Final Project

Kate Kang, Hannah Cheren, Devashi Ghoshal, Gary Shi

2021 4 26

Introduction

Hundreds of thousands of movies are released each year, but what factors determine what makes certain movies more successful than others? We decided to explore this question more in-depth by taking a look at a data set of movies. We decided to look at two variables that we thought reflect how successful a movie is: the gross revenue and the movie's score. We then decided to look at a third variable, the movies' budget, and compare it to the two variables above to see if this had anything to do with how successful a movie is.

What we mean by saying that "we're using these variables to determine a movie's success" is that we are assuming that movies with higher gross revenues are more successful than movies with lower gross revenues, and movies that receive a higher user score are more successful than movies with a lower user score.

Question of Interest

Is there a positive, linear relationship between a movie's budget and its respective gross revenue or score received?

Thesis Statements

Although we believe that the size of a movie's budget could be important to a movie's success, we do not think we will see a relationship between the budget and the gross revenue/the score in our analysis. In other words, we do not believe there's a relationship between the budget and (what we defined above as) a movie's success.

Background

Description of Data and Variables

We obtained this data set from Kaggle, a website that contains free databases for users to analyze. We chose the data set titled "Movie Industry: Three decades of movies." The movies were released during the years 1986-2016 and there are 220 movies listed per year - which makes 6820 movies total in the data set. The movies in the dataset are pulled from countries all over the world (45 different countries). The data was scraped from IMDb, an online database that contains information related to the film industry.

Below is a more in-depth description of the variables we used and how we used them in our analysis:

- budget: The budget of a movie (some movies didn't have budgets, so a "0" was entered. We later decided to get rid of these rows, which we explain later) (dollar)
- gross: The gross revenue the movie made in total (dollar)
- score: On the IMDb website, where the data was scrapped from, users can put in a score on a scale of 1 to 10. We assumed that a lower score means that the movie objectively isn't so good, and a higher score means that the movie is objectively good. (1-10 scale)

Citation of Data

Grijalva, D. (2017, October 05). Movie industry. Retrieved April 12, 2021, from https://www.kaggle.com/danielgrijalvas/movies (https://www.kaggle.com/danielgrijalvas/movies)

Background Information

We did not feel that we needed any additional background information to understand the data, the question we posed, or how the data relates to the question asked.

Unusual Factors

An unusual factor that we noticed was that for the ‘budget’ variable column, the creator of the dataset would enter a 0 if the movie did not have a budget. For this reason, we decided to omit all of the rows that how a 0 entered in the budget column. We did this because the question we posed asks whether or not a movie’s budget relates to its gross income/score, so we want only movies with budgets as part of our analysis. Movies that don’t have specified budgets are useless to us for this analysis, which is why we can justify why we got rid of these rows.

Intention with Report

To keep our Analysis section organized, we will split it into two parts. The first part will show the analysis done on the variables ‘budget’ and ‘gross.’ The second part will show the same analysis, but done on the variables ‘budget’ and ‘score’ instead. Keep this format in mind when viewing the Analysis section.

For the rest of the report, we intend to answer the question - “is a movie’s budget related to a movie’s success in terms of its gross revenue or the score it receives?” To do this, we first want to calculate the correlation coefficient and create a graph of the regression between the ‘budget’ and ‘revenue’ data and then do the same thing with the ‘budget’ and ‘score’ data. We do this because we want to see the strength of the relationship between the respective variables. We then want to learn more about the linear relationship of the variables, respectively. To do this, we will plot ‘budget’ vs. ‘gross’ and then ‘budget’ vs. ‘score’ on a scatterplot and create a least-squares line and create a linear regression model from this plot so we can visualize the relationship in algebraic terms. We then want to create a plot of the residuals to check that a linear model is an appropriate model to use for both relationships.

Movie Data Set

Here is the head of our data.

```
##                        name   budget     gross score
## 1              Stand by Me  8000000  52287414   8.1
## 2 Ferris Bueller's Day Off  6000000  70136369   7.8
## 3                  Top Gun 15000000 179800601   6.9
## 4                   Aliens 18500000  85160248   8.4
## 5  Flight of the Navigator  9000000  18564613   6.9
## 6                  Platoon  6000000 138530565   8.1
```

