Final Project: Maximizing Book Profits

Author: Mikea Mullins

Discussants: Kaggle

Introduction

My roommate and dear friend is a wonderful writer, and post-graduation she plans to write professionally and eventually publish a novel. To advance that goal she is majoring in English, but she has some concerns about her ability to live on an English-major writer's salary. I want to alleviate some of her worries, so this report will explore which aspects of book publishing are the best predictors of a book's total sales revenue. I will then be able to advise her on which aspects of book publication are the most important for her to focus on in order to increase her income.

I will use Josh Murrey's data set Book Sales and Ratings, uploaded to Kaggle on 12/5/23 at https://www.kaggle.com/datasets/thedevastator/books-sales-and-ratings/data.

There are five other analyses linked to this data, all uploaded within the last three days. Two of these use regression, but both are predicting number of sales rather than gross sales and use different variables to form their predictions. I am not using this data in another class or research project.

Results

Data wrangling: Variables that correlate with gross sales

I began by omitting the single case with missing values and then applied a log transformation to the gross.sales variable to minimize heteroscedasticity. In general this data was relatively clean from the outset. I used Cook's distance to assess for outliers and did not find any.

The variables used in this analysis are:

Book_ratings_count: The number of ratings readers' gave a book

gross.sales: A book's total sales revenue publisher.revenue: The publisher's revenue from selling a book

sale.price: A book's sale price

sales.rank: A book's rank based on sales performance

units.sold: The number of books sold

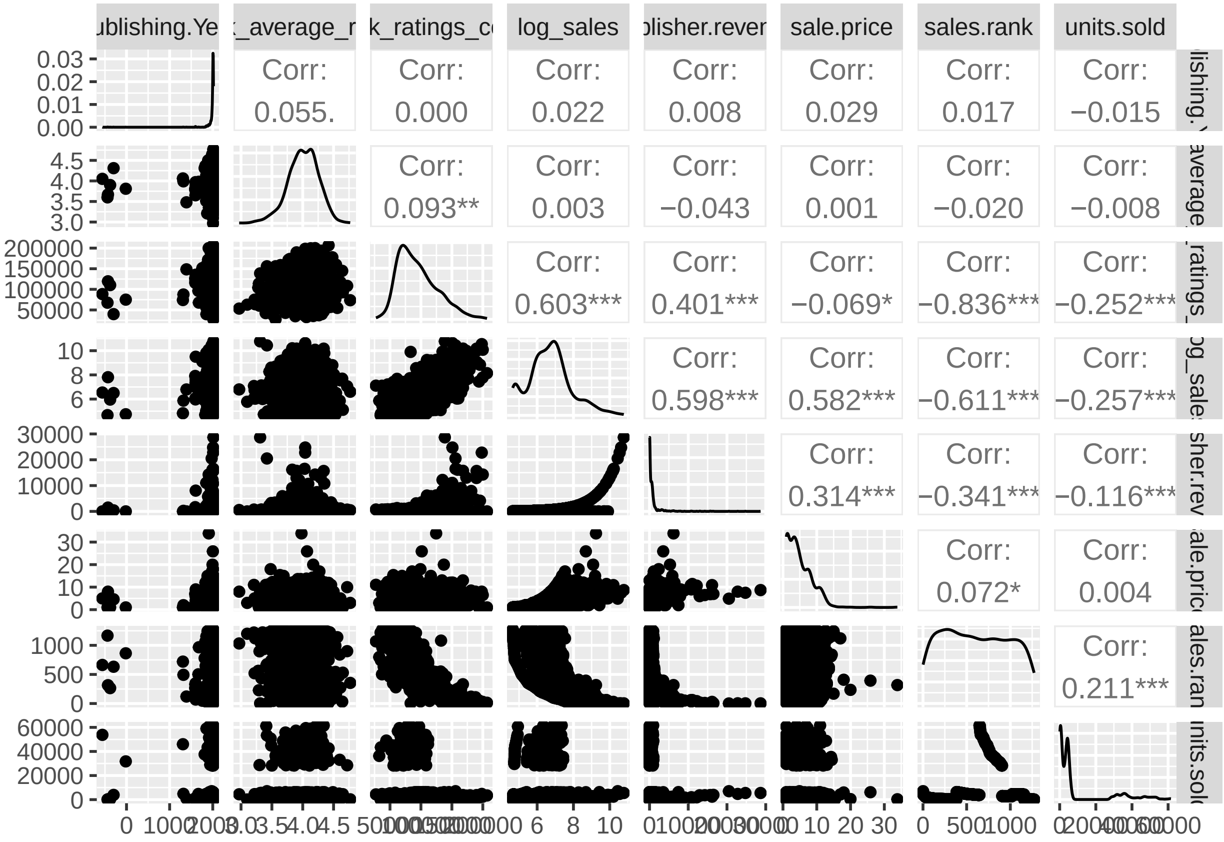
Visualize the data: Correlation between numeric variables and visualization of categorical variable

