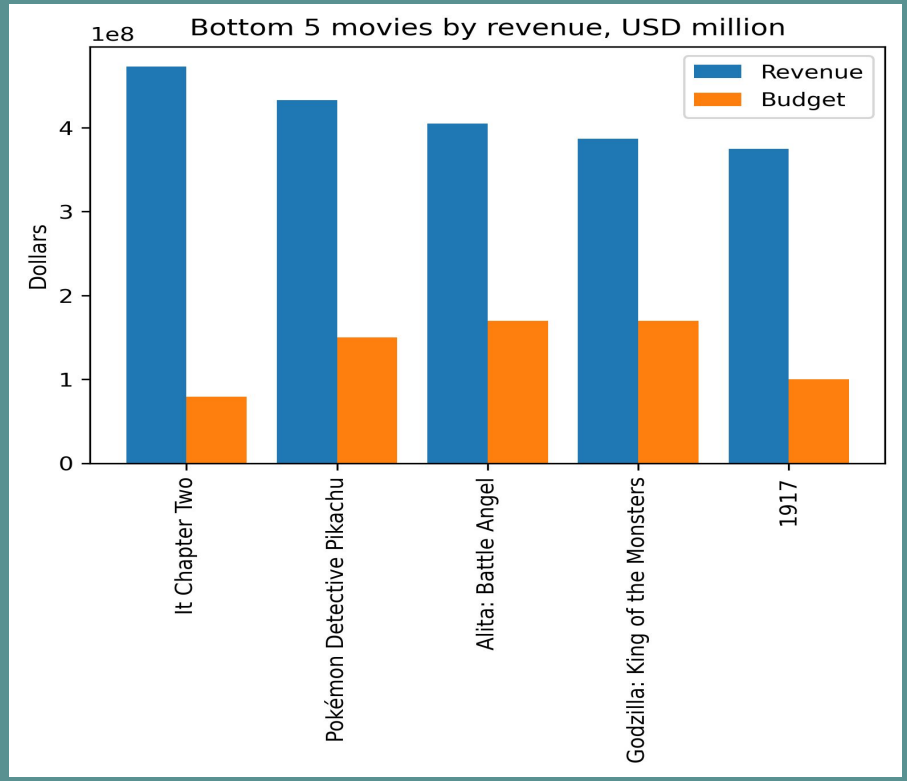
Data Analysis: Budget vs. Revenue

The correlation coefficient between Revenue and Budget is 0.52

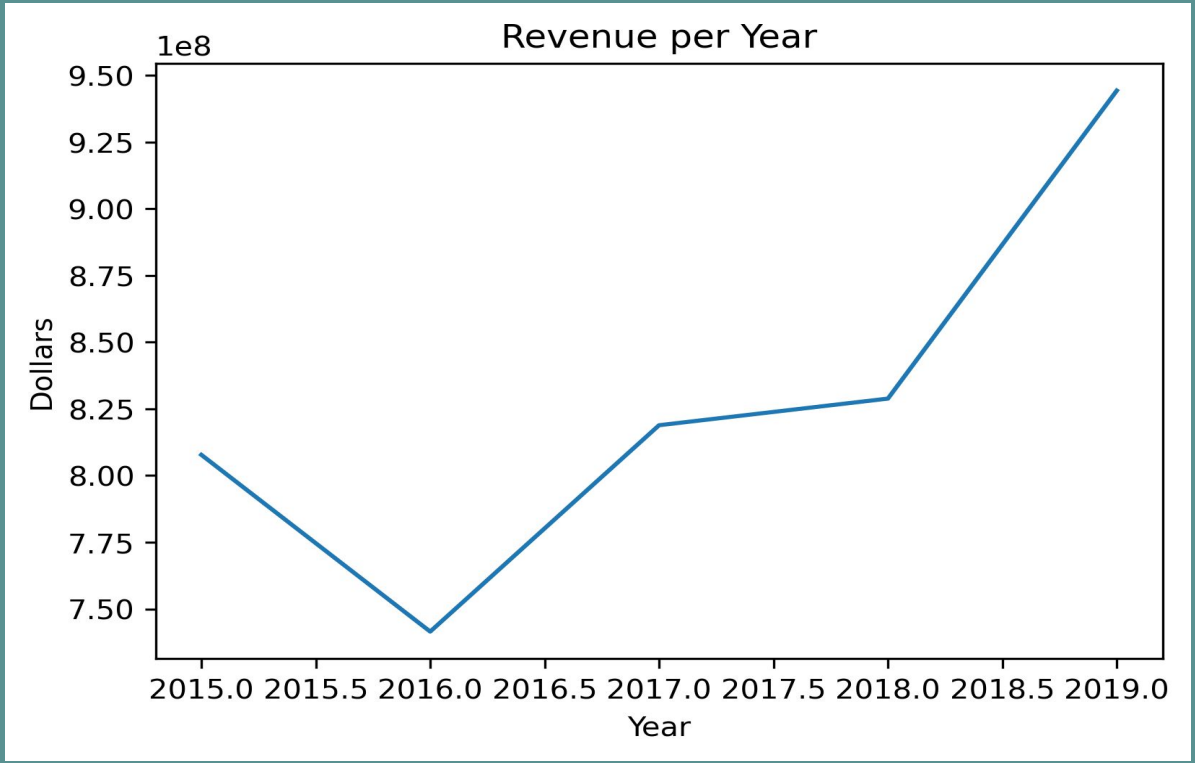


Budget vs. Revenue





Budget vs. Revenue



Data Analysis: Revenue vs. Ratings

In [4]:
```
# Clean Data
clean_df = ratings_df[["Title","Budget","Revenue","Rotten Tomatoes","TMDb","IMDb","Metascore"]]

clean_df.head(20)
```

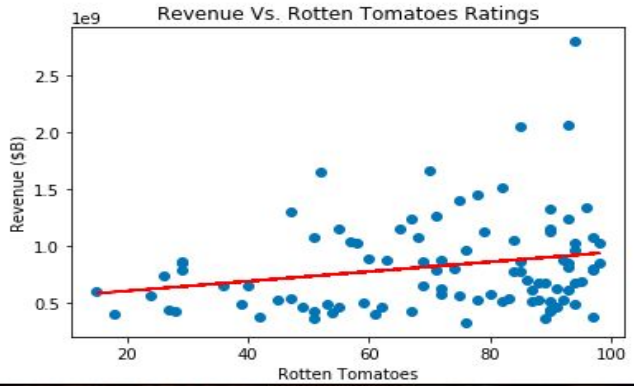
Out[4]:

	Title	Budget	Revenue	Rotten Tomatoes	TMDb	IMDb	Metascore
0	Star Wars: The Force Awakens	245000000	2068223624	93.0	7.4	7.9	80.0
1	Jurassic World	150000000	1671713208	70.0	6.7	7.0	59.0
2	Furious 7	190000000	1515047671	82.0	7.3	7.1	67.0
3	Avengers: Age of Ultron	250000000	1405403694	75.0	7.3	7.3	66.0
4	Minions	74000000	1156730962	55.0	6.4	6.4	56.0
5	Spectre	245000000	880674609	63.0	6.5	6.8	60.0
6	Inside Out	175000000	857611174	98.0	7.9	8.1	94.0
7	Mission: Impossible - Rogue Nation	150000000	682330139	94.0	7.2	7.4	75.0
8	The Hunger Games: Mockingjay - Part 2	160000000	653428261	69.0	6.9	6.6	65.0
9	The Martian	108000000	630161890	91.0	7.7	8.0	80.0
10	Fifty Shades of Grey	4000000	571006128	24.0	5.9	4.1	46.0
11	Cinderella	95000000	543514353	83.0	6.8	6.9	67.0

