Descriptive Statistics of Key Variables, Interpretation of Mean & Skewness

Descriptive Statistics

Statistic	N	Mean	Std. Dev.	Median	Pctl(25)	Pctl(75)	IQR
box_office_revenue	250	201.964	137.082	187.046	77.490	303.494	226.004
movie_budget	250	81.003	35.276	84.766	58.688	105.967	47.279
audience_score	250	64.023	16.694	62.993	51.986	76.599	24.613

Table 1.1 Summary Statistics for box_office_revenue, movie_budget, audience_score

Interpretation of Mean:
The average box office revenue at theatres for movies in this sample is $201.964m, a baseline estimate for the total box office collections. Similarly, the mean production budget for the films in the sample is $81.003m. The average audience score for movies in this sample is 64.023 out of 100, the rating on Rotten Tomatoes.

Symmetry:

The median box office revenue for films in this sample is $187.046m, which is less than the mean of $201.964m. The distance between the third quartile (Q3) from the median ($116.448m) is slightly larger than the distance between the first quartile and the median ($109.556m). These two factors suggest that box office revenue is positively skewed.

The median movie budget is $84.766m, which is slightly higher than the mean of $81.003m. The distance between Q1 and the median ($26.078m) is larger than the distance between Q3 and the median ($21.201m). Consequently, these observations imply a negative skew in movie budgets.

The median audience score is 62.993 out of 100, slightly lower than the mean audience score of 64.023 out of 100. The distance between Q3 and the median (13.606 points) is larger than the distance between Q1 and the median ($11.007m). As a result, we can infer that there is a positive skew in audience scores on Rotten Tomatoes.

Confidence Intervals for Means

95% CI($\mu_{box\_office\_revenue}$) = [184.972, 218.957] ($m)
95% CI(μ_{movie_budget}) = [76.631, 85.376] ($m)
95% CI($\mu_{audience\_score}$) = [61.954, 66.093] (out of 100)

