UNIT 3

USE CASE 3 - Making Inferences from Data

The Business of Online Reviews: How many exclamation points are too many?!!!

The internet is anything but transparent, especially when it comes to online shopping. It could be a phone charger with 12,000 reviews, a volume entirely out of proportion to its low-stakes decision to buy. Or the decision to book a hotel—the 2,000 reviews show 4.8 stars, but a stay greets travelers with cold showers or broken TVs.

How can a system supposedly designed to help people navigate infinite choices online be so defective?

It's an indecisive shopper's nightmare, and a flawed system fueled by greed and desperation carried out with so-called fake reviews. Experts track categories of deceit to untangle the creative webs of lies crafted online that translate to direct influence on global online spending, estimated at $152 billion annually. In countries that prize e-commerce the most, fake reviews influence $791 billion in online spending every year in the U.S., another $6.4 billion in Japan, $5 billion in the UK, $2.3 billion in Canada, and $900 million in Australia (World Economic Forum, 2021).

Self-reported data by sites including TripAdvisor, Yelp, TrustPilot, and Amazon estimate that 4% of online reviews are false. However, researchers say that number rises from 16 to 40% when accounting for all reviews posted across e-commerce sites. It's a marketplace in plain sight where money exchanges with blatantly fraudulent activity. "This suggests that there must be a sufficiently large group of lay Internet users who—for reasons that could range from malicious to benign—are devoted to writing reviews based on imagination rather than genuine post-purchase experiences" (Banerjee & Chua, 2023).

Online retailers started taking notice—and legal action—to regain credibility with consumers and fend off the more extensive networks of unscrupulous offenders who run pay-for-posting rings as brokers between sellers and individuals willing to write fraudulent reviews. Amazon recently sued four companies it accused of piloting a scheme that generated fake reviews for products across its platform, with three firms "employing" an estimated 350,000 reviewers to produce 5-star write-ups (Kleinman, 2022).

Data unearthed by researchers count fake online reviews well into the tens of millions. One site, Trustpilot—a company dedicated to hosting consumer reviews with 167 million of them currently circulating for everything from insurance to jewelry—reported removing more than 2 million reviews deemed to be fake, accounting for 5% of its reviews in a year (Povich, 2022).

Research shows most people count on reviews when weighing a purchase, with one Market Monitoring Survey about misleading online practices in Europe finding that 71% of consumers consider reviews important when researching places to stay on vacation. It also found that most websites scoured by regulators did not have consumer protections highlighting incentivized reviews, for instance, written after exchanging money and/or goods (European Commission, 2021).

Retailers report hiring investigators and using specialized software that uses Natural Language Processing such as Fakespot, Thereviewindex, ReviewMeta, and OpenAI/GPT-2 to root out reviews with repetitive diction, hyperbole and superlatives, and poor grammar and spelling, all signs of potentially bot-created content. Antitrust regulators in the UK launched an investigation in 2021 into whether Amazon and Google had done enough to eliminate fake reviews (Morris, 2022).

In the U.S., the Federal Trade Commission (FTC) Act prohibits deceptive practices, asking platforms themselves to do more to intervene on behalf of wronged consumers. Tactics include improving detection technology, sharing details about bad actors and fraudulent patterns, and providing greater access to outside researchers. The FTC says its reach is limited in scope because of other laws in place to protect the free flow of communications (Gaynor et al., 2022).

Still, humans, not machines, create fake and misleading content, driving the most concern. People hired by the brokers are advised to make their writing seem more realistic by eliminating exclamation points and avoiding easily detected words, including "treat," "excellent," and "perfect," language linguistics experts have flagged as most likely to be in fake reviews (Harris, 2022).

Rating manipulations often hit the smallest and fledgling businesses the hardest, with a 1-star review on Yelp resulting in an estimated 5 to 9% decline in revenue (Salminen et al., 2022).

The result can be harsh for service-oriented companies like family-run restaurants and local plumbers who rely on reviews to drive customers to them. With food and vacation stays, seemingly nothing satisfies a sense of entitlement to get something for free. Businesses themselves feed into a broken model by comping higher-end goods and services—$100 worth of chicken wings or an upgraded suite with a view—in exchange for the promise of a 5-star review or rushing patrons into completing a review on a tablet as the manager looks on (Cramer, 2023).

The realm of fake reviews blurs into the up-and-coming marketplace of influencers and bloggers who act as product billboards for a price. Their reviews are sunny and effective; "micro-influencers," those with strong local followings within a metro area, drive up sales the most (Harris, 2022).

While business owners are becoming savvier in handling the carousel of reviews that determine success, consumers can protect themselves by refining their baseline "persuasion knowledge," which means understanding the characteristics of fake posts. Experts recommend a get-informed approach to look deeper than aggregate scores and analyze potential text for one-sidedness, exaggeration, a personal selling style, and generic descriptions that all signal possible sham reviews (Filho et al., 2023).

Unit 3 explores Mathematical Common Core Standards:

Common CORE Standards

Interpreting Categorical and Quantitative Data (S-ID)

- Summarize, represent, and interpret data on a single count or measurement variable
- Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve

Making Inferences and Justifying Conclusions (S-IC)

- Understand and evaluate random processes underlying statistical experiments.
    - Understand statistics as a process for making inferences about population parameters based on a random sample from that population.
    - Decide if a specified model is consistent with results from a given data-generating process

Milestone 1 - Fake Reviews

Introduction

The business client for this use case is the Fair Business Commission (FBC), a nonprofit social enterprise established in 2019 to educate businesses and the public on core themes that support ethical, equitable, responsible, and sustainable business practices.

Bad actors are individuals or groups who intend to cause harm.

ROLE

Your role in milestone 1 is as an FBC intern tasked to support data gathering, assessing, and reporting tasks that contribute to informing FBC and collaborating with business communities to educate on offensive and defensive strategies that drive fair business practices.

In Milestone 1, you will explore fake review identification approaches to infer and draw conclusions. First, the following terms inform applying inferences and conclusions from data. An article, Creating and detecting fake reviews of online products, offers examples to help describe the terms.

Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, Bernard J. Jansen, Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, Volume 64, 2022, 102771, ISSN 0969-6989, https://doi.org/10.1016/j.jretconser.2021.102771. (https://www.sciencedirect.com/science/article/pii/S0969698921003374)

Terms

Fake Product Reviews

Fake product reviews - are also referred to as "opinion spam." Fraudulent actors create fake reviews by

paying human content creators to write authentic-appearing reviews and

utilizing computer-generated **text generation algorithms** to automate fake review creation.

Fake reviews can degrade the credibility of online reviews as they praise or attack brands. Most online marketplace algorithms use reviews to determine a product's ranking among other products in the same category. This equates to increased or decreased visibility of products shown to potential customers, resulting in unfair competition.

Text Generation

Text generation - including natural language processing (NLP) and machine learning (ML) software automates fake reviews generated at low cost, at scale.

Incentivized Reviews

Incentivized reviews - refer to reviews obtained by a marketing campaign, like influencer endorsements or offering customers something in exchange for a review. For influencers, reputation is essential, and they typically avoid faking a review. There may be more "truth" in incentivized reviews; however, they are a higher-cost competitive advantage not affordable for many small businesses.

Descriptive statistics	Describing the data with measures of center, spread, or shape.
Inferential statistics	Drawing conclusions on how data relate to one another and predict center, spread, or shape.
Parameters	When center, spread, and relative frequency measure a population.
Statistics	When center, spread, and relative frequency measure a sample.

Inferences From Data

We all make inferences, based on evidence and reasoning, on what to buy, watch, eat, sell, or make, cook, and build. A friend invites you to a concert, a family member recommends a pizza, or you see reviews on a new multiplayer game.

Inferences from this data (evidence) are different from **conclusions from the data.** The two can interweave with different contexts or interpretations.

Inference	Conclusion
something that you consider true, based on the information you have	something you determine after thinking about all the information you have

Conclusions typically require a deeper understanding of a position, an opinion, or a judgment. Your task is to recognize each separately and where they blend as you consider gathering, assessing, and reporting data.

Example Salimen, et al. **inferences from data**:

- *Our results indicate that human accuracy in detecting fake reviews is only slightly higher than random chance. In other words, the generator can fool humans.* (p. 11)
- *The performance is more inconsistent with shorter reviews, most likely because short reviews contain less information for the model to judge – consider a review such as "Ok, it works great!"* (p. 8)
- *Since fake reviews have detectable but nuanced patterns, it seems reasonable that machines would be better at this task than people and would agree with other machines more often than people.* (p. 9)
- *Results suggest that it is becoming increasingly difficult for consumers to distinguish high-quality products from low-quality ones based on online reviews. This debacle can be considered one of the most prominent risks for e-commerce, mainly because trust has an elevated role in electronic marketplaces.* (p. 12)

Research Questions

Studies use research questions to address an issue or a problem, which, through analysis and interpretation of data, is answered in the study's conclusion.

For example, the research questions in the Salminen, et al. 2021 study are:

> **RQ1:** How realistic (i.e., high-quality) are reviews that current text generation algorithms produce? (i.e., can text generation fool humans?)
>
> **RQ2:** (a) Can a machine detect a fake review generated by another machine?
> (b) Does a machine classifier do so better than a human?)
>
> **RQ3:** Can a machine detect a fake review generated by humans? (p. 2)

These questions address the capability of current technology to generate fake reviews that harm preserving the credibility of the marketplace.

Methodology

Methodology - is **how** researchers investigate their research questions. For the three research questions in the Salminen, et al. study, the researchers created and measured their fake reviews developed based on mirroring and editing pre-existing real reviews. They experimented with different technologies and determined which approach supported inferences from data from which they can draw conclusions.

The Salimen, et al. **conclusions from data**:

> *Detection of fake reviews is a problem for researchers, e-commerce sites, and firms engaged in online business. Our results indicate that current text generation methods yield fake reviews that appear so realistic that it is challenging for a human to detect them. Fortunately, machine learning classifiers do much better in this regard, with almost perfect accuracy in detecting reviews generated by other machines, implying "machines can fight machines" in the battle against fake reviews.*

Count Data

Fake and real reviews represent **count data** - non-negative integers representing the number of times a discrete (uses distinct, countable values) event is observed. Likes, stars, and upvotes are examples of count data.

A **single count variable** is discrete because it consists of non-negative integers.
Have you ever seen -3 stars on a product? 4.5 stars works because it is a positive integer (whole numbers aren't required).

- If the review of your favorite movie on a streaming service rates 4.6 stars, that's a single count variable

- If you expand your review variable from stars for one movie to randomly selecting 1,000 movies from one streaming service, the ratings are **discrete random variables** because although the number of movies escalates, the total number of data observations is limited to one streaming service and controlled by the range of ratings (1-5 stars)

He, S., Hollenbeck, B., & Proserpio, D. (2022). The Market for Fake Reviews. Marketing Science, Vol.41(5), p. 896-921, http://dx.doi.org/10.2139/ssrn.3664992

In an alternative approach (He, et al. 2022), **inferences from data** informed:

Companies buying fake reviews for their products are highly clustered in the product-reviewer network due to their reliance on common reviewers

This clustering allows detection with high accuracy

Human reviewers have strong incentives to evade detection; they strive to write fake reviews that mirror organic reviews (avoiding particular words or phrases)

Based on the inferences from data, He, et al. developed their sampling methodology. To analyze fake review behavior on Amazon, they began by collecting data from the private Facebook groups where sellers buy reviews (23 groups from March-October 2020). (pp. 6-7)

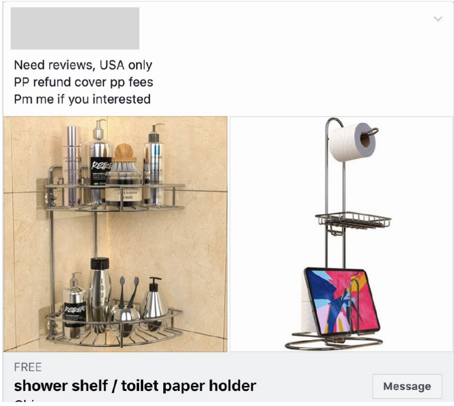
- Groups average 16,000 members and post 568 fake review requests per day per group

- Sellers post product pictures and review requests, after which the reviewer and seller communicate by Facebook message

- Most reviewers are compensated by refunding the cost of the product via a PayPal transaction after the five-star review is posted, plus the PayPal fee, sales tax, and occasionally a commission. Reviewers buying the product means the review is listed as a "Verified Purchase" review

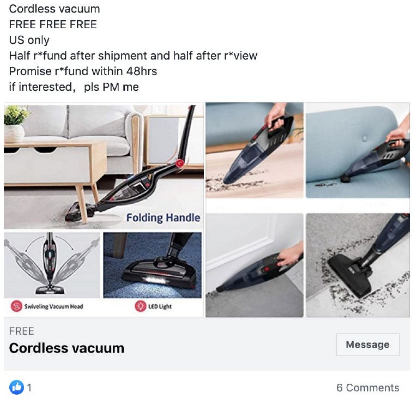
- Amazon sellers buy only positive reviews, likely because buying fake negative reviews to hurt competitors is more costly (buying multiple competitor products) and because positive reviews more directly increase sale

Example Facebook review request (p. 9)



Need reviews, USA only
PP refund cover pp fees
Pm me if you interested

FREE
shower shelf / toilet paper holder



Cordless vacuum
FREE FREE FREE
US only
Half r*fund after shipment and half after r*view
Promise r*fund within 48hrs
if interested, pls PM me

FREE
Cordless vacuum

After identifying a sampling of approximately 1,500 unique products from the Facebook groups, He, et al. collected reviews and ratings for each of the products daily. Noting the rating, product ID, review text, photos, and helpful votes.

