Quantitative Exploratory Data Analysis

Learning Objectives

- Identify and create appropriate summary statistics and plots given a data set or research question
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, inter-quartile range, coefficient of determination, regression line slope
- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers)
- Use scatterplots to assess the relationship between two quantitative variables
- Find the correlation coefficient
- Plot a scatterplot
- Estimate slope using the correlation coefficient
- Find the estimated line of regression using summary statistics and R Linear Model Output
- Understand what the slope coefficient represents
- Understand what the coefficient of determination is

Movies Released in 2016

A data set was collected on Movies released since 1916 to 2016. Here is a list of some of the variables collected on these movies.

Year: Year the movie was released *Budget*: The amount of money (in US $) budgeted for the production of the movie *Revenue*: The amount of money (in US $) the movie made after release *Duration*: The length of the movie (in minutes) *Content Rating*: Rating of the movie (G, PG, PG-13, R, Not Rated) *IMDb Score*: User rating score from 1 to 10 *Genre*: Category the movie falls into *Movie Facebook Likes*: Number of likes a movie receives on Facebook

Vocabulary Review

1. What are the observational units in this data set?

2. Which of the above listed variables are categorical?

3. Which of the above listed variables are quantitative?

Summarizing a single quantitative variable

- Center: There are two measures of center used for quantitative data: the **mean** or the average and the **median**. The mean is found by adding up all the values in the data set and dividing the sum by the sample size. The median is the 50th percentile of the ordered data set.
- Spread: The spread is also referred to as variability. Again there are two measures of spread used to describe the data: the **standard deviation** and the **Interquartile Range or IQR**. The standard deviation measures the average distance of each observation from the mean. The IQR is found by subtracting the value of the first quartile from the third quartile and measures the middle 50% of the data. $IQR = Q_3 - Q_1$

The favstats function gives the summary statistics for a quantitative variable. Here we have the summary statistics for 'IMDb'.

```
favstats(moviesa$imdb_score)
```

```
##  min  Q1 median  Q3 max     mean        sd  n missing
##  3.4 5.9    6.6 7.1 8.2 6.459016 0.9218418 61       0
```

4. Give the values for the two measures of center.

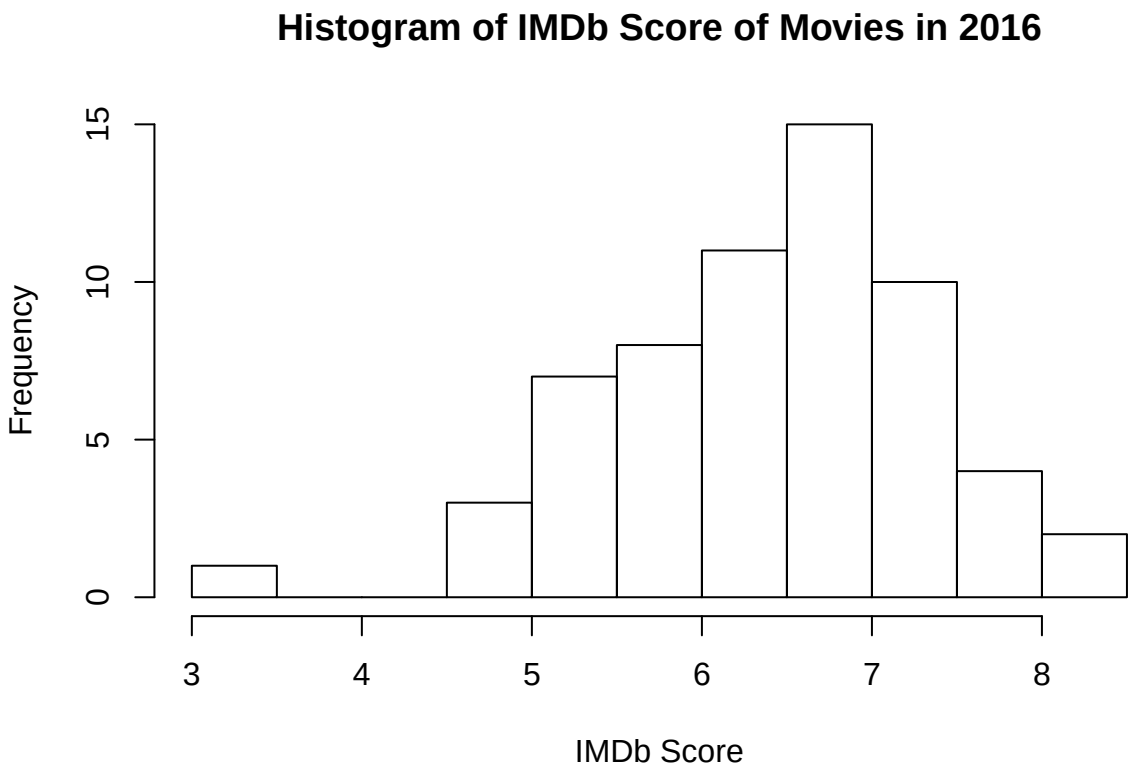
5. Calculate the IQR.

