Unit 2 homework sample solution

DKU Stats 101 Fall 2025 Session 2

Anonymous

2025-11-16

Books on Amazon.

Introduction

Question 1: Describing your data (10 points)

1a. Where is this data from?

For this dataset, describe the data according to the five Ws & *how* defined in the textbook Chapter 1.2. What are some possible problems with the *who* and *what* of the dataset?

The dataset you are using for this assignment is a subset of the original dataset that can be found here.

Who: Books listed on Amazon between 1997 and 2023 that have at least 200 reviews and are priced at $5 USD or more. Each observation represents a single book (title/edition).

What: Metadata and sales details for each book, including its title, author, publisher, publication year, format (hardcover, paperback, or ebook), page count, language, ISBNs, category and sub-category, average rating (1–5 scale), number of ratings, price (USD), and basic physical attributes such as dimensions and weight.

When: Data covers books published from 1997 to 2023 and was compiled and released on Kaggle in 2023.

Where: Scraped from Amazon's public book listings and user reviews across various genres and publishers.

Why: To examine factors that influence book prices, ratings, and popularity, and to serve as a teaching dataset for data analysis practice.

How: Collected through web-scraping by open-source contributors and cleaned for educational use on Kaggle with accompanying tutorial scripts.

Possible problems: For *Who*, the dataset mainly includes popular books with many reviews, which may underrepresent niche, newly published, or low-selling titles. For *What*, some variables such as price and rating may be outdated or inconsistently recorded across listings, and the category or sub-category labels may vary in format or completeness.

1b. What are the variable types?

For the following variables, please make a table.

One column should be the variable name, the second should be the variable type as defined in the textbook Chapter 1.3, and the third the units of the variable (if applicable).

- `title`
- `published_date`
- `format`
- `page_count`
- `isbn_10`
- `category`
- `average_rating`
- `price`
- `features_text`

Table 1: Variable Types

Variable	Type	Units (if applicable)
`title`	Identifier	None
`publisher_date`	Categorical	None
`format`	Categorical	None
`page_count`	Quantitative	Pages
`isbn_10`	Identifier	None
`average_rating`	Quantitative	Rating points (1–5)
`price`	Quantitative	USD
`features_text`	Categorical	None

Question 2: Association (20 points)

Using the `mutate()` verb as described in the DataCamp lab, make a new variable called `age` that subtracts the year 2025 from the book's `publish_year`. Please display your code using `#| echo: true` code block option.

```r
amazon_book <- amazon_book %>%
  mutate(age = 2025 - publisher_year)
```

2a. Investigating `age` vs. `average_rating`

Using the Think-Show-Tell framework from the textbook (example on page 213), please examine the relationship *in association terms* between `age` and `average_rating`. How strongly are they associated?

Note: for this question and all other Think sections in the homework, you do not need to report the W's of the data (you have already completed this in Q1)

> Think
>
> The variable `age` measures how many years have passed since a book was published, and `average_rating` represents the mean reader rating on a 1–5 scale. Both variables are quantitative. Given that this is a modern dataset with many recently published books, I expect age to have a moderate negative association with average_rating. Newer books may receive more attention and review activity on Amazon due to recent visibility, marketing, or genre trends, while older books may accumulate fewer new ratings over time.
>
> Because both variables are quantitative and we want to explore the direction and strength of their association, a scatterplot is the most appropriate display. In addition, computing the correlation coefficient will help quantify the degree of linear association between age and average_rating.
>
> Show

Table 2: Correlation between age and average_rating

variable_1	variable_2	correlation
age	average_rating	-0.014

Figure 1: Scatterplot of age vs. average_rating

Figure 1 displays a scatterplot of age versus average_rating. The points are widely scattered with no clear upward or downward pattern. Average ratings cluster tightly between 4.0 and 5.0 across all ages, and books of different ages appear to have similar rating levels. The fitted line is nearly flat, indicating little visible linear trend.

Table 1 reports the correlation between the two variables, which is close to zero (–0.014), consistent with the weak and patternless appearance of the scatterplot.

Tell

The scatterplot and correlation indicate that age and average_rating have essentially no linear association. Ratings remain tightly clustered between 4.0 and 5.0 regardless of how old a book is, and the fitted line shows almost no slope. The correlation of –0.014 confirms that the relationship is extremely weak.

These findings suggest that, in this dataset, book age does not meaningfully influence average rating. Amazon ratings tend to be consistently high overall, and

readers often rate books based on content quality or personal preference rather than how long ago the book was published. As a result, both older and newer books receive similarly high evaluations, producing very little variation across publication years. Although the correlation is technically negative, the value is far too small to indicate any real strength, which contradicts my expectation of a moderate negative association.

2b. Investigating `age` vs. `rating_number`

Using the Think-Show-Tell framework from the textbook, please examine the relationship *in association terms* between `age` and 'rating_number". How strongly are they associated?

Think

The variable `age` measures how many years have passed since a book was published, and `rating_number` represents the total number of user ratings the book has received on Amazon. Both variables are quantitative. Because this dataset includes many recently published books and reader activity on Amazon is generally higher for newer titles, I expect age to have a moderate negative association with rating_number. As books get older, they may receive fewer new ratings, while newer books tend to attract more attention and accumulate larger rating counts.

Before examining their association, it is helpful to first inspect the distribution of rating_number to understand its overall spread and skewness. Rating counts on Amazon can vary widely, with many books receiving only a small number of ratings while a few very popular titles receive thousands. Because of this potential skewness, viewing the distribution first provides useful context for interpreting the association. After reviewing the distribution, a scatterplot of age versus rating_number is the most appropriate display for assessing the direction and strength of their relationship. Computing the correlation coefficient will further quantify the linear association between the two variables.

Show

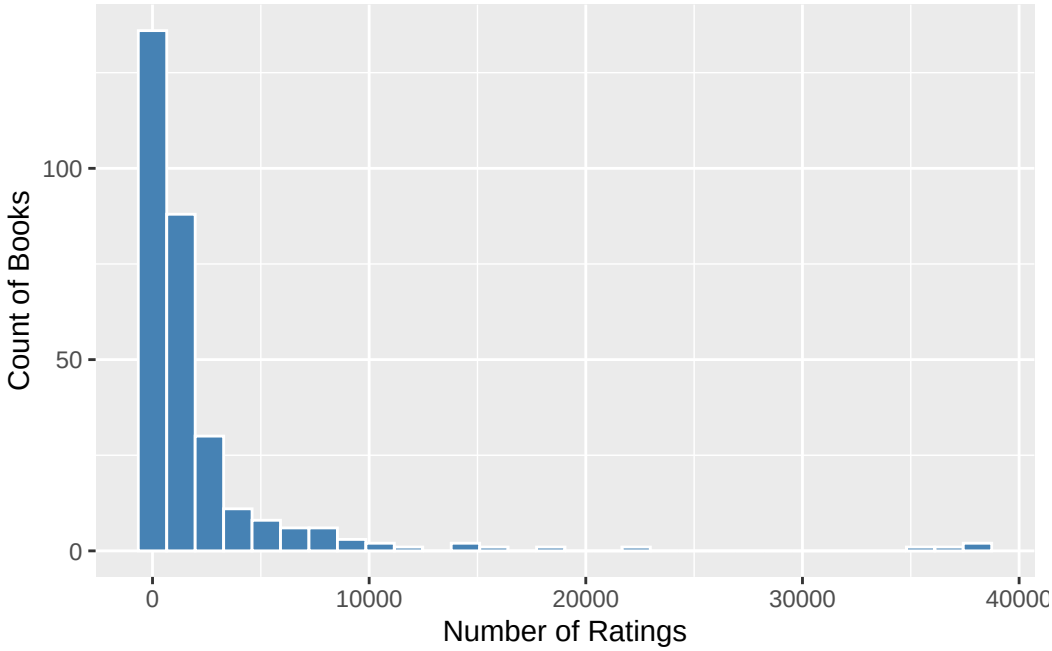

Figure 2: Histogram of rating_number (raw scale)

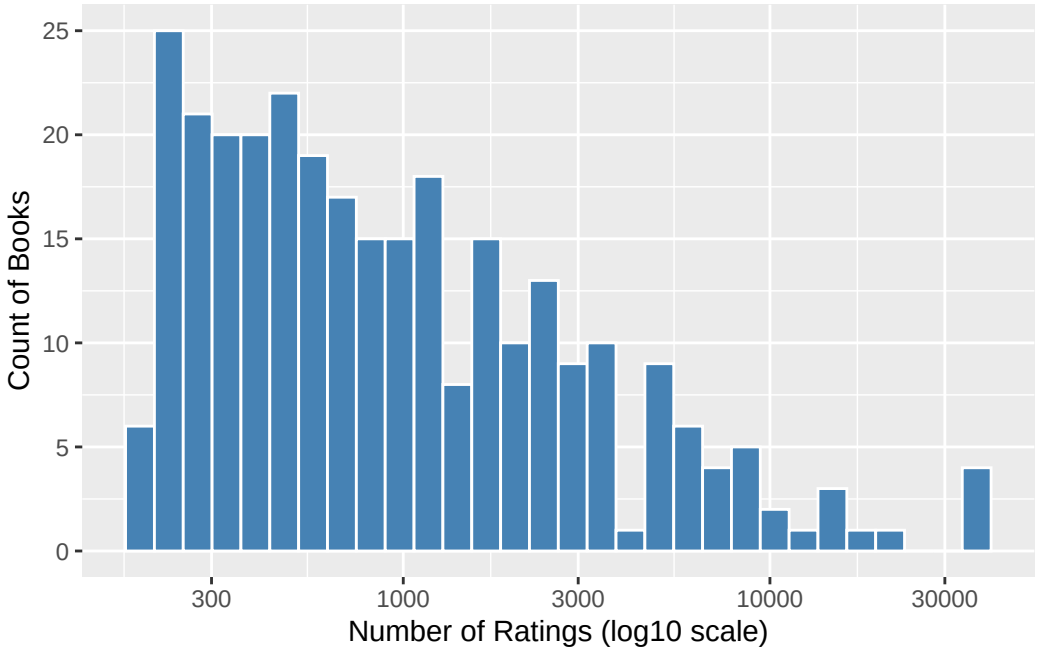

Figure 3: Histogram of rating_number (log10 scale)

Table 3: Correlation between age and log10(rating_number)

variable_1	variable_2	correlation
age	log10(rating_number)	-0.065

Figure 4: Scatterplot of age vs. log10(rating_number)

The distribution of *rating_number* is highly right-skewed, with most books receiving only a small number of ratings and a few titles having very large counts. The log10-transformed histogram spreads the values out but still shows a right-skewed pattern.