In addition, they collected daily review data for 2,714 competitor products to serve as a comparison set. For each product buying fake reviews, He, et al. selected two products that appeared most frequently in the search results seven days before and after the date of the product's first Facebook post. The comparison set is then in the same subcategory and has a similar search rank before the fake reviews were posted.

Count data **inferences from data** (pp. 3-4):

- In the weeks after purchasing fake reviews, the number of reviews posted per week roughly doubled
- The average rating and share of five-star reviews increased substantially, as did search position and sales rank
- The increase in average ratings was short-lived, with ratings falling back to the previous level within two to four weeks
- But the increase in the weekly number of reviews, sales rank, and position in search listings remained substantially higher more than four weeks later
- Sales began to fall significantly right after the fake review campaign ended
- New products with few reviews, which might be using fake reviews efficiently to solve the cold-start problem (new product launched on Amazon), see a larger increase in sales initially and a similar drop-off afterward
- Amazon deleted a very large share, half of their reviews, but the deletions occur with an average lag of over 100 days, allowing sellers to benefit from the short-term boost in ratings, reviews, and sales

The **top 15 categories and subcategories** in the He, et al. study sample are noted in the table from highest number to lowest number in each column. Third-party sellers sold most products:

Category	N (number)	Subcategory	N (number)
Beauty & Personal Care	193	Humidifiers	17
Health & Household	159	Teeth Whitening Products	15
Home & Kitchen	148	Power Dental Flossers	14
Tools & Home Improvement	120	Sleep Sound Machines	12
Kitchen & Dining	112	Men's Rotary Shavers	11
Cell Phones & Accessories	81	Vacuum Sealers	11
Sports & Outdoors	77	Bug Zappers	10
Pet Supplies	62	Electric Back Massagers	10
Patio, Lawn & Garden	59	Light Hair Removal Devices	9
Electronics	57	Outdoor String Lights	9
Baby	42	Cell Phone Charging Stations	8
Office Products	30	Electric Foot Massagers	8

The study's count data observations reports mean and adds additional median details. Descriptive statistics describe data with measures of center, and the spread from center. As you know, mean, median, and mode are different measures of center in a data set (central tendency of a probability distribution). In previous classes you learned:

Mean	Median	Mode
add up the values in the data set and divide by the number of values	create a list of values in the data set, in numerical order, and identify which value appears in the middle of the list. If two are in the middle, average them to obtain the median	identify the most common number in a data set

If you draw conclusions from the data to predict center, spread, or shape, that's inferential statistics.

Tasks

In this class, you will utilize mean, median, and mode to make specific estimates. Please discuss and respond to the following questions, based on study data observations (p. 12):

Ratings on Amazon are count data. The following chart notes the study's count and the mean.

	Count	Mean
Displayed Rating		
Fake Review Products	1,315	4.4
All Products	203,480	432
Number of Reviews		
Fake Review Products	1,425	183.1
All Products	203,485	451.4
Price		
Fake Review Products	1,425	33.4
All Products	236,542	44.7
Sponsored		
Fake Review Products	1,425	.1
All Products	236,542	.1
Keyword Position		
Fake Review Products	1,425	21.4
All Products	236,542	28.2
Age (days)		
Fake Review Products	1,305	229.8
All Products	153,625	757.8
Sales Rank		
Fake Review Products	1,300	73,292.3
All Products	5,647	89,926.1

1. Products purchasing fake reviews had, at the time of their first Facebook post, relatively high product ratings. The mean rating is 4.4 stars, and the median is 4.5 stars, which are both higher than the average ratings of competitor products.

   Why, in this context, are the mean and median important considerations? How do the ratings below four stars contribute to your understanding of the study observations?

   ____

   ____

   ____

2. With a mean age of 229 days, the products collecting fake reviews are not generally new to Amazon. Of the 1,500 products observed, only 94 solicit fake reviews in their first month. What does the mean age of products collecting fake reviews suggest?

   ____

   ____

   ____

3. Fake review products charge a higher median price than their competitors, but there are far fewer high-priced products among the fake review products than among competitors. What does it mean that fake review products charge a higher median price than their competitors, even though their mean price is lower?

   ____

   ____

   ____

Almost none of the sellers in these markets were well-known brands. Brand name sellers may still be buying fake reviews via other (more private) channels, or they may avoid buying fake reviews altogether to avoid damage to their reputation.

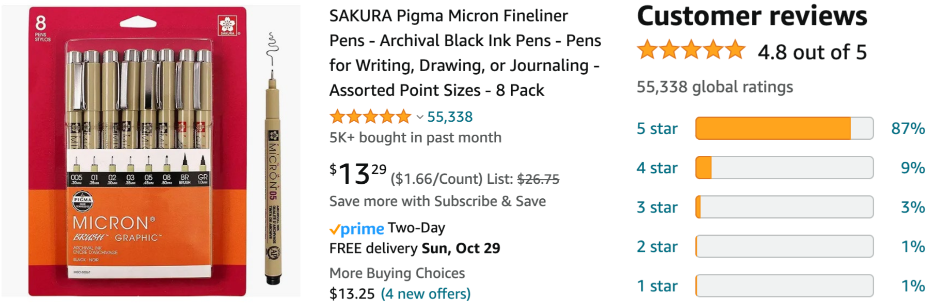
A sampling of the study's **conclusions from data** related to the data you assessed includes:

- The market for fake Amazon product reviews, which take place in private Facebook groups, features millions of products
- Soliciting reviews on Facebook is highly effective at improving several sellers' outcomes
- The effects are often short-lived; the boost in sales does not lead to a positive self-sustaining relationship between organic ratings and sales, and both fall significantly when fake review recruiting ends
- Rating manipulation is not used efficiently by sellers to solve a product cold-start problem
- Fake review recruiters eventually see a large decrease in ratings and an increase in one-star reviews
- Rating manipulation likely harms honest sellers and the platform's reputation
- Fake review firms continuously improve manipulation strategies
- Amazon claims to have spent over $500 million in 2019 alone and employed over 8,000 people to reduce fraud and abuse on its platform
- Amazon deletes large numbers of reviews, and deletions are well-targeted; however, there is a long lag before these reviews are deleted that does not eliminate the short-term profits or the consumer the harm they cause

Milestone 1 Check-in Meeting

Per your teacher's instructions, meet in small teams to assess and discuss the following five Milestone 1 questions:

1. The Amazon item below has 55,338 reviews that add up to an average of 4.8 stars.



SAKURA Pigma Micron Fineliner Pens - Archival Black Ink Pens - Pens for Writing, Drawing, or Journaling - Assorted Point Sizes - 8 Pack

★★★★★ 55,338

5K+ bought in past month

$13.29 ($1.66/Count) List: ~~$26.75~~

Save more with Subscribe & Save

✓prime Two-Day

FREE delivery **Sun, Oct 29**

More Buying Choices

$13.25 (4 new offers)

Customer reviews

★★★★★ 4.8 out of 5

55,338 global ratings

5 star	87%
4 star	9%
3 star	3%
2 star	1%
1 star	1%

Out of 55,338 (so far) reviews, with the distribution of stars you see, will most of the data be close to the mean or scattered from the mean? What do you estimate the SD will approximate?

____

____

____

2. Go to the Amazon page for this Sakura Pen product or something close to it. Then, go to the website https://reviewmeta.com/ to copy and paste the Amazon URL into their text box to analyze Amazon product reviews and filter out reviews their algorithm detects may be unnatural. Press to update/refresh if requested.

   How did the ReviewMeta.com algorithm change the review?

____

____

____

3. Go to Amazon.com and select a product you recently purchased, whether or not you bought it at Amazon. How many reviews are there? How many 5, 4, 3, 2, and 1 stars?

____

____

____

4. Go to the website https://reviewmeta.com/ to copy and paste your product's Amazon URL into their text box to analyze Amazon product reviews and filter out reviews their algorithm detects that may be unnatural. Press to update/refresh if requested. How did the ReviewMeta.com algorithm change the review?

____

____

____

5. What can you infer from the data explored in the He, et al. study and from your Amazon reviews exploration?

____

____

____

6. Are you prepared to recommend conclusions from Milestone 1? What data inferences inform your conclusion, or what data are needed to draw a conclusion from data?

____

____

____

____

Milestone 2

Public free datasets are common. Businesses often access datasets for insight into everything from customer perception to marketing effectiveness to the weather at their locations. In Milestone 2, you will use public dataset data and tools to practice inferences toward conclusions from data.

The essential question is:

What do the data tell us?

Terms

Target Population

Research studies are usually carried out on sample of subjects rather than whole populations. The most challenge of fieldwork is drawing a random sample from the **target population** to which the results of the study may be generalized across the target population. The task is so difficult that some sampling bias occurs in almost all studies.

There are three types of populations available to draw samples from include:

Literal Population	Virtual Population	Metaphorical Population
A literal population is an identifiable group- like movie and other product reviews. Or people, cars, houses, etc.	A virtual population is when every measurement will return a slightly different answer- like the weather, heart rates, humidity, etc., data change constantly.	A metaphorical population is when there is no larger population- like what the long-term effect of EV batteries to the environment or the influence of robotics on the workforce. Events that can occur but haven't yet.

A target population is the population you want to learn something about, from any of the three population types available. The target population of Amazon movies is all the movies in the dataset.

Sample Surveys

A **sample survey** is a method for collecting data from or about a population so that inferences about the population can be obtained from the population sample.

A sample survey may study the attitudes of individuals in a population toward a particular subject. In Amazon movies, respondents submitted 1–5-star reviews for movie.

With sample surveys, it's essential to select survey respondents that are representative of your target population. This first step is crucial because a quality sample ensures your results' validity. Once respondents are identified, you can use online surveys, interviews, or other methods to ask open- and closed-ended questions from your sample. Their responses give you raw data. Any inferences from a sample refer only to the target population from which the sample was selected. After collecting data, you can make estimates about your target population using statistics.

The **target population** is the group the researcher hopes to understand. The **experimentally accessible** population is the group that a researcher can access to measure. A problem occurs if the experimentally accessible population is not representative of the target population.

In practice it's difficult for a random sample in your target population to be identified and surveyed. A random number generator (you will use with Google Sheets) works well to identify the sample; however, you may need to reach out to respondents in multiple ways like web forms, email, phone calls, etc. With the Amazon movies example, respondents self-selected whether they wanted to review the movie. This creates validity challenges for several reasons. Why do some people reply to review and others do not? People who watch a movie and decide to review it are not randomly selected.