6. Report the value of the standard deviation and interpret this value in context of the problem.

Displaying a single quantitative variable

To plot quantitative variables, we can use a dotplot, histogram or boxplot. To create a histogram the variable is broken into bins on a set width. Each bin plots the frequency of each variable for a certain bin. The following code creates a histogram of the variable 'IMBd Score'. Notice that the bin width is 0.5. For example the first bin consists of the number of movies in the data set with an IMBd score of 3 to 3.5. It is important to note that a movie with a IMBd score of 5 will fall into the bin for 5 - 5.5. Visually this shows us the range of IMBd scores for Movies released in 2016.

```
hist(moviesa$imdb_score, #dataset name and variable
     main = "Histogram of IMDb Score of Movies in 2016", #title for plot
     xlab = "IMDb Score") #label for x axis
```



7. Which range of IMBb scores have the highest frequency?

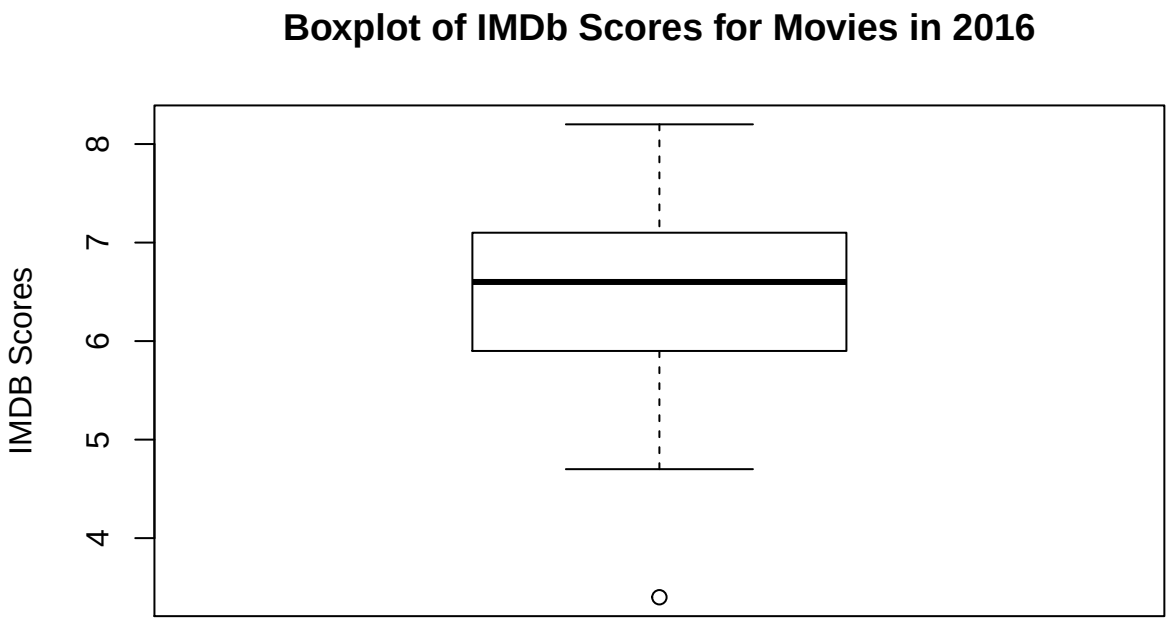
When comparing distributions we will look at four characteristics:

- Shape: A distribution that has approximately the same trailing off in both tails is considered a symmetric distribution. One in which the data trails off to the right is a right-skewed distribution. If the data trails off to the left it is considered a left-skewed distribution.
- Center: Mean or Median
- Spread: Standard Deviation or IQR
- Outliers: Outliers are values less than $Q_1 - 1.5 * IQR$ and greater than $Q_3 + 1.5 * IQR$. On the box plot they are displayed as a single dot beyond the whisker.

The boxplot is created using the five number summary: * Minimum value * Quartile 1 (Q1) - the value at the 25th percentile * Median - the value at the 50th percentile * Quartile 3 (Q3) - the value at the 75th percentile * Maximum value

Here is a boxplot of IMDb scores for movies in 2016.

```
boxplot(moviesa$imdb_score,
       main = "Boxplot of IMDb Scores for Movies in 2016",
        ylab="IMDB Scores")
```



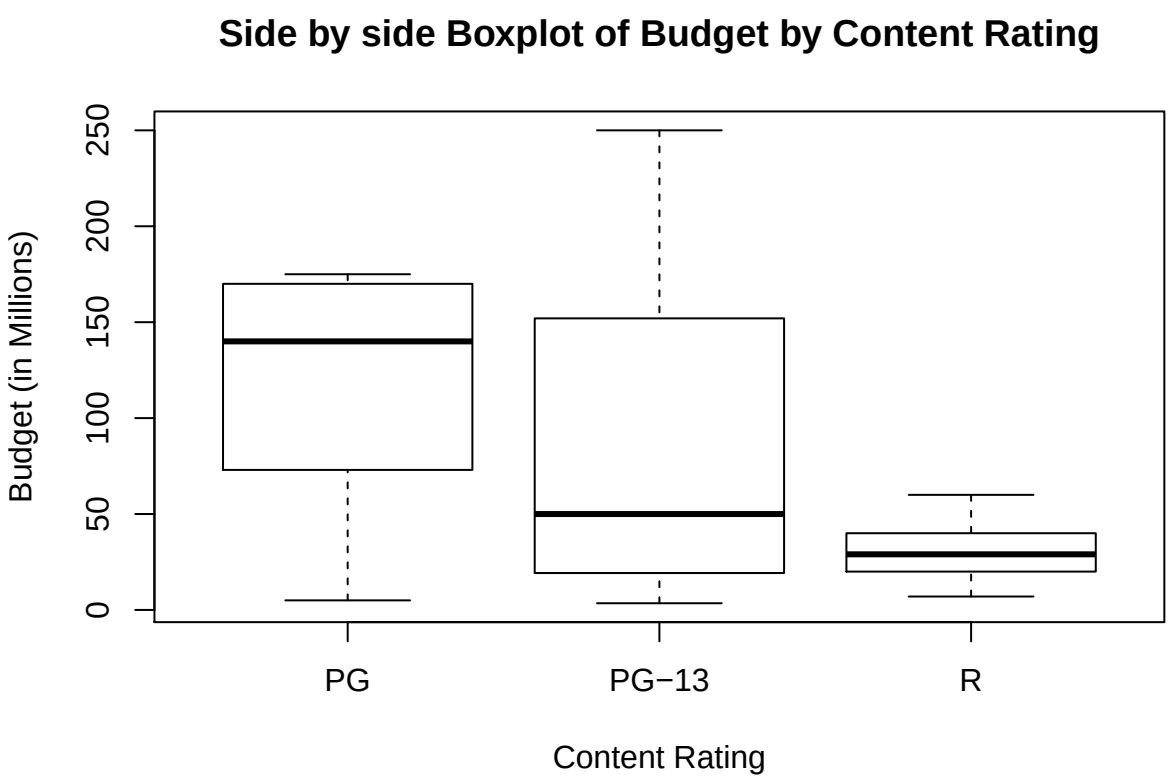
8. Using the two plots above, what is the shape of the distribution of IMDb scores?