Density Function and Skewness of Box Office Revenue

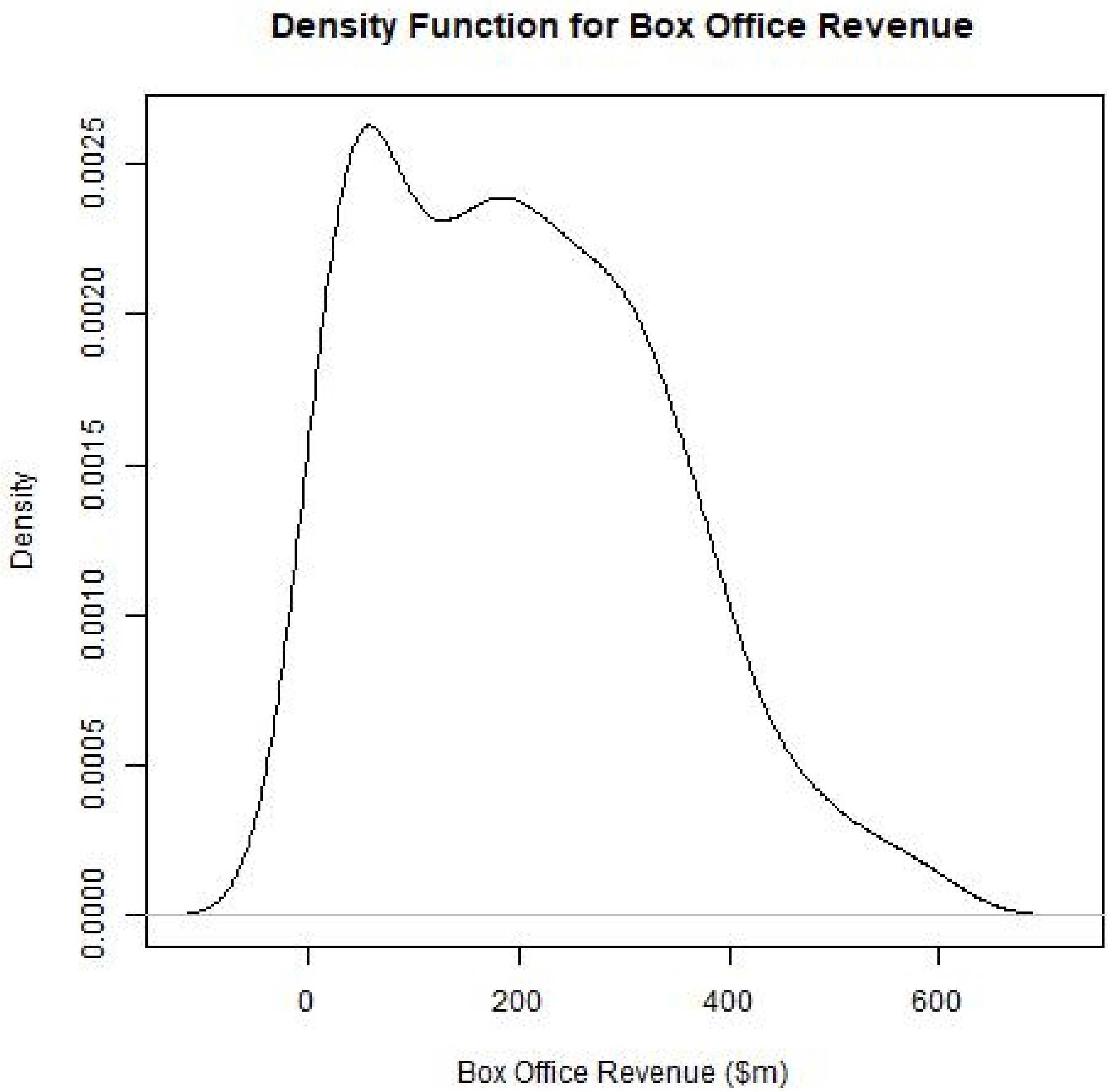

Figure 3.1 Density Function for Box_Office_Revenue

The density function (Figure 3.1) for box office revenue shows an evident positive skew as the values tail off to the right. Economically, we can observe that the minimum box office revenue is 0, however, because films can be superhits or blockbusters, they can generate colossal box office revenues. These revenues may greatly exceed the median, resulting in a positive skew.

High Budget vs Low Budget Films Density Functions

Henceforth, we shall define high budget films as the subset of data for which movie_budget is greater than or equal to the median movie_budget for the sample data. Low budget films are all the films in the sample for which the movie_budget is less than the median movie_budget.

From Figure 4.1, we can observe that the box office revenue of high budget films tends to be higher than low budget firms, represented with a higher density for higher values of box office revenue. Furthermore, there is a lower density for high budget movies for lower values of box office revenue when compared to low budget movies. This is due to the fact that a majority of the high budget density is shifted further to the right compared to the low budget density. This suggests a positive relationship between the box office revenue and the movie budget.

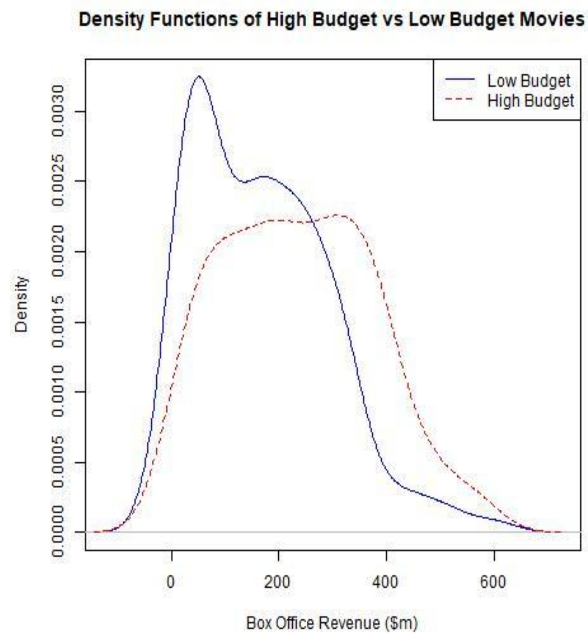

Figure 4.1 - Density Functions for box_office_revenue for High and Low Budget Films

The mean revenue for high budget movies is $235.738m whereas, the mean revenue for low budget movies is $168.190m, reflecting a large disparity in revenue. A potential explanation for this is that higher budget films have more money to invest in a higher quality production team, recognisable actors and more advertising for the movies, enabling it to reach a wider audience. A consequence of reaching a wider audience may be more box office revenue as greater public interest may surround the film.