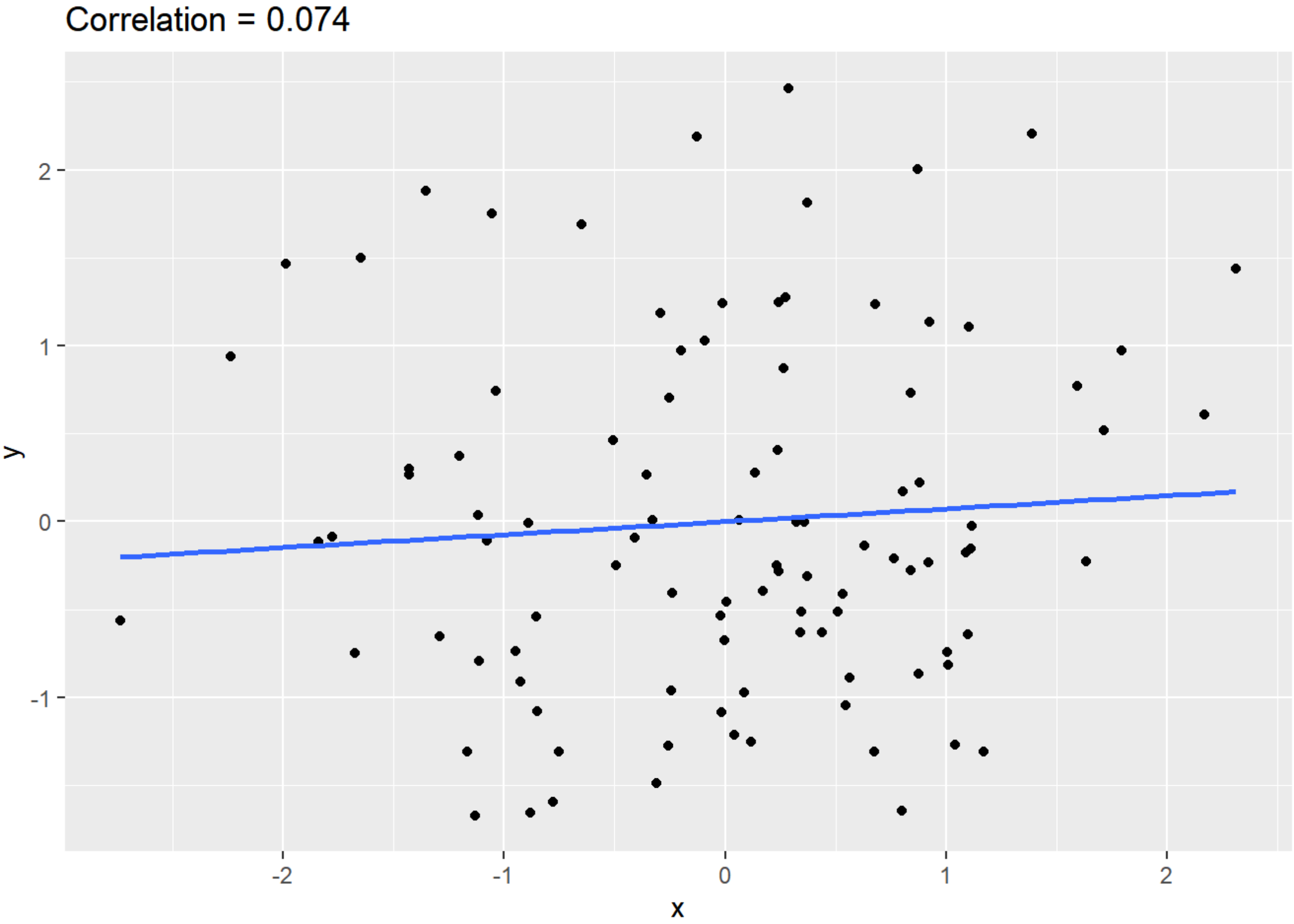
Analysis

Budget vs. Score

We want to view thew relationahip between budget and score and how they associate with one another. We do this by taking a look at the correlation between the two.

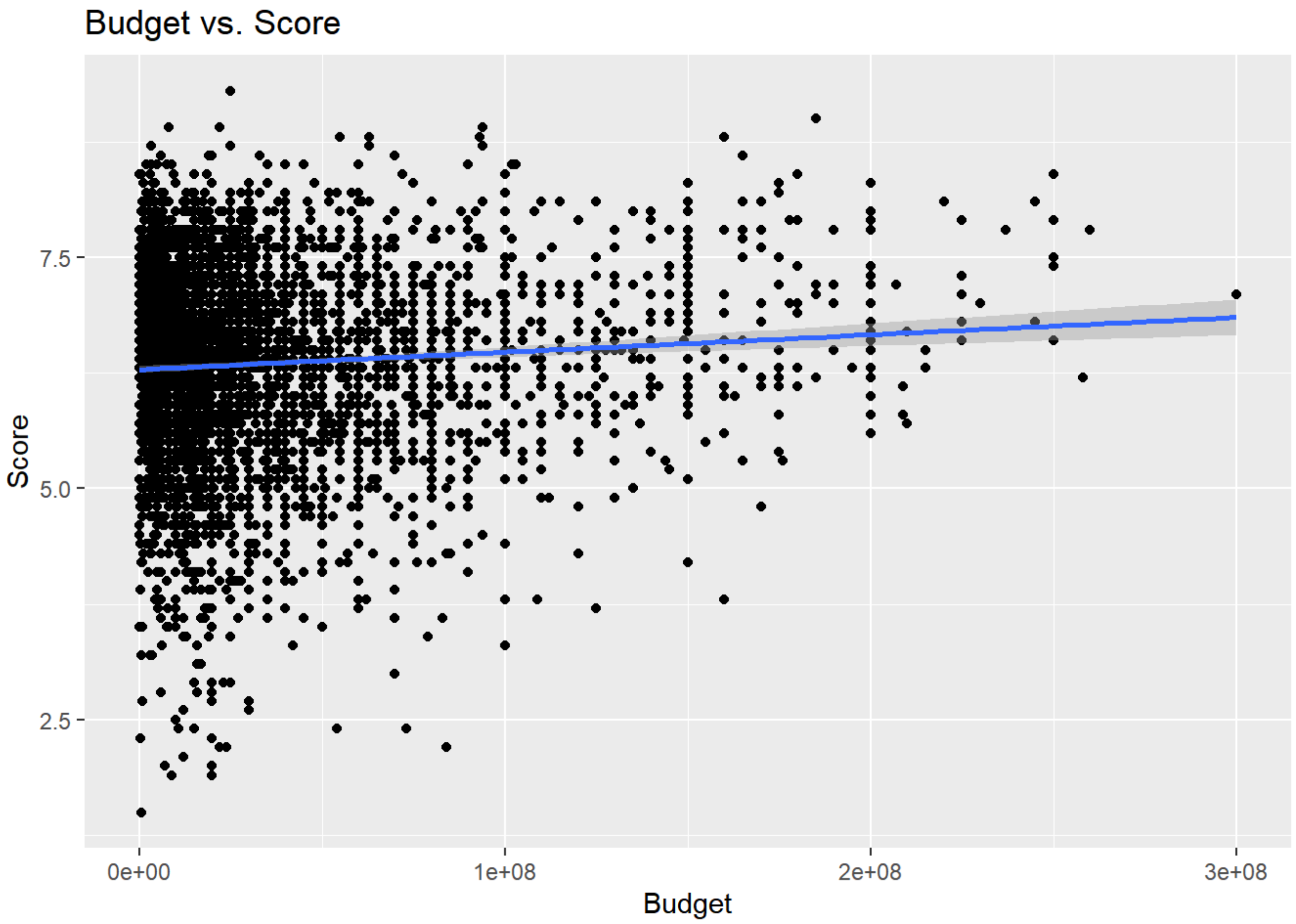
```
##             r
## 1 0.07450023
```