Displaying a Single Categorical and Single Quantitative Variable

Side by side plots are created from a single categorical and single quantitative variable. We can create side by side boxplots, side by side histograms, and side by side dotplots.

The boxplot of 'Budget' in millions by 'Content rating' is plotted using the code below. This plot helps to compare the budget for different levels of content rating.

```
boxplot(budget_mil~content_rating #response~explanatory
        , data=moviesa, main = "Side by side Boxplot of Budget by Content Rating",
        xlab = "Content Rating", ylab = "Budget (in Millions)")
```



9. Compare PG and PG-13 movies using the four characteristics for comparing distributions.

Association between two Quantitative Variables

To plot two quantitative variables we will use a scatterplot. If there is a clear relationship between these variables the variables are associated. We will analyze four characteristics of scatterplots to assess this relationship.

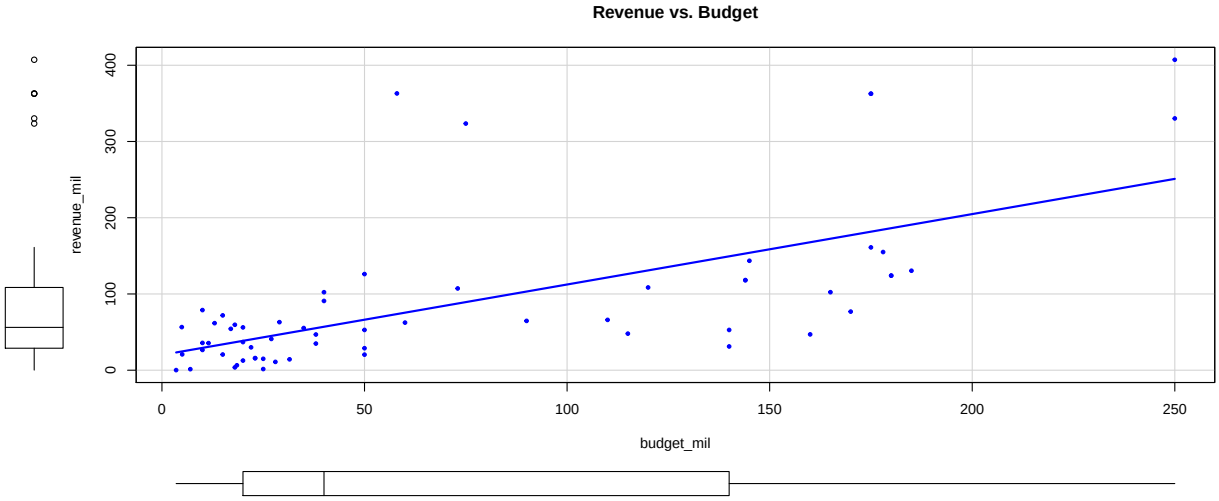
- Form: Is the relationship linear or non-linear.
- Direction: As the x variable increases is the y variable increasing (positive relationship) or decreasing (negative relationship).
- Strength: How strong is the association between the variables? We will look at the measure of correlation to assess the strength. When assessing the scatterplot we will see how close the individual data points are to the line of regression.
- Unusual Observations or Outliers: We will also assess any data points that do not fit the overall pattern of the data. These are called outliers or unusual observations.

We will look at the relationship between 'Budget' and 'Revenue' for movies released in 2016.