Hypothesis Test for the Difference in Means

Hypotheses:

$$H0: \boldsymbol{\mu}_{(\boldsymbol{box\ office\ revenue\ if\ high\ budget} = \mathbf{1})} = \boldsymbol{\mu}_{(\boldsymbol{box\ office\ revenue\ if\ high\ budget} = \mathbf{0})}$$

$$H1: \boldsymbol{\mu}_{(\boldsymbol{box\ office\ revenue\ if\ high\ budget} = \mathbf{1})} \neq \boldsymbol{\mu}_{(\boldsymbol{box\ office\ revenue\ if\ high\ budget} = \mathbf{0})}$$

Sample mean for box office revenue if high budget = 1 is $235.738m
Sample mean for box office revenue if high budget = 0 is $168.190m
Estimate difference in means (high budget = 1 - high budget = 0) is $67.548m

95% CI ($\boldsymbol{\mu}_{(\boldsymbol{revenue|high\ budget} = \mathbf{1})} - \boldsymbol{\mu}_{\boldsymbol{revenue|high\ budget} = \mathbf{0})}$) is between [34.385, 100.712] ($m)

Note: Revenue refers to box office revenue.

t-statistic = 4.012 > 1.960 (critical t-value for hypothesis tests at the 5% level of significance)
p-value = $8.006 * 10^{-5} < 0.05$ (the significance level for the hypothesis test)

Both the t-statistic and the p-value lead us to reject the null hypothesis that the mean box office revenues are the same regardless of whether movie budget is high or not, implying a statistically significant difference in means.

The % change in the conditional mean of box office revenue when going **from** high_budget = 0 films **to** high_budget = 1 films is shown below:

$$\%\,Change = (235.7383 - 168.1898)/168.1898 = 40.162\%\ increase$$

(computed using the conditional means for box_office_revenue for high and low budget films) This positive percent change suggests a positive relationship between movie budget and box office revenue as, generally, high budget films correspond with more box office revenue.

Scatter Plot & Single Linear Regression

From question 4, we observed a higher mean revenue for high budget movies as opposed to movies with a lower budget, suggesting a positive relationship between movie budget and revenue. Additionally, from question 5, the result of the hypothesis test shows that the means of high budget and lower budget films are statistically different. This means that the coefficient of the regression line for movie budget is non-zero from question 5 and positive from question 4. Therefore, this aligns with what we see in the scatterplot in figure 6.1, as the coefficient of movie budget is 1.299.

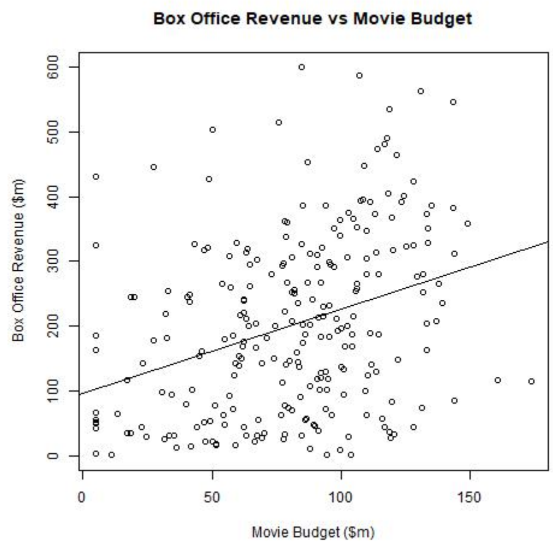

Figure 6.1 - Single Linear Regression between box_office_revenue and movie_budget

Single Linear Regressions

Regression	$\widehat{\beta_0}$	$\widehat{\beta_1}$	SE($\widehat{\beta_0}$)	SE($\widehat{\beta_1}$)
Regression 1 (Box Office Revenue vs Movie Budget)	96.745	1.299	20.541	0.233
Regression 2 (Box Office Revenue vs Audience Score)	45.009	2.452	32.922	0.498

Table 7.1 - Coefficient Estimates and Corresponding Standard Errors for Regression 1 and 2

One standard deviation of movie_budget (independent variable) is $35.276m. The estimated increase in box office revenue is $1.299m ($\widehat{\beta_1}$ for regression 1) for every increase of $1m in movie budget. This means an increase of \$35.276 m in the movie budget would be expected to correspond to an increase of \$45,823,524 in the box office revenue.

An increase in the audience score by 1 units corresponds to an increase in box office revenue by \$2.452m ($\widehat{\beta_1}$ for regression 2), therefore, an increase of 20 units of audience score corresponds to an increase of \$49.04m in box office revenue.