Experiments

An **experiment** deliberately imposes a treatment to observe the response. Experiments involves researchers manipulating at least one independent variable (explanatory variable of interest) under controlled conditions, and they measure the changes in the dependent variable (response).

The experiment study aim is to identify an association with independent variables and changes in the dependent variable; association does not equate to causation. Saying X (independent variable) causes Y (dependent variable) does not mean every time X occurs, Y also occurs. Or that Y only occurs if X does. Experiments assess whether X increases the proportion of times that Y happens.

In a randomized experimental design, objects or individuals are randomly assigned to an experimental group. Using randomization is the most reliable method of creating treatment groups, without potential biases or judgments.

In a **completely randomized design**, objects or subjects are assigned to groups at random.

If the study identifies differences among groups of subjects or objects within an experimental group, in a **block design**, experimental subjects are first divided into a block (age, location, gender, etc.) then are randomly assigned to a group, followed by randomizing within the group.

Observational Studies

Observational studies collect information or look at data that was already collected, without changing existing conditions. The researcher observes and assesses (without intervening). Three types of observational studies include:

Cohort studies	Case-control studies	Cross-sectional studies
a group linked in some way. One or more samples (called cohorts) are followed and evaluations toward an outcome are conducted. Researchers compare what happens to the data in a cohort to data outside the cohort.	identifying an existing "case" and a group without the problem "controls" and comparing the data. A case-control study starts with an outcome then traces back through data.	collect data from many different individuals at one point in time. In cross-sectional research, you observe variables without influencing them. You can describe characteristics that exist, but not to determine cause-and-effect between different variables. Used to make inferences about possible relationships or to gather preliminary data to support future study.

A major issue of observational evidence is that it is known to have limited internal validity as it is subject to both bias and confounding. A confounding variable is a factor other than the one being studied that is associated with the dependent variable and the independent variable. A confounding variable can distort or mask the effects of another variable.

Standard Deviation

Standard deviation (SD) measures the spread of a data distribution.

It measures the typical distance between each data point and the mean.

It shows how much variation there is from the average (mean).

a **low SD** indicates the data points tend to be close to the mean

a **high SD** indicates the data are scattered out over a large range of values

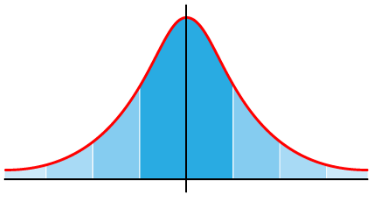
The formula for standard deviation depends on whether the data is being considered a population of its own or the data is a sample representing a larger population.

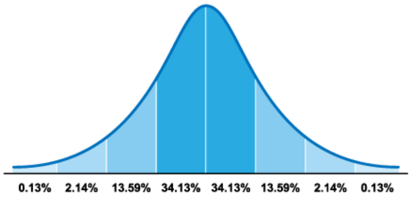
The purpose of mean and SD is to describe the sample so the study's findings can be generalized to other contexts like the study, for example, fake reviews on hotel, food, and other product ecommerce sites.

Measures of central tendency are the mean, median, and mode

Measures of dispersion are the range, SD, and interquartile range

On the left, you see the mean, shown with the vertical line on the normal distribution (normal curve). On the right, you see the same normal distribution and the relationships between the mean and SD.





SD is a measure of dispersion to describe the sample; if you think of the distance from the mean as a positive number, SD tells you how far from the mean the average data are.

The mean tells you what the average value is

The SD tells you what the average scatter of values is around the mean

mean ± 1 SD includes 68.3% of the population
mean ± 2 SD includes 95.5% of the population
mean ± 3 SD includes 99.7% of the population

Published tables of the area under the normal curve support calculating the probability of finding a value at any distance from the mean when the distance from the mean is expressed in terms of the SD.

Tasks

With fake reviews, a SD indicates the average variability in your data set (how spread out your data are). You will use Google Sheets (or a preferred spreadsheet program) to assess the data.

Tables

Using data tables makes it easy to focus on one or two variables. Results are easy to read and share. Tables help you to condense, organize, and make sense of data, that would be difficult to see with hundreds, and even thousands of pages of data which form the basis of most qualitative studies.

Working with a Calculator

Please go to the website

https://www.StandardDeviationCalculator.io/mean-calculator

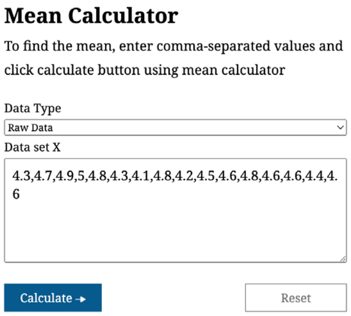
This table shows the first 15 movies listed in an Amazon movie reviews database.

Title	Movie_Rating	No_of_Ratings
Totally Killer	4.3	323
Guy Ritchie's The Covenant	4.7	13268
A Million Miles Away	4.9	1126
Kelce	5	570
Despicable Me 3	4.8	31813
Those Who Wish Me Dead	4.3	7403
Renfield	4.1	9259
The Proposal	4.8	52086
Black Adam	4.2	22762
A Thousand and One	4.5	1317
Pokémon Detective Pikachu	4.6	52376
Top Gun: Maverick	4.8	107727

DC League of Super-Pets	4.6	8156
Aquaman	4.6	87453
Dungeons & Dragons: Honor Among Thieves	4.4	16970

STEP 1

Enter each movie rating, separated by a comma (no space), into the Mean Calculator and select Calculate



Mean Calculator

To find the mean, enter comma-separated values and click calculate button using mean calculator

Data Type: Raw Data

Data set X: 4.3,4.7,4.9,5,4.8,4.3,4.1,4.8,4.2,4.5,4.6,4.8,4.6,4.6,4.4,4.6

What result do you have for

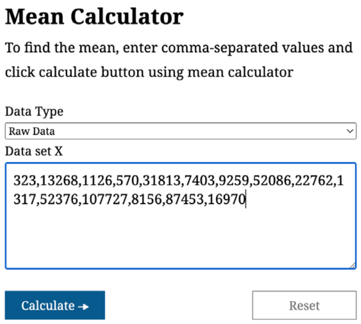
Mean ______________________________ Total ______________________________

Smallest ____________________________ Largest ____________________________

What is the Median? __________________ What is the Mode? __________________

STEP 2

Reset and enter the values for the number of ratings



Mean Calculator

To find the mean, enter comma-separated values and click calculate button using mean calculator

Data Type: Raw Data

Data set X: 323,13268,1126,570,31813,7403,9259,52086,22762,1317,52376,107727,8156,87453,16970

What result do you have for

Mean ________________________________ Total ________________________________

Smallest ______________________________ Largest ______________________________

What is the Median? ____________________ What is the Mode? ____________________

STEP 3

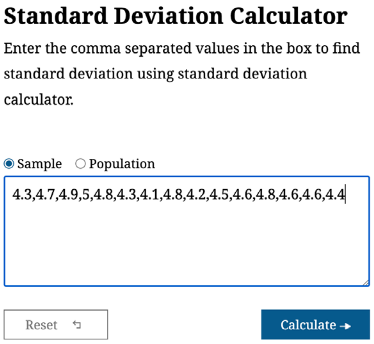
Change your URL to https://www.StandardDeviationCalculator.io/

Normal distribution drives from large numbers of small influences; for example, thousands of reviews within the 4–5-star range. Normal distribution is characterized by its **mean of expectation** and its standard deviation (SD) - the measure of spread from the mean.

Mean and SD are **statistics** that describe a set of parameters when describing a population. How many SD a data point is from the mean is called the **z-score**.

STEP 4

Enter each movie rating, separated by a comma (no space), into the SD Calculator, select Sample, and select Calculate



Standard Deviation Calculator

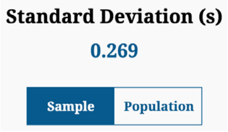
Enter the comma separated values in the box to find standard deviation using standard deviation calculator.

● Sample ○ Population

4.3,4.7,4.9,5,4.8,4.3,4.1,4.8,4.2,4.5,4.6,4.8,4.6,4.6,4.4

STEP 5

On the right side of the screen, at the top of the grayed box, you will see the SD

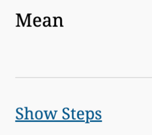


Standard Deviation (s)

0.269

STEP 6

Toward the bottom of the grayed box, select the Show Steps link.

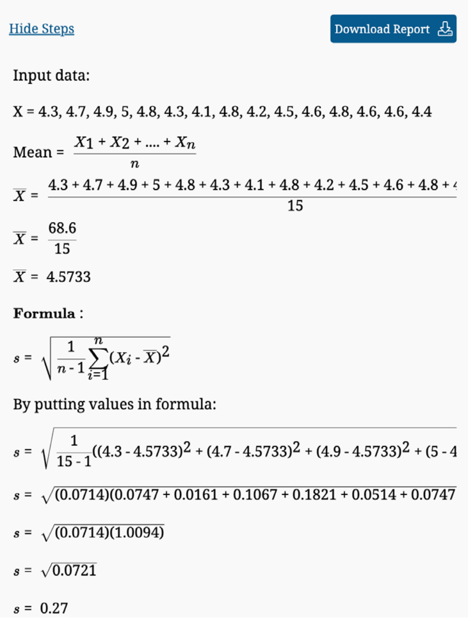


Mean

You see your Input data and how your values were divided by the total number of inputs to inform the mean.

STEP 7

Next, you see the formula for SD and the calculation to s, the sample SD.



Input data:

X = 4.3, 4.7, 4.9, 5, 4.8, 4.3, 4.1, 4.8, 4.2, 4.5, 4.6, 4.8, 4.6, 4.6, 4.4

$$\text{Mean} = \frac{X_1 + X_2 + ... + X_n}{n}$$

$$\overline{X} = \frac{4.3 + 4.7 + 4.9 + 5 + 4.8 + 4.3 + 4.1 + 4.8 + 4.2 + 4.5 + 4.6 + 4.8 +}{15}$$

$$\overline{X} = \frac{68.6}{15}$$

$$\overline{X} = 4.5733$$

Formula :

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

By putting values in formula:

$$s = \sqrt{\frac{1}{15-1}((4.3 - 4.5733)^2 + (4.7 - 4.5733)^2 + (4.9 - 4.5733)^2 + (5 - 4}$$

$$s = \sqrt{(0.0714)(0.0747 + 0.0161 + 0.1067 + 0.1821 + 0.0514 + 0.0747}$$

$$s = \sqrt{(0.0714)(1.0094)}$$

$$s = \sqrt{0.0721}$$

$$s = 0.27$$

1. From entering the data, you know there were no 3, 2, or 1 stars/votes. You know the smallest and the largest movie rating score. Utilizing the mean, median, mode, and sample SD in the workplace relates to understanding how to apply data to a goal, problem, or prediction.

   What do they tell you about whether the movies are recommended or not?

   ____

   ____

   ____

2. Whether a movie is good or not may be interesting when you select a movie to watch; however, business goals are typically more complex. To determine if a movie is good, looking at one movie review, noted as stars on the streaming platform, is probably enough. Your task at the Fair Business Commission is to look for signals about fake reviews.

   How might utilizing the mean, median, mode, and sample SD infer whether there may be fake reviews in the sample set of Amazon movies?

   ____

   ____

   ____

3. If you were to calculate the sample SD of the number of ratings, would it add value to inferring if there may be fake reviews? Not all data are described well utilizing normal distribution. How would the sample SD of the movie ratings compare with the sample SD of the number of ratings contribute to your goal?

   ____

   ____

   ____

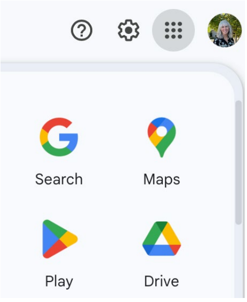
The first 15 movie ratings supported using count data to infer your goal to consider the possibility of fake reviews in the Amazon movie dataset. The actual dataset has over 2,000 lines. You would not be expected to key over 2,000 values into a calculator, much less tens of thousands of values; a spreadsheet supports your efforts.