Using a Scatterplot to Assess the Association between Quantitative Variables

This shows a scatterplot of 'Budget' as a predictor of 'Revenue' (note: both variables are measures in "millions of dollars".

```
par(mfrow=c(1,2))
scatterplot(revenue_mil~budget_mil, data=moviesa,pch=20,main="Revenue vs. Budget", smooth=F)
```



9. Assess the four features of the scatterplot that describe this relationship.

- Form (linear, non-linear)

- Direction (positive, negative)

- Strength

- Unusual Observations or Outliers

10. Does there appear to be an association between 'Budget' and 'Revenue'? Explain.

Correlation

Correlation measures the strength and the direction between two quantitative variables. The closer the value of correlation to + or - 1 the stronger the linear relationship. Values close to zero indicate a very weak linear relationship between the two variables. The following output shows a correlation matrix between several pairs of quantitative variables.

```
moviesb <- moviesa[,c("budget_mil", "revenue_mil", "duration", "imdb_score", "movie_facebook_likes")]
round(cor(moviesb,use="pairwise.complete.obs"),4)
```

```
##                       budget_mil revenue_mil duration imdb_score
## budget_mil                1.0000      0.6466   0.5274     0.3081
## revenue_mil               0.6466      1.0000   0.2516     0.4876
## duration                  0.5274      0.2516   1.0000     0.2362
## imdb_score                0.3081      0.4876   0.2362     1.0000
## movie_facebook_likes      0.6481      0.6710   0.5619     0.3462
##                       movie_facebook_likes
## budget_mil                          0.6481
## revenue_mil                         0.6710
## duration                            0.5619
## imdb_score                          0.3462
## movie_facebook_likes                1.0000
```

11. Using the output above, which two variables have the strongest correlation?

12. What is the value of correlation between 'Budget' and 'Revenue'?

13. Based on the value of correlation what would the sign of the slope be? Positive or negative? Explain.

14. Does your answer to question 13 match the direction you choose in question 9?

Slope

The favstats code in R gives the five number summary (min, Q1, median, Q3, max), as well as the mean and the standard deviation. The slope of the least squares line can be estimated by $b_1 = \frac{s_y}{s_x} R$

```
favstats(moviesa$budget_mil)
```

```
##  min Q1 median  Q3 max     mean       sd  n missing
##  3.5 20     40 140 250 74.06393 69.42953 61       0
```

```
favstats(moviesa$revenue_mil)
```

```
##       min       Q1   median       Q3      max     mean       sd  n missing
##  0.123777 28.83712 56.15409 108.5218 407.1973 88.44888 99.17766 61       0
```

15. What is the standard deviation for 'Budget'? What is the standard deviation for 'Revenue'? Use proper notation.

16. Calculate the slope of the least squares line between 'Budget' and 'Revenue'.

The slope measures the change in y for each increase in x by 1. In other words, as the x variable increases by 1 unit, the y variable changes (increase/decreases) by the value of slope.

17. Interpret the value of slope in context of the problem.

We can use the **point-slope** form of a line to find the least squares line using the estimate for slope and the mean of x and y. $y - \bar{y} = b_1(x - \bar{x})$

18. Identify the least squares line using the summary statistics provided.

The linear model function in R can also give us the summary of the least squares fit. The estimate for (Intercept) is the y-intercept for the line of least squares and the estimate for budget is the value of b_1, the slope.

```
revenueLM <- lm(revenue_mil~budget_mil,data=moviesa)
summary(revenueLM)
```

```
##
## Call:
## lm(formula = revenue_mil ~ budget_mil, data = moviesa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.85  -34.95  -13.40   18.52  289.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.0362    14.3458   1.397    0.168
## budget_mil    0.9237     0.1419   6.511 1.81e-08
##
## Residual standard error: 76.29 on 59 degrees of freedom
## Multiple R-squared:  0.4181, Adjusted R-squared:  0.4083
## F-statistic:  42.4 on 1 and 59 DF,  p-value: 1.806e-08
```

19. Does the least squares line found in Question 18 match the output from the linear model? Write out this equation.

20. Using the least squares line from Question 19, predict the revenue for a movie with a budget of 165 million.

Residuals:

The model we are using assumes the relationship between the two variables follows a straight line. The residuals are the errors, or the part that hasn't been modeled by the line.

$$\text{Data} = \text{Model} + \text{Residual}$$

$$\text{Residual} = \text{Data - Model}$$

$$e_i = y_i - \hat{y}_i$$

21. The movie, *Independence Day: Resurgence*, had a budget of 165 million and revenue of 102.315 million. Find the residual for this movie.