We conduct hypothesis tests for each regression testing whether each coefficient (B_0 & B_1) is significantly different from 0 at the 5% level of significance.

Hypothesis Test	β_0	β_1
Regression 1	$H_0: \beta_0 = 0$ $H_1: \beta_0 \neq 0$ t-statistic = 4.710 > 1.960 p-value = $4.13 * 10^{-6} < 0.05$ Both the t-statistic and the p-value lead us to reject the null hypothesis that β_0 is 0, meaning the intercept term for regression 1 **is** significantly different from 0.	$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$ t-statistic = 5.585 > 1.96 p-value = $6.12 * 10^{-8} < 0.05$ Both the t-statistic and the p-value lead us to reject the null hypothesis that β_1 is 0, meaning the slope term for regression 1 **is** significantly different from 0.
Regression 2	$H_0: \beta_0 = 0$ $H_1: \beta_0 \neq 0$ t-statistic = 1.367 < 1.960 p-value = 0.173 > 0.05 Both the t-statistic and the p-value lead us to not rejecting the null hypothesis that β_0 is 0, meaning the intercept term for regression 2 **is not** significantly different from 0.	$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$ t-statistic = 4.926 > 1.96 p-value = $1.53 * 10^{-6} < 0.05$ Both the t-statistic and the p-value lead us to reject the null hypothesis that β_1 is 0, meaning the slope term for regression 2 **is** significantly different from 0.

Table 7.2 - Hypothesis Tests testing whether each coefficient in regressions 1 and 2 is different from 0

Multiple Linear Regression

```
Residuals:
    Min      1Q  Median      3Q     Max 
-249.93  -98.18  -11.76   90.40  394.18 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)   
(Intercept)     44.9722    32.2985   1.392  0.16506   
movie_budget     0.9420     0.2884   3.267  0.00124 **
audience_score   1.2602     0.6094   2.068  0.03968 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 128.6 on 247 degrees of freedom
Multiple R-squared:  0.1269,  Adjusted R-squared:  0.1198 
F-statistic: 17.94 on 2 and 247 DF,  p-value: 5.302e-08
```

Output 8.1 The multiple linear regression results assuming box_office revenue is the dependent variable and movie_budget and audience_score are the independent variables

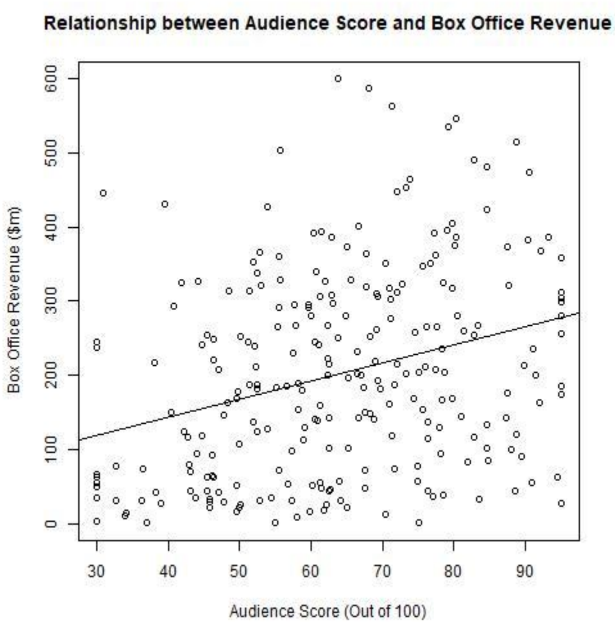

Figure 8.1 - Scatter plot with linear regression between box_office_revenue and audience_score.