First I will visualize the correlation between pairs of numeric variables. This will help me decide which variables to investigate further as predictors of gross sales.

```r
library(dplyr)
library(ggplot2)
library(GGally)
library(car)

books <- read.csv("Books_Data_Clean.csv")

# apply log transformation to gross.sales to minimize
# heteroscedasticity
log_books <- books %>%
    mutate(log_sales = log(gross.sales)) %>%
    na.omit
# visualize correlation between pairs of numeric data
ggpairs(log_books %>%
    select(Publishing.Year, Book_average_rating, Book_ratings_count,
        log_sales, publisher.revenue, sale.price, sales.rank,
        units.sold))
```

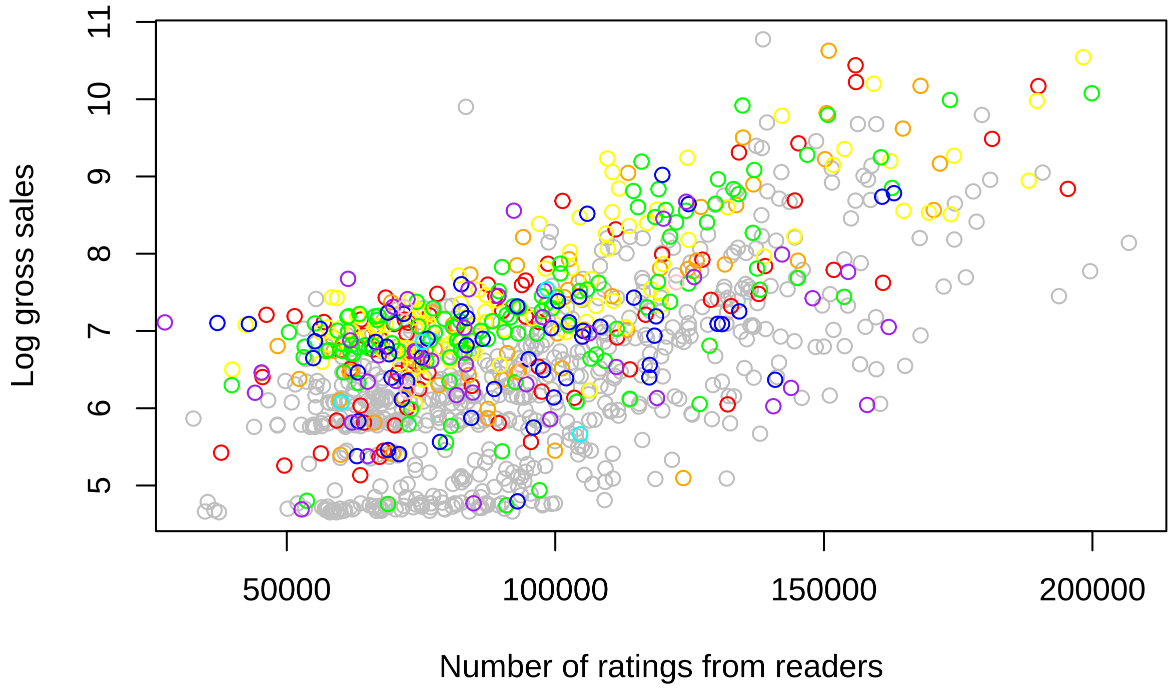


The ggpairs plot indicates that there is strong correlation between gross sales, publisher revenue, sale price, sales rank, units sold, and number of book ratings.