Revenue vs. Rotten Tomatoes

```
x_values = clean_df["Rotten Tomatoes"]
y_values = clean_df["Revenue"]
(slope, intercept, rvalue, pvalue, stderr) = linregress(x_values, y_values)
regress_values = x_values * slope + intercept
line_eq = "y = " + str(round(slope,2)) + "x + " + str(round(round(intercept,2)))
plt.scatter(x_values,y_values)
plt.plot(x_values,regress_values,"r-" )
plt.annotate(line_eq,(6,0),fontsize=12,color="blue")
plt.title('Revenue Vs. Rotten Tomatoes Ratings')
plt.xlabel('Rotten Tomatoes')
plt.ylabel('Revenue ($B)')
print(f"The r-squared is: {rvalue}")
plt.savefig("Images/RottenTomates_Revenue.png")

plt.show()
```

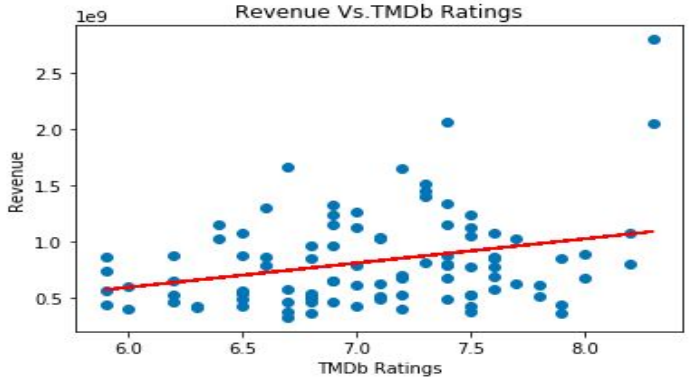
```
The r-squared is: 0.22594909665880133
```



Revenue vs. The Movie Database

```
x_values = clean_df["TMDb"]
y_values = clean_df["Revenue"]
(slope, intercept, rvalue, pvalue, stderr) = linregress(x_values, y_values)
regress_values = x_values * slope + intercept
line_eq = "y = " + str(round(slope,2)) + "x + " + str(round(round(intercept,2)))
plt.scatter(x_values,y_values)
plt.plot(x_values,regress_values,"r-" )
plt.annotate(line_eq,(6,0),fontsize=12,color="blue")
plt.title('Revenue Vs.TMDb Ratings')
plt.xlabel('TMDb Ratings')
plt.ylabel('Revenue')
print(f"The r-squared is: {rvalue}")
plt.savefig("Images/TMDb_Revenue.png")

plt.show()
```

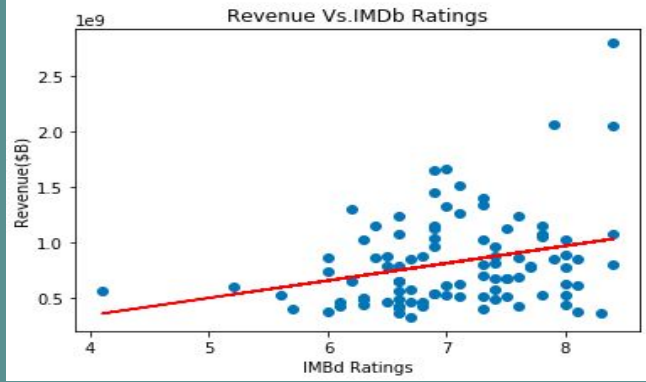
```
The r-squared is: 0.3004891560154141
```



Revenue vs. IMDb

```
x_values = clean_df["IMDb"]
y_values = clean_df["Revenue"]
(slope, intercept, rvalue, pvalue, stderr) = linregress(x_values, y_values)
regress_values = x_values * slope + intercept
line_eq = "y = " + str(round(slope,2)) + "x + " + str(round(round(intercept,2)))
plt.scatter(x_values,y_values)
plt.plot(x_values,regress_values,"r-" )
plt.annotate(line_eq,(6,0),fontsize=12,color="blue")
plt.title('Revenue Vs.IMDb Ratings')
plt.xlabel('IMBd Ratings')
plt.ylabel('Revenue($B)')
print(f"The r-squared is: {rvalue}")
plt.savefig("Images/IMBd_Revenue.png")

plt.show()
```