Our r value is 0.074. The correlation coefficient is positive but it is very close to 0. As a result, we can say that there is a very weak positive relationship between Budget and Score.



The correlation plot supports our r value as the slope of the line is positive and not very steep.

Another way we visualize the relationship between the budget and score is to plot them against each other. Although this does not represent a causal relationship, this does show how they relate to each other between the movies in our dataset. We will use the least squares regression line to take a look at the linear relationship.



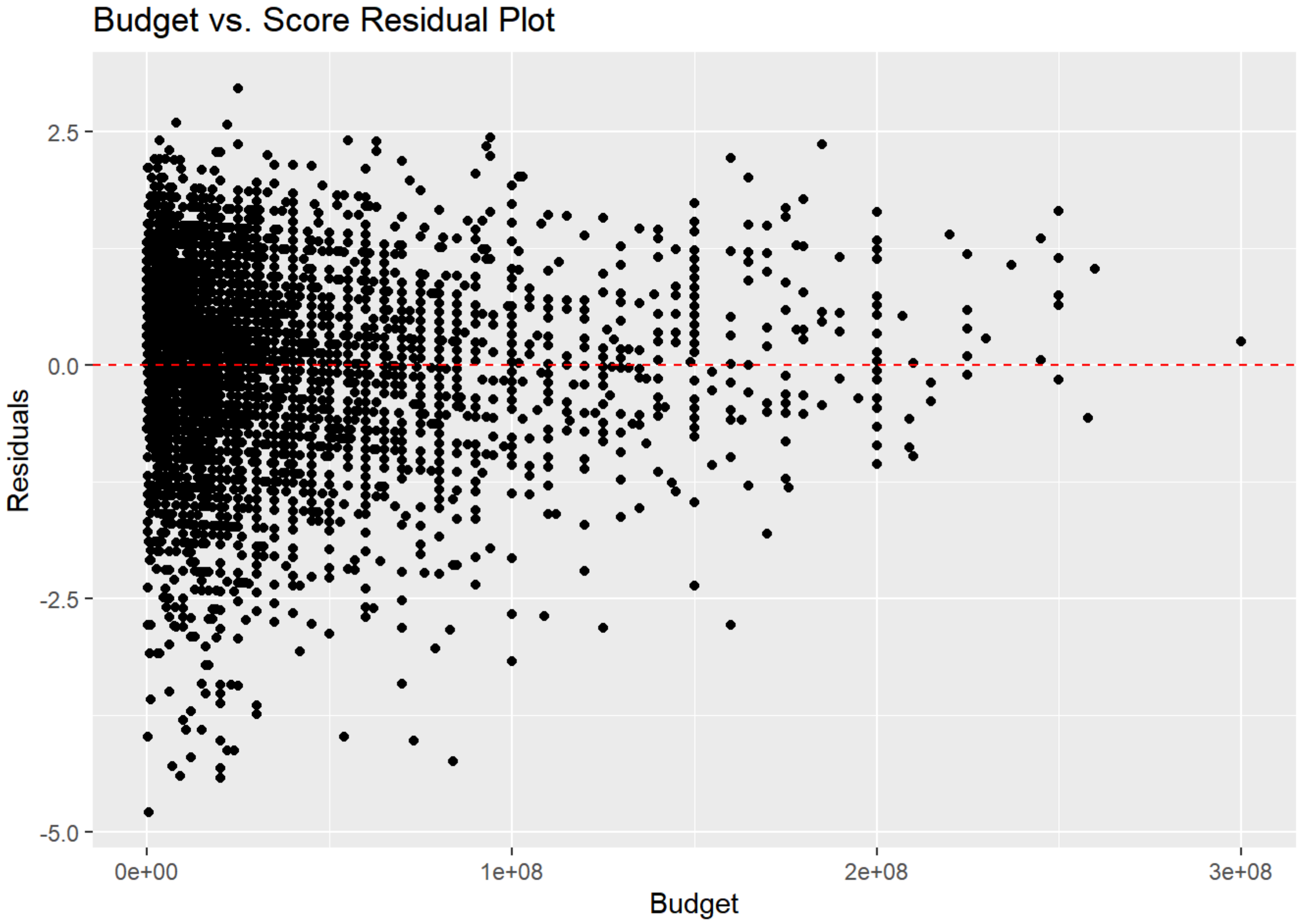
```
##  (Intercept)           x 
## 6.287974e+00 1.878127e-09
```