Next I will investigate whether there is a connection between a book's publisher and its gross sales.

```
publisher_data <- log_books %>% filter(Publisher %in% c("HarperCollins Publishers", "Amazon Digital Ser

plot(log_sales ~ Book_ratings_count, data = filter(publisher_data, Publisher == "Amazon Digital Service
points(log_sales ~ Book_ratings_count, data = filter(publisher_data, Publisher == "HarperCollins Publis
points(log_sales ~ Book_ratings_count, data = filter(publisher_data, Publisher == "Hachette Book Group"
points(log_sales ~ Book_ratings_count, data = filter(publisher_data, Publisher == "Penguin Group (USA) I
points(log_sales ~ Book_ratings_count, data = filter(publisher_data, Publisher == "Random House LLC"), c
points(log_sales ~ Book_ratings_count, data = filter(publisher_data, Publisher == "Simon and Schuster Di
points(log_sales ~ Book_ratings_count, data = filter(publisher_data, Publisher == "Macmillan"), col = "
points(log_sales ~ Book_ratings_count, data = filter(publisher_data, Publisher == "HarperCollins Publis
points(log_sales ~ Book_ratings_count, data = filter(publisher_data, Publisher == "HarperCollins Christ
```



The data overlaps, which makes it difficult to assess visually whether there is a correlation between a book's publisher and its total revenue.