In the scatterplot of age versus log10(rating_number), the points are widely scattered with no strong visible pattern. The fitted line slopes slightly downward, but variation remains large across all ages. The correlation is –0.065, indicating a very weak negative linear association.

Tell

The results show that age and rating_number have only a very weak negative association. Although the fitted line slopes slightly downward and the correlation is –0.065, rating counts vary widely at every age, and both new and old books can have either very high or very low engagement.

This weak relationship likely reflects the fact that rating activity depends more on a book's popularity, genre, and visibility rather than its publication year. Older

books may continue to attract readers if they are classics or widely recommended, while some newer books may receive limited attention. Because these factors dominate over age, the linear association remains minimal.

Overall, the direction of the association matches my expectation, but the strength does not. The relationship is far weaker than the moderate negative association I expected.

2c. Thinking about your results

Consider the results of 2a. and 2b. together. What can we understand about how books are rated on Amazon from this information? What do you think explains the relationships you have identified?

Answers will vary here, good quality effort to interpret investigation of this question is required.

Question 3: Simple regression (20 points)

3a. Investigating `price` vs. `average_rating`

Using the Think-Show-Tell framework from the textbook, please examine how the price of a book is related to the rating of a book.

Think

The variable price measures how much a book costs in U.S. dollars, and average_rating is the mean reader rating on a 1–5 scale. Both are quantitative, and I will treat price as the explanatory variable and average_rating as the response.

I expect a moderate to strong positive association between the two. More expensive books are often higher-quality editions or specialized academic texts, which may better meet reader expectations and receive higher ratings. In contrast, cheaper mass-market editions may lead to more mixed reviews.

Before examining this relationship, it is useful to look at the distribution of price, which is typically right-skewed on Amazon due to many low-priced paperbacks and fewer costly specialty books. A histogram helps assess this skewness and whether a transformation is needed. After that, a scatterplot of price versus average_rating with a fitted regression line is the most appropriate display. Finally, summarizing the regression output will quantify the direction and strength of the linear association.

Show

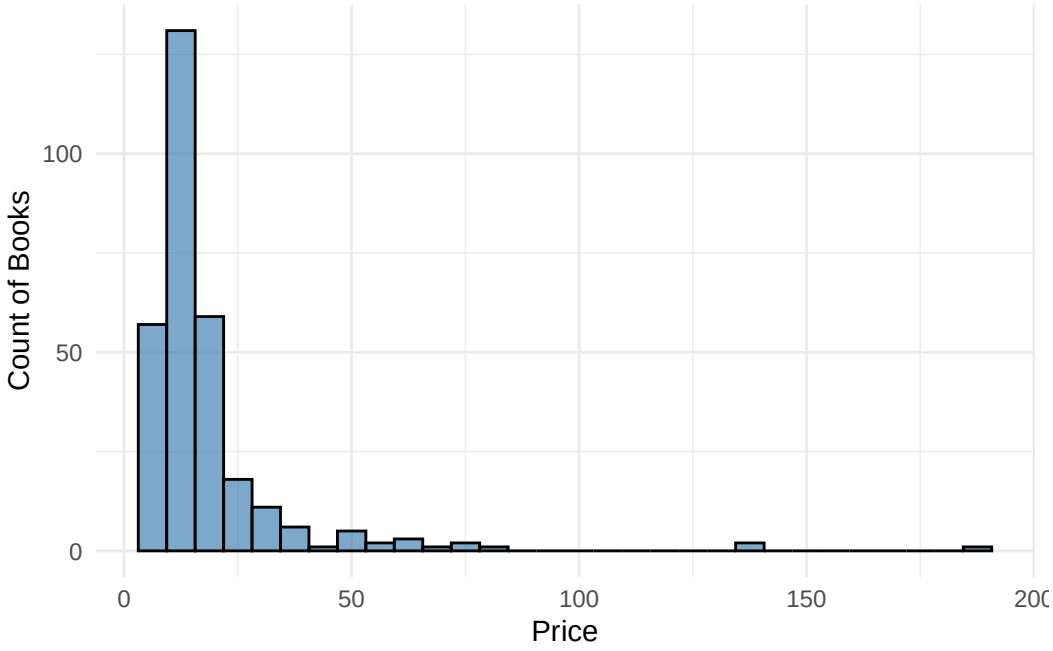

Figure 5: Histogram of Book Prices

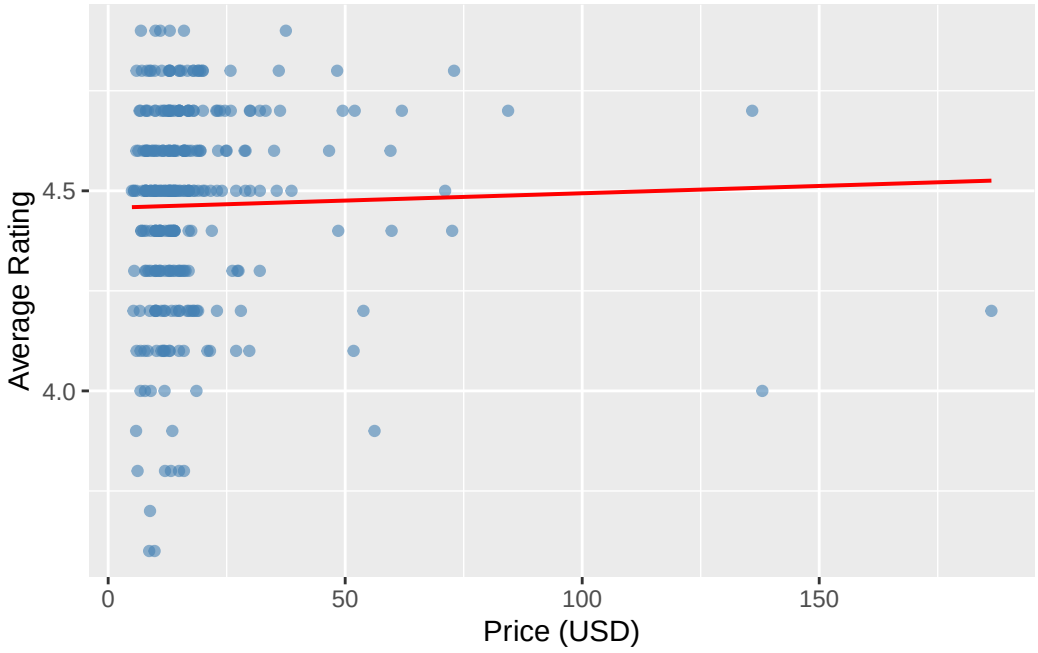

Figure 6: Scatterplot of price vs. average_rating with fitted regression line

Table 4: Simple Regression model: average_rating ~ log10(price)

	avg_rating ~ log10(price)
Intercept	4.338
	(0.069)
log10(Price in USD)	0.108
	(0.058)
Num.Obs.	300
R2	0.011
R2 Adj.	0.008
AIC	25.6
BIC	36.8
Log.Lik.	−9.823
F	3.455
RMSE	0.25

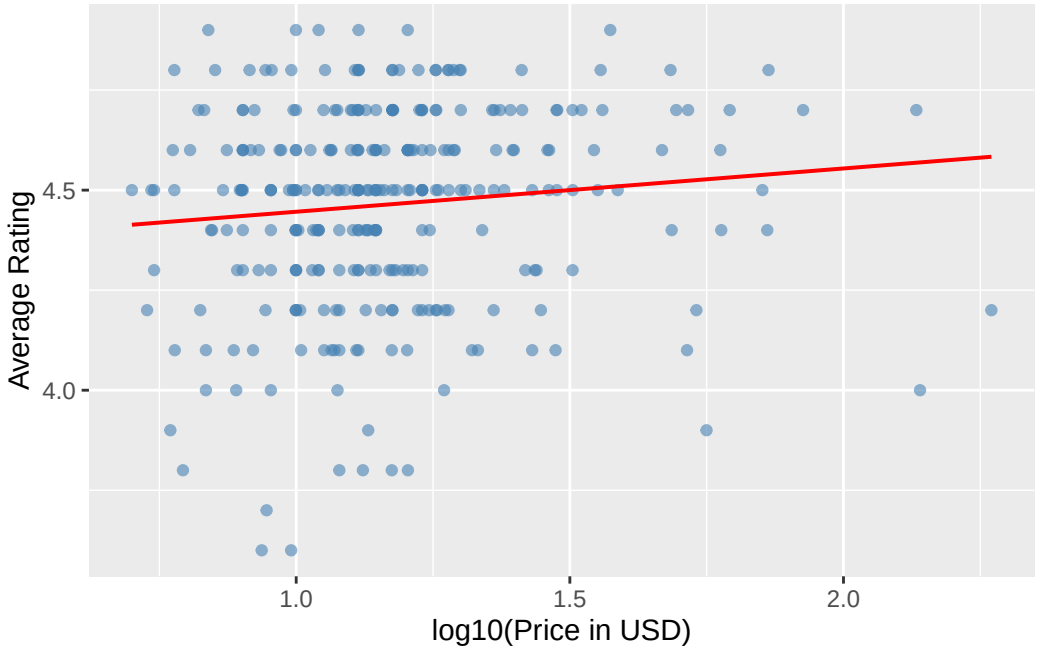

Figure 7: Scatterplot of log10(price) vs. average_rating with fitted regression line

The histogram of price shows a strong right-skew, with most books priced under $30 and only a few costly books extending the upper tail. This skewness also appears in the initial scatterplot of price versus average_rating, where the heavy clustering of low-priced books makes the overall trend difficult to see. After applying a log10 transformation to price, the scatterplot becomes more balanced, and the fitted regression line reveals a slight positive slope.

