**Business Problem**

Often, you will hear that this movie underperformed or overperformed their statistics. But that concept is so vague, I wanted to try to look to clarify and expand on how key aspects and performance stats compare with one another and which movie would be considered a failure, just through the numbers. I wanted to look at how the popularity of a movie is a reflection on the movie’s revenue performance.

**Background/History**

For this project, I chose to use the TMDB (the movie data base) as it has a comprehensive and complete collection of both financial aspects of movies, but also their popularity, which makes this database an excellent tool in comparing monetary success and viewership success. Having grown up watching movies all my life, I find that there are several films that I really enjoyed that have poor popularity ratings. There are also those movies that many have seen and liked but have been considered box office failures.

I want to get a handle on this distinction. Are these situations outliers or is there a pattern of disconnect between a movie’s monetary considerations and its popularity with the masses? This is my goal to find out. I used a Pearson correlation matrix to find variable correlation and a linear regression model to compare the popularity and the profit of different movies.

For my data set, I found a list of approximately 10,000 different movies from a Kaggle source, listed below. This database contains popularity, budget, revenue, title, cast, website, director, tagline, keywords, overview, runtime, genre, production companies, release date, vote count, vote average, release year, and adjusted budget/revenue totals. I chose to focus on the former three topics, popularity, budget, and revenue, as they are most often used in reference to measure a movie success. I also added a fourth topic, profit, which is revenue-budget. I want to find correlation between the different variables then find the regression of popularity compared to profit.

To be specific, there are three different things that I will be measuring. First, I will be looking for the correlations between the budget and the revenue of the film. We often find that different movies either outperform or underperform their budget, but I want to see if there is truly a correlation between these two numbers is or if they are more independent than most think. The next topic under consideration is the correlation between the budget of the film and its popularity. In other words, is it true that high budget films lead to more popular films? Lastly, I wanted to create a logistic regression to create a true idea of whether a movie is successful. I plan to do this using the profit of a movie (Revenue – Budget) As compared to its popularity.

**Data Explanation (Data Prep/Data Dictionary/etc)**

I used R code for my data, as I felt it was the best tool to visualize my data.

*library(ggplot2)*

*getwd()*

*setwd("~/Documents/DSC 680")*

*##SET DIRECTORY*

*library(readr)*

*df <-read\_csv("tmdb-movies.csv")*

*View(df)*

*##IMPORT DATA*

*df1<- df[, c('popularity','budget','revenue')]*

*View(df1)*

*##REDUCED VARIABLES TO THOSE WE WANT*

*profit <- df1$revenue-df1$budget*

*profit*

*##CREATED A NEW VARIABLE, JOINT EDUCATION LEVEL*

*df1$profit <- profit*

*View(df1)*

*##INSERT BACK INTO DF*

*corr <- round(cor(df1),2)*

*View(corr)*

*##FINDS CORRELATIONS OF VARIABLES*

*library(reshape2)*

*melted\_corr <- melt(corr)*

*View(melted\_corr)*

*##MELTS DATA TO BE ABLE TO PUT INTO HEAT MAP*

*ggplot(data = melted\_corr, aes(x=Var1, y=Var2, fill=value)) + geom\_tile(color = "white")+*

*scale\_fill\_gradient2(low = "blue", high = "red", mid = "white",*

*midpoint = 0, limit = c(-1,1), space = "Lab",*

*name="Pearson\nCorrelation") + geom\_text(aes(Var2, Var1, label = value), color = "black", size = 4) +*

*theme(*

*axis.title.x = element\_blank(),*

*axis.title.y = element\_blank(),*

*panel.grid.major = element\_blank(),*

*panel.border = element\_blank(),*

*panel.background = element\_blank(),*

*axis.ticks = element\_blank(),*

*legend.justification = c(1, 0),*

*legend.direction = "horizontal")+*

*guides(fill = guide\_colorbar(barwidth = 7, barheight = 1,*

*title.position = "top", title.hjust = 0.5))*

*##GRAPHED MELTED DF INTO HEAT MAP*

*install.packages("dplyr")*

*install.packages("caTools")*

*install.packages("ROCR")*

*##INSTALLING PACKAGES REQUIRED FOR LINEAR REGRESSION*

*library(dplyr)*

*library(caTools)*

*##loading packages*

*profit\_lm <- lm(profit ~ popularity, data = df1)*

*summary(profit\_lm)*

*##LINEAR REGRESSION MODEL*

*plot(profit\_lm)*

*##Check for homoscedasticity*

*profit\_graph<-ggplot(df1, aes(x=profit, y=popularity))+geom\_point()+*

*geom\_smooth(method="lm", col="red")+theme\_bw() +*

*labs(title = "Movie profits VS popularity",*

*x = "Profit",*

*y = "Popularity")*

*profit\_graph*

*##PLOT THE DATA WITH LINEAR REGRESSION MODEL*

![Chart, treemap chart

Description automatically generated](data:application/pdf;base64,)

![Chart, scatter chart

Description automatically generated](data:application/pdf;base64,)![Chart, scatter chart

Description automatically generated](data:application/pdf;base64,)