Analyses: Using multiple regression to model a book's gross profits

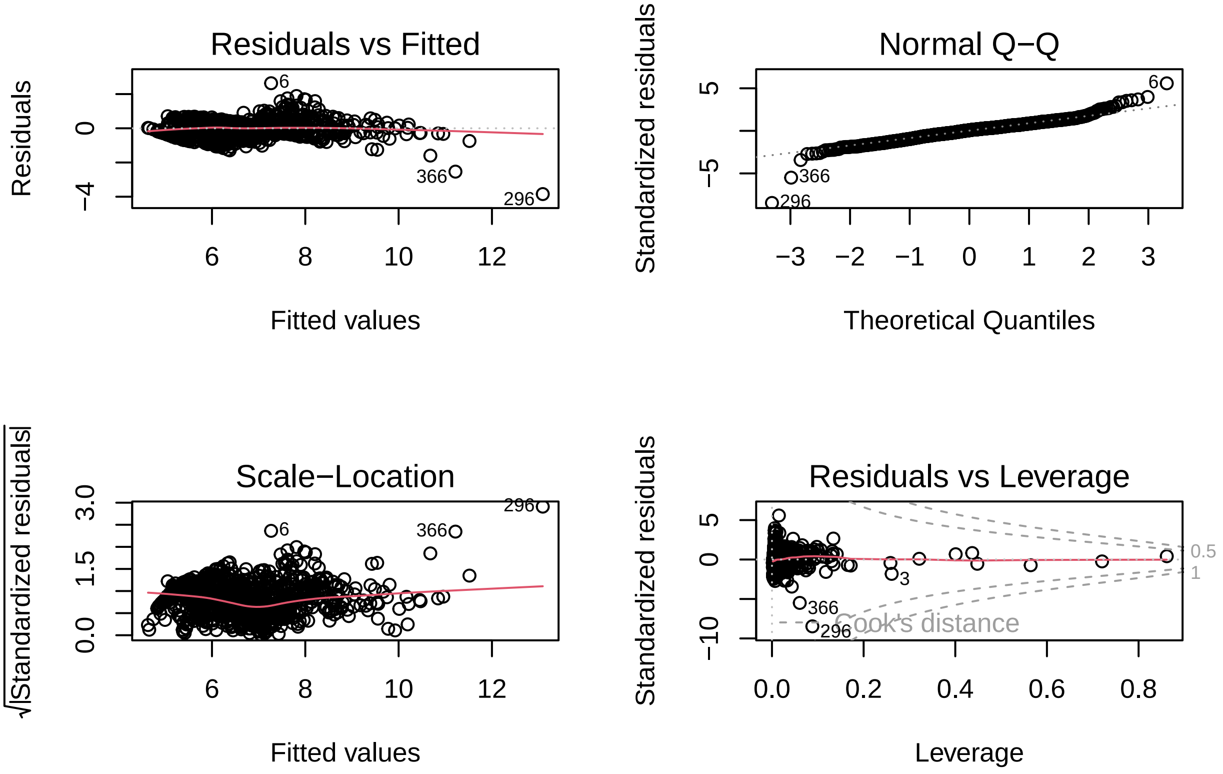
```
library(plotly)

# model for multiple regression Publisher
lm_fit <- lm(formula = log_sales ~ Book_ratings_count + publisher.revenue +
    sale.price + sales.rank + units.sold + Book_ratings_count *
    Publisher, data = publisher_data)

summary(lm_fit)
```

```
##
## Call:
## lm(formula = log_sales ~ Book_ratings_count + publisher.revenue +
##     sale.price + sales.rank + units.sold + Book_ratings_count *
##     Publisher, data = publisher_data)
##
## Residuals:
##     Min       1Q  Median       3Q      Max
## -3.8465 -0.2641  0.0323  0.2842  2.6342
##
## Coefficients:
##                                                                     Estimate
## (Intercept)                                                        5.577e+00
## Book_ratings_count                                                 9.845e-06
## publisher.revenue                                                  1.061e-04
## sale.price                                                         1.817e-01
## sales.rank                                                        -1.185e-03
## units.sold                                                        -7.707e-06
## PublisherHachette Book Group                                       3.792e-01
## PublisherHarperCollins Christian Publishing                        2.088e+00
## PublisherHarperCollins Publishers                                  3.704e-02
## PublisherHarperCollins Publishing                                 -2.004e-02
## PublisherMacmillan                                                 5.106e-01
## PublisherPenguin Group (USA) LLC                                   4.718e-01
## PublisherRandom House LLC                                          4.847e-01
## PublisherSimon and Schuster Digital Sales Inc                      4.801e-01
## Book_ratings_count:PublisherHachette Book Group                   -2.838e-06
## Book_ratings_count:PublisherHarperCollins Christian Publishing   -2.500e-05
## Book_ratings_count:PublisherHarperCollins Publishers               3.957e-07
## Book_ratings_count:PublisherHarperCollins Publishing               1.138e-06
## Book_ratings_count:PublisherMacmillan                             -5.156e-06
## Book_ratings_count:PublisherPenguin Group (USA) LLC               -4.953e-06
## Book_ratings_count:PublisherRandom House LLC                      -4.669e-06
## Book_ratings_count:PublisherSimon and Schuster Digital Sales Inc -4.934e-06
##                                                                   Std. Error
## (Intercept)                                                        1.318e-01
## Book_ratings_count                                                 9.664e-07
## publisher.revenue                                                  8.776e-06
## sale.price                                                         5.207e-03
## sales.rank                                                         7.223e-05
## units.sold                                                         9.812e-07
## PublisherHachette Book Group                                       2.045e-01
## PublisherHarperCollins Christian Publishing                        1.155e+00
## PublisherHarperCollins Publishers                                  1.758e-01
## PublisherHarperCollins Publishing                                  1.069e+00
## PublisherMacmillan                                                 2.311e-01
## PublisherPenguin Group (USA) LLC                                   1.621e-01
## PublisherRandom House LLC                                          1.616e-01
## PublisherSimon and Schuster Digital Sales Inc                      2.342e-01
## Book_ratings_count:PublisherHachette Book Group                    2.023e-06
## Book_ratings_count:PublisherHarperCollins Christian Publishing    1.335e-05
## Book_ratings_count:PublisherHarperCollins Publishers               1.761e-06
## Book_ratings_count:PublisherHarperCollins Publishing               1.154e-05
## Book_ratings_count:PublisherMacmillan                              2.322e-06
## Book_ratings_count:PublisherPenguin