We also calculated the parameters of the linear model. The fitted model is:

$$(\text{score}) = 6.287974 + 1.878127 \times 10^{-9}(\text{budget})$$

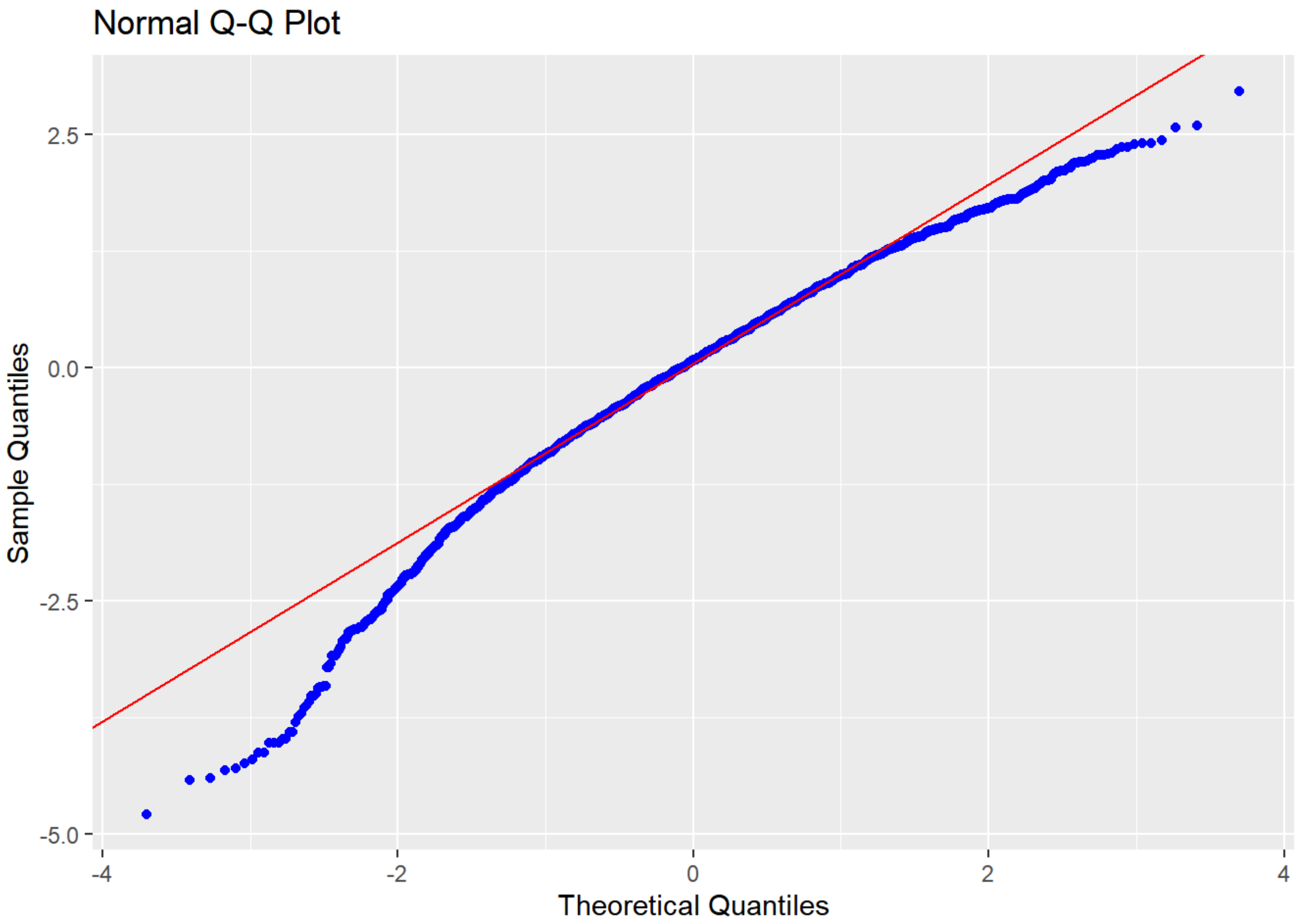
For a dollar increase in budget there is a 1.87×10^{-9} increase in the predicted value of score.

So far, we computed this linear regression model under 4 assumptions: linearity, homoscedasticity, independence and normality. In order to check if a linear model fits the data, we can take a look at the residual plot and check for any patterns that might suggest that a linear model does not fit. This also would check the homoscedasticity assumption of a linear regression model.



The residuals seem to be randomly scattered about y = 0, although there is some clustering of values with a lower budget. This suggests that the linear model is appropriate.

Further, we can use a QQ-plot to evaluate the normality assumption of a linear model.



