Movie Statistics Project

Explore Netflix data with your new understanding of summary statistics!

Overview

In this project, you'll practice using summary statistics from real data.

You will:

- interpret summary statistics in context
- make choices about which summary statistics are appropriate based on the situation
- answer free-response questions and compare your answers to a sample solution available after you submit your answer
- answer toggle questions that can be clicked on to toggle an explanation or solution
- explore some really interesting data!

As an added bonus, some toggle prompts are labeled *"Just for Fun"* and don't have to be answered. These prompts provide extra interesting details about the data for your enjoyment!

Motivation

You just got a very cool job in the film industry! Your first assignment is to do some research on the content produced by streaming services in the last few years. You're putting together information for a report on the kinds of films being produced as well as any patterns that might be worth exploring further.

Dataset

You decide to start your research by exploring some data about films and documentaries produced by Netflix. The dataset you'll be using is a modified version of one found on the website Kaggle. Your dataset includes 503 films with the following variables:

- title: title of the film
- genre: genre of the film
- language: primary language of the film
- year: year the film premiered
- runtime: length of the film in minutes
- score: film rating of 1 to 10 (worst to best) from the website IMDb

Individual Variables

Language

You decide to start with the language variable. The table that follows gives the count of films in each language.

Using the table, try answering the following individual questions. Then use your answers to write up a brief summary about the primary languages of the films.

▼ There are clearly a lot of films that have English as their primary language. Of the 503 films, what proportion have English as their primary language? (Click to Toggle Correct Answer)

There are 360 films with English as the primary language. To get the proportion, we divide 360 by the total of 503 films: 360 ÷ 503 = 0.72. About 0.72 of the films have English as their primary language.

▼ What is the ratio of English-language films to films in a single language that is NOT English? (Click to Toggle Correct Answer)

We know there are 360 English-language films, but we have to do a little work to find the number of other single-language films. We can add the four categories for Spanish, Hindi, French, and "other single" (29 + 27 + 15 + 51 = 122). Or we can subtract the English and "multiple" categories from the total (503 - 360 - 21 = 122). This means the ratio of English films to non-English single-language films is 360 to 122. Since 360 ÷ 122 is 2.95, this means there are almost 3 English films for every non-English film.

▼ What proportion of the films has multiple primary languages? (Click to Toggle Correct Answer)

There are 21 films that have multiple primary languages. Dividing 21 by 503 gives a proportion of 0.04.

Free response

Based on the questions you've just answered, write a few sentences summarizing what you found out about film languages. Give all proportions as percentages to help readers of your research report better understand your findings.

Your response

From the films and documentaries produced by Netflix, we found that 72% of the films have English as their primary language, for each movie in English, there are almost 3 movies in other languages, and about 4% of the films and documentatires have multiple languages.