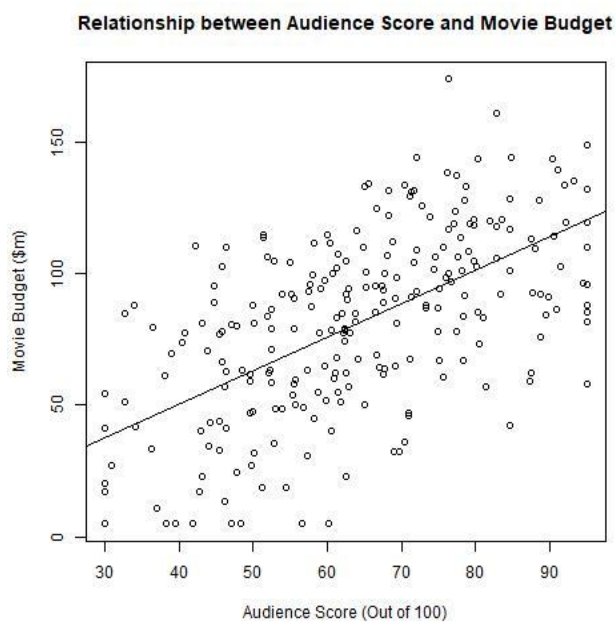

Figure 8.2 - Scatter plot with linear regression between box_office_revenue and audience_score.

These two scatterplots (Figures 8.1 and 8.2) with single linear regressions suggest that audience_score has a positive relationship with both box_office_revenue and movie_budget. This highlights a positive relationship between box_office_revenue and movie_budget, reinforcing the findings in Q7, with audience_score a potential confounding variable. However, as a consequence of movie_budget and audience_score being positively correlated, the single linear regression model in Q7 is biased due to movie_budget and the error term being correlated. This leads to the OLS estimator assumption of independence being violated as a result of omitted variable bias. Therefore, the bias term is positive and non-zero which was inferred from Figure 8.1 and Figure 8.2.

The bias term being positive and non-zero means that the coefficient of movie_budget found in Q7 has a positive bias, therefore, being overestimated. This provides a plausible explanation to why the magnitude of the coefficient has decreased from 1.299 in Q7 to 0.942 when implementing the multiple linear regression. On the other hand, the sign of the coefficient corresponding to movie_budget remains the same because the magnitude of the bias term is not large enough to outweigh the true population coefficient.

Multiple Linear Regression with High Budget and Low Budget Films

Multiple Linear Regression 1: Dependent Variable: Box Office Revenue, Independent Variables: Movie Budget (High Budget = 0) and Audience Score. Regression results are shown below:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-174.40  -92.55  -33.94   67.20  391.80 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)  
(Intercept)                       44.0084    42.0383   1.047   0.2972  
movie_budget[high_budget == 0]     0.8908     0.5097   1.748   0.0830 .
audience_score[high_budget == 0]   1.3804     0.8034   1.718   0.0883 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 120.9 on 122 degrees of freedom
Multiple R-squared:  0.08229, Adjusted R-squared:  0.06724 
F-statistic: 5.469 on 2 and 122 DF,  p-value: 0.005311
```

Output 9.1 The multiple linear regression results for the high_budget = 0 data subset

Multiple Linear Regression 2: Dependent Variable: Box Office Revenue, Independent Variables: Movie Budget (High Budget = 1) and Audience Score. Regression results are shown below:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-250.81 -103.38   -2.21   94.82  359.82 

Coefficients:
```

```
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -15.3296    85.9377  -0.178   0.8587
movie_budget[high_budget == 1]        1.5232     0.7192   2.118   0.0362 *
audience_score[high_budget == 1]      1.1742     0.9263   1.268   0.2073
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136.9 on 122 degrees of freedom
Multiple R-squared:  0.06692, Adjusted R-squared:  0.05162
F-statistic: 4.375 on 2 and 122 DF,  p-value: 0.01462
```

Output 9.2 The multiple linear regression results for the high budget = 1 data subset

As we are now utilising a multiple linear regression model, we can now isolate the effects of movie_budget (high and low budget films) and how it relates with box_office_revenue to guide decision making. We can observe from Output 9.1 and Output 9.2 that the coefficient estimate for movie_budget for high budget films is $1.523m for every $1m additional invested, compared to $0.891m for every $1m additional invested in low budget films. We can use these coefficient estimates for movie_budgets to decide which movie (the high or the low budget film) should be allocated an additional $1m budget given these coefficient estimates are independent of audience score.

Investing $1m into a low budget film isn't a worthwhile investment given we can only expect $0.891m extra box office revenue (less than the money invested) and since the expected extra box office revenue is less for the low budget film as opposed to the high budget film.
From these coefficient estimates, there is a difference of $0.632m, reflecting the estimated forgone revenue associated with funding the extra $1m into the low budget film, rather than the high budget film. This may be due to high budget films spending the extra allocation of funds for increased advertising and distribution of films (movie screens), which low budget films typically lack given their scale.