Group (USA) LLC                1.600e-06
## Book_ratings_count:PublisherRandom House LLC                       1.630e-06
## Book_ratings_count:PublisherSimon and Schuster Digital Sales Inc  2.425e-06
##                                                                   t value
## (Intercept)                                                        42.304
## Book_ratings_count                                                 10.187
## publisher.revenue                                                  12.094
## sale.price                                                         34.886
## sales.rank                                                       -16.403
## units.sold                                                        -7.855
## PublisherHachette Book Group                                       1.854
## PublisherHarperCollins Christian Publishing                        1.808
## PublisherHarperCollins Publishers                                  0.211
## PublisherHarperCollins Publishing                                 -0.019
## PublisherMacmillan                                                 2.209
## PublisherPenguin Group (USA) LLC                                   2.911
## PublisherRandom House LLC                                          2.999
## PublisherSimon and Schuster Digital Sales Inc                      2.050
## Book_ratings_count:PublisherHachette Book Group                   -1.403
## Book_ratings_count:PublisherHarperCollins Christian Publishing    -1.874
## Book_ratings_count:PublisherHarperCollins Publishers               0.225
## Book_ratings_count:PublisherHarperCollins Publishing               0.099
## Book_ratings_count:PublisherMacmillan                             -2.221
## Book_ratings_count:PublisherPenguin Group (USA) LLC               -3.095
## Book_ratings_count:PublisherRandom House LLC                      -2.865
## Book_ratings_count:PublisherSimon and Schuster Digital Sales Inc  -2.035
##                                                                  Pr(>|t|)
## (Intercept)                                                       < 2e-16 ***
## Book_ratings_count                                                < 2e-16 ***
## publisher.revenue                                                 < 2e-16 ***
## sale.price                                                        < 2e-16 ***
## sales.rank                                                        < 2e-16 ***
## units.sold                                                       9.88e-15 ***
## PublisherHachette Book Group                                      0.06400 .
## PublisherHarperCollins Christian Publishing                       0.07090 .
## PublisherHarperCollins Publishers                                 0.83317
## PublisherHarperCollins Publishing                                 0.98505
## PublisherMacmillan                                                0.02739 *
## PublisherPenguin Group (USA) LLC                                  0.00368 **
## PublisherRandom House LLC                                         0.00277 **
## PublisherSimon and Schuster Digital Sales Inc                     0.04064 *
## Book_ratings_count:PublisherHachette Book Group                   0.16099
## Book_ratings_count:PublisherHarperCollins Christian Publishing    0.06126 .
## Book_ratings_count:PublisherHarperCollins Publishers              0.82228
## Book_ratings_count:PublisherHarperCollins Publishing              0.92150
## Book_ratings_count:PublisherMacmillan                             0.02659 *
## Book_ratings_count:PublisherPenguin Group (USA) LLC               0.00202 **
## Book_ratings_count:PublisherRandom House LLC                      0.00426 **
## Book_ratings_count:PublisherSimon and Schuster Digital Sales Inc  0.04211 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4756 on 1047 degrees of freedom
## Multiple R-squared:  0.8389, Adjusted R-squared:  0.8356
## F-statistic: 259.6 on 21 and 1047 DF,  p-value: < 2.2e-16
```

```
# diagnostic plots
par(mfrow = c(2, 2))
plot(lm_fit)
```



```r
# maximum Cook's distance
max(cooks.distance(lm_fit))
```

```
## [1] 0.3152226
```