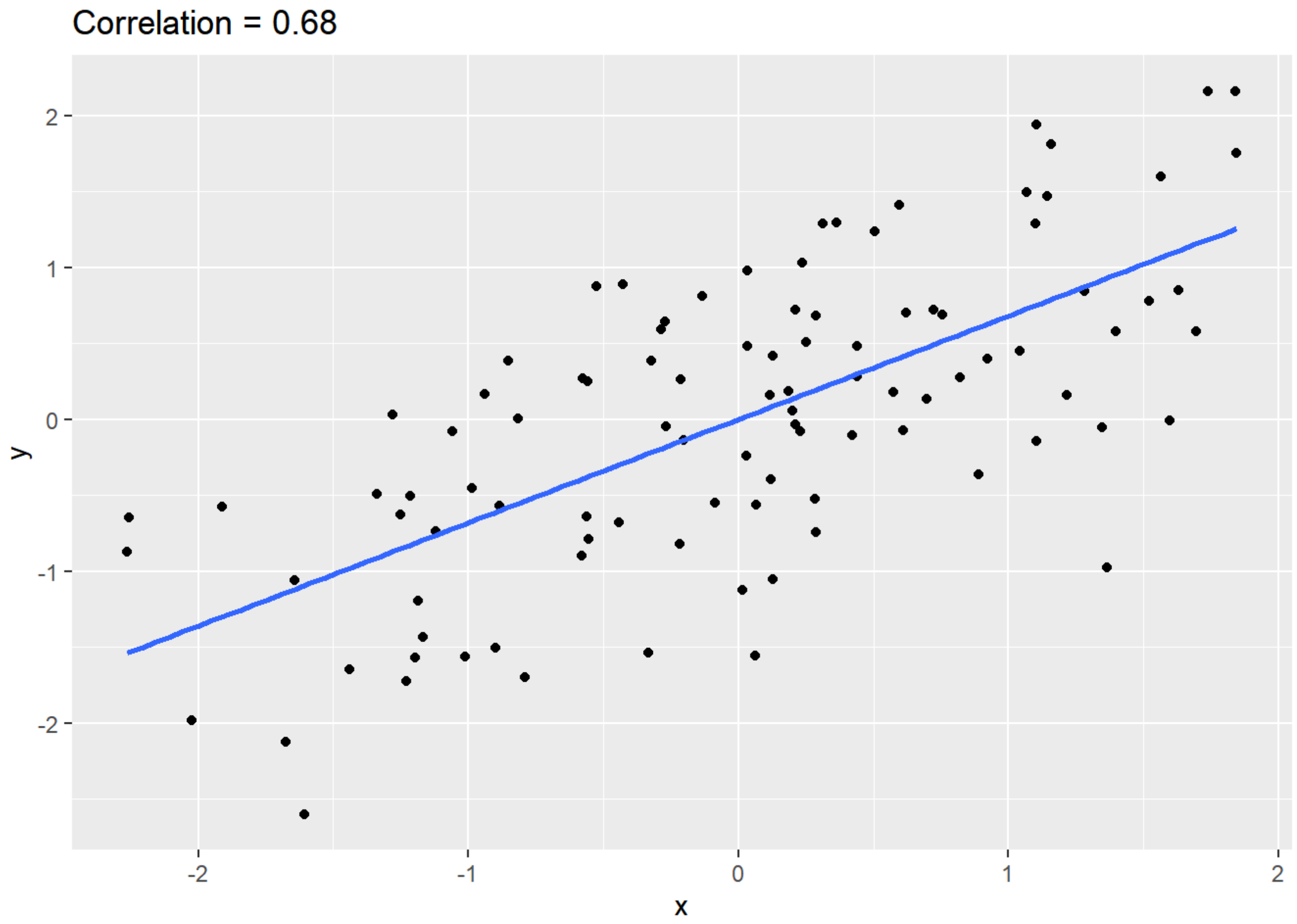
The points mostly follow the reference line and follow a relatively linear trend. As a result, we can assume normality and confirm that a linear model is appropriate.

Budget vs. Gross Income

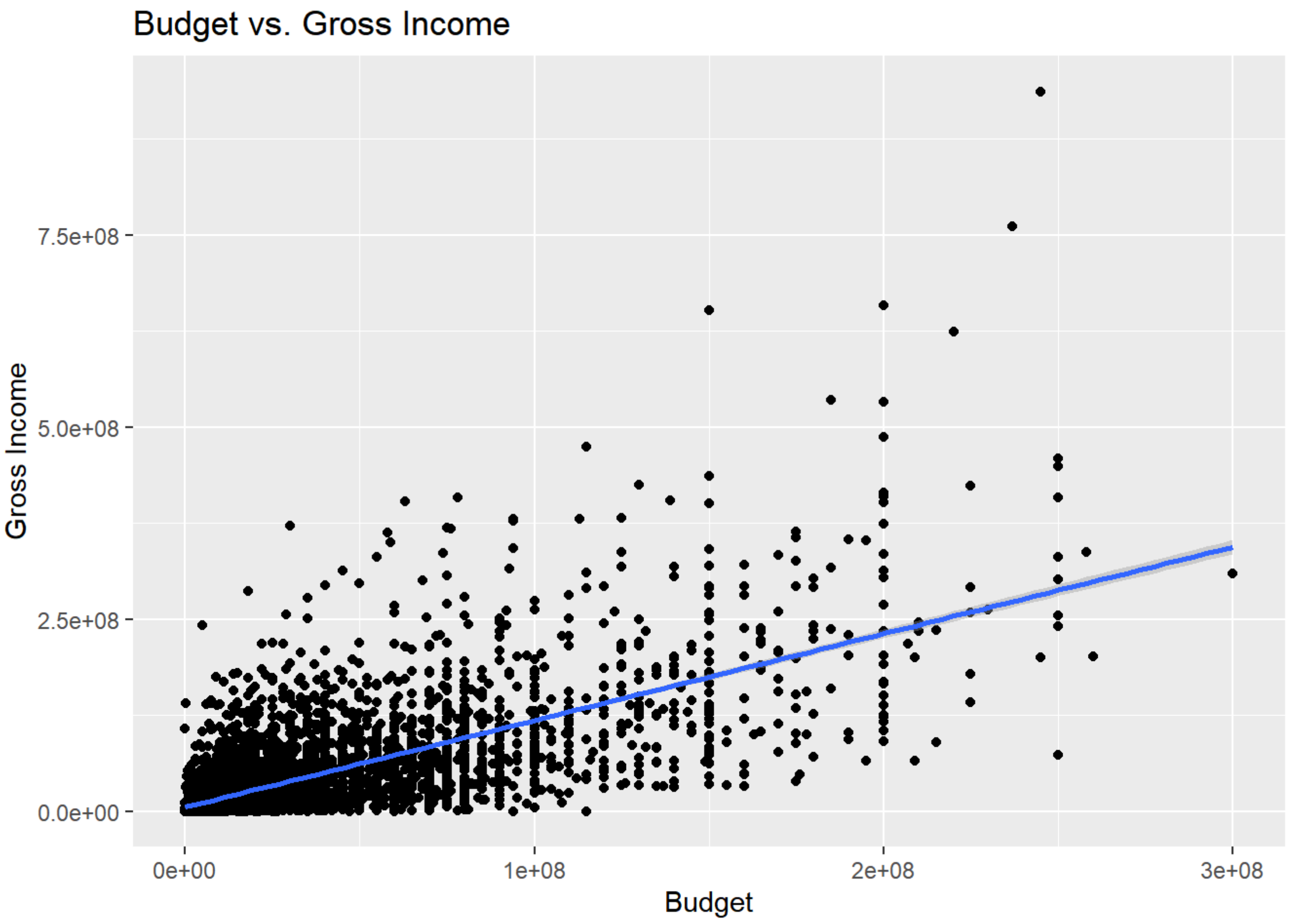
To examine the relationship between Budget and Gross Income, we computed the correlation coefficient r and plotted

```
##           r
## 1 0.6801097
```