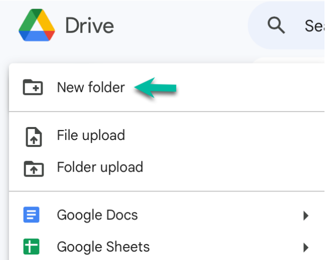
In your collaboration with the Fair Business Commission, statistical sampling methods better inform risks to fair business practices.

Working with a Spreadsheet

STEP 1

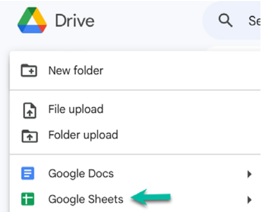
Please open the Gmail account you created in Unit 1 for the Google Earth project. Select Google Drive. Create a New folder and name it IM3_unit3

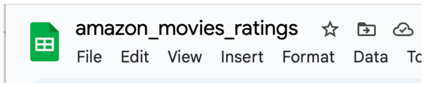




STEP 2

Create a new Google Sheet and name it amazon_movies_ratings

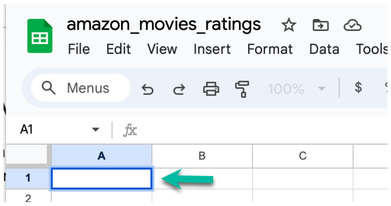




Note: Your Google Sheet file name is different from your imported file name, to reduce entry errors.

STEP 3

Click on Cell A1

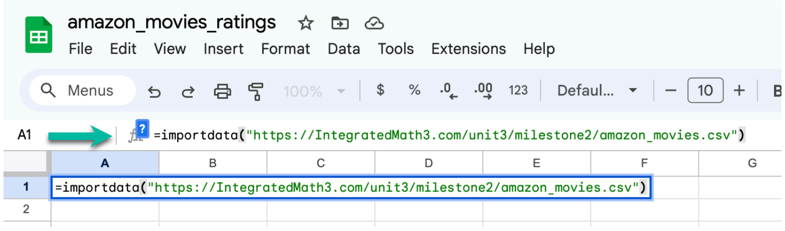


STEP 4

Next, you will key the following into the function field with A1 selected

=importdata("https://IntegratedMath3.com/unit3/milestone2/amazon_movies.csv")

- =importdata tells Sheets to import the dataset in the .csv file
- (__:__) - range sets the set of values whose value you want to calculate- first cell:last cell
- " and the ending " are always needed when importing from a website
- https:// means the site is secure
- IntegratedMath3.com/ is the book's website
- unit3/ is this unit
- milestone2/ is this unit project
- amazon_movies.csv are the data you're importing



```
=importdata("https://IntegratedMath3.com/unit3/milestone2/amazon_movies.csv")
```

STEP 5

Select Enter (return on a Mac). The word Loading appears for a moment, and the data are entered into your Google Sheet.

Look at the data. The first column is the identifier of that record of data. Additional columns include the movie Title, Movie Rating, and Number of Ratings. The first 15 are familiar; you keyed them into the Calculator portion of this Task.

When working with data, you will identify which fields (columns) are relevant to your task and which are not. In the workplace, you will import all data and select which to work with based on your goal. It's good to keep a full dataset intact and pull out what you need rather than reduce from the set because once you develop inferences toward conclusions, often additional questions come up. The foundation of central tendency can be calculated in Google Sheets.

How many rows are in the dataset amazon_movies_ratings? (scroll down to view)

____

Mean

STEP 6

In Google Sheets, scroll to the end of the document, visually checking for empty rows in and at the end of the document.

What is the location of the last data cell in field (column) Movie_Rating? ____________________

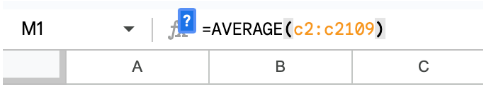
If your document ends at row 2,109, you're ready for Step 7 (1 header row and 2,108 data rows). If there are additional rows at the end that were added in downloading, please delete those empty rows.

STEP 7

Click on cell M1. With M1 selected, in the function field key

=AVERAGE(c2:c2109)

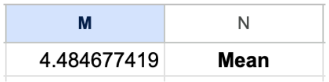
The AVERAGE function asks for all data, located in the range from C2 through C2109 to be included in calculating the mean.



M1 =AVERAGE(c2:c2109)

STEP 8

Key Enter (return on a Mac). The mean calculates and outputs in cell M1. In N1, key Mean, then bold and center the Mean label.



M	N
4.484677419	**Mean**

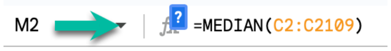
The number of decimal places to keep or omit depends on your goal with the data. Stars are numbers that cluster close together; 2 or 3 decimal places are informative. More than that tend to be overlooked by peers and customers. No decimal places would not work because the number 4.0 offers little to no data for inference.

Median

STEP 9

In Google Sheets, click on cell M2. With M2 selected, in the function field key

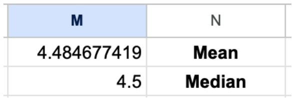
=MEDIAN(c2:c2109)



```
M2    =MEDIAN(C2:C2109)
```

STEP 10

Key Enter (return on a Mac). The median calculates and outputs in cell M2. In N2, key Median, then bold and center the Median label.



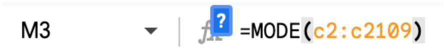
M	N
4.484677419	**Mean**
4.5	**Median**

Mode

STEP 11

In Google Sheets, click on cell M3. With M3 selected, in the function field key

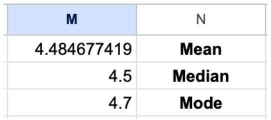
=MODE(c2:c2109)



```
M3    =MODE(c2:c2109)
```

STEP 12

Key Enter (return on a Mac). The mode calculates and outputs in cell M3. In N3, key Mode, then bold and center the Mode label.



M	N
4.484677419	**Mean**
4.5	**Median**
4.7	**Mode**

Your goal is inferences from this data. You know that:

Mean is the arithmetic average of the 2,108 (2,109- row 1 labels) movie ratings

Median is the value at the middle of the list after arranging them by increasing order

Mode is the most frequent value in the list

What do your data infer in the difference between mean, median, and mode from your 15 calculator entries and the 2,108 entries?

____

____

____

Standard Deviation (SD)

For a given set of values, standard deviation measures how much a specific value varies from the mean of the set of values.

There are two ways to calculate SD in Google Sheets:

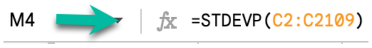
=STDEV(range) calculates the SD of a **sample**

=STDEVP(range) calculates the SD of a **population**

STEP 13

Click on cell M4. With M4 selected, in the function field key

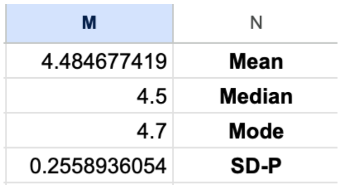
=STDEVP(c2:c2109)



M4 fx =STDEVP(C2:C2109)

STEP 14

Key Enter (return on a Mac). The SD calculates and outputs in cell M4. In N4, key SD, then bold and center the SD-P label.

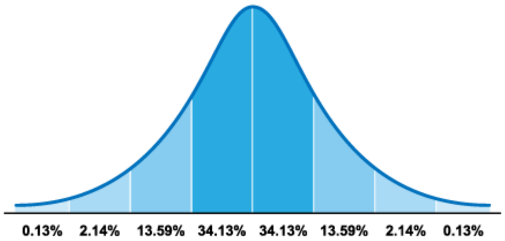


M	N
4.484677419	**Mean**
4.5	**Median**
4.7	**Mode**
0.2558936054	**SD-P**

SD is a single number that summarizes the variability in a dataset. It represents the typical distance between each data point and the mean. Smaller values indicate the data points cluster closer to the mean- their data values are relatively consistent. Higher values indicate the values spread out further from the mean. For this reason, SD is the most widely used measure of variability.

The Amazon movie review data SD of 0.256 (rounded to 3 decimals) is small. You could have guessed that by looking at how close the reviews are, especially with zero 1, 2, and 3-star ratings. Thinking again about your task with the Fair Business Commission, they are not concerned with whether people like a movie, they wonder about how reviews may shift business unfairly.

Looking back to an earlier image of normal distribution



0.256 = 25.6% - reporting what you expected, the movie rating SD is close to the mean

mean ± 1 SD includes 68.3% of the population
mean ± 2 SD includes 95.5% of the population
mean ± 3 SD includes 99.7% of the population

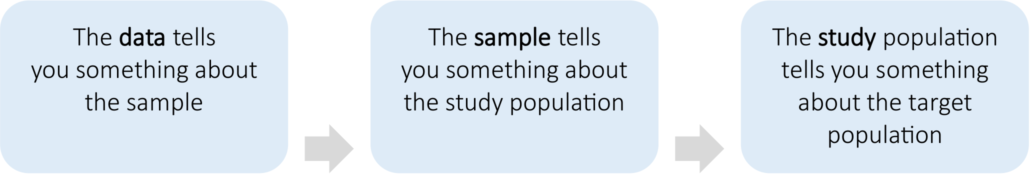
Low variability may equate to consistent quality. On the other hand, low variability raises concerns about why a movie (or any product) is so narrowly rated. For example, select one great movie from the Amazon list you have seen, and one you found boring or didn't finish watching. Ask your peers in class who saw the movies and how many stars, from 1-5, they give them. Ask your family. Ask a couple of friends away from school. If you were to calculate the mean of all responses (your population), would the movie rate a mean of 4.5? People have different tastes in movies and watch on days when many other things are going on, or they're tired or hungry; many elements go into a movie review that could lead to 3- and 2-star ratings. For the Fair Business Commission, sometimes great is too good. That's where data inference comes in and questions how and where fair business practice stumbles.

STEP 15

Please close your amazon_movie_ratings Google Sheet.

Random Sampling

Your goal is for the data to be reliable, having low variability from sample to sample.



The **data** tells you something about the sample

The **sample** tells you something about the study population

The **study** population tells you something about the target population

Going from sample to study population depends on the quality of the study, also known as **internal validity.** Does your sample accurately reflect what is going on in the group you study?

Going from study population to target papulation requires **external validity**. If external validity is established, it means that the findings can be generalizable to similar individuals or populations.

You'll learn more about internal and external validity soon; for now, the best way to avoid bias is with random sampling. It's common to have 10,000+ rows in a public dataset. It's also common to have 100 with no way to survey the full population due to logistics like location or access.

Random sampling:

- also known as probability sampling, is a method that allows for the randomization of a sample selection
- is designed to remove bias
- is not typically representative of the population; variations are referred to as sampling errors
- helps to cancel the effects of unobserved factors

There are four primary random (probability) sampling **methods**:

simple random sampling

systematic sampling

stratified sampling

cluster sampling

Presented in a future unit with Probability

Simple random sampling randomizes the selection of a small segment from a whole population, with an equal and fair probability of being chosen. The simple random sampling method is a convenient and simple sample selection technique.

STEP 1

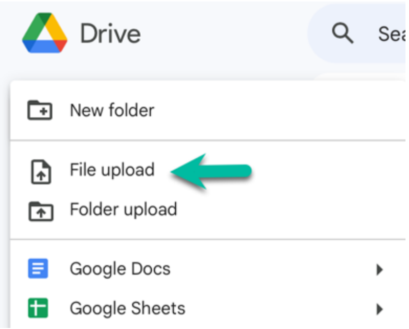
Please open a new browser tab and key the link below into the URL locator (Note- use Google Chrome since you're working with Google Sheets):

https://integratedmath3.com/unit3/milestone2/amazon_movies_random_sampling.csv

A .csv file named amazon_movies_random_sampling.csv will download. You're downloading this time because Google Sheets does not allow data sorting on imported spreadsheets.

STEP 2

Open your Google Drive folder IM3_unit3. Select New > File upload > and upload the amazon_movies_random_sampling.csv file, it may be in your Downloads folder.