The model captures 83.59% of the total sum of squares of gross sales.

The diagnostic plots indicate that the model is linear, normal, homoscedastic, and without high-leverage points.

Additionally, the maximum Cook's distance is less than .5, which reaffirms that there are no high-leverage points.

Conclusion

The best predictors of gross sales are publisher revenue, sale price, units sold, and the number of ratings a book receives from readers. The first three of these are not particularly useful–it makes intuitive sense that selling more books and increasing sale price would increase total revenue, and of course the publisher's revenue increases when gross profits increase.

This leaves books ratings. There is a clear correlation between the number of ratings a book receives and its gross profit, so I will tell my writer friend that her goal should be to increase the number of reviews her readers leave. She could do this by engaging in self-promotion on social media and encouraging her readers to rate her book, or by paying people to write reviews for her. Additionally, adding a book's publisher to the model as an interaction term shows that the best publisher for increasing gross profits is Amazon Digital Services, Inc., so my friend should try to publish with Amazon.

Reflection

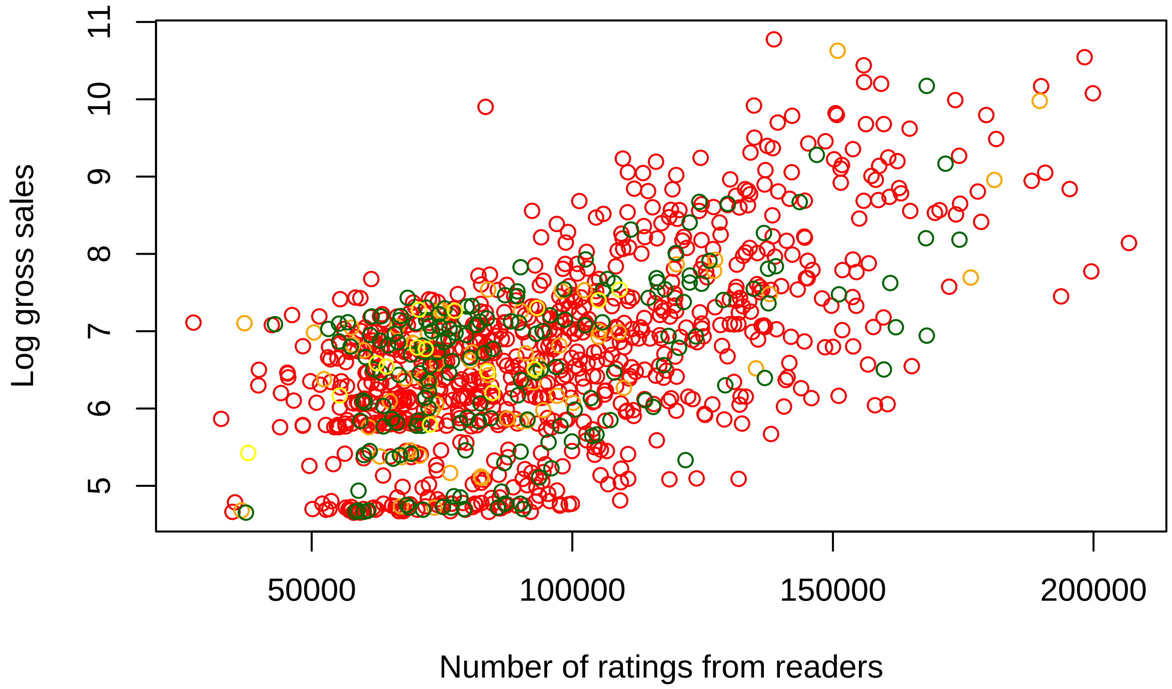
The most difficult part of this project was choosing a data set. I wasn't sure exactly what to look for, and the data set I chose initially ended up being too difficult for me to use. I ended up trying several data sets before choosing on this one. Once I settled on a data set, though, the project went pretty smoothly. I enjoyed analyzing and looking for patterns in the data. I spent about twelve hours on this project.

Appendix

Appendix: Data Wrangling

I considered using genre as a categorical variable in addition to or instead of publishers, but I did not find evidence that genre significantly impacted gross book sales.

```r
# visualization of data filtered by genre
genre_data <- log_books %>% filter(genre %in% c("genre fiction", "fiction", "nonfiction", "children")) %
plot(log_sales ~ Book_ratings_count, data = filter(genre_data, genre == "genre fiction"), col = "red", 
points(log_sales ~ Book_ratings_count, data = filter(genre_data, genre == "fiction"), col = "orange")
points(log_sales ~ Book_ratings_count, data = filter(genre_data, genre == "nonfiction"), col = "darkgre
points(log_sales ~ Book_ratings_count, data = filter(genre_data, genre == "children"), col = "yellow")
```