Our answer

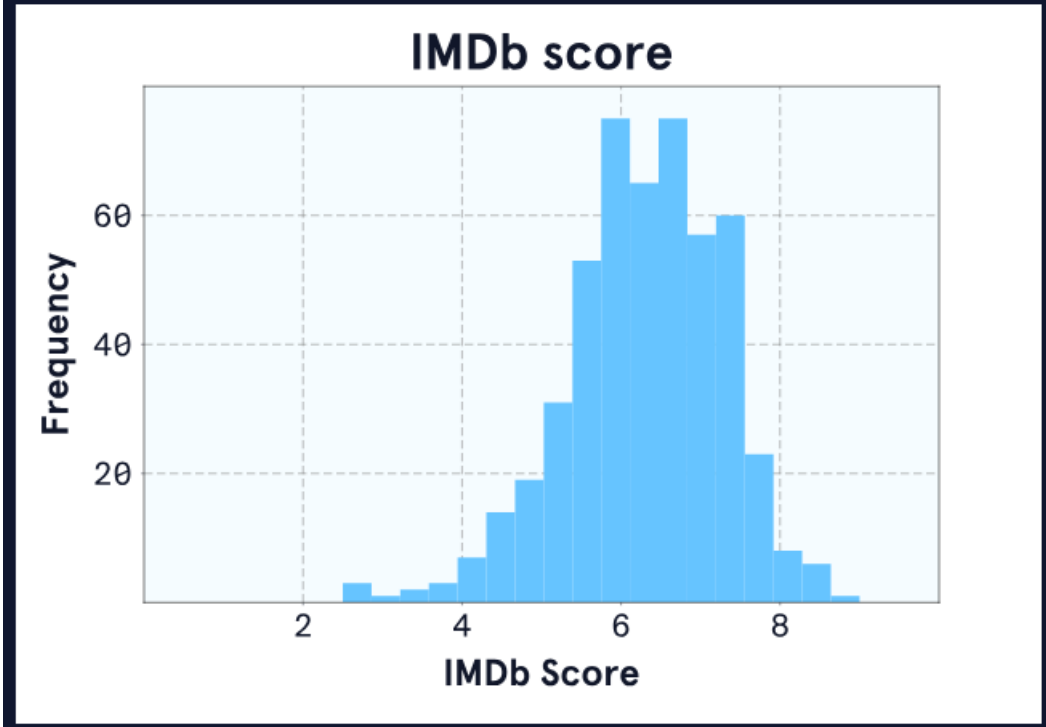
The majority of the Netflix films in the dataset have English as their primary language. With 360 English-language films out of a total of 503, this is about 72% of the films. Further, there are about 3 English-language films for every non-English single-language film. Only about 4% of films have multiple primary languages.

▼ *Just for Fun:* About 4% of the films have multiple primary languages. Find out which genres tend to have this interesting feature. (Click to Toggle the Information)

Of the 21 films with multiple primary languages, 16 are documentaries. This makes a lot of sense if the documentary makers are using a different language than the documentary subjects.

IMDb Score

You are excited to take a look at all the IMDb scores! Are most films rated around the same score? Are there some extremely low or high scores?



Mean: 6.3

Standard Deviation: 1.0

Free response

Using the plot and summary statistics, describe the distribution of IMDb scores.

Your response

The distribution of IMDb Scores is left skewed, the average score is around 6.1 with about 62% of the movies and films, having that score.

Our answer

The distribution of IMDb scores is mostly symmetrical in a bell shape, indicating a normal distribution. There are a couple of very low scores, but they are not far from the rest of the distribution, so they may not be extreme enough to be considered outliers. Since the distribution is fairly symmetrical, we can rely on the mean of 6.3 to give us a good idea of what a typical IMDb rating is. With a standard deviation of 1, we know there is some variation in scores, but most scores fall between 4 and 8 on the 1-10 scale.

▼ ***Just for Fun:*** **Find out which films got the highest and lowest IMDb scores! (Click to Toggle the Information)**

▼ *Just for Fun:* **Find out which films got the highest and lowest IMDb scores! (Click to Toggle the Information)**

Highest Scoring Films

TITLE	GENRE	LANGUAGE	YEAR	RUNTIME	SCORE
David Attenborough: A Life on Our Planet	Documentary	English	2020	83	9.0
Emicida: AmarElo - It's All For Yesterday	Documentary	other single	2020	89	8.6
Springsteen on Broadway	Other	English	2018	153	8.5
Taylor Swift: Reputation Stadium Tour	Other	English	2018	125	8.4
Ben Platt: Live from Radio City Music Hall	Other	English	2020	85	8.4

Lowest Scoring Films

TITLE	GENRE	LANGUAGE	YEAR	RUNTIME	SCORE
Enter the Anime	Documentary	multiple	2019	58	2.5
Dark Forces	Horror/Thriller	Spanish	2020	81	2.6
The App	Action/Sci-Fi	other single	2019	79	2.6
The Open House	Horror/Thriller	English	2018	94	3.2
Kaali Khuhi	Other	Hindi	2020	90	3.4

Genre

Your colleague is helping you with your research. They will be writing up the summary for the `genre` variable.