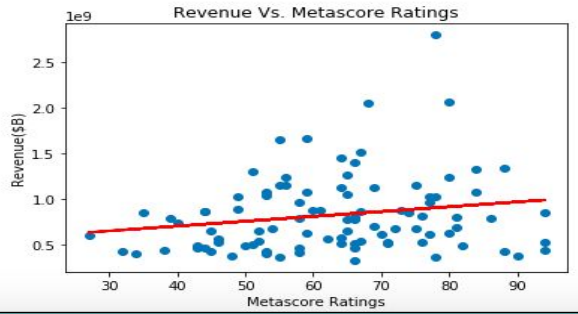
```
The r-squared is: 0.2831761177840927
```



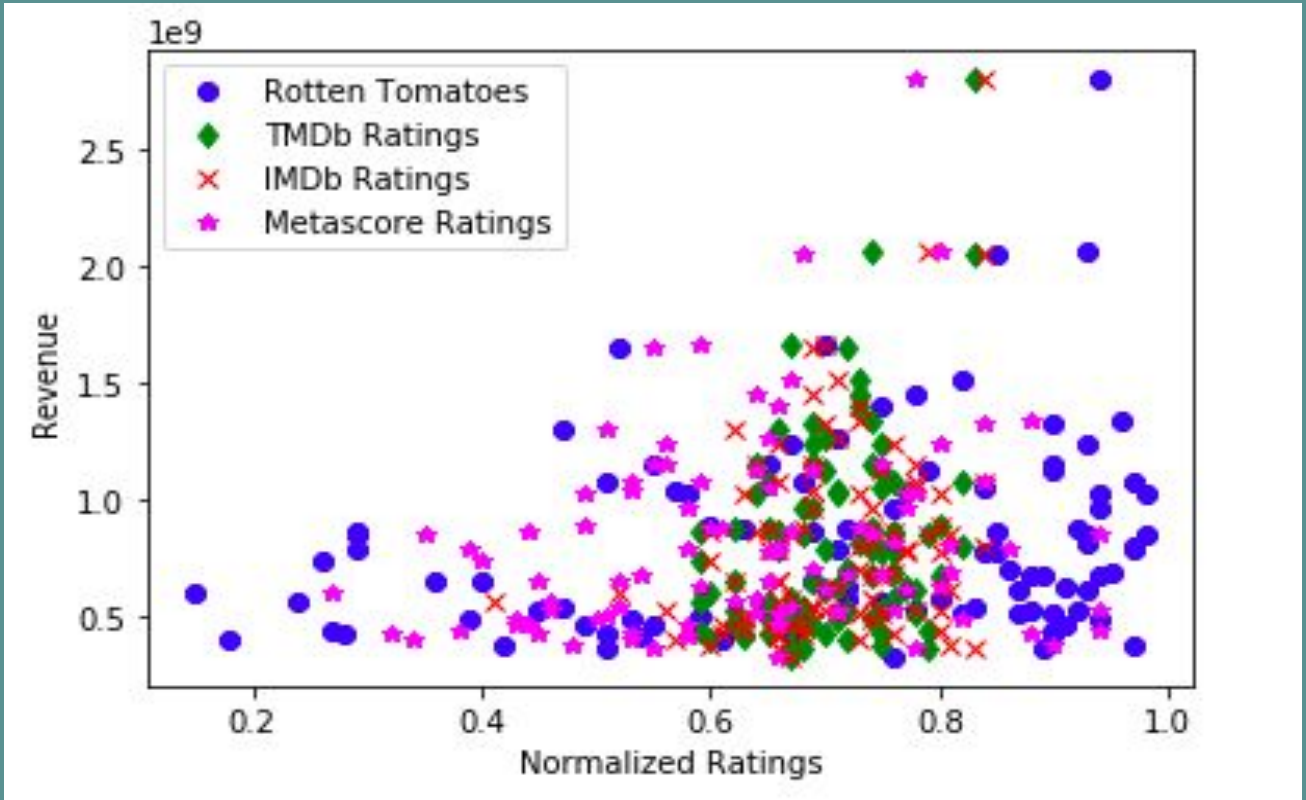
Revenue vs. Metacritic

```
x_values = clean_df["Metascore"]
y_values = clean_df["Revenue"]

mask = ~np.isnan(x_values) & ~np.isnan(y_values)
slope, intercept, r_value, p_value, std_err = linregress(x_values[mask], y_values[mask])

# (slope, intercept, rvalue, pvalue, stderr) = linregress(x_values, y_values)
regress_values = x_values * slope + intercept
line_eq = "y = " + str(round(slope,2)) + "x + " + str(round(round(intercept,2)))

plt.scatter(x_values,y_values)
plt.plot(x_values,regress_values,color="red",linewidth=2)
plt.annotate(line_eq,(6,0),fontsize=14,color="blue")
plt.title('Revenue Vs. Metascore Ratings')
plt.xlabel('Metascore Ratings')
plt.ylabel('Revenue($B)')
print(f"The r-squared is: {rvalue}")
plt.savefig("Images/Metascore_Revenue.png")
# print(slope, intercept, rvalue, pvalue, stderr)
# print(x_values)
# print(y_values)
plt.show()
```

```
The r-squared is: 0.2831761177840927
```