Our r value is 0.68 and since the sign is positive and value is relatively close to 1, we say there is a moderate strong positive linear relationship between Budget and Gross Income.



Also, from the correlation plot, we see that the sign of the slope matches with the sign of the r, indicating the positive relation and the points are relatively clustered around the line indicating moderate positive linear relationship between Budget and Gross Income.



To learn more about this linear relationship, we plotted Budget vs. Gross Income dot plot along with the least squares line which minimizes the sum of squared residuals. Again, the slope of the least squares line is positive.

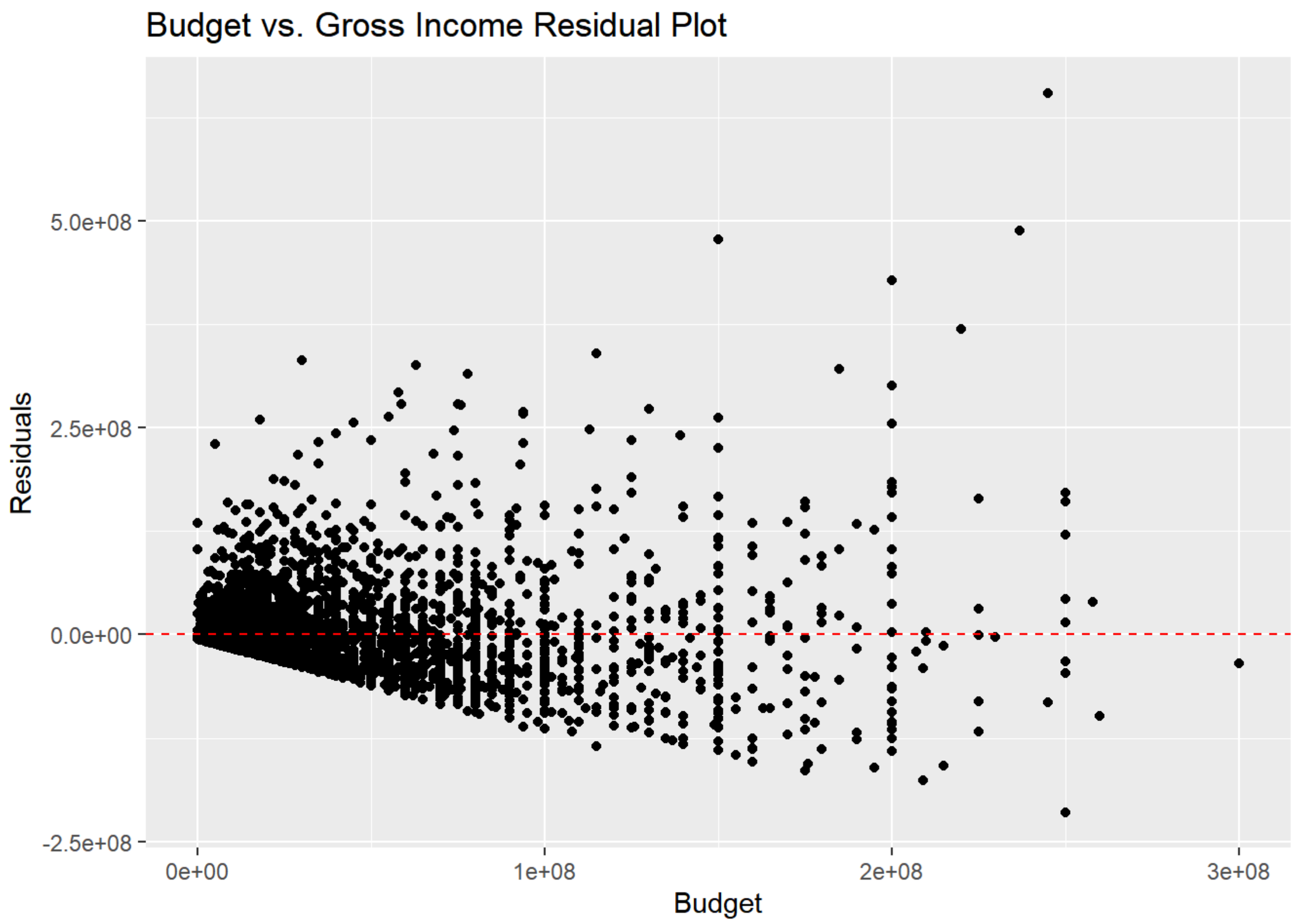
```
##  (Intercept)            x 
## 5.371441e+06 1.128629e+00
```