The regression output is consistent with these visuals: the slope for log10(price) is 0.108, indicating only a very small increase in rating as price increases. The R^2 value of 0.011 shows that price explains about 1% of the variation in ratings, confirming that the association between price and average_rating is extremely weak.

Tell

The regression results show that price has only a very small effect on a book's average rating. After transforming price to a log scale for a clearer pattern, the fitted line indicates that more expensive books tend to have slightly higher ratings, but the difference is tiny. For example, increasing the price by a factor of 10 (such as from $10 to $100) is associated with only about a 0.1 point increase in average rating on a 1–5 scale. In everyday terms, this difference is so small that most readers would not notice it.

The intercept represents the predicted rating for a book priced at $1, which is about 4.34. This matches the overall pattern that most books, regardless of price, tend to receive ratings between 4.0 and 5.0.

These findings only partly match my expectation. I expected more expensive books to receive noticeably higher ratings because of higher quality or production value. While the relationship is technically positive, the effect is extremely weak. The very low R^2 value (about 1%) shows that price explains almost none of the variation in ratings. In this dataset, readers generally give high ratings to books whether they are cheap or expensive, this likely reflects how Amazon ratings work: most reviews are written by readers who already enjoy the book's genre or author, and are therefore more inclined to rate positively.

3b. Checking model fit

Make use of all the tools described in the textbook to assess model fit in the `Think again` section - if it is necessary to revise your model, do it in the `Think again` section. Then state any updated conclusions in the `Revising conclusions` section.

Think again

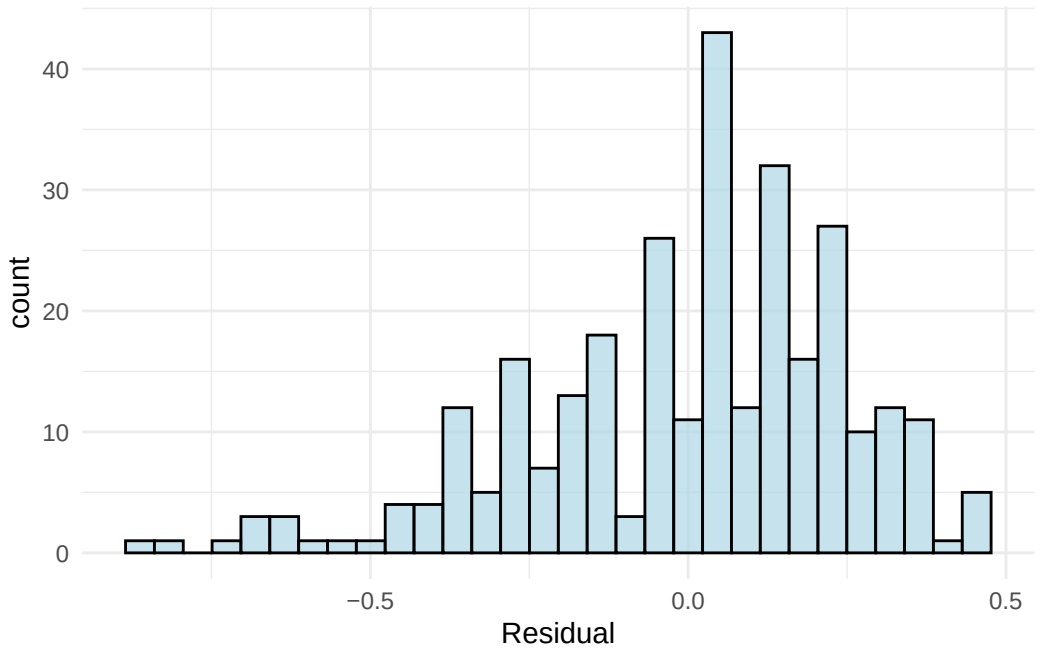

Figure 8: Histogram of residuals from the regression of average_rating on log10(price)

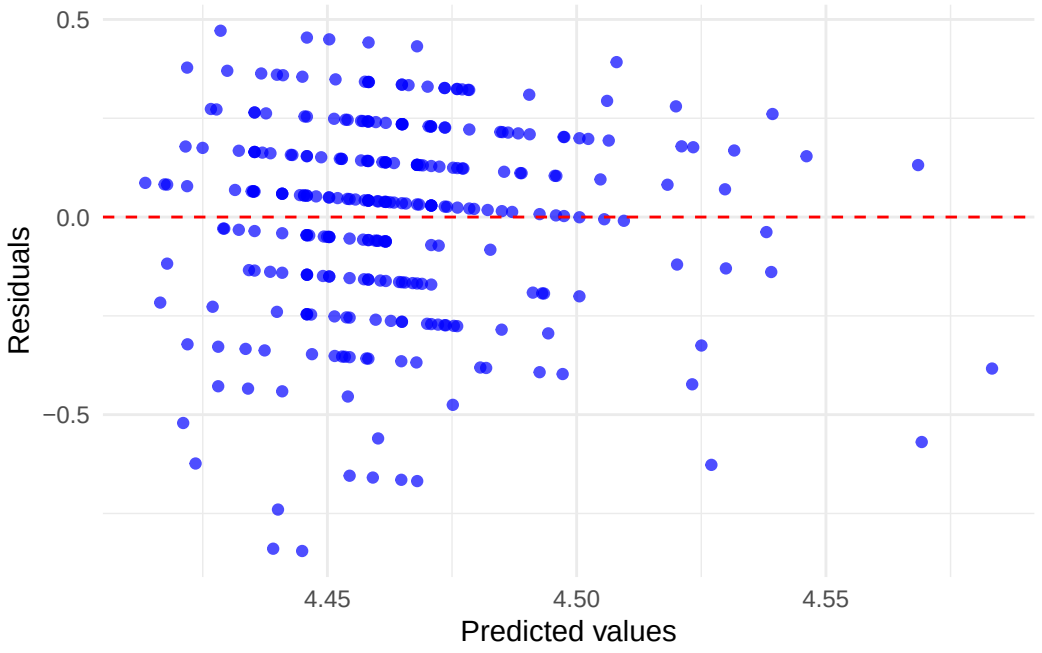

Figure 9: Scatterplot of Residuals vs. Fitted Values for the Price–Average Rating Model

The diagnostic plots do not reveal any major issues with the linear regression model. The histogram of residuals shows slight left-skewness, which is expected given that average ratings are constrained within the 1–5 range and tend to cluster near the upper end. The residuals-versus-fitted plot does not display noticeable curvature or systematic patterns, and the points remain reasonably centered around zero, with only mild variation in spread. Since there is no strong evidence of nonlinearity or other assumption violations, the simple linear model using log10(price) continues to be appropriate, and no model revision is necessary.

Revising conclusions

Based on the diagnostic checks, the overall interpretation of the model does not change. Even after accounting for the skewed price distribution and checking the residuals, the association between log(price) and average_rating remains extremely weak. The fitted line shows only a very small positive slope, and the model explains about 1% of the variation in ratings. This reinforces the earlier conclusion that price is not a meaningful predictor of how highly a book is rated in this dataset.

3c. Investigating `price` vs. `item_weight`

Similar to 3a. and 3b., fully analyze the relationship between price and the weight of the book.

Think

In this question, I treat item_weight as the explanatory variable and price as the response variable. Book weight reflects production characteristics such as page count, binding type, and material quality, which can be related to how expensive a book is to produce. Because heavier books often use more materials, I expect the regression model to show a positive association, meaning that higher item_weight is typically linked with higher price.

Both variables are quantitative, and I will use the cleaned `weight_grams` variable for analysis. Before examining the association, it is helpful to look at the distribution of `weight_grams`, since physical product weights can vary widely and are often strongly right-skewed. After reviewing the distribution, the most appropriate display for assessing the relationship is a scatterplot of price versus weight, potentially with a log transformation if skewness is severe. A fitted regression line and the correlation coefficient will then help quantify the direction and strength of the linear association.

Show

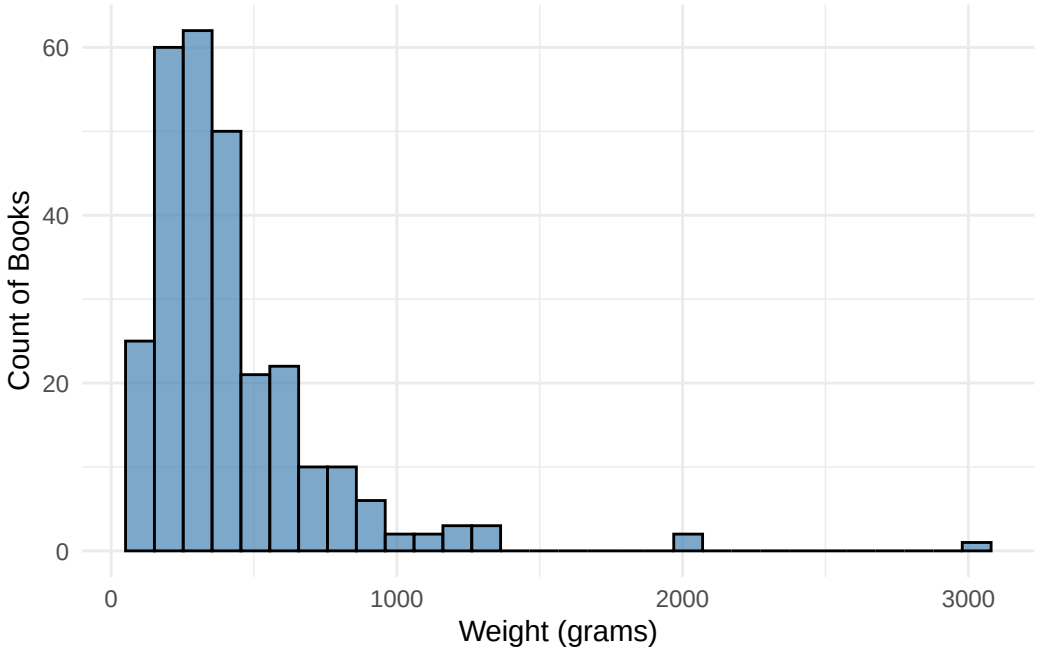

Figure 10: Histogram of Book Weights (in grams)