STEP 3

Name your spreadsheet amazon_movies_sampling_methods

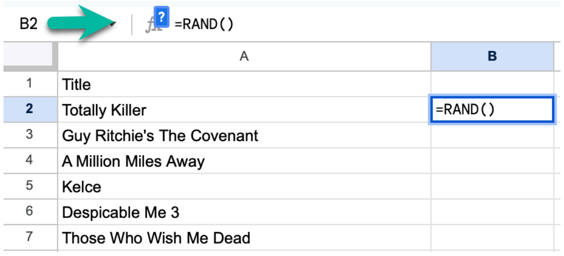
Note: Your Google Sheet file name is different from your imported file name, to reduce entry errors.

You'll see the field Title is the same as your original table. You don't need the rest of the data for the purpose of practicing sampling methods. Expand the width of Column A if you'd like to view the full movie title by selecting the line separating heading A and B, and dragging it to the right, far enough for the movie title to show in full.

STEP 4

Click on cell B2. With B2 selected, in the function field key

=RAND ( )

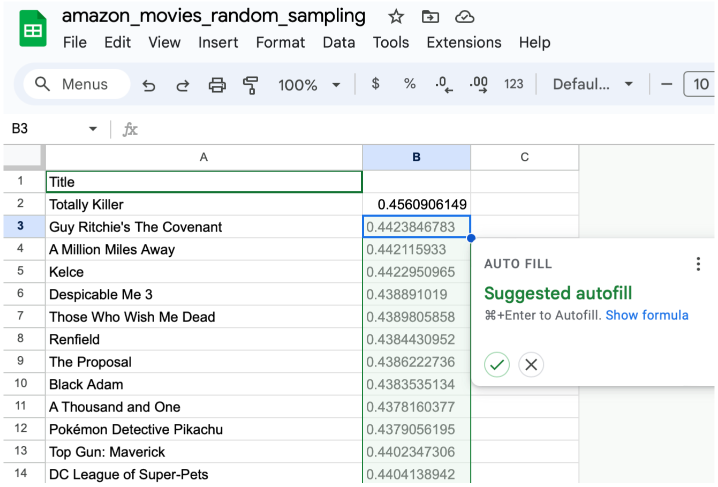


B2 =RAND()

	A	B
1	Title	
2	Totally Killer	=RAND()
3	Guy Ritchie's The Covenant	
4	A Million Miles Away	
5	Kelce	
6	Despicable Me 3	
7	Those Who Wish Me Dead	

STEP 5

Key Enter (return on a Mac). Select the green checkbox in Suggested autofill (you may need to click on it twice).



amazon_movies_random_sampling
File Edit View Insert Format Data Tools Extensions Help
Menus 100% $ % .0 .00 123 Defaul... 10

B3

	A	B	C
1	Title		
2	Totally Killer	0.4560906149	
3	Guy Ritchie's The Covenant	0.4423846783	
4	A Million Miles Away	0.442115933	
5	Kelce	0.4422950965	
6	Despicable Me 3	0.438891019	
7	Those Who Wish Me Dead	0.4389805858	
8	Renfield	0.4384430952	
9	The Proposal	0.4386222736	
10	Black Adam	0.4383535134	
11	A Thousand and One	0.4378160377	
12	Pokémon Detective Pikachu	0.4379056195	
13	Top Gun: Maverick	0.4402347306	
14	DC League of Super-Pets	0.4404138942	

AUTO FILL

Suggested autofill

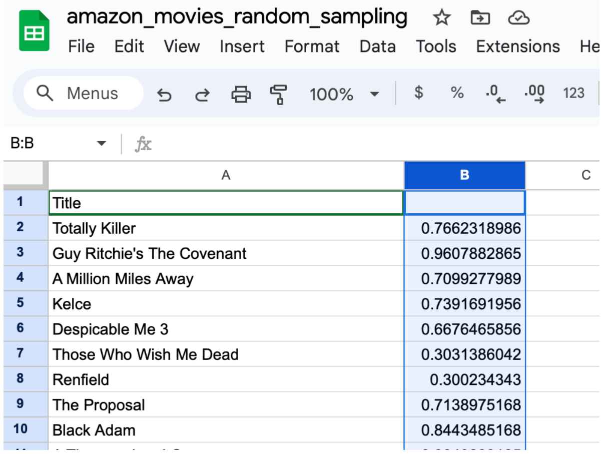
⌘+Enter to Autofill. Show formula

✓ ✕

If you make a mistake or want to redo =RAND(), each time you'll see a different set of numbers. =RAND() is a random number generator; output will be random each time you call the function.

STEP 6

Highlight column B by clicking on the B header cell.

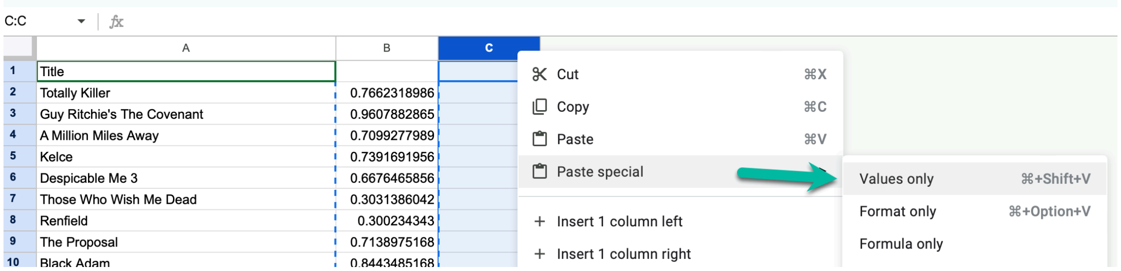


	A	B	C
1	Title		
2	Totally Killer	0.7662318986	
3	Guy Ritchie's The Covenant	0.9607882865	
4	A Million Miles Away	0.7099277989	
5	Kelce	0.7391691956	
6	Despicable Me 3	0.6676465856	
7	Those Who Wish Me Dead	0.3031386042	
8	Renfield	0.300234343	
9	The Proposal	0.7138975168	
10	Black Adam	0.8443485168	

STEP 7

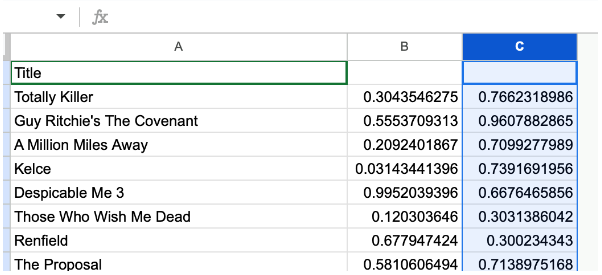
Ctrl+C (Cmd+C on a Mac) to copy all values in column B.

Right-click on cell **C1** and choose **Paste Special > Paste values only**



	A	B	C
1	Title		
2	Totally Killer	0.7662318986	
3	Guy Ritchie's The Covenant	0.9607882865	
4	A Million Miles Away	0.7099277989	
5	Kelce	0.7391691956	
6	Despicable Me 3	0.6676465856	
7	Those Who Wish Me Dead	0.3031386042	
8	Renfield	0.300234343	
9	The Proposal	0.7138975168	
10	Black Adam	0.8443485168	

The values in column B will change, your purpose is only to generate random numbers.



A	B	C
Title		
Totally Killer	0.3043546275	0.7662318986
Guy Ritchie's The Covenant	0.5553709313	0.9607882865
A Million Miles Away	0.2092401867	0.7099277989
Kelce	0.03143441396	0.7391691956
Despicable Me 3	0.9952039396	0.6676465856
Those Who Wish Me Dead	0.120303646	0.3031386042
Renfield	0.677947424	0.300234343
The Proposal	0.5810606494	0.7138975168

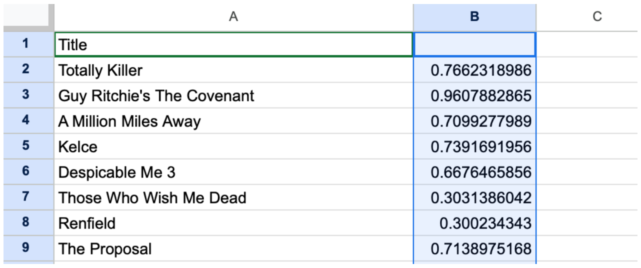
STEP 8

Click on the C header to select all of column C

STEP 9 — Ctrl+X (PC) or Cmd+X (Mac) to cut

STEP 10 — Click on B to select all of column B

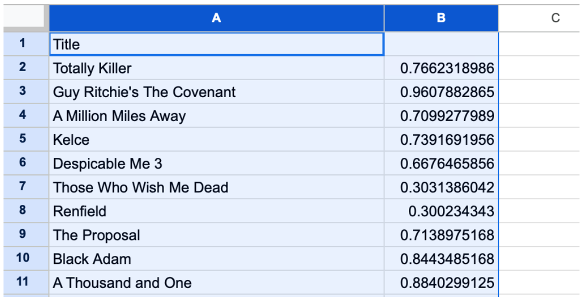
STEP 11 — Ctrl+V (PC) or Cmd+V (Mac) to paste column C over the top of Column B



	A	B	C
1	Title		
2	Totally Killer	0.7662318986	
3	Guy Ritchie's The Covenant	0.9607882865	
4	A Million Miles Away	0.7099277989	
5	Kelce	0.7391691956	
6	Despicable Me 3	0.6676465856	
7	Those Who Wish Me Dead	0.3031386042	
8	Renfield	0.300234343	
9	The Proposal	0.7138975168	

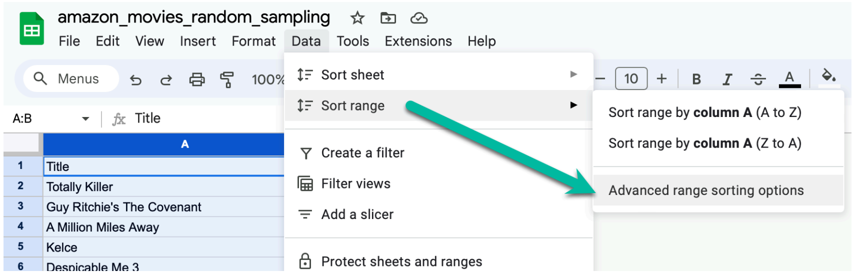
If you have an error or get tangled up, repeating the process a second time is fast.

STEP 12 — Slowly drag your mouse over column headings A and B to select both columns. Or select column A and holding shift, right arrow across to select column B



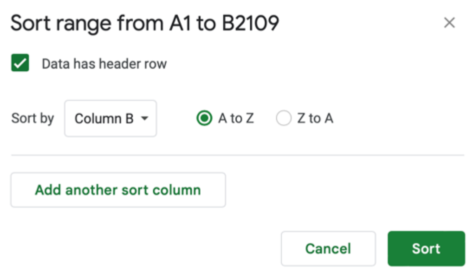
	A	B	C
1	Title		
2	Totally Killer	0.7662318986	
3	Guy Ritchie's The Covenant	0.9607882865	
4	A Million Miles Away	0.7099277989	
5	Kelce	0.7391691956	
6	Despicable Me 3	0.6676465856	
7	Those Who Wish Me Dead	0.3031386042	
8	Renfield	0.300234343	
9	The Proposal	0.7138975168	
10	Black Adam	0.8443485168	
11	A Thousand and One	0.8840299125	

STEP 13 — Select the Data tab along the top ribbon > Sort Range > Advanced range sorting options



STEP 14

Check Data has a header row so that row 1, with Title doesn't become a part of the data sort. Sort by Column B, from A to Z

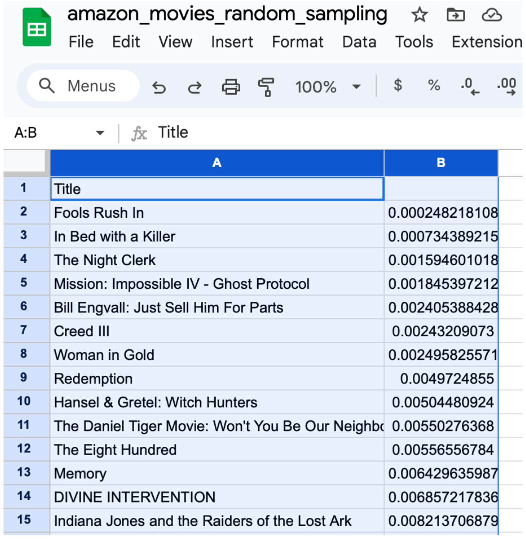


Sort range from A1 to B2109

☑ Data has header row

Sort by Column B ● A to Z ○ Z to A

Note: If your sort range is longer than your 2109 Amazon movies lines, scroll to the bottom of your spreadsheet and delete any extra empty rows.



amazon_movies_random_sampling

A:B fx Title

	A	B
1	Title	
2	Fools Rush In	0.000248218108
3	In Bed with a Killer	0.000734389215
4	The Night Clerk	0.001594601018
5	Mission: Impossible IV - Ghost Protocol	0.001845397212
6	Bill Engvall: Just Sell Him For Parts	0.002405388428
7	Creed III	0.00243209073
8	Woman in Gold	0.002495825571
9	Redemption	0.0049724855
10	Hansel & Gretel: Witch Hunters	0.00504480924
11	The Daniel Tiger Movie: Won't You Be Our Neighbo	0.00550276368
12	The Eight Hundred	0.00556556784
13	Memory	0.006429635987
14	DIVINE INTERVENTION	0.006857217836
15	Indiana Jones and the Raiders of the Lost Ark	0.008213706879

Here's what happened:

You set up a random number generator associated with your raw data, Title. You can't set up a random number generator on multiple fields (columns) because the number associates with each datum (singular for data).