Linear models with genre as a predictor variable:

```
# comparing linear models with genre vs publisher as categorical variables
lm_genre <- lm(formula = log_sales ~ Book_ratings_count + genre, data = genre_data)
summary(lm_genre)
```

```
##
## Call:
## lm(formula = log_sales ~ Book_ratings_count + genre, data = genre_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4389 -0.6414  0.0446  0.6661  3.4680
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.939e+00  2.518e-01  19.618   <2e-16 ***
## Book_ratings_count  2.256e-05  9.141e-07  24.679   <2e-16 ***
## genrefiction       -2.978e-01  2.697e-01  -1.104    0.270
## genregenre fiction -3.857e-01  2.447e-01  -1.577    0.115
## genrenonfiction    -3.879e-01  2.526e-01  -1.536    0.125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9364 on 1064 degrees of freedom
## Multiple R-squared:  0.3651, Adjusted R-squared:  0.3627
## F-statistic:   153 on 4 and 1064 DF,  p-value: < 2.2e-16
```

```r
lm_publisher <- lm(formula = log_sales ~ Book_ratings_count + Publisher, data = publisher_data)
summary(lm_publisher)
```

```
##
## Call:
## lm(formula = log_sales ~ Book_ratings_count + Publisher, data = publisher_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7745 -0.5895  0.0751  0.5074  3.8473
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 4.197e+00  8.217e-02  51.073
## Book_ratings_count                          2.228e-05  7.916e-07  28.149
## PublisherHachette Book Group                9.140e-01  1.064e-01   8.588
## PublisherHarperCollins Christian Publishing  4.514e-01  4.089e-01   1.104
## PublisherHarperCollins Publishers           8.234e-01  1.023e-01   8.051
## PublisherHarperCollins Publishing           7.077e-01  4.088e-01   1.731
## PublisherMacmillan                          5.230e-01  1.315e-01   3.976
## PublisherPenguin Group (USA) LLC            1.190e+00  8.518e-02  13.965
## PublisherRandom House LLC                   9.480e-01  8.149e-02  11.634
## PublisherSimon and Schuster Digital Sales Inc 5.988e-01  1.139e-01   5.258
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## Book_ratings_count                           < 2e-16 ***
## PublisherHachette Book Group                 < 2e-16 ***
## PublisherHarperCollins Christian Publishing    0.2698
## PublisherHarperCollins Publishers           2.19e-15 ***
## PublisherHarperCollins Publishing             0.0837 .
## PublisherMacmillan                          7.49e-05 ***
## PublisherPenguin Group (USA) LLC             < 2e-16 ***
## PublisherRandom House LLC                    < 2e-16 ***
## PublisherSimon and Schuster Digital Sales Inc 1.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8149 on 1059 degrees of freedom
## Multiple R-squared:  0.5215, Adjusted R-squared:  0.5174
## F-statistic: 128.2 on 9 and 1059 DF,  p-value: < 2.2e-16
```

```r
# linear model with both genre and publisher as categorical variables
publisher_and_genre_data <- log_books %>% filter(Publisher %in% c("HarperCollins Publishers", "Amazon D

lm_combo <- lm(formula = log_sales ~ Book_ratings_count + Publisher + genre + Book_ratings_count * genr
summary(lm_combo)
```

```
##
## Call:
## lm(formula = log_sales ~ Book_ratings_count + Publisher + genre +
##     Book_ratings_count * genre + Book_ratings_count * Publisher,
##     data = publisher_and_genre_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9794 -0.5878  0.0835  0.5055  3.8580
##
## Coefficients:
##                                                                     Estimate
## (Intercept)                                                        4.500e+00
## Book_ratings_count                                                 2.540e-05
## PublisherHachette Book Group                                       3.421e-01
## PublisherHarperCollins Christian Publishing                        1.853e+00
## PublisherHarperCollins Publishers                                  7.280e-01
## PublisherHarperCollins Publishing                                  1.132e+00
## PublisherMacmillan                                                 1.937e+00
## PublisherPenguin Group (USA) LLC                                   1.288e+00
## PublisherRandom House LLC                                          8.570e-01
## PublisherSimon and Schuster Digital Sales Inc                      1.424e+00
## genrefiction                                                      -5.561e-01
## genregenre fiction                                                -4.002e-01
## genrenonfiction                                                   -1.316e-01
## Book_ratings_count:genrefiction                                   -1.027e-06
## Book_ratings_count:genregenre fiction                             -2.078e-06
## Book_ratings_count:genrenonfiction                                -5.749e-06
## Book_ratings_count:PublisherHachette Book Group                    6.020e-06
## Book_ratings_count:PublisherHarperCollins Christian Publishing   -1.594e-05
## Book_ratings_count:PublisherHarperCollins Publishers               1.134e-06
## Book_ratings_count:PublisherHarperCollins Publishing              -3.967e-06
## Book_ratings_count:PublisherMacmillan                             -1.515e-05
## Book_ratings_count:PublisherPenguin Group (USA) LLC               -9.449e-07
## Book_ratings_count:PublisherRandom House LLC                       1.061e-06
## Book_ratings_count:PublisherSimon and Schuster Digital Sales Inc -8.807e-06
##                                                                   Std. Error
## (Intercept)                                                        9.142e-01
## Book_ratings_count                                                 1.158e-05
## PublisherHachette Book Group                                       