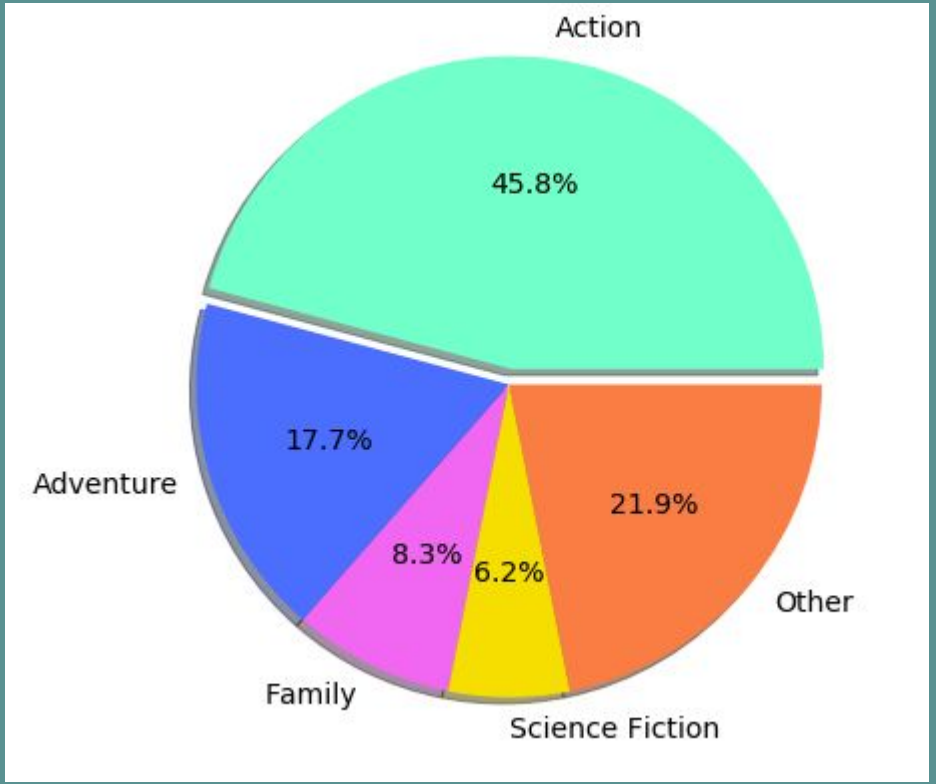


Revenue vs. All Ratings



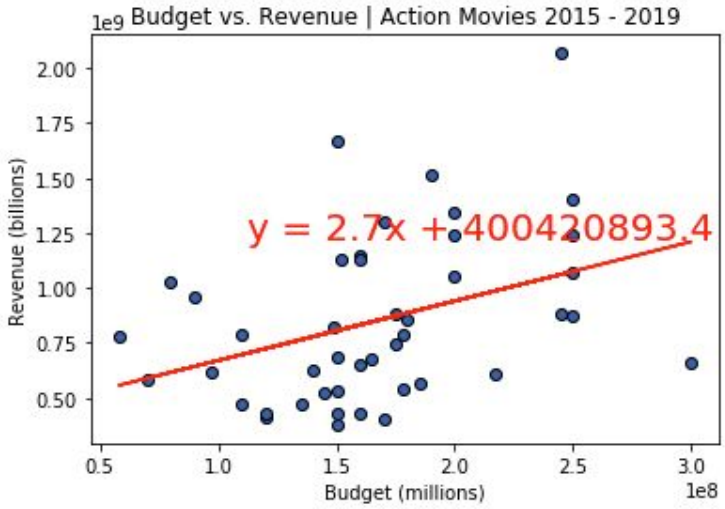
Data Analysis: Genre Breakdown

- **Action**: 44 movies
- **Adventure**: 17 movies
- **Family**: 8 movies
- **Sci-Fi**: 6 movies
- **Other**: 21 movies

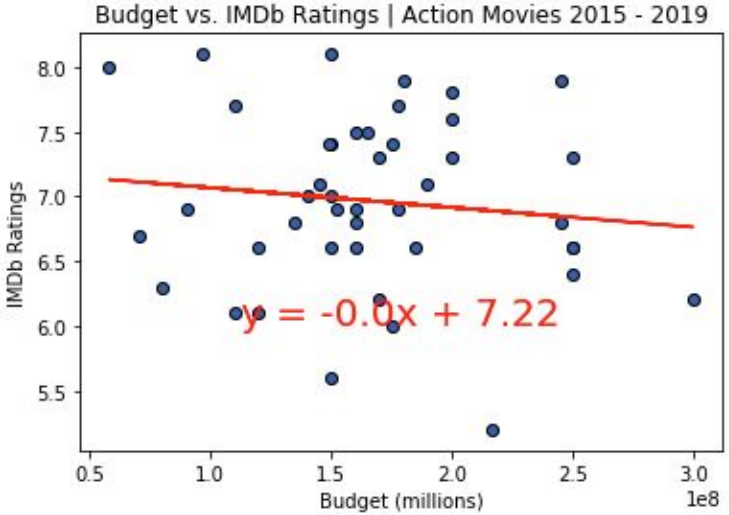


