
0.1 Question 3d

We analyzed the data, but something seems odd. Upon closer look, there are many negative values for `profit`. For example, the movie `102 Dalmatians` looks to have lost around $18M, but it was a widely successful film! What may account for this issue? Think about how we constrained our data from the start of the problem.

The dataset may not include all sources of revenue. For example, international box office, home video sales, streaming deals, merchandise, etc. If we only have domestic (USA) revenue, some movies may appear unprofitable even if they made significant money worldwide. Some movies have reported budgets that include marketing costs, while others only report production budgets. This can make some movies seem more expensive than they were. The query that created movie_gross might have filtered out certain types of revenue data, leading to an underestimation of total earnings.

0.2 Question 4b

What do you notice about the summary values generated in `earnings_summary`? We can represent the five-number summary graphically using a box plot. Identify two properties about the boxplot of the data. (You do not need to explicitly create a boxplot, but think about how the summary statistics would be distributed in a boxplot.)

Hint: Think in terms of about concepts from statistics like spread, modality, skew, etc. and how they may apply here.

Type your answer here, replacing this text.

The difference between the median (2.3M) and the 75th percentile (20M) is much larger than the difference between the 25th percentile (166K) and the median.