![Chart, scatter chart

Description automatically generated](data:application/pdf;base64,)![Chart, line chart

Description automatically generated](data:application/pdf;base64,)

![Chart, scatter chart

Description automatically generated](data:application/pdf;base64,)

summary(profit\_lm)

Call: lm(formula = profit ~ popularity, data = df1)

Residuals:

Min 1Q Median 3Q Max

-1.484e+09 -1.857e+07 -3.148e+06 7.173e+06 1.986e+09

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -14068309 857754 -16.40 <2e-16 \*\*\*

popularity 60741709 720277 84.33 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75090000 on 10864 degrees of freedom

Multiple R-squared: 0.3956, Adjusted R-squared: 0.3956

F-statistic: 7112 on 1 and 10864 DF, p-value: < 2.2e-16

**Analysis**

The first visual was the correlation. As you can see from the correlation heat map, all correlations are positive and mid to strong correlation. Below is a chart of the correlations:

|  |  |  |
| --- | --- | --- |
| Variable 1 | Variable 2 | Correlation factor |
| Popularity | Budget | 0.55 |
| Popularity | Revenue | 0.66 |
| Popularity | Profit | 0.63 |
| Budget | Revenue | 0.73 |
| Budget | Profit | 0.57 |
| Revenue | Profit | 0.98 |

Let’s look at some takeaways from these correlations. First, we see the obvious that revenue and profit are about as strongly correlated as possible, which makes complete sense. The more you earn, the higher the profit. The next topic, one which I will be investigating further later, is the correlation between Profit and Popularity. You can see that it has a 0.63 correlation. It is a positive correlation, but I would consider it just below a strong correlation (0.7 and above can be considered strong). The last main takeaway from is that the budget of a film and its popularity came in last on terms of correlation strength at 0.55. That is still closer to strong than weak, but it is lower than the other correlations.

The next four visuals look at the residuals and their strengths. First though, let’s talk about residuals. “When you perform simple linear regression (or any other type of regression analysis), you get a line of best fit. The data points usually don’t fall *exactly*on this regression equation line; they are scattered around. A residual is the vertical distance between a data point and the regression line. Each data point has one residual. They are:

* Positive if they are above the regression line,
* Negative if they are below the regression line,
* Zero if the regression line actually passes through the point,

As residuals are the difference between any data point and the regression line, they are sometimes called “**errors**.” Error in this context doesn’t mean that there’s something wrong with the analysis; it just means that there is some unexplained difference. In other words, the residual is the error that isn’t explained by the regression line.

**Residual = Observed value – predicted value e = y – ŷ”**

Each of the visuals looks at the residuals and look to see how much variance there is in the residuals. What we see most residuals are hover right around the line of best fit however, where the fit become weaker is in the larger numbers. The variance grows as the movies with the profit and popularity grows. So, the high popularity movies might not have such a high profit, or the high profit movies might not have a concurrent popularity.

This pattern continues with the linear regression model, where we see most of the movies scattered near the line of best fit closer to the start of the line. The pattern continues, when you move forward on the line, as further you go, the further the movies data is from the line of best fit.

**Conclusion**

What we see here is a duality, most/ average movies compared to the exceptional movies. We see that movies with high popularity or/and high profits veered off and did not follow the expected ratio, while most movies followed the line of best fit, where the popularity of the movie follows proportionally to the profits. This tells us that those outliers I mentioned in the background, those “successful” movies that have low popularity or the popular movies with low ratings are present only if one of the variables takes off and is far above the normal, while the other variable stays consistent with the linear regression line.

The other focus of this analysis was looking at the correlations of the variables. The two correlations we wanted to focus on were budget compared to popularity as well as budget versus revenue. We found that the Pearson correlation factors were 0.55 and 0.73 respectively. Both are positive and are on the stronger side, but the correlation of budget and revenue is much more correlated as compared to budget and popularity. It turns out that monetary relations of movies correlate stronger than that of its popularity.

**Assumptions/limitations/Challenges/Ethical Assessment**

There are a few different considerations when looking at this project. First off, this is a purely numerical, populist ranking of the movies. It does not take into consideration the actual cinematic topics, but rather focused solely on the numerical performance of the films. Also, my numbers that I plan to use reward populist firms. While these films tend to have the highest profit and popularity, they may pander to audiences as compared to other smaller films, which have more meaning and depth, but are less appealing to all. Next, I am using the iMDB rating, which is just one measure of popularity. It does not factor in other movie ranking sites, such as rotten tomatoes, which may have different measures of popularity. Finally, I am reliant on third hand figures of budget and revenue, which may not be the most accurate data, for all I know.

I don’t foresee too many issues. One issue that may pop up is the use of profit in the linear regression. It is a different calculation that I will need to add on and will result in negative totals. I worry that these negative numbers will affect the linear function. Another issue is that the database is missing some revenues, or a movie might not have made any revenue (it got scraped in production). The concern is whether to include zero revenue films into consideration, or, if not, which films are removed.

**References**

* https://www.kaggle.com/code/muhammetgamal5/tmdb-5000-movies/notebook
* https://www.statisticshowto.com/residual/