Action Movies

```
The r-squared value is: 0.37337653559717365
The slope is: 2.7003163127413945
The Y-Intercept is: 400420893.40433884
```

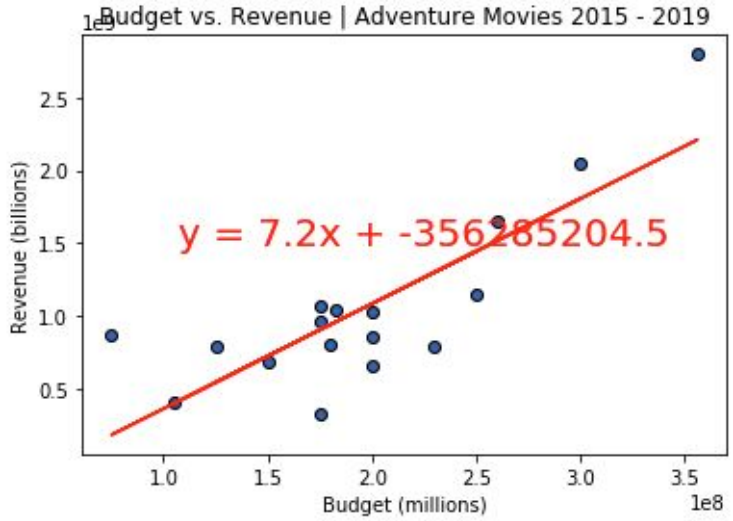



```
The r-squared value is: -0.11947908242985701
The slope is: -1.5215015183096764e-09
The Y-Intercept is: 7.219688628247154
```