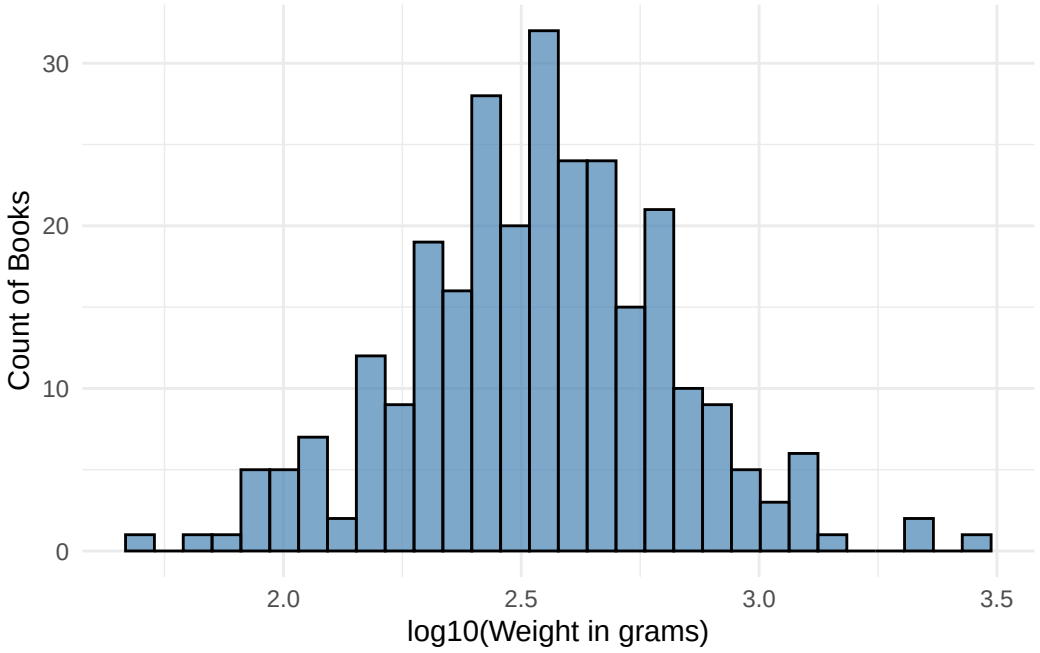

Figure 11: Histogram of log10(Book Weight in grams)

Table 5: Simple Regression Model: log10(price) ~ log10(weight_grams)

	log10(price) ~ log10(weight)
Intercept	0.440
	(0.134)
log10(Weight in grams)	0.289
	(0.053)
Num.Obs.	279
R2	0.098
R2 Adj.	0.094
AIC	1976.0
BIC	1986.9
Log.Lik.	0.177
F	29.991
RMSE	0.24

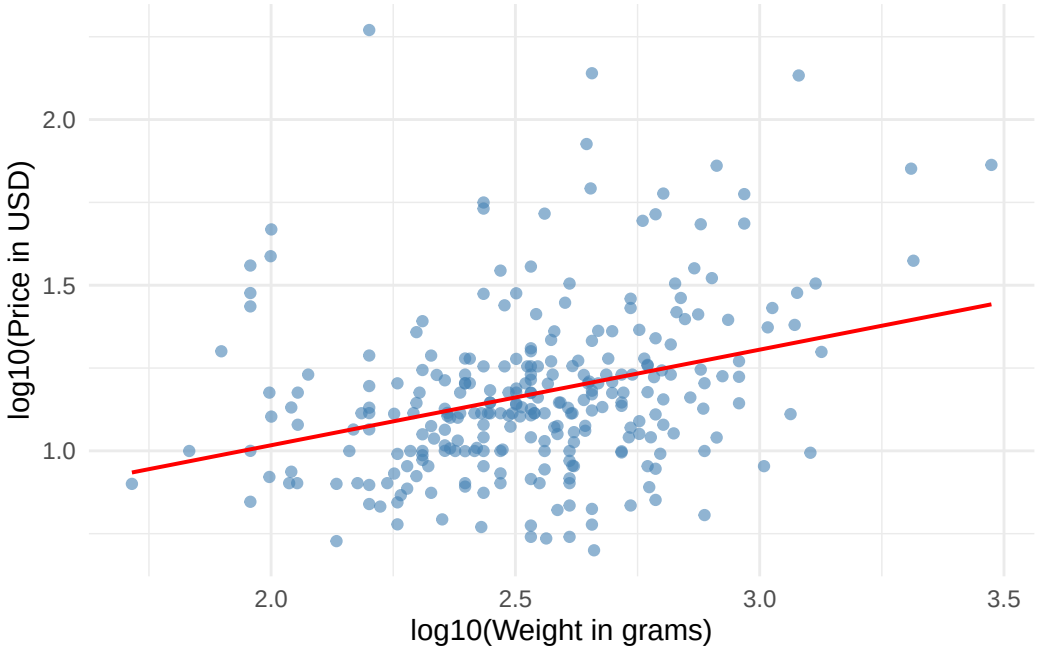

Figure 12: Scatterplot of log10(Price) vs. log10(Weight)

The histogram of weight_grams is strongly right-skewed, with a few unusually heavy books acting as clear outliers. After applying a log10 transformation, the distribution becomes much more symmetric and approximately bell-shaped, making it more appropriate for linear regression.

The log–log scatterplot of log10(weight) versus log10(price) reveals a clearer positive pattern. The log transformation reduces the influence of extreme values in both variables, allowing the upward trend to appear more consistently. The fitted regression line shows a visible positive association between the two log-transformed variables.

Tell

The regression results show a clear positive association between book weight and price on the log–log scale. The slope of 0.289 indicates that for every 10-fold increase in weight, the expected price increases by 0.289 log10 units, which corresponds to roughly a 1.94× increase in actual price (because $10^{0.289} \approx 1.94$). In practical terms, heavier books tend to be more expensive, consistent with the idea that larger or higher-material books cost more to produce.

The intercept of 0.440 represents the predicted log10(price) when weight is 1 gram. Converting this back gives a predicted price of about 2.75 USD ($10^{0.440} \approx 2.75$) hich aligns with the idea that extremely light books, such as thin pamphlets, tend to be very inexpensive.

The R^2 of 0.098 shows that weight explains about 10% of the variation in log10(price). This indicates that while weight is meaningfully related to price, many other features, such as genre, publication format, branding, and demand, also contribute to pricing.

Overall, the results support my expectation: heavier books tend to have higher prices. However, the modest R^2 suggests that the relationship is real but not strong, reflecting the many additional factors that influence book pricing beyond weight alone.

Think again

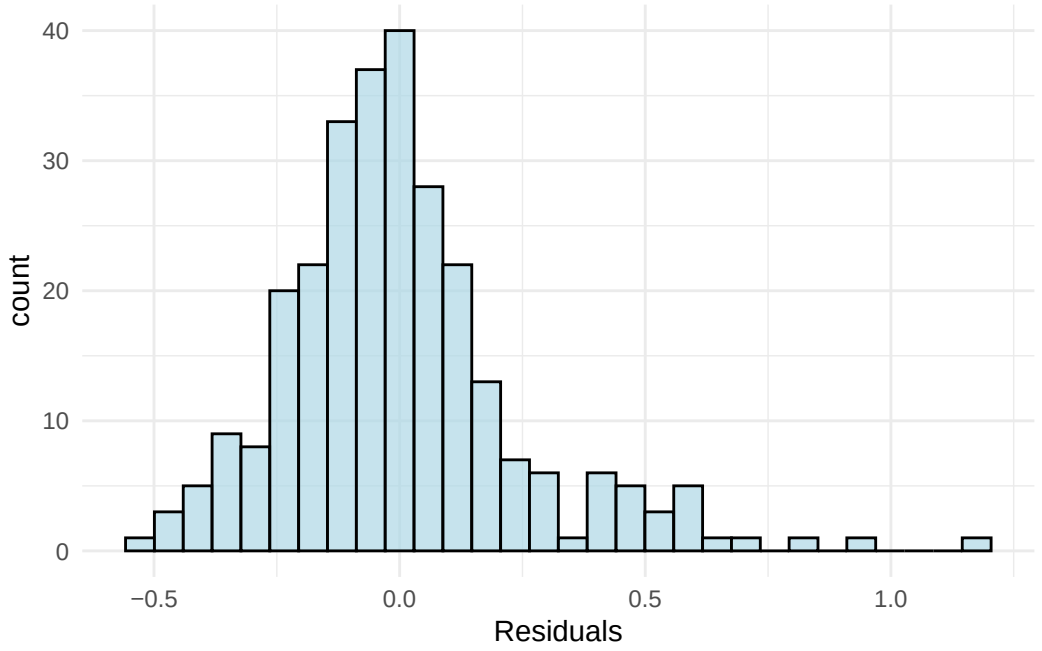

Figure 13: Histogram of residuals from the regression of log10(weight) on log10(price)