3.376e-01
## PublisherHarperCollins Christian Publishing                        1.966e+00
## PublisherHarperCollins Publishers                                  2.884e-01
## PublisherHarperCollins Publishing                                  1.822e+00
## PublisherMacmillan                                                 3.878e-01
## PublisherPenguin Group (USA) LLC                                   2.624e-01
## PublisherRandom House LLC                                          2.653e-01
## PublisherSimon and Schuster Digital Sales Inc                      3.977e-01
## genrefiction                                                       9.647e-01
## genregenre fiction                                                 9.162e-01
## genrenonfiction                                                    9.344e-01
## Book_ratings_count:genrefiction                                    1.203e-05
## Book_ratings_count:genregenre fiction                              1.159e-05
## Book_ratings_count:genrenonfiction                                 1.175e-05
## Book_ratings_count:PublisherHachette Book Group                    3.285e-06
## Book_ratings_count:PublisherHarperCollins Christian Publishing    2.270e-05
## Book_ratings_count:PublisherHarperCollins Publishers               2.855e-06
## Book_ratings_count:PublisherHarperCollins Publishing               1.967e-05
## Book_ratings_count:PublisherMacmillan                              3.916e-06
## Book_ratings_count:PublisherPenguin Group (USA) LLC                2.578e-06
## Book_ratings_count:PublisherRandom House LLC                      2.668e-06
## Book_ratings_count:PublisherSimon and Schuster Digital Sales Inc  4.122e-06
##                                                                  t value
## (Intercept)                                                        4.922
## Book_ratings_count                                                 2.194
## PublisherHachette Book Group                                       1.013
## PublisherHarperCollins Christian Publishing                        0.942
## PublisherHarperCollins Publishers                                  2.524
## PublisherHarperCollins Publishing                                  0.621
## PublisherMacmillan                                                 4.996
## PublisherPenguin Group (USA) LLC                                   4.910
## PublisherRandom House LLC                                          3.231
## PublisherSimon and Schuster Digital Sales Inc                      3.580
## genrefiction                                                      -0.576
## genregenre fiction                                                -0.437
## genrenonfiction                                                   -0.141
## Book_ratings_count:genrefiction                                   -0.085
## Book_ratings_count:genregenre fiction                             -0.179
## Book_ratings_count:genrenonfiction                                -0.489
## Book_ratings_count:PublisherHachette Book Group                    1.833
## Book_ratings_count:PublisherHarperCollins Christian Publishing    -0.702
## Book_ratings_count:PublisherHarperCollins Publishers               0.397
## Book_ratings_count:PublisherHarperCollins Publishing              -0.202
## Book_ratings_count:PublisherMacmillan                             -3.868
## Book_ratings_count:PublisherPenguin Group (USA) LLC               -0.367
## Book_ratings_count:PublisherRandom House LLC                       0.398
## Book_ratings_count:PublisherSimon and Schuster Digital Sales Inc  -2.137
##                                                                  Pr(>|t|)
## (Intercept)                                                      9.93e-07 ***
## Book_ratings_count                                               0.028451 *
## PublisherHachette Book Group                                     0.311148
## PublisherHarperCollins Christian Publishing                      0.346267
## PublisherHarperCollins Publishers                                0.011753 *
## PublisherHarperCollins Publishing                                0.534548
## PublisherMacmillan                                               6.85e-07 ***
## PublisherPenguin Group (USA) LLC                                 1.06e-06 ***
## PublisherRandom House LLC                                        0.001274 **
## PublisherSimon and Schuster Digital Sales Inc                    0.000360 ***
## genrefiction                                                     0.564449
## genregenre fiction                                               0.662362
## genrenonfiction                                                  0.888024
## Book_ratings_count:genrefiction                                  0.931989
## Book_ratings_count:genregenre fiction                            0.857726
## Book_ratings_count:genrenonfiction                               0.624846
## Book_ratings_count:PublisherHachette Book Group                  0.067116 .
## Book_ratings_count:PublisherHarperCollins Christian Publishing   0.482528
## Book_ratings_count:PublisherHarperCollins Publishers             0.691205
## Book_ratings_count:PublisherHarperCollins Publishing             0.840210
## Book_ratings_count:PublisherMacmillan                            0.000117 ***
## Book_ratings_count:PublisherPenguin Group (USA) LLC              0.714022
## Book_ratings_count:PublisherRandom House LLC                     0.690824
## Book_ratings_count:PublisherSimon and Schuster Digital Sales Inc 0.032841 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 0.8063 on 1045 degrees of freedom
## Multiple R-squared:  0.5377, Adjusted R-squared:  0.5275
## F-statistic: 52.84 on 23 and 1045 DF,  p-value: < 2.2e-16
```