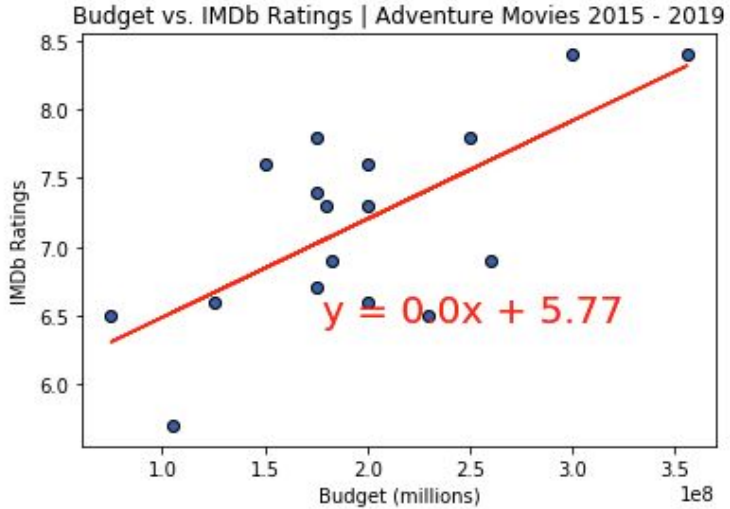


Adventure Movies

```
The r-squared value is: 0.8201201313625246
The slope is: 7.203430377221383
The Y-Intercept is: -356285204.5024823
```

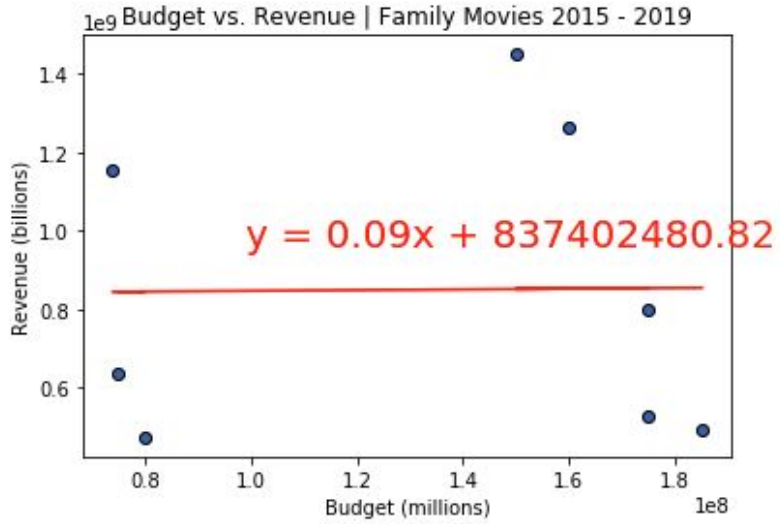



```
The r-squared value is: 0.685034012888139
The slope is: 7.159200894938582e-09
The Y-Intercept is: 5.770319306576475
```

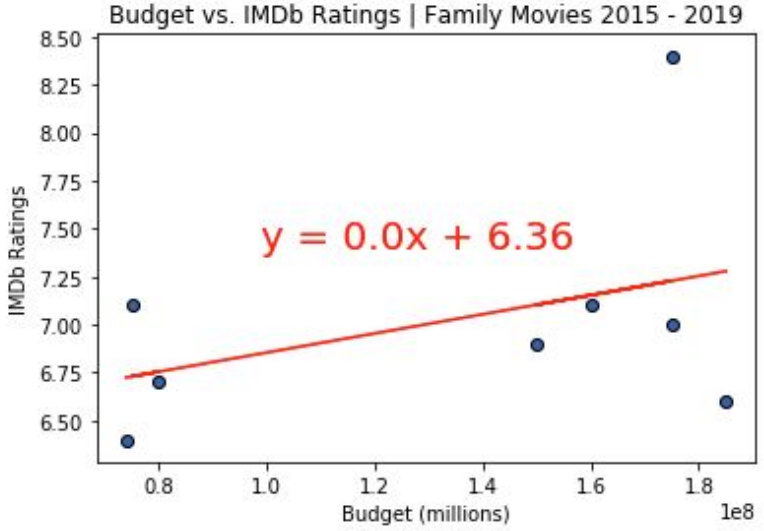


Family Movies

```
The r-squared value is: 0.011794927394165025
The slope is: 0.09288263259627623
The Y-Intercept is: 837402480.8239499
```

