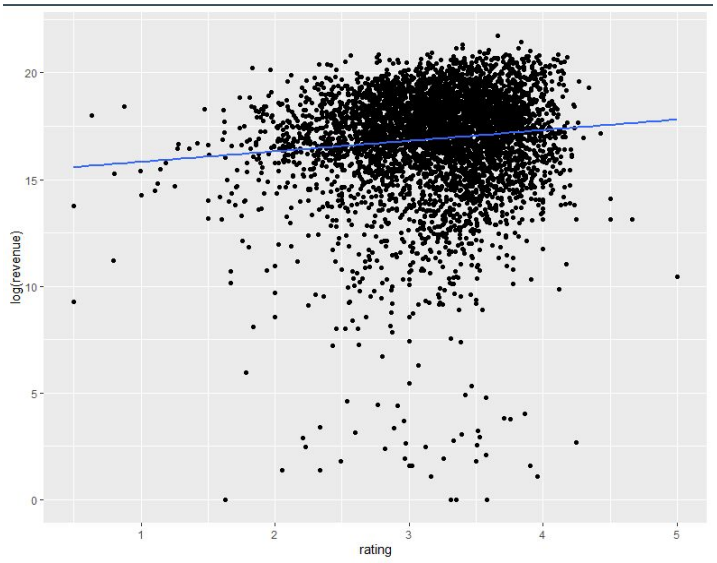
Mikko Pajula - Programmatic Business Analytics



Plot: By looking at the relationship between the average rating and movie revenue through a scatter plot, it is hard to say that revenue is dependant on the average rating. The fitted regression line though, indicates a small positive relationship and the data points seem to be more densely packed in the top right and spreading out slightly towards the bottom left, this could be an indication of a heteroskedasticity problem. There doesn't seem to be significant outliers or influential points in the data.

```
Coefficients:
(Intercept)       rating
    15.3287       0.4954
```

Coefficients: The slope coefficient is 0.4954, which means when the average rating of a movie increases by 1, the movie revenue increases in logarithmic scale by 0.4954. To untransform this into a percent change, we do the calculation (exp(0.4954) – 1) * 100 = 64%. So according to our linear model with rating increasing by 1 point, revenue would increase by 64%. By using the logarithm, we preserved the model as linear, even though the effective relationship seems to be non-linear indicated by the very large untransformed coefficient.

```
Residuals:
     Min       1Q   Median       3Q      Max
-17.1037  -0.9454   0.4681   1.6049   4.6045

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.32866    0.22910  66.907  < 2e-16 ***
rating       0.49537    0.07106   6.971 3.56e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.531 on 4791 degrees of freedom
Multiple R-squared:  0.01004,   Adjusted R-squared:  0.009836
F-statistic:  48.6 on 1 and 4791 DF,  p-value: 3.564e-12
```

Model fit: The residual standard error indicates that the difference between the observed revenue and our model's prediction is around 2.5, which is fairly large considering the average revenue is 18.4 on a logarithmic scale. Looking at the R-squared value, according to our model, average rating explains only around 1% of the variability in revenue. According to the F-test, we do observe that the model is statistically significant with p-value of 3.56e-12, meaning that ratings would be correlated with revenue.

Conclusion: The model's combination of statistical significance and low R-squared means that it indicates a correlation between the average rating and revenue, but the ratings do not explain much of the changes in revenue. Our linear model could be a bad fit for the data, since the effective relationship seems to be non-linear. There were also signs of heteroskedasticity, the ratings correlating with the variance of revenue. In order to mitigate this problem we could calculate more heteroskedasticity robust standard errors.