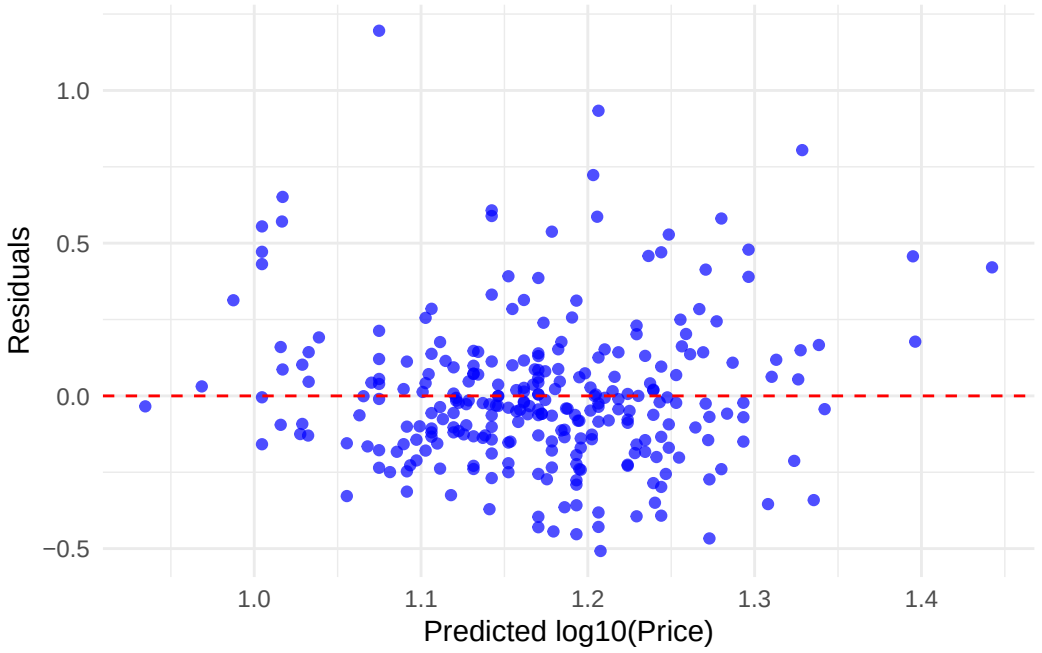

Figure 14: Residuals vs. Fitted Values for the log10(price) ~ log10(weight) Model

Revising conclusions

The diagnostic plots suggest that the regression assumptions are reasonably satisfied. The residual histogram is mostly symmetric with only a few larger positive values, and the residuals–fitted plot does not show clear curvature or changing spread. While a few points have larger residuals, they do not indicate any systematic problem. Therefore, no revision to the model is needed.

The main conclusion also stays the same: log10(weight) and log10(price) show a positive but moderate association. With an R^2 of about 0.10, weight accounts for roughly 10% of the variation in book prices—meaning heavier books tend to cost more, but many other factors influence price as well.

3d. Thinking about your results

What can we learn about how price is determined in these two investigations? Do the results surprise you? What lurking variables do you think could be at work here, if any?

Answers will vary here, good quality effort to interpret investigation of this question is required

Question 4: Multiple regression (30 points)

4a. Investigating `average_rating` vs. `price`, `rating_number`, and `age`

Using the Think-Show-Tell framework from the textbook, please examine, using a multiple regression model, how `average_rating` relate to `price` and `rating_number`. Make use of all the tools described in the textbook to assess model fit in the `Think again` section - if it is necessary to revise your model, do it in the `Think again` section. Then state any updated conclusions in the `Revising conclusions` section.

Think

The variable *average_rating* is the mean reader rating on a 1–5 scale, and I will treat it as the response variable. The predictors in this multiple regression model are *price* and *rating_number*. All three variables are quantitative.

Based on earlier results, I expect both log10(price) have only small positive associations with average_rating, and I expect age and log10(rating_number) to have little negative to no relationship. Since Amazon ratings are consistently high and tightly clustered, the overall model is likely to explain only a small portion of the variation in average ratings.

From earlier analyses, both price and rating_number are strongly right-skewed, so I will continue using their log10-transformed versions in the multiple regression model. In contrast, age has not yet been examined in this context, so it is helpful to first look at the distribution of age to determine whether a transformation is needed. Then, I will compute a correlation matrix for average_rating, log10(price), log10(rating_number), and age to evaluate their linear relationships. After reviewing these preliminary checks, I will fit the multiple regression model and examine the regression summary to assess the direction and strength of each predictor's association with average_rating.

Show

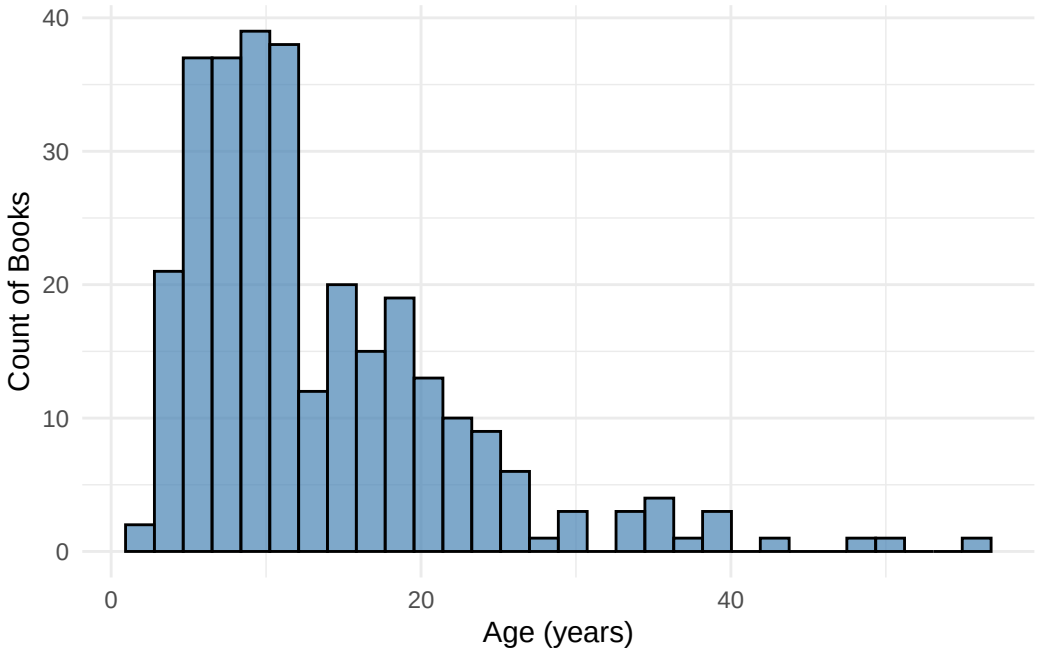

Figure 15: Histogram of book age

Figure 16: Correlation plot of average_rating, log10(price), log10(rating_number), and log10(age)

The histogram of age shows a clear right-skewed distribution, with most books published within the last 5–20 years and a small number of much older titles. Because of this skewness, applying a log10 transformation to age is appropriate before fitting the multiple regression model.

The correlation heatmap indicates that all three predictors have very weak correlations with average_rating. The predictors are also only weakly correlated with one another, suggesting no concerns about multicollinearity. Given these patterns, using log-transformed predictors together in a multiple regression model is reasonable.

Tell

The multiple regression model shows that log10(price), log10(rating_number), and log10(age) together explain only a very small share of the variation in average_rating. All three coefficients are close to zero, and none of the predictors show a strong association once they are included in the model simultaneously. This fits the overall pattern seen throughout the dataset: Amazon book ratings are highly concentrated between 4.0 and 5.0, leaving little room for predictors to explain meaningful differences.

Price remains weakly positive, while both rating_number and age show slight negative associations which align with my expectation, but their magnitudes are

Table 6: Multiple Regression Model on average_rating

	average_rating ~ log10(price) + log10(rating_number) + log10(age)
Intercept	4.505
	(0.129)
log10(Price in USD)	0.106
	(0.058)
log10(Number of Ratings)	−0.042
	(0.029)
log10(Age in years)	−0.038
	(0.054)
Num.Obs.	297
R2	0.020
R2 Adj.	0.010
AIC	27.7
BIC	46.1
Log.Lik.	−8.825
F	1.989
RMSE	0.25

very small. Overall, the model confirms that average ratings on Amazon do not vary much with price, popularity, or age, consistent with the idea that most books receive uniformly high evaluations regardless of these characteristics.