22. Did the line of regression overestimate or underestimate the revenue for this movie?

Coefficient of Determination

The coefficient of determination, R^2, can also be used to describe the strength of the linear relationship between two quantitative variables. R^2 describes the amount of variation in the response that is explained by the least squares line with the explanatory variable.

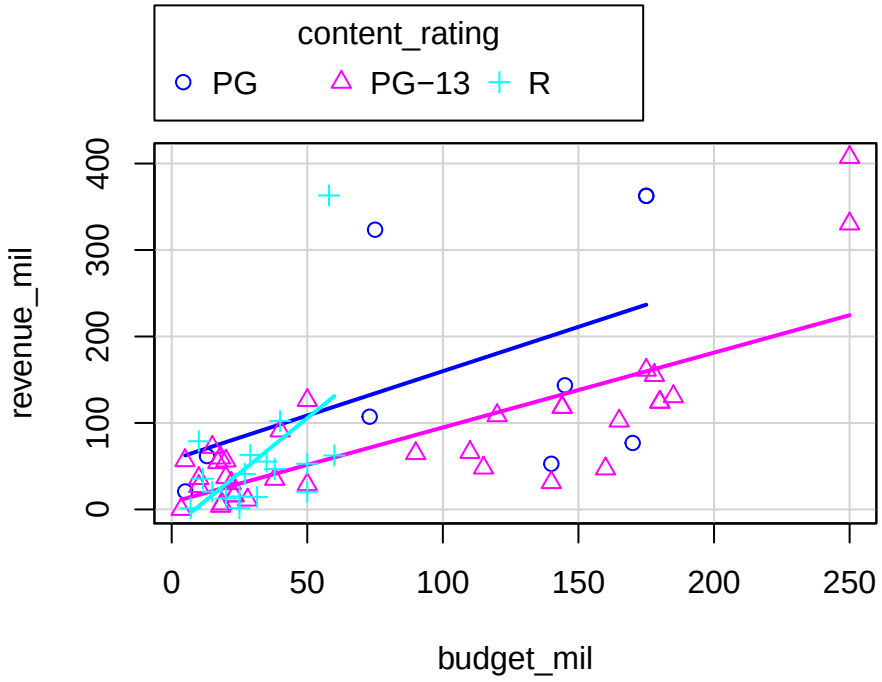
23. Calculate the coefficient of determination between 'Budget' and 'Revenue'.

24. Interpret the coefficient of determination in context of the problem.

Multiple Linear Regression

Let's look at this linear relationship between 'Budget' and 'Revenue' for different ratings.

```
scatterplot(revenue_mil~budget_mil|content_rating, data = moviesa, smooth=F)
```