Check Data has a header row so that row 1, with Title doesn't become a part of the data sort. Sort by Column B, from A to Z.

A few things to note:

- Google Sheets does not perform =RAND() from an imported spreadsheet because of changes that can happen with the import origination file causing a constant redo of the random generation.
- The sample size (in this case movies) you will select from a population varies, based on organizational protocols, common practice, etc.
- For most studies, this method satisfied randomness. If this were a graduate school mathematics study, you'd use more sophisticated software to select a random sample.

Conroy (2018) details:

- the ideal sample size is one that collects sufficient data to have a good chance of measuring what you set out to measure
- Key issues: will the sample be representative of the population? Will the sample be precise enough?
- An unrepresentative sample will result in biased conclusions, and the bias cannot be eliminated by taking a larger sample
- The larger the sample, the smaller the margin of uncertainty (confidence interval) around the results
- The more something varies, the bigger your sample needs to be to achieve the same degree of certainty (p. 4)

Conroy, Ronán. (2018). The RCSI Sample size handbook. 10.13140/RG.2.2.30497.51043.

A quick, common guide for **practice studies** is:

A sample size minimum of 100

A maximum sample size of 10% of the population

Since there is little variance among ratings of Amazon movies, your sample size will be the first 100 rows in the spreadsheet, after random sampling was applied. If you decided on 10%, your sample size would be 210 rows. Using a random number generator satisfies your commitment to avoid bias in your recommendations for the Fair Business Commission.

Please select the first 100 rows in your random sample. Compare your rows with another student's. Why are your first 100 different than their first 100?

____

____

____

STEP 15

Please close your amazon_movies_random_sampling Google Sheet.

Systematic sampling is probability sampling where researchers select specific data from a whole population. Selection often follows a predetermined regular interval (k). The systematic sampling method is comparable to simple random sampling and can be less complicated to conduct.

In random sampling, you used a random number generator. With systematic sampling, you will select data at intervals, like every 10th product or person in the study. Systematic sampling is stronger when applied after random sampling; however, it can be applied without random sampling when there is a low risk of data manipulation.

For example, a company's internal study. Since the Fair Business Commission expects data tampering, you will add systematic sampling to your random sampling.

Your data population count is ________________________________

For systematic sampling, you will still want 100 in your sample. Based on your Amazon movies data population, and a 100-row sample, what will be your predetermined regular interval (k)?

The predetermined regular interval is every ___th (for example every 8th or 15th) to result in 100. Your (k) can be any number; think about dividing what you need by the available population.

____

____

____

Please list what will be the first 5 of your 100, drawn from your predetermined regular interval. If you decided on every 9th, your (k) will be the 9th, 18th, 27th, 36th, 45th row movie.

1. ____

2. ____

3. ____

4. ____

5. ____

Please select the first 100 rows in your random sample. Compare your rows with another student's. Why are your first 100 different than their first 100?

____

____

____

Compare your five with two other students. Are there any overlaps? Discuss and list how this approach supports the information the Fair Business Commission goals. How does systematic sampling help you to select a sample from the population?

____

____

____

Stratified sampling includes partitioning a population into subclasses with distinctions and variances. The stratified sampling method allows the researcher to make more reliable and informed conclusions by confirming that each respective subclass has been represented in the selected sample.

STEP 1

Open your IM3_unit3 Google Drive. Create a new spreadsheet and name it amazon_movies_stratified_sampling

STEP 2

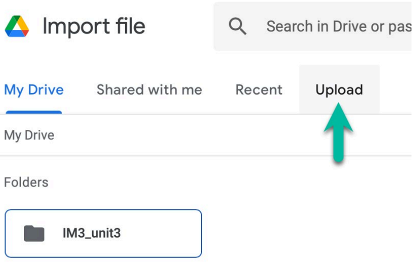
Please open a new browser tab and key the link below into the URL locator (Note- use Google Chrome since you're working with Google Sheets)

https://integratedmath3.com/unit3/milestone2/amazon_movies.csv

A .csv file named amazon_movies.csv will download. You previously imported this .csv file. You're downloading it this time because Google Sheets does not allow data sorting on imported spreadsheets.

STEP 3

In the ribbon, select File > Import, and Upload

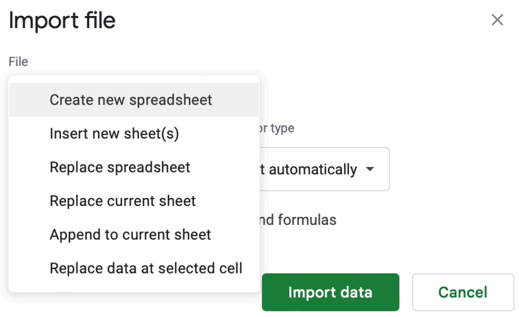


STEP 4

Browse for or drag your amazon_movies.csv file to import

If there's an issue with managing files, you might see this popup. If so, select Create new spreadsheet and then name it amazon_movies_stratified_sampling.

If you have a few data points from practicing, you can select Replace spreadsheet to import the dataset over your practice file.



Looking at the data, how might partitioning a population into subclasses offer information for you to report on?

____

____

How might sampling by the following fields offer insight for inference data on assessing fake reviews?

____

____

Please discuss and list a question for each:

Example: Number of Ratings

Does larger or fewer numbers of ratings affect the movie rating mean? If so, how far away from the mean (SD)?

Release Year ____

____

MPAA Rating ____

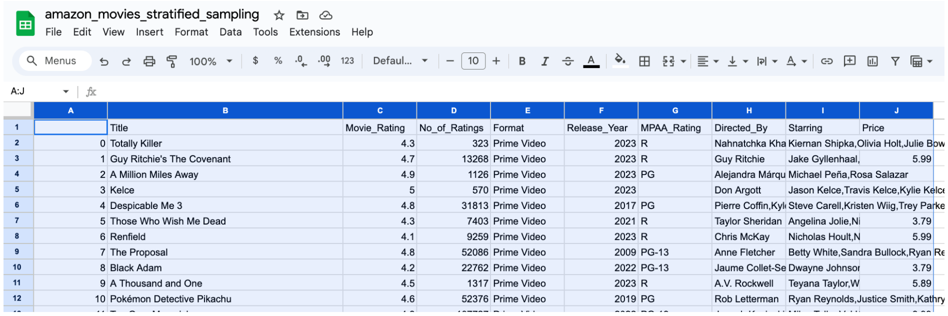
____

Select one of the three (number of ratings, release year, or MPAA rating) you infer may contribute to narrowing in on a **subclass** of data to study. Subclasses are also referred to as **strata**. The following steps address the example, Number of Ratings. You will follow the same Google Sheets steps with this or one of the other two strata.

Note: Whenever you sort data in any spreadsheet, all the fields (columns) need to be selected. If not, your one column sorts and all others remain the same. For example, if you sort only column F, the Release_Year, that column will sort the columns associated with it will not sort.

STEP 5

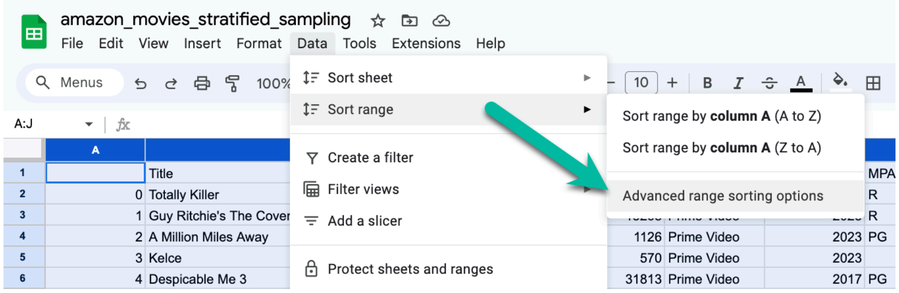
Select column A and slowly drag to select A-J or use Shift+right-arrow to select all columns


amazon_movies_stratified_sampling
File Edit View Insert Format Data Tools Extensions Help
Menus
100%
Defaul...
10
A:J

	A	B	C	D	E	F	G	H	I	J
1		Title	Movie_Rating	No_of_Ratings	Format	Release_Year	MPAA_Rating	Directed_By	Starring	Price
2	0	Totally Killer	4.3	323	Prime Video	2023	R	Nahnatchka Kha	Kiernan Shipka,Olivia Holt,Julie Bow	
3	1	Guy Ritchie's The Covenant	4.7	13268	Prime Video	2023	R	Guy Ritchie	Jake Gyllenhaal,	5.99
4	2	A Million Miles Away	4.9	1126	Prime Video	2023	PG	Alejandra Márqu	Michael Peña,Rosa Salazar	
5	3	Kelce	5	570	Prime Video	2023		Don Argott	Jason Kelce,Travis Kelce,Kylie Kelce	
6	4	Despicable Me 3	4.8	31813	Prime Video	2017	PG	Pierre Coffin,Kyl	Steve Carell,Kristen Wiig,Trey Parke	
7	5	Those Who Wish Me Dead	4.3	7403	Prime Video	2021	R	Taylor Sheridan	Angelina Jolie,Ni	3.79
8	6	Renfield	4.1	9259	Prime Video	2023	R	Chris McKay	Nicholas Hoult,N	5.99
9	7	The Proposal	4.8	52086	Prime Video	2009	PG-13	Anne Fletcher	Betty White,Sandra Bullock,Ryan Re	
10	8	Black Adam	4.2	22762	Prime Video	2022	PG-13	Jaume Collet-Se	Dwayne Johnsor	3.79
11	9	A Thousand and One	4.5	1317	Prime Video	2023	R	A.V. Rockwell	Teyana Taylor,W	5.89
12	10	Pokémon Detective Pikachu	4.6	52376	Prime Video	2019	PG	Rob Letterman	Ryan Reynolds,Justice Smith,Kathry	

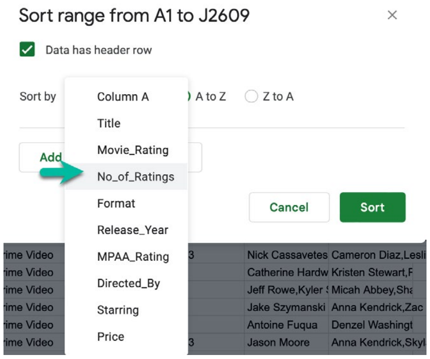
STEP 6

From the ribbon, select Data > Sort range > Advanced range sorting options


amazon_movies_stratified_sampling
File Edit View Insert Format Data Tools Extensions Help
Menus
100%
A:J
Sort sheet
Sort range
Create a filter
Filter views
Add a slicer
Protect sheets and ranges
Sort range by column A (A to Z)
Sort range by column A (Z to A)
Advanced range sorting options
10

STEP 7

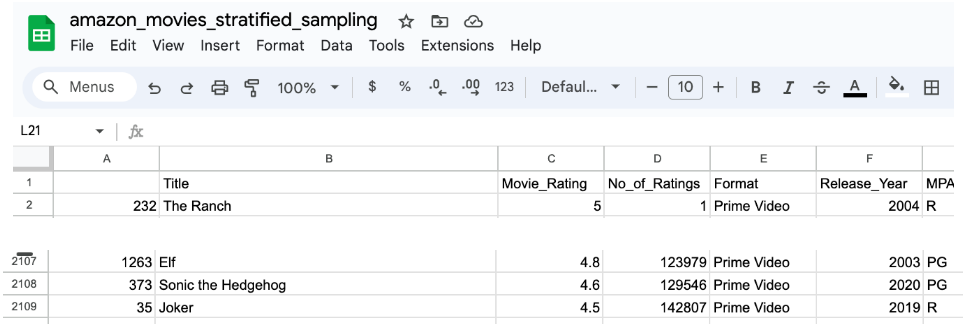
Select Data has header row > No_of_ratings (or your field)



Sort range from A1 to J2609

☑ Data has header row

Your sort starts with the lowest number of ratings (1) and ends with the largest number of ratings (142,807). Stratified sampling subclasses would, in this case, identify ranges that may better inform patterns with ratings.



amazon_movies_stratified_sampling

	A	B	C	D	E	F	
1		Title	Movie_Rating	No_of_Ratings	Format	Release_Year	MPA
2	232	The Ranch	5	1	Prime Video	2004	R
2107	1263	Elf	4.8	123979	Prime Video	2003	PG
2108	373	Sonic the Hedgehog	4.6	129546	Prime Video	2020	PG
2109	35	Joker	4.5	142807	Prime Video	2019	R

What subgroups would you consider for:

Release Year ____

____

MPAA Rating ____

____

STEP 8

Please close your amazon_movies_stratified_sampling Google Sheet

Data Cleaning

In addition to determining sampling methodologies for working with data, it's essential to assess the quality of your data. In the case of Amazon movie reviews, the data are provided by scraping the web. Web scraping, web harvesting, or web data extraction extract data from websites. Scraping takes place by the company, in this case Amazon, by data science organizations, and others. The challenge is although scraping is an efficient option to assess content, the nature of web data is that quality varies.

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

When you work with large amounts of data in a spreadsheet, expect duplicate records and missing data. Whether through human error or robots that put them there, duplicates and missing data affect your workflows, documentation, and data analysis.

You can quickly remove duplicates in Google Sheets; however, first determine if an automated cleanup or a manual cleanup will best support your efforts.

STEP 1

In your Google Drive, open amazon_movies.csv

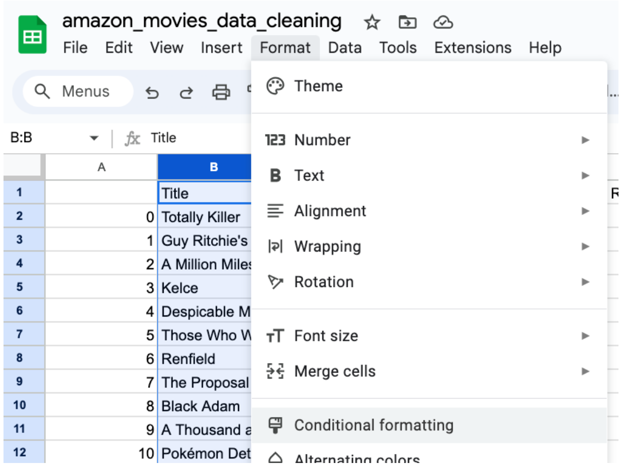
STEP 2

In your ribbon, select File > Make a copy and name the copy amazon_movies_data_cleaning

STEP 3

Highlight column B Title to check for duplicate information

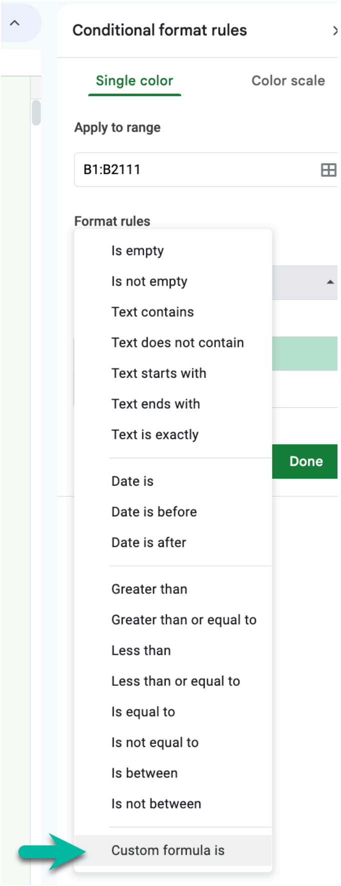
Select Format > Conditional Formatting



	A	B
1		Title
2	0	Totally Killer
3	1	Guy Ritchie's
4	2	A Million Miles
5	3	Kelce
6	4	Despicable M
7	5	Those Who W
8	6	Renfield
9	7	The Proposal
10	8	Black Adam
11	9	A Thousand a
12	10	Pokémon Det

STEP 4

From the Conditional format rules window that appears, click the dropdown menu under Format rules, and select Custom formula is



Conditional format rules

Single color | Color scale

Apply to range

B1:B2111

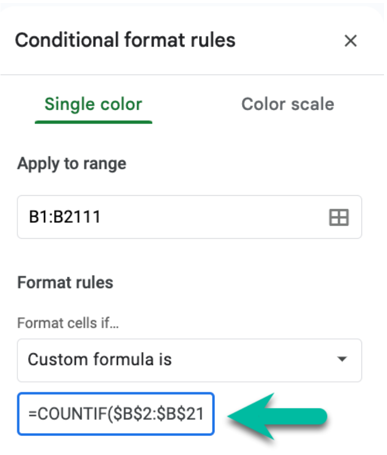
Format rules

STEP 5

Enter a custom duplicate checking formula in the Value or formula bar.

In this example, we're looking for duplicates in cells B2:B15, so the custom formula is

=COUNTIF($B$2:$B$2109,B2)>1



Conditional format rules

Single color | Color scale

Apply to range

B1:B2111

Format rules

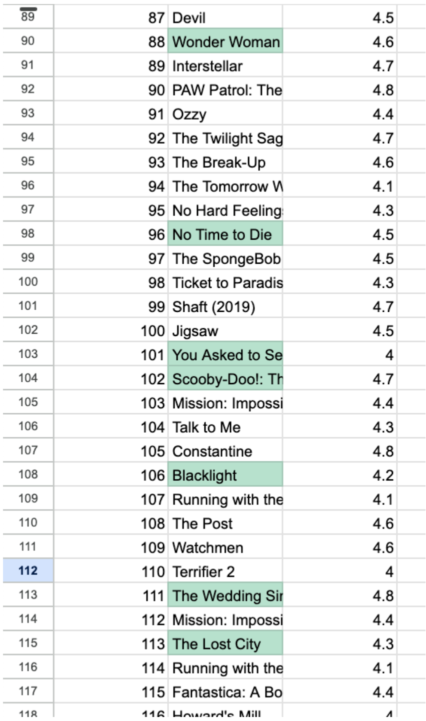
Format cells if…

Custom formula is

=COUNTIF($B$2:$B$21

STEP 6

Select Done. Scroll down and you'll see duplicates in green. The next question is whether they are duplicates or if the two are different and should be added together

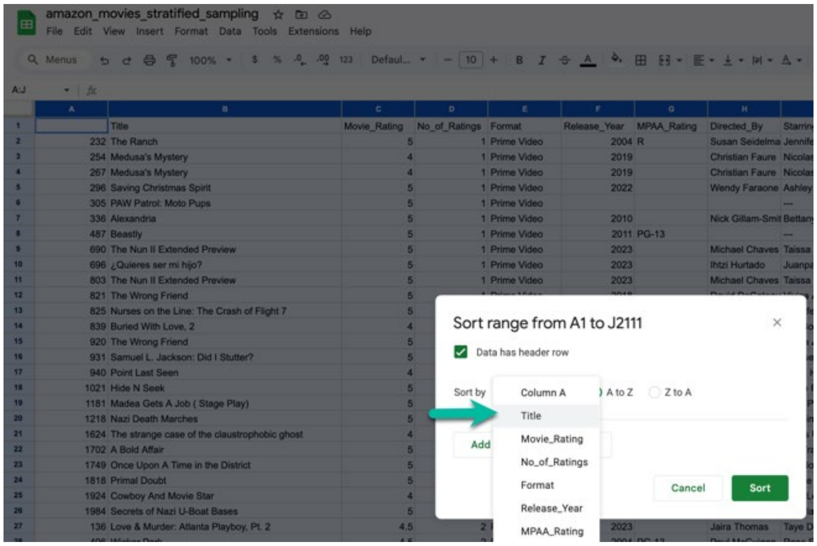


89	87	Devil	4.5
90	88	Wonder Woman	4.6
91	89	Interstellar	4.7
92	90	PAW Patrol: The	4.8
93	91	Ozzy	4.4
94	92	The Twilight Sag	4.7
95	93	The Break-Up	4.6
96	94	The Tomorrow W	4.1
97	95	No Hard Feeling	4.3
98	96	No Time to Die	4.5
99	97	The SpongeBob	4.5
100	98	Ticket to Paradis	4.3
101	99	Shaft (2019)	4.7
102	100	Jigsaw	4.5
103	101	You Asked to Se	4
104	102	Scooby-Doo!: Th	4.7
105	103	Mission: Impossi	4.4
106	104	Talk to Me	4.3
107	105	Constantine	4.8
108	106	Blacklight	4.2
109	107	Running with the	4.1
110	108	The Post	4.6
111	109	Watchmen	4.6
112	110	Terrifier 2	4
113	111	The Wedding Sir	4.8
114	112	Mission: Impossi	4.4
115	113	The Lost City	4.3
116	114	Running with the	4.1
117	115	Fantastica: A Bo	4.4
118	116	Howard's Mill	4

It's not unusual to assume you have a duplicate and it's fast to enter a different function to delete. However, more often you will want to take a closer look at the data first. This is called a manual clean versus an automated clean.

STEP 7

Select all columns. Sort by Title



amazon_movies_stratified_sampling
File Edit View Insert Format Data Tools Extensions Help
Menus
100%
Defaul...
10
A:J

Sort range from A1 to J2111

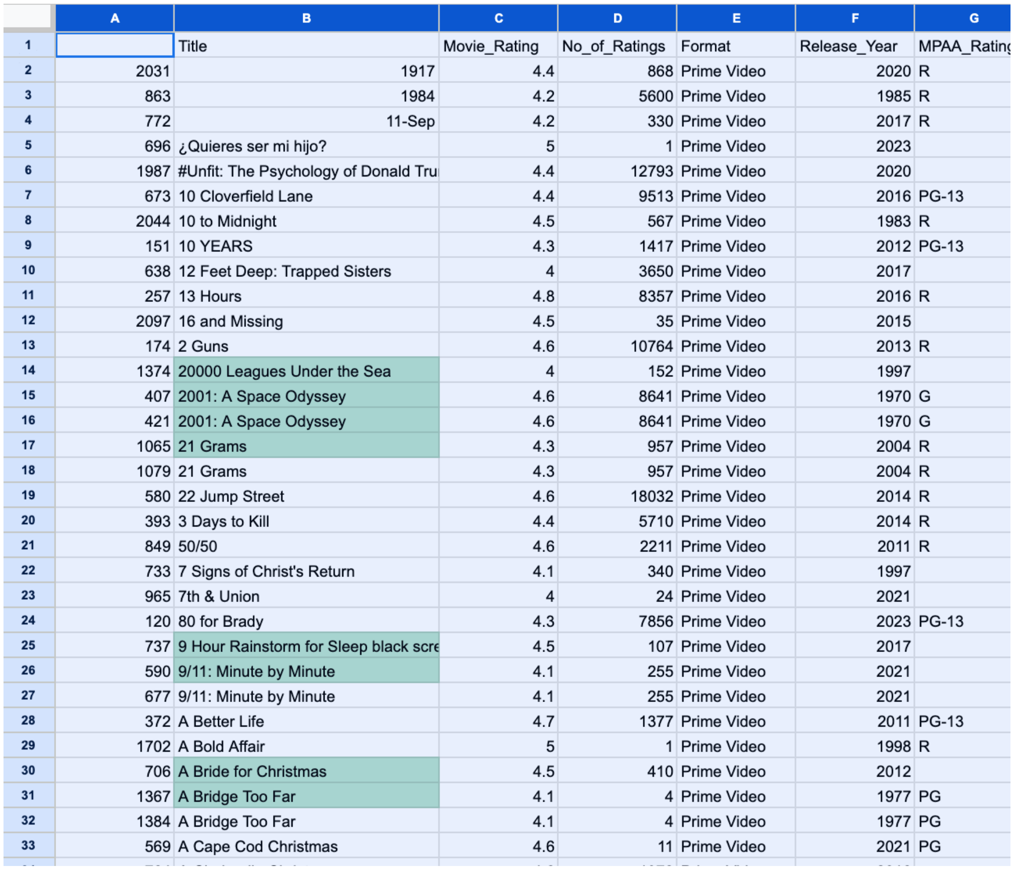
☑ Data has header row

Sort by: Title (Column A, Title, Movie_Rating, No_of_Ratings, Format, Release_Year, MPAA_Rating) ● A to Z ○ Z to A

Add

Cancel | Sort

	A	B	C	D	E	F	G	H	
1		Title	Movie_Rating	No_of_Ratings	Format	Release_Year	MPAA_Rating	Directed_By	Starrin
2	232	The Ranch	5	1	Prime Video	2004	R	Susan Seidelma	Jennife
3	254	Medusa's Mystery	4	1	Prime Video	2019		Christian Faure	Nicolas
4	267	Medusa's Mystery	4	1	Prime Video	2019		Christian Faure	Nicolas
5	296	Saving Christmas Spirit	5	1	Prime Video	2022		Wendy Faraone	Ashley
6	305	PAW Patrol: Moto Pups	5	1	Prime Video				—
7	336	Alexandria	5	1	Prime Video	2010		Nick Gillam-Smit	Bettany
8	487	Beastly	5	1	Prime Video	2011	PG-13		—
9	690	The Nun II Extended Preview	5	1	Prime Video	2023		Michael Chaves	Taissa
10	696	¿Quieres ser mi hijo?	5	1	Prime Video	2023		Ihtzi Hurtado	Juanpa
11	803	The Nun II Extended Preview	5	1	Prime Video	2023		Michael Chaves	Taissa
12	821	The Wrong Friend	5						
13	825	Nurses on the Line: The Crash of Flight 7	5						
14	839	Buried With Love, 2	4						
15	920	The Wrong Friend	5						
16	931	Samuel L. Jackson: Did I Stutter?	5						
17	940	Point Last Seen	4						
18	1021	Hide N Seek	5						
19	1181	Madea Gets A Job (Stage Play)	5						
20	1218	Nazi Death Marches	5						
21	1624	The strange case of the claustrophobic ghost	4						
22	1702	A Bold Affair	5						
23	1749	Once Upon A Time in the District	5						
24	1818	Primal Doubt	5						
25	1924	Cowboy And Movie Star	4						
26	1984	Secrets of Nazi U-Boat Bases	5						
27	136	Love & Murder: Atlanta Playboy, Pt. 2	4.5	2		2023		Jaira Thomas	Taye D



	A	B	C	D	E	F	G
1		Title	Movie_Rating	No_of_Ratings	Format	Release_Year	MPAA_Ratinc
2	2031	1917	4.4	868	Prime Video	2020	R
3	863	1984	4.2	5600	Prime Video	1985	R
4	772	11-Sep	4.2	330	Prime Video	2017	R
5	696	¿Quieres ser mi hijo?	5	1	Prime Video	2023	
6	1987	#Unfit: The Psychology of Donald Tru	4.4	12793	Prime Video	2020	
7	673	10 Cloverfield Lane	4.4	9513	Prime Video	2016	PG-13
8	2044	10 to Midnight	4.5	567	Prime Video	1983	R
9	151	10 YEARS	4.3	1417	Prime Video	2012	PG-13
10	638	12 Feet Deep: Trapped Sisters	4	3650	Prime Video	2017	
11	257	13 Hours	4.8	8357	Prime Video	2016	R
12	2097	16 and Missing	4.5	35	Prime Video	2015	
13	174	2 Guns	4.6	10764	Prime Video	2013	R
14	1374	20000 Leagues Under the Sea	4	152	Prime Video	1997	
15	407	2001: A Space Odyssey	4.6	8641	Prime Video	1970	G
16	421	2001: A Space Odyssey	4.6	8641	Prime Video	1970	G
17	1065	21 Grams	4.3	957	Prime Video	2004	R
18	1079	21 Grams	4.3	957	Prime Video	2004	R
19	580	22 Jump Street	4.6	18032	Prime Video	2014	R
20	393	3 Days to Kill	4.4	5710	Prime Video	2014	R
21	849	50/50	4.6	2211	Prime Video	2011	R
22	733	7 Signs of Christ's Return	4.1	340	Prime Video	1997	
23	965	7th & Union	4	24	Prime Video	2021	
24	120	80 for Brady	4.3	7856	Prime Video	2023	PG-13
25	737	9 Hour Rainstorm for Sleep black scre	4.5	107	Prime Video	2017	
26	590	9/11: Minute by Minute	4.1	255	Prime Video	2021	
27	677	9/11: Minute by Minute	4.1	255	Prime Video	2021	
28	372	A Better Life	4.7	1377	Prime Video	2011	PG-13
29	1702	A Bold Affair	5	1	Prime Video	1998	R
30	706	A Bride for Christmas	4.5	410	Prime Video	2012	
31	1367	A Bridge Too Far	4.1	4	Prime Video	1977	PG
32	1384	A Bridge Too Far	4.1	4	Prime Video	1977	PG
33	569	A Cape Cod Christmas	4.6	11	Prime Video	2021	PG

There are several data issues to assess:

Line 4: Is 11-Sep a movie title or an error? ____

Look at lines 14-17. What about line 18, it's not highlighted. Are they all duplicates, or which is a duplicate?

____

____

What is your recommendation for cleaning lines 25-27?

____

____

What is your recommendation for cleaning lines 30-32?

____

____

What is your recommendation for cleaning lines 60-63?

____

____

What is your recommendation for cleaning lines 134-139?

____

____

You can see the amount of time it will take to address all 2,108 data points. And why it's important to not automate a match and delete without reviewing. Another challenge is missing data points. MPAA_Rating missing data on lines 5, 6, 10, 12, etc. would need to be researched and manually entered.

When a dataset has a clear set of duplicates, in Google Sheets:

STEP 8

Click any cell that contains data

Then, select the Data tab > Data cleanup > Remove duplicates

Your dataset is now ready for future analysis

You practiced working with sampling methods and assessing data quality for missing or duplicate data points. In the Unit 3 Deliverable, you will apply what you practiced to a different dataset.

Milestone 2 Check-in Meeting

Per your teacher's instructions, meet in small teams to assess and discuss the following Milestone 2 questions:

1. What study sample(s) did you practice and what are their differences?

   ____

   ____

2. What are some of the challenges of an observational study like Amazon movies?

   ____

   ____

3. How can you tell what percentage of the numbers in your study are close to the average and how close?

   ____

4. What have you observed about the Amazon movies dataset that will contribute to you forming conclusions about fake reviews?

   ____

Unit 3 Deliverable

It's important to apply what you practice to different contexts. The Unit 3 Deliverable dataset addresses product reviews and the steps of:

- Data cleaning
- Random sampling
- Inference from data
- Measurement

Your teacher will supply you with a dataset, assessment rubric, and an instruction sheet to guide you through developing a recommendation for a current Fair Business Commission Practice project.

In Unit 4 your business use case changes from online reviews to an entirely different industry sector, semiconductor chips. You will apply what you learned in Unit 3 and expand using statistics to draw inferences and justify conclusions. The purpose of changing use cases is to develop your problem-solving skills utilizing data to inform decision-making across multiple business types.

Works Cited

Banerjee, S., & Chua, A. (2023). Understanding online fake review production strategies. *Journal of Business Research, 156*, 113534. https://doi.org/10.1016/j.jbusres.2022.113534

Collins, J. (2022, June 5). *4.1 million-square-foot warehouse in California will be Amazon's biggest ever*. The Seattle Times. https://www.seattletimes.com/business/4-1-million-square-foot-warehouse-in-california-will-be-amazons-biggest-ever/

Cramer, M. (2023, January 25). *How Sites Like Tripadvisor and Yelp Are Fighting Fake Reviews*. The New York Times. https://www.nytimes.com/2023/01/25/travel/fake-review-investigators.html

European Commission. (2021). Sweeps. https://commission.europa.eu/live-work-travel-eu/consumer-rights-and-complaints/enforcement-consumer-protection/sweeps_en#ref-2021--sweep-on-online-consumer-reviews

Filho, M.C., Rafael, D.N., Barros, L.S., & Mesquita, E. (2023). Mind the fake reviews! Protecting consumers from deception through persuasion knowledge acquisition. *Journal of Business Research*, *156*, 113538. https://doi.org/10.1016/j.jbusres.2022.113538

Gaynor, A., Levine, S., & Fair, L. (2022, September 13). *What companies – and platforms – can do to help stop fake post-for-pay reviews*. Federal Trade Commission. https://www.ftc.gov/business-guidance/blog/2022/09/what-companies-and-platforms-can-do-help-stop-fake-post-pay-reviews

Giacobone, B. (2023, February 17). *Truckers waste over a billion hours sitting in traffic every year, making shipping even more expensive. See the 10 worst bottlenecks.* Business Insider. https://www.businessinsider.com/worst-highway-shipping-bottlenecks-trucks-traffic-congestion-freight-2023-2

Hagman, C. (2023, January 25). *The Inland Empire benefits from the warehouse industry.* San Bernardino Sun. https://www.pressenterprise.com/2023/01/25/the-inland-empire-benefits-from-our-warehouses/

Harris, C. G. (2022). Detecting fraudulent online Yelp reviews using K-L divergence and linguistic features. *Procedia Computer Science*, *204*, 618-626. https://doi.org/10.1016/j.procs.2022.08.075

Harris, J. (2022, September 11). *$10000 for one Instagram post? How food influencers can make or break restaurants.* Los Angeles Times. https://www.latimes.com/food/story/2022-09-11/food-influencers-ethics-fees-charged

Jozsa, E. (2023, March 28). *U.S. National Industrial Report March 2023*. CommercialEdge. https://www.commercialedge.com/blog/national-industrial-report/

Kaneko, A. (2022, May 2). *Warehouses Pave Over Historic Dairy Lands in Ontario and Chino*. KCET. https://www.kcet.org/shows/earth-focus/warehouses-pave-over-historic-dairy-lands-in-ontario-and-chino

King, K. (2023, March 28). *Rapid Warehouse Growth Sparks Local Resident Backlash Across the U.S*. The Wall Street Journal. https://www.wsj.com/articles/rapid-warehouse-growth-sparks-local-resident-backlash-across-the-u-s-53502618

Kleinman, Z. (2022, May 6). *Amazon targets review firms with legal action*. BBC. https://www.bbc.com/news/technology-61348521

Morris, C. (2022, July 19). *Inside the thriving business of writing fake Amazon reviews.* Fast Company. https://www.fastcompany.com/90770579/amazon-fake-reviews-facebook-groups-lawsuit

Newton, J., & Osborn, J. (2023, February 23). *Inland Empire warehouse fallout spans class, racial divides*. CalMatters. https://calmatters.org/commentary/2023/02/inland-empire-warehouse-class-