Think again

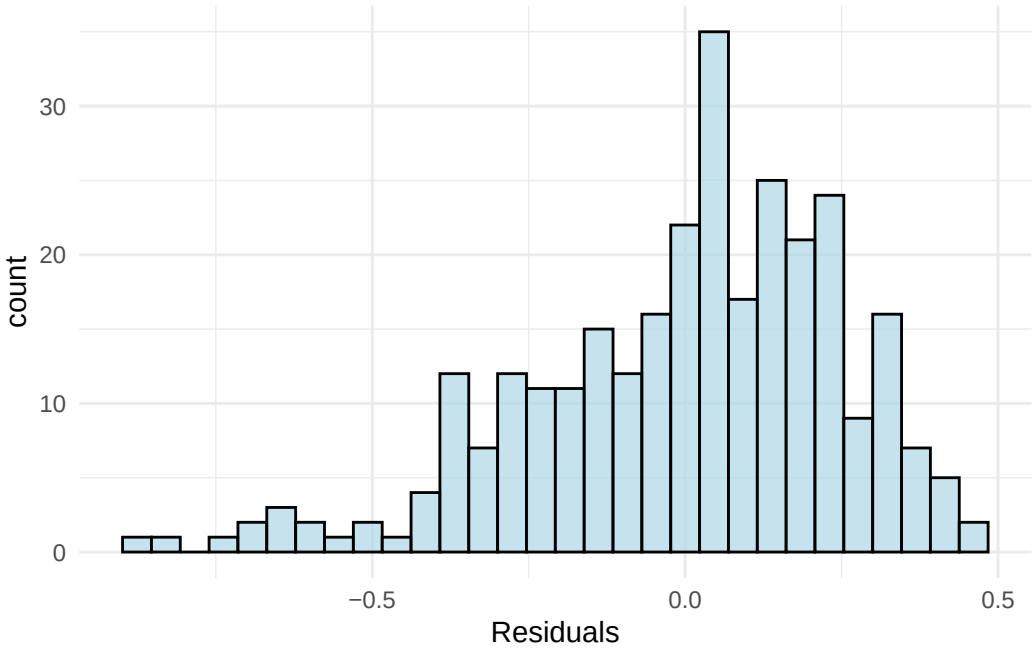

Figure 17: Histogram of residuals from the multiple regression of average_rating

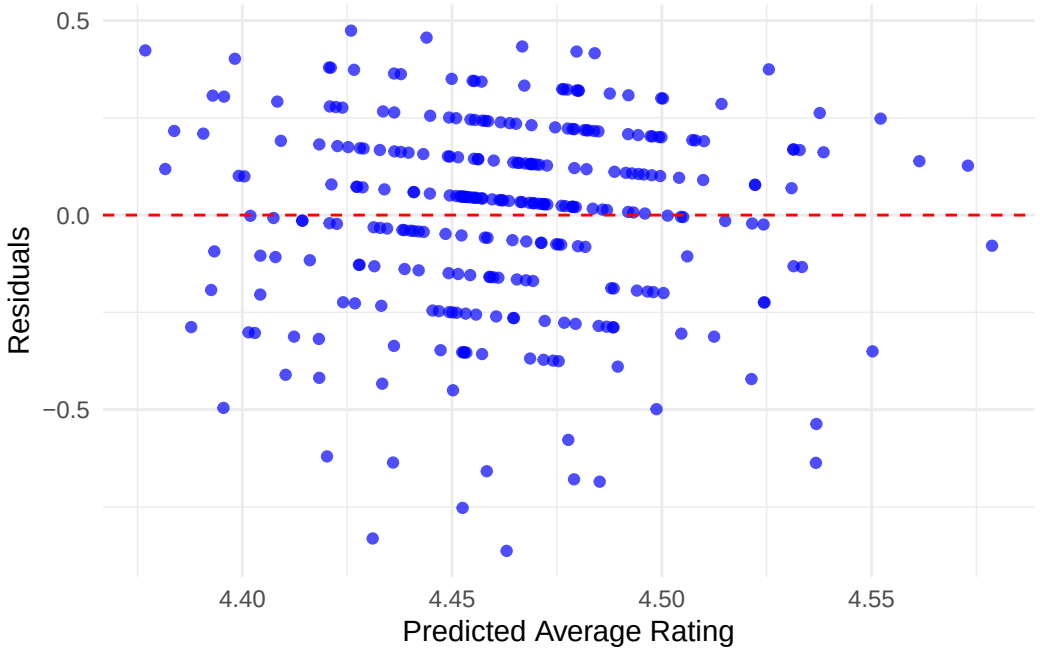

Figure 18: Residuals vs. Fitted Values for the average_rating

The diagnostic plots do not reveal any major problems with the multiple regression model. The histogram of residuals is roughly centered around zero, showing a mild left-skew but no extreme outliers or heavy tails. This shape is reasonable given that average ratings are tightly clustered between 4 and 5, leaving limited room for symmetric residual behavior.

In the residuals-versus-fitted plot, the points appear widely scattered with no strong curvature or obvious pattern. The spread of the residuals is fairly constant across fitted values, and the points are centered around zero. This suggests that the linearity and constant-variance assumptions are reasonably met.

Revising conclusions

The diagnostic plots do not show major violations of the regression assumptions, so there is no need to change the model. The main conclusion remains that log10(price), log10(rating_number), and log10(age) have only very small effects on average_rating, and together they explain only about 2% of its variation.

In the context of Amazon books, this makes sense: ratings are heavily "ceilinged" around 4–5 stars, and reviews are usually written by readers who already like the book, author, or genre. As a result, differences in price, popularity, or publication age do not translate into large differences in the average star rating. The model therefore suggests that, in this dataset, average_rating is more a reflection of

Table 7: Coefficient Estimates for the Multiple Regression Model

term	estimate	std.error	statistic	p.value
(Intercept)	4.5046578	0.1287671	34.9829776	0.0000000
log10(price)	0.1060536	0.0584260	1.8151793	0.0705187
log10(rating_number)	-0.0422904	0.0287909	-1.4688800	0.1429381
log10(age)	-0.0375537	0.0535753	-0.7009521	0.4838892

Table 8: Model Summary Statistics for the Multiple Regression Model

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	devianc
0.019963	0.0099285	0.2509644	1.989434	0.1156495	3	-8.825287	27.65057	46.11923	18.4540

general reader enthusiasm than of measurable book characteristics like price, rating count, or age.

4b. Interpreting coefficients of 4a. model

Carefully interpret your coefficients from 4a. What do they mean? Are there any lurking variables here?

Think

For this part, I will focus on interpreting the coefficients from the multiple regression model. I will using `tidy()` and a `glance()` to summarize each predictor's estimate and summarize model performance . The goal is to interpret each slope as the expected change in average_rating associated with a 1-unit increase in the corresponding log10 predictor, holding the other predictors constant. I will also think about possible lurking variables that are not in the model.

Show

The intercept of 4.50 represents the predicted average rating for a book when all predictors are at 1 on the log scale, that is, when price = 1 USD, rating_number = 1, and age = 1 year. This value is consistent with the overall pattern that most books on Amazon tend to have ratings around 4–5 stars.

The coefficient for log10(price) is 0.106, meaning that a tenfold increase in price (for example, from $10 to $100) is associated with only a 0.11-point increase in predicted average rating on the 1–5 scale, holding other variables constant. This effect is small.

The coefficient for log10(rating_number) is –0.042, meaning that a tenfold increase in the number of ratings (e.g., from 10 to 100, or 100 to 1000) is associated with a 0.04-point decrease

in predicted average rating, holding other variables constant. The effect is very small and not practically meaningful.

The coefficient for log10(age) is –0.038, meaning that a tenfold increase in age (e.g., from 1 year to 10 years old, or from 2 to 20 years old) is associated with a 0.04-point decrease in predicted average rating, holding other variables constant. This effect is also very small.

The model's R^2 is 0.020, meaning that only about 2% of the variation in average_rating is explained by price, rating_number, and age together. This confirms that the predictors have very limited explanatory power, and most differences in ratings are driven by factors not captured in the model.

> Tell
>
> The multiple regression results show that all three predictors have only very small associations with average_rating. Each slope is close to zero, and even a tenfold change in any predictor leads to only minimal changes in the predicted rating. The R^2 value is about 2%, indicating that the model explains almost none of the variation in average ratings. Overall, the fitted model suggests that Amazon book ratings remain highly consistent and are only weakly related to price, popularity, or age.
>
> Because the predictors in the model explain so little, it is likely that important factors influencing average_rating are not included. Possible lurking variables include book genre, author reputation, marketing exposure, editorial quality, or reader selection effects (e.g., fans of a genre being more likely to leave positive reviews). These omitted factors may play a much larger role in shaping ratings than price, rating_count, or age.

4c. Add the variable `category`

Now add the variable `category` to your model and analyze the relationship similar to what you did in 4a.

> Think
>
> In this part, I add **`category`**, a categorical variable representing each book's subject area, to the multiple regression model for predicting average_rating. Since category reflects meaningful differences across genres (e.g., children's books, fiction, technical manuals, and textbooks), I expect category to contribute additional explanatory power. Different genres often receive systematically different types of reviews: some categories may attract more enthusiastic readers, while others (such as technical or academic books) may receive more moderate ratings. Therefore, I expect category to shift the predicted rating by different amounts depending on the group.

Before fitting the model, it is helpful to explore the distribution of average_rating across categories using a boxplot, since this allows visual comparison of differences in central tendency and variability across groups. After that, I will extend the previous regression model by adding category as a factor predictor and examine the regression output. The goal is to see whether category meaningfully improves the model's explanatory power, whether certain categories have consistently higher or lower ratings, and how the inclusion of this categorical variable changes the interpretation of the model.

Show

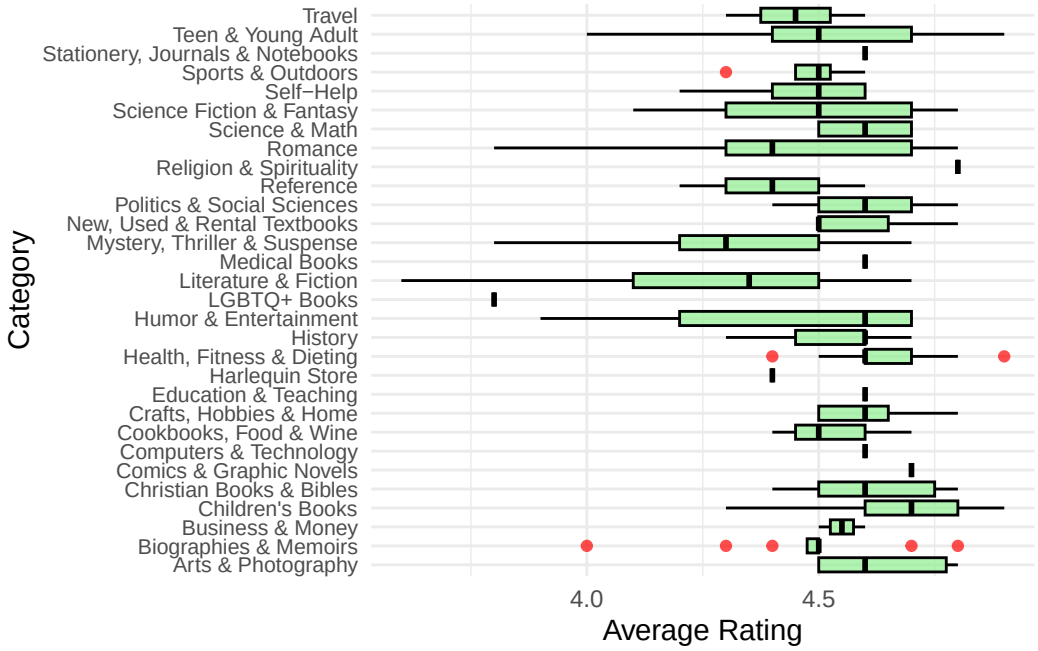

Figure 19: Boxplot of Average Rating by Category

The boxplot of average_rating across categories shows visible variation among genres, though the overall rating range remains narrow (mostly between 4.0 and 4.8). Several categories, such as Children's Books, Science & Math, and Self-Help, cluster around higher median ratings, while categories like Biographies & Memoirs, History, and Literature & Fiction show slightly lower central values or wider spreads. Some categories display very thin boxes or appear as a single horizontal line, indicating that only a few books fall into those groups (e.g., Harlequin Store, Stationery, Journals & Notebooks, Medical Books).