We also calculated the parameters of the linear model. The fitted model is:

$$(\text{gross income}) = 5.371441 \times 10^6 + 1.128629(\text{budget})$$

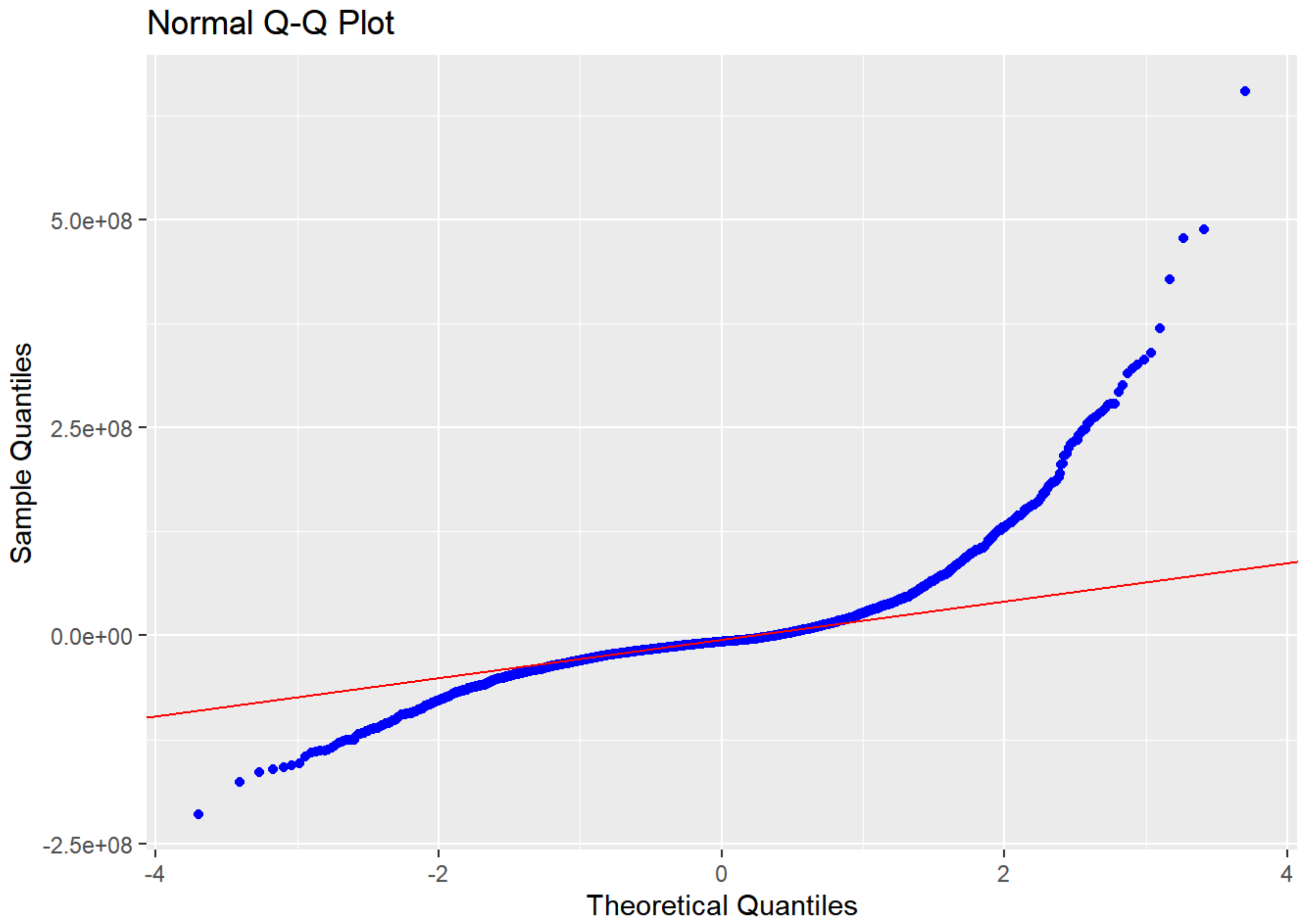
For a dollar increase in budget there is 5.37×10^6 dollar increase in the predicted value of gross income.

We again need to check for 4 assumptions. We check for linearity and homoscedasticity, we plotted Budget vs. Residuals.



The residuals do not appear to be randomly scattered about y = 0. In fact, there is a funneling effect when Revenue is small, suggesting a linear model is not appropriate.

Further, we used a QQ-plot to evaluate the normality assumption.



Since not all points are within the reference line and there clearly is a non linear trend, we cannot assume normality. There seems to be a slight cubic trend.

Discussion

A note before the analysis: when we refer to the relationship between variables, we are not talking about a causal relationship, we are strictly talking about a correlated relationship. Correlation does not imply causation, so that’s not what we’re referring to in terms of variables having a relationship between them.

From our data analysis, we were partially able to answer our initial question, which was: “Is there a positive, linear relationship between a movie’s budget and its respective gross revenue or score received?” Our thesis statement was, while we think the budget of a movie might be important in determining a movie’s success, we did not think we would see a relationship between the budget and gross revenue or the budget and its score.