Free response

Which summary statistics might your colleague include in their summary of the film genres?

Your response

The productions with highest scores belong to the documentary gender in English or another single language, and to other genders, mainly music concerts. The horror movies, and sci-fi movies have very low scores, as well as the documentaries in multiple languages.

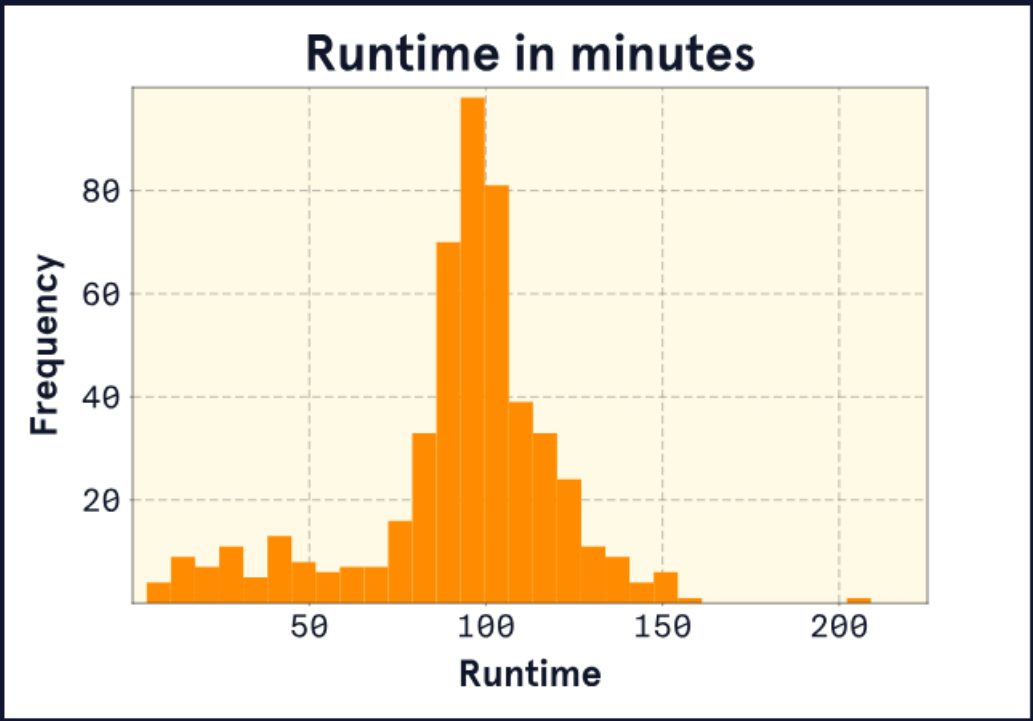
Our answer

Genre is a categorical variable — this variable gives information about a quality of the films that is non-numeric. We can describe categorical variables using ***frequencies, proportions, and ratios***.

Our colleague might create a table showing the different genre categories, the count of films in each category (frequency), and the percentage of the total this count represents (proportion). Our colleague might also compare counts of genres to one another using ratios.

Runtime

Your colleague used analytics software to create a summary of the `runtime` variable. The software program has a default setting for numeric variables that outputs a distribution plot, the mean, and the standard deviation. The analytics software produced the following plot and statistics for the runtimes:



Mean: 92.5

Standard Deviation: 28.4

Free response

Based on the distribution plot, what concerns might you have with the default summary statistics used by the analytics software?

Your response

There are outliers, that are highly affecting the mean.

Our answer

There are two aspects of this distribution plot that might lead to concern about using the mean and standard deviation:

1. The distribution is left-skewed — it has a long tail of low values on the left side. These values might influence the mean to be lower.
2. There is a single high value of just above 200 minutes. This value might be an outlier that influences the mean to be higher.

Free response

Which alternative statistics could your colleague use in this case?

Your response

To use IQR

Our answer

We could use statistics that are more ***robust*** to outliers and skewness, such as the median and interquartile range (IQR). The median is the middle value and the IQR is the range of the middle 50% of the data (Q3 - Q1).