When category is added to the regression model, many of the estimated category coefficients differ modestly from the reference group, consistent with the visual differences in the boxplot. No category shows an extremely large effect, but several genres with lower medians (e.g.,

Literature & Fiction, Mystery, Thriller & Suspense, Humor & Entertainment) have negative coefficients. The model's R^2 increases to about 0.35, indicating that category explains a substantial share of the variation in average ratings compared to models that only included numerical predictors.

Tell

The results generally align with my expectations. Adding category substantially improves the model's explanatory power, raising R^2 from about 0.02 to roughly 0.35, which suggests that genre meaningfully helps explain differences in average ratings. Several categories show noticeable shifts relative to the reference group, consistent with the idea that different types of books attract different audiences and reviewing behaviors.

While many coefficients are modest, categories with consistently lower medians in the boxplot (such as Literature & Fiction or Mystery, Thriller & Suspense) also receive negative estimates in the model, reinforcing the visual patterns. At the same time, some coefficients are difficult to interpret because several genres have very small sample sizes, which is evident from the very narrow or single-line boxplots. Overall, including category confirms that genre plays a much larger role in rating patterns than price, popularity, or age alone. However, the large number of sparse categories makes the full model hard to explain clearly. In the Think again section, I will consider combining low-count genres into an "Other" group and examine how this simplification changes the model fit and interpretation.

Think again

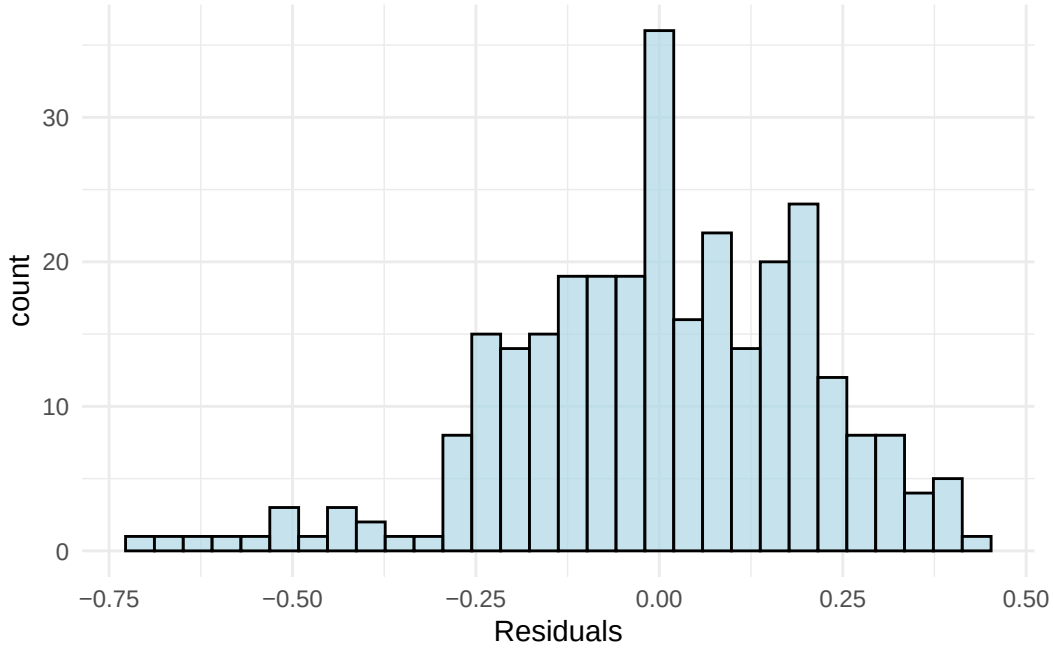

Figure 20: Histogram of residuals from the model with category

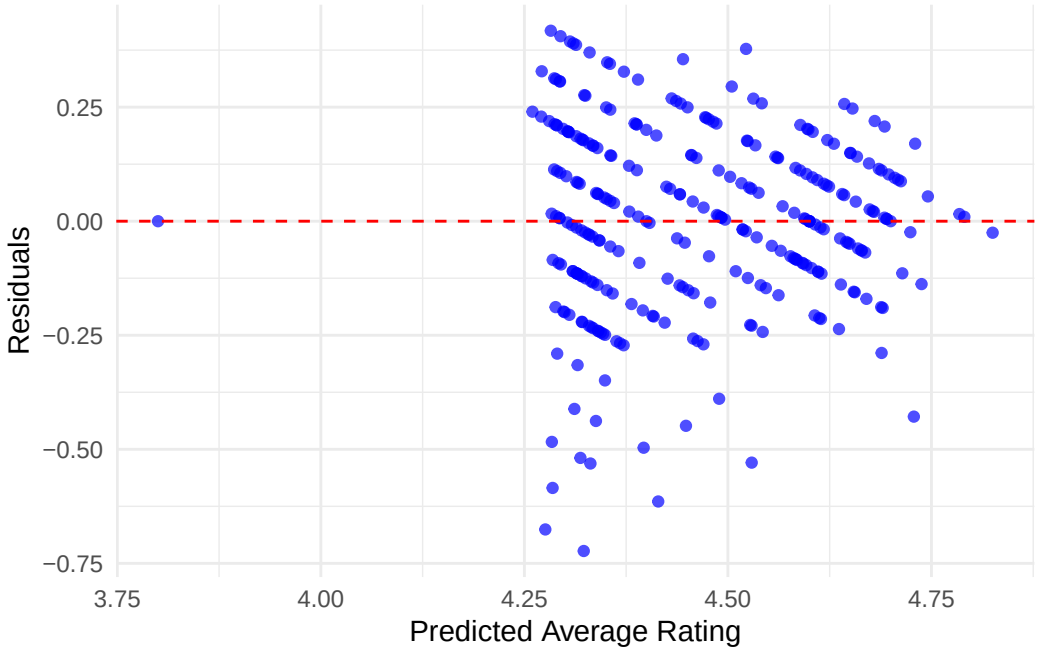

Figure 21: Residuals vs. Fitted Values for the model with category

The residual histogram is roughly centered near zero but remains somewhat right-skewed, reflecting that average ratings are bounded above and clustered at the high end of the 1–5 scale. The residuals-versus-fitted plot shows a noticeable banding pattern because many categories share similar fitted values, but there is no strong curvature or systematic increase in spread. Overall, the diagnostics do not indicate major violations of linear regression assumptions, though the pattern in residuals suggests that the categorical structure, and the compressed scale of ratings, limits how well a linear model can capture the variation.

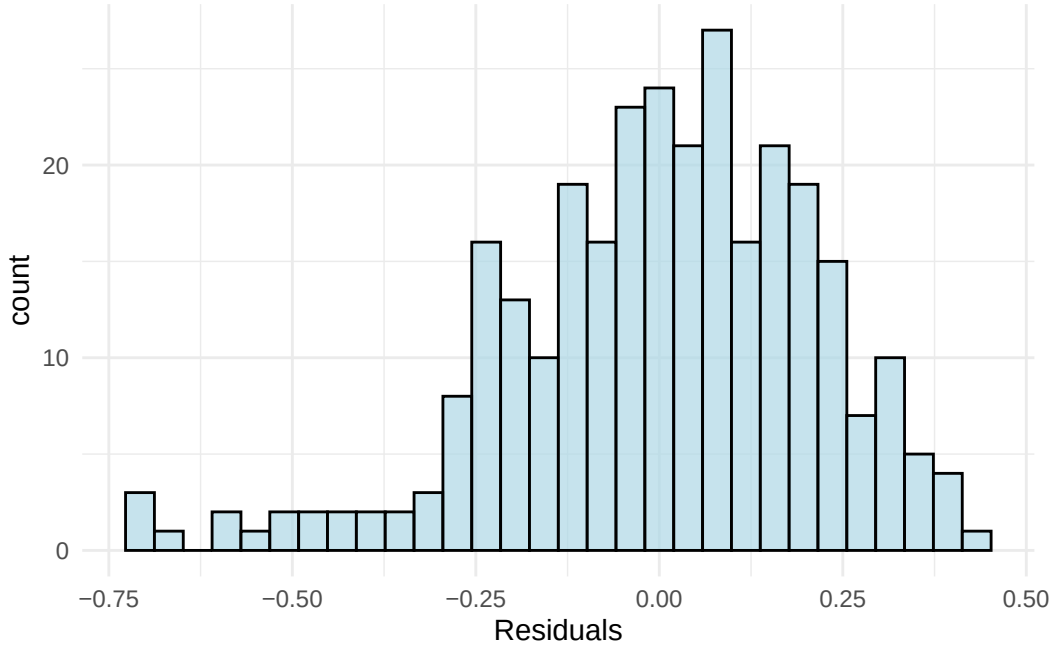

Figure 22: Histogram of residuals from the simplified-category multiple regression model

Table 9: Summary Statistics for Each Category

category	n	mean_rating	sd_rating
Comics & Graphic Novels	1	4.700	NA
Computers & Technology	1	4.600	NA
Education & Teaching	1	4.600	NA
Harlequin Store	1	4.400	NA
LGBTQ+ Books	1	3.800	NA
Medical Books	1	4.600	NA
Stationery, Journals & Notebooks	1	4.600	NA
	2	4.400	0.566
Business & Money	2	4.550	0.071
Travel	2	4.450	0.212
Cookbooks, Food & Wine	3	4.533	0.153
New, Used & Rental Textbooks	3	4.600	0.173
Reference	3	4.400	0.200
Religion & Spirituality	4	4.800	0.000
Sports & Outdoors	4	4.475	0.126
Humor & Entertainment	5	4.420	0.356
Science & Math	5	4.600	0.100
Self-Help	5	4.460	0.167
Arts & Photography	6	4.633	0.151
Crafts, Hobbies & Home	7	4.600	0.115
Health, Fitness & Dieting	9	4.633	0.150
Romance	9	4.433	0.300
Christian Books & Bibles	11	4.627	0.142
History	11	4.527	0.142
Biographies & Memoirs	12	4.475	0.196
Politics & Social Sciences	12	4.600	0.141
Teen & Young Adult	12	4.533	0.246
Science Fiction & Fantasy	14	4.479	0.242
Children's Books	28	4.693	0.151
Mystery, Thriller & Suspense	44	4.323	0.201
Literature & Fiction	80	4.326	0.264