We first found the correlation coefficient for the relationship between ‘budget’ vs. ‘score’ and then created a scatterplot using these variables. We found the coefficient to be r = 0.074, which indicated there is a weak, positive, linear relationship between the ‘budget and ’score’ variables. We created the scatterplot to visualize this correlation and overlayed a trend line so we could distinguish the pattern of the graph. We confirmed there is a weak, positive relationship between the two variables as the slope of the line was positive but was not very steep. We then decided that we wanted to visualize the linear model differently, so we created another scatter plot using the ‘budget’ and ‘score’ variables and overlayed a least-squares line. We once again saw that the relationship between ‘budget’ and ‘score appeared to be positive. We then created a fitted model for this relationship with ’budget’ as our explanatory variable and ‘score’ as our response variable. We again wanted to view this relationship in algebraic terms and accompany the visual representations of this linear relationship. We saw that for every dollar increase in the budget, the score increased by 1.87×10^{-9}.

Up to this point in our analysis, we were assuming that a linear model was appropriate to use to model this relationship. So, we wanted to create a residual plot to double-check if using a linear model to model this relationship was appropriate. In the ‘budget’ vs. ‘score’ residual plot, we saw that the points were randomly scattered this time. This led us to conclude that using a linear model to model the relationship between ‘budget’ and ‘score’ is appropriate. We did note some clustering around low points of the ‘budget’ variable, but this pattern did not seem as obvious as the one we saw between ‘budget’ and ‘gross,’ so we say that a linear model is appropriate for the ‘budget’ vs. ‘score’ relationship. When working with a linear model, we assume that the relationship between the given variables is linear, homogeneous, independent, and normal. The residual model helped us to confirm linearity and homogeneity, and we assume from what we know about the data that the two variables are independent of each other. To test for normality, we again created a QQ-plot using the ‘budget’ and ‘score’ variables. In this plot, we saw that most of the points were within the reference line and followed a relatively linear trend. Therefore, we can assume that the relationship between these variables is normal. All in all, for the relationship between ‘budget’ and ‘score,’ we confirmed that a linear model is an appropriate model to use and therefore we can use the correlation coefficient that we calculated above to claim that there is a weak, positive relationship between these variables. It is unlikely that a movie’s budget greatly affects the score the movie receives because the correlation is very weak (close to 0).

We then found the correlation coefficient for the budget vs. the gross revenue and then the budget vs. the score and plotted each of these graphs, respectively. For the budget vs. the gross revenue, we found the coefficient to be r = 0.68, which indicates there is a moderately strong, positive, linear relationship between these two variables. Then, we created a scatterplot to visualize the correlation and overlayed a trend line to better see the pattern. We again saw a positive relationship between the two variables. We again decided to explore this linear relationship even further and created another scatter plot with the ‘budget’ and ‘gross’ variables and overlayed a least-squares line, and we saw once again that there appeared to be a positive relationship between these two variables. We then created a fitted model with ‘budget’ as our explanatory variable and ‘gross’ as our response variable. This helped us to view the relationship in algebraic terms to accompany the visual representations. We saw that for every dollar increase in the budget, the gross revenue increased by 5.37×10^6 dollars.

As with the previous relationship we had been investigating, we assumed that a linear model was appropriate to use to model the relationship between ‘budget’ and ‘gross.’ So, we wanted to create a residual plot to double-check if using a linear model to model this relationship was appropriate. We created a residual plot to see if using a linear model to model this relationship was appropriate in this case. After creating a plot of the residuals for ‘budget’ vs. ‘gross,’ we saw that the points were not randomly scattered - there appeared to be a funneling effect that starts when the gross revenue is small, and then fans out. We, therefore, concluded that a linear model is not appropriate to use to model this relationship. We had also been assuming that the relationship between these two variables is normal, so we decided to create a QQ-plot to either confirm or deny this assumption. In this plot, we saw that not all of the points were within the reference line, meaning that we could not see a clear linear trend between these two variables, and therefore we could not assume that the relationship between them is normal. The pattern that we saw in the QQ-plot looks like there might be a cubic relationship between the variables ‘budget’ and ‘gross.’ When one is working with a linear model, one must assume that the relationship between the given variables is linear, homogeneous, independent, and normal. With this relationship, we could not assume normality or linearity, and therefore using a linear model is not appropriate to model this relationship. Therefore, we cannot use the correlation coefficient that we calculated above to claim what the strength of the relationship is since correlation requires a linear model to calculate.

From our analysis, we found that we cannot make any claims regarding the strength of the relationship between the ‘budget’ and ‘gross’ variables because this relationship is not linear. We were, however, able to conclude that the strength between the ‘budget’ and ‘score’ variables is weak and positive since we proved that using a linear model was appropriate in this case. Therefore, we cannot fully answer our original thesis statement because we don’t have all the answers. We did find that, as we predict, there is not a strong relationship between the ‘budget’ and ‘score,’ but we are unable at this point able to comment on the strength between the ‘budget’ and ‘score’ variables.

Potential Short-Comings

A potential shortcoming of our analysis was that we only used one dataset to answer our question. This could be the reason why the relationship between 'budget' vs. 'gross' was not linear and therefore why we were unable able to come to a conclusion about the strength of this relationship. This could also be why the correlation between the 'budget' and 'score' variables was so weak. If we had more data, it might help to either strengthen the results we got from this analysis or would show us different results. Either way, more data would allow for a more accurate analysis.

Another shortcoming is that we only used one type of model in our analysis (linear model). If we had used a cubic or quadratic relationship to model the relationship between 'budget' and 'gross' we might have been able to yield better results and therefore be able to answer whether or not these variables are related to each other.

Potential future directions for additional work:

New Data:

- Merge multiple datasets to get more data for computing similar/the same analyses (to potentially get more confirming results)

New Questions:

- Is there a relationship between any of the variables and 'gross' or 'score'? If so, which ones? How strong are those relationships and can we used them to predict how successful a movie is?
- Which country/production company makes the most revenue from the film?
- How do runtime and budget change over time?