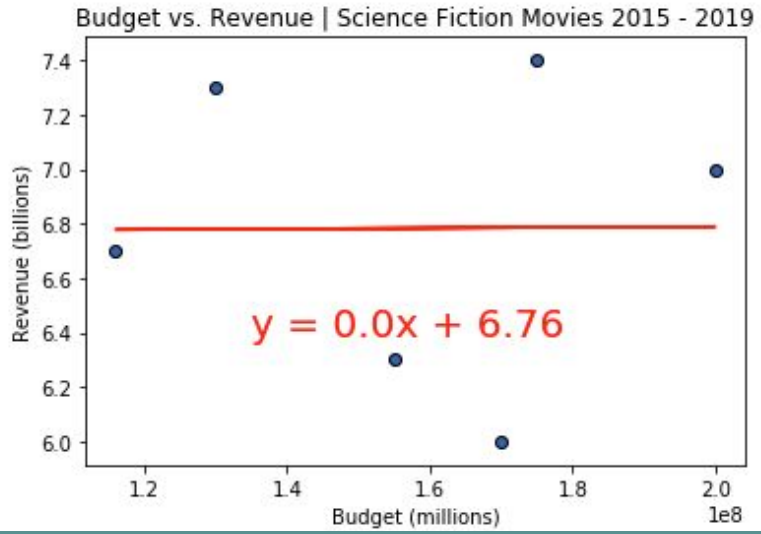



```
The r-squared value is: 0.40240847592801
The slope is: 4.987715715004587e-09
The Y-Intercept is: 6.355399165260635
```

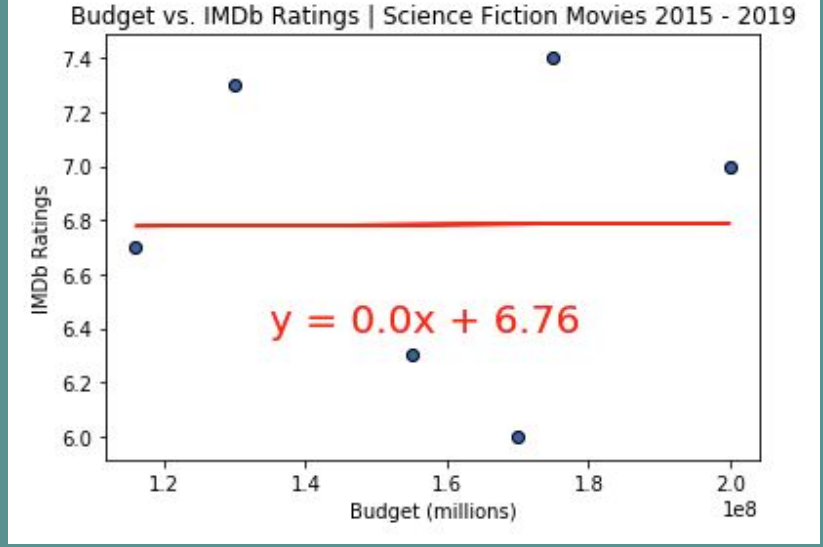


Science Fiction

```
The r-squared value is: 0.007771008316512648
The slope is: 1.4025245441795344e-10
The Y-Intercept is: 6.761220196353436
```




```
The r-squared value is: 0.007771008316512648
The slope is: 1.4025245441795344e-10
The Y-Intercept is: 6.761220196353436
```

Conclusions

- Budget has a positive effect on revenue while revenue has a positive, but less significant, effect on ratings.
  - Budget vs. Revenue | R-Value: .52
  - Revenue vs. Ratings | R-Value: Between .2 & .3

- Broken down by genre, the correlations start to tell us even more for adventure films.
  - Adventure | Budget vs. Revenue | R-Value: .82
  - Adventure | Budget vs. Ratings| R-Value: .68

Follow-Up Potential Analysis

- Analysis on Awards
- Look deeper into the genres
- Expand the sample dataset to include more years and movies

Q&A