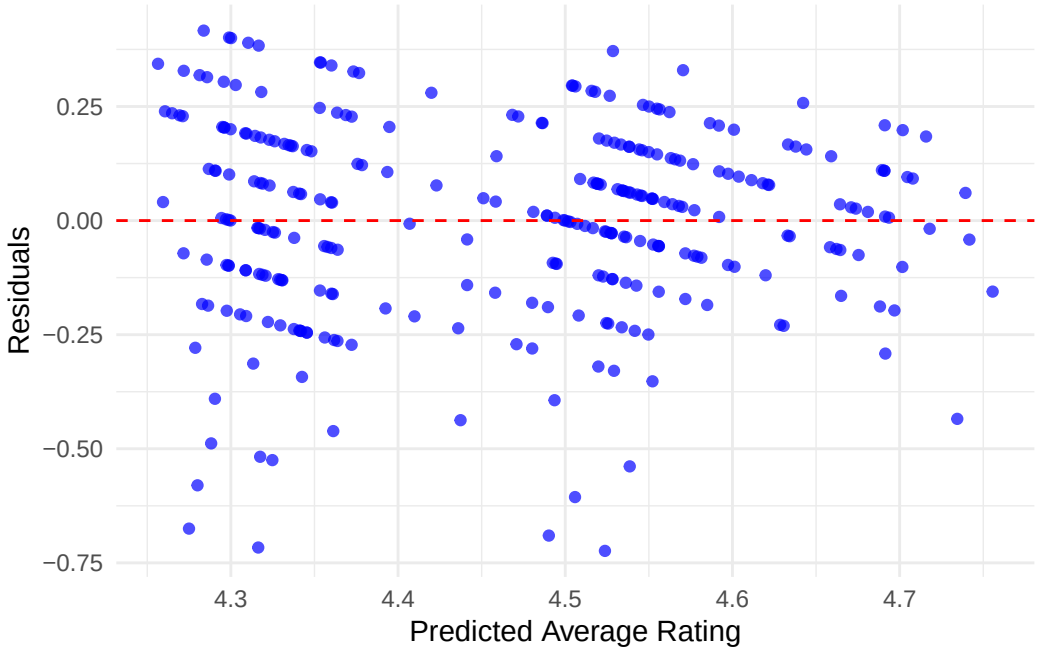

Figure 23: Residuals vs. Fitted values for the simplified-category model

Revising conclusions

The residual diagnostics for the original model (using all categories) suggest that its assumptions are not fully met. The histogram of residuals, while roughly unimodal, shows a left tail of large negative outliers. This observation directly matches the residual-versus-fitted plot, which shows one prominent outlier. This point acts as a significant influential observation, likely driven by a rare category with an unusually low rating. The residual–fitted pattern also has a faint "striped" structure, a common side-effect of a categorical predictor with many levels dominating the prediction. Overall, this outlier violation indicates that the original model's coefficients are likely unstable.

Comparing the original and recoded category models reveals a crucial trade-off between model fit and model validity. The original model's higher R^2 (0.35) appears to be artificially inflated by its overfitting to that single influential outlier. The recoded model, which collapses rare categories into a single "Other" group, successfully resolves this violation. Its residual-versus-fitted plot is much robust, showing no severe outliers and a more evenly distributed residual cloud. While this simplification lowers the adjusted R^2 to 0.250, this is an acceptable trade-off for gaining a more robust, stable, and generalizable model. Retaining the original model would mean accepting coefficients that are skewed by a single, rare group. Therefore, the simplified model is a better choice, as its results provide a more

Table 10: Coefficient Estimates for the Multiple Regression Model (With Category)

term	estimate	std.error	statistic	p.value
(Intercept)	4.2686649	0.1395032	30.5990477	0.0000000
log10(price)	0.0998897	0.0537658	1.8578654	0.0642303
log10(rating_number)	0.0378069	0.0270092	1.3997765	0.1626794
log10(age)	-0.0241742	0.0477784	-0.5059656	0.6132761
category_recodeChildren's Books	0.2321989	0.0759769	3.0561786	0.0024566
category_recodeChristian Books & Bibles	0.1558718	0.0916780	1.7002085	0.0901942
category_recodeHistory	0.0577710	0.0916686	0.6302157	0.5290638
category_recodeLiterature & Fiction	-0.1532333	0.0677546	-2.2615921	0.0244842
category_recodeMystery, Thriller & Suspense	-0.1473544	0.0715579	-2.0592317	0.0403901
category_recodeOther	0.0690998	0.0694522	0.9949253	0.3206255
category_recodePolitics & Social Sciences	0.1219343	0.0891934	1.3670767	0.1726900
category_recodeScience Fiction & Fantasy	0.0017409	0.0865693	0.0201104	0.9839695
category_recodeTeen & Young Adult	0.0796155	0.0912854	0.8721604	0.3838629

trustworthy and valid interpretation of the predictors.

4d. Reinterpret your coefficients

Carefully re-interpret your coefficients from 4c and compare them to 4b. What do they mean? Any new lurking variables to consider?

Think

In this section, I will reinterpret the coefficients from the expanded multiple regression model, which adds the `category_recode` variable to the numeric predictors. Using the `tidy()` and `glance()` summaries, I will compare this model's estimates to those from the previous multiple regression model.

The primary goal is to analyze how the coefficients for `log10(price)`, `log10(rating_number)`, and `log10(age)` have changed now that we are controlling for book genre. Additionally, I will interpret the `category_recode` coefficients themselves, which represent the predicted shift in `average_rating` for each genre relative to the baseline category. Finally, I will consider any important lurking variables that might still be missing from the model, even after accounting for genre.

Show

Table 11: Model Summary Statistics for the Multiple Regression Model

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance
0.2802071	0.2495776	0.2173138	9.148279	0	12	38.35246	-48.70492	2.912737	13.31754

The intercept of 4.27 represents the predicted average rating for a book in the "Biographies & Memoirs" category (the baseline) when its price is $1, it has 1 rating, and it is 1 year post-publication. This is lower than the original model's intercept (4.50), which represented the average for all books. This is expected, as the new intercept is now specific to one category.

The coefficient for log10(price) is 0.10, meaning a tenfold increase in price (e.g., from $10 to $100) is associated with a 0.10-point increase in predicted average rating, holding other variables constant. This effect is small and very similar to the original model's coefficient (0.106), suggesting that controlling for category did not meaningfully change the small, positive association between price and rating.

The coefficient for log10(rating_number) is now 0.038, meaning a tenfold increase in the number of ratings (e.g., 100 to 1000) is associated with a 0.04-point increase in predicted average rating, holding other variables constant. This is the most notable change, as this coefficient was negative (–0.042) in the previous model. This flip suggests the original model was suffering from omitted variable bias; the negative effect was likely an artifact of not accounting for category. In both models, however, the effect is very small and not statistically significant.

The coefficient for log10(age) is –0.024, meaning a tenfold increase in age (e.g., 1 to 10 years post-publication) is associated with a 0.02-point decrease in predicted average rating, holding other variables constant. This effect is even smaller than in the original model (–0.038) and remains negligible.

The category_recode coefficients show the predicted shift in rating compared to the "Biographies & Memoirs" baseline:

1. "Children's Books" has a coefficient of 0.232. This means, holding all other variables constant, Children's Books are predicted to have an average rating that is 0.232 points higher than Biographies & Memoirs.
2. "Literature & Fiction" has a coefficient of –0.153. This means its predicted rating is 0.153 points lower than the Biographies & Memoirs baseline, all else being equal.
3. "Science Fiction & Fantasy" has a coefficient of 0.0017 (p=0.98), indicating its predicted rating is virtually identical to the baseline, there is no statistically significant difference between these two genres.

Finally, the model's R-squared is 0.280, meaning that about 28% of the variation in `average_rating` is explained by all predictors together. This is a massive improvement from the original model's R-squared (0.020). This confirms that `category` has substantial

explanatory power, and the original model was missing the single most important predictor of ratings.

> Tell
>
> The multiple regression results, after adding `category` as a predictor, show a massive improvement in explanatory power. The model's R-squared jumped from 2% to 28%, confirming that book genre, a suspected lurking variable in the previous model, is a significant predictor of average ratings.
>
> Despite this improvement, the other predictors remain weak. After controlling for category, the effects of `log10(price)`, `log10(rating_number)`, and `log10(age)` are still negligible and mostly not statistically significant. This suggests that once genre is known, a book's price, popularity, or age offer very little additional information about its rating.
>
> Because the model still leaves 72% of the variation unexplained, it is clear that other important factors are missing. Possible lurking variables could now include author reputation, marketing exposure, editorial quality, or whether the book is part of a popular series. These omitted factors may play a much larger role than the numeric predictors and even add nuance to the broad category effects.

4e. Thinking about your results

Consider the results of 4a.-4d. together. What can we learn about the ratings of books on Amazon? How did your conclusions change from 3d.? Why do you think they changed?

> Answers will vary here, good quality effort to interpret investigation of this question is required

Question 5: Your own investigation (20 points)

5a. Selecting your own question

Develop your own model of `average_rating`. Use the Think-Show-Tell procedure to conduct your investigation. Think deeply about what your result means and interpret your coefficients carefully.

> Think
>
> Show
>
> Tell

Think again

Revising conclusions

Answers will vary here, good quality effort to interpret investigation of this question is required

5b. In summary

Sum up everything that you have learned from questions 1-5. Do not simply repeat/rephrase your previous results but try to say something larger that synthesizes the results together to draw a more meaningful general conclusion.

Need to think deeply about what information this dataset provides for full points