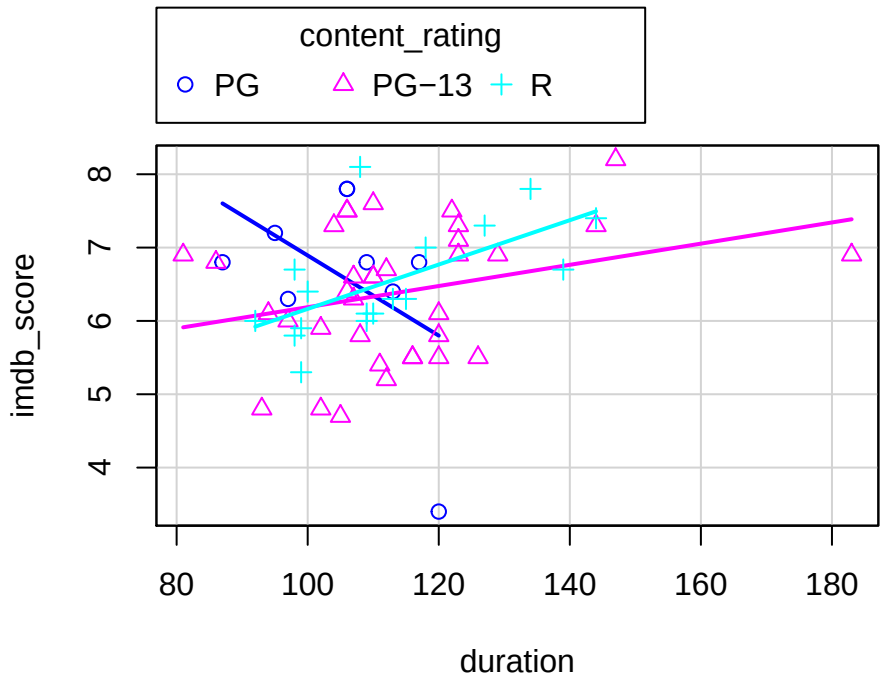


25. Does the relationship between 'Budget' and 'Revenue' differ among the different content ratings? Explain.

```r
imdbLM <- lm(imdb_score~duration, data=moviesa)
summary(imdbLM)
```

```
##
## Call:
## lm(formula = imdb_score ~ duration, data = moviesa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.16127 -0.46127  0.03949  0.64720  1.69565
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.992111   0.794099   6.287 4.31e-08
## duration    0.013076   0.007003   1.867   0.0668
##
## Residual standard error: 0.9033 on 59 degrees of freedom
## Multiple R-squared:  0.05579,    Adjusted R-squared:  0.03979
## F-statistic: 3.486 on 1 and 59 DF,  p-value: 0.06685
```

```r
scatterplot(imdb_score~duration|content_rating, data = moviesa, smooth=F)
```

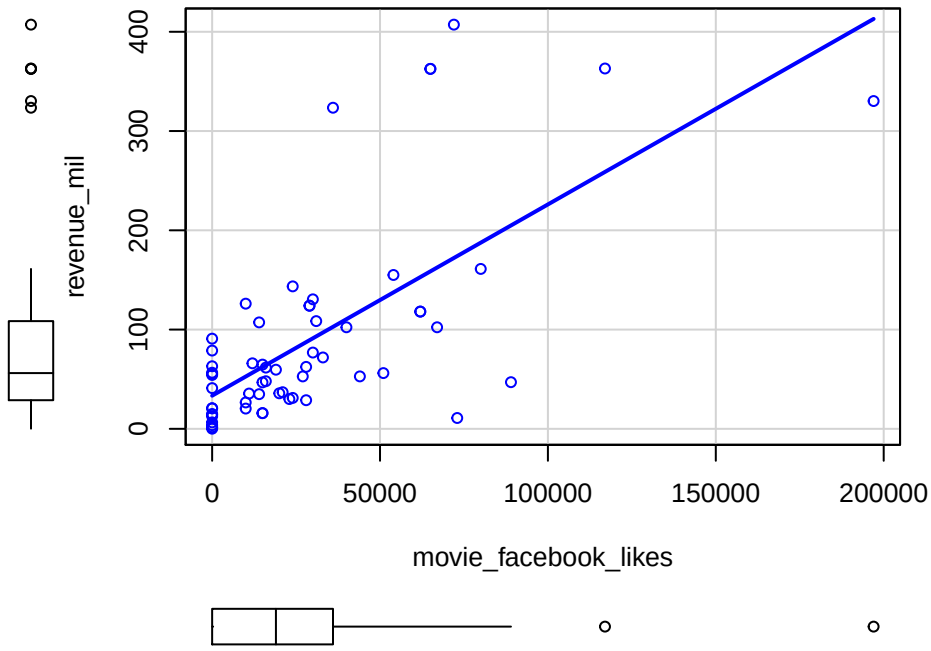



```r
imdbLM2 <- lm(imdb_score~duration+content_rating, data=moviesa)
summary(imdbLM2)
```