▼ ***Just for Fun:*** **Based on the distribution plot, what is your guess for the median and IQR of runtimes? Find out if your guess was close! (Click to Toggle the Median and IQR)**

Mean: 92.5

Standard Deviation: 28.4

The mean describes a typical runtime as in the low 90s. The standard deviation describes the distribution as having wide variability, with runtimes an average of almost 30 minutes different than the mean. These measurements are not wrong, but they don't help us do a good job of summarizing what we're seeing in the distribution.

Median: 97.0

IQR: 21.8

In contrast, the median describes a higher runtime as most typical. The low IQR indicates that half the values aren't very far from the center value. These descriptions better match the large number of values near 100 that we see in the distribution plot.

Since the mean is less than the median, it seems like the left-skew is more influential on the mean than the high potential outlier is.

▼ ***Just for Fun:*** **Find out which movie has that really high runtime. (Click to Toggle the Information)**

TITLE	GENRE	LANGUAGE	YEAR	RUNTIME	SCORE
The Irishman	Drama	English	2019	209	7.8

This film is about 3 and a half hours long!

Relationships

Next, you want to explore some relationships between variables. Aggregating data across genres and languages may give you some insights.

IMDb Score by Genre

You want to know if any genres have particularly high IMDb scores, so you look at a table of means and standard deviations of IMDb scores for each genre.

GENRE	MEAN	STD DEV
Action/Sci-Fi	5.7	1.0
Animation	6.6	0.9
Comedy	5.8	0.8
Documentary	7.0	0.8
Drama	6.3	0.8
Horror/Thriller	5.6	1.0
Other	6.4	1.0
Romance/ Romantic Comedy	5.9	0.6

Free response

Describe what you learn about IMDb scores across genres from the means and standard deviations in the table.

Your response

The productions with the highest scores are documentaries, and the productions with the lowest scores are the Horror/Thriller movies.

Our answer

Most of the mean and standard deviation pairs are not far from the overall mean and standard deviation of all IMDb scores (6.3 and 1.0). However, there are a couple of patterns that stand out.

- The "Romance/Romantic Comedy" genre has the lowest standard deviation at 0.6. This may indicate this genre was pretty consistent in getting scores close to the mean of 5.9.
- The "Action/Sci-Fi" and "Comedy" genres had similar mean scores to "Romance/Romantic Comedy" but a wider spread of scores.
- The "Documentary" genre had the highest mean IMDb score. Since its standard deviation isn't particularly large, this may indicate Netflix documentaries tended to rate well fairly consistently.

Runtime by Language

You are wondering if there are any differences in the length of films across languages.

LANGUAGE	MEAN	STD DEV
English	92	29
Spanish	93	25
Hindi	115	17
French	91	13
other single	95	24
multiple	73	36

Free response

Describe what you learn about film runtimes across languages from the means and standard deviations in the table.

Your response

Hindi productions tend to be longer and productions in multiple languages ten to be shorter.

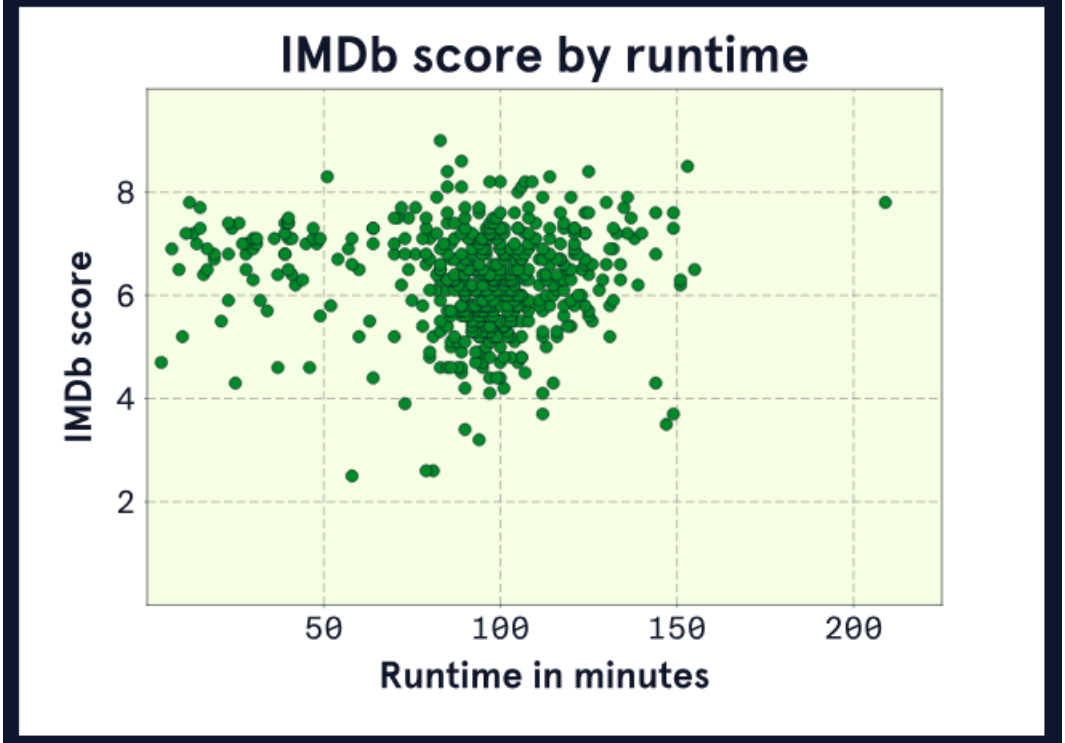
Our answer

There are some interesting differences among the means and standard deviations in the table.

- English, Spanish, French, and "other single" films all have around the same average runtime. However, French films have about half the standard deviation of the others. There is a lot less variability in French film runtimes.
- Hindi films have the highest mean runtime at 115 minutes. The low standard deviation indicates that Hindi films in our dataset are consistently a relatively long runtime.
- The lowest mean runtime belongs to the "multiple" language category. This is also the category with the highest standard deviation. We know the runtime distribution had a long tail of low values, so perhaps this category contains a lot of those very low runtimes that pull the mean down and widen the standard deviation.

Runtime and IMDb Score

For your final exploration, you've just added the following plot and summary statistic to your report.



Correlation Coefficient: 0.92

Your colleague is concerned there may be an error. The plot and the correlation coefficient don't match.

Free response

What does the plot tell you about the relationship between runtime and IMDb score?

Your response

The productions withi a runtime about 100 minutes have higher scores.

Our answer

The plot does not show any linear relationship between the two variables. The plot mainly shows a cloud of points that aren't close to the shape of a line. Lower runtimes aren't associated with particularly low or high IMDb scores. Higher runtimes aren't associated with particularly low or high IMDb scores.

Most of the films have runtimes between 50 and 150 minutes with varying IMDb scores between about 6.0 and 8.0 across those runtimes in no particular pattern.

Free response

What does the correlation coefficient tell you about the relationship between runtime and IMDb score?

Your response

There is almost no correlation.

Our answer

The correlation coefficient of 0.92 indicates a very strong, positive linear relationship between runtimes and IMDb score. Shorter films have lower IMDb scores and longer films have higher IMDb scores. Since 0.92 is so close to 1, we conclude this pattern holds very strongly with little variation.

▼ ***Just for Fun:*** **The correlation coefficient was incorrect. What do you think the true correlation coefficient is? Find out if your guess is correct! (Click to Toggle the Information)**

The true correlation coefficient is -0.04. The sign indicates a negative relationship, but the value is so close to zero that we should conclude there's really no linear relationship between the variables.

Final Thoughts

You finished your first movie research report! You did a lot of exploring and learned a lot about the different features of Netflix content. These findings could be the final step or the starting point of a deeper analysis. Great job!