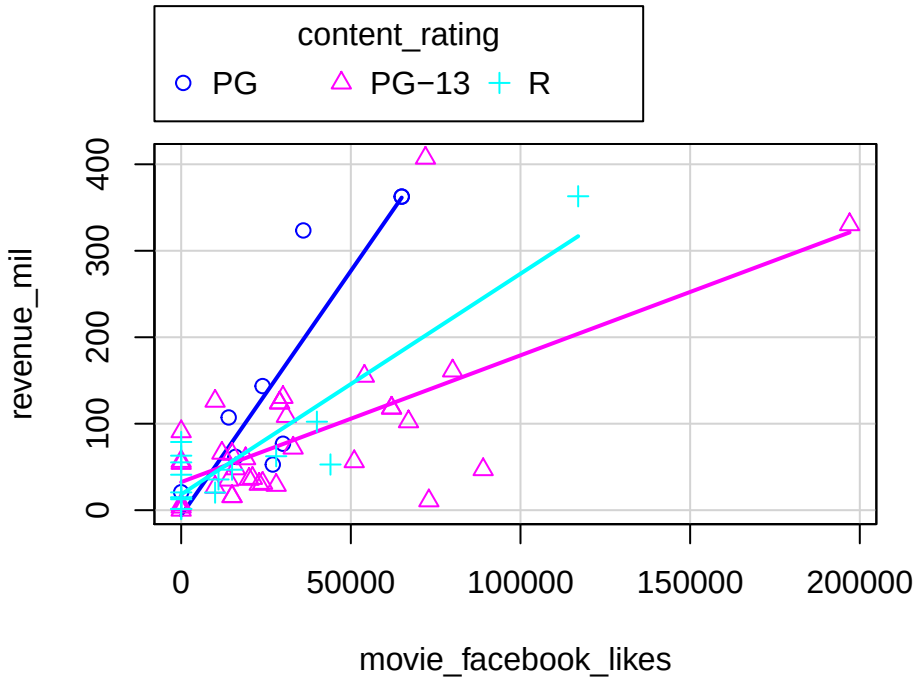
```
##
## Call:
## lm(formula = imdb_score ~ duration + content_rating, data = moviesa)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.3935 -0.4670  0.0355  0.5530  1.6222
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.093303   0.815093   6.249 5.61e-08
## duration              0.014169   0.007166   1.977   0.0529
## content_ratingPG-13 -0.319178   0.345441  -0.924   0.3594
## content_ratingR     -0.145690   0.378736  -0.385   0.7019
##
## Residual standard error: 0.9108 on 57 degrees of freedom
## Multiple R-squared:  0.07252,    Adjusted R-squared:  0.02371
## F-statistic: 1.486 on 3 and 57 DF,  p-value: 0.2281
```

```
scatterplot(revenue_mil~movie_facebook_likes, data = moviesa, smooth=F)
```




```
scatterplot(revenue_mil~movie_facebook_likes|content_rating, data = moviesa, smooth=F)
```




```
facebookLM <- lm(revenue_mil~movie_facebook_likes+content_rating, data=moviesa)
summary(facebookLM)
```

```
##
## Call:
## lm(formula = revenue_mil ~ movie_facebook_likes + content_rating,
##     data = moviesa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -146.79  -28.24  -11.31   32.17  251.43
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.083e+02  2.404e+01   4.508 3.31e-05
## movie_facebook_likes    1.937e-03  2.615e-04   7.407 6.66e-10
## content_ratingPG-13    -9.204e+01  2.541e+01  -3.622 0.000623
## content_ratingR        -8.050e+01  2.829e+01  -2.846 0.006149
##
## Residual standard error: 67.94 on 57 degrees of freedom
## Multiple R-squared:  0.5541, Adjusted R-squared:  0.5307
## F-statistic: 23.62 on 3 and 57 DF,  p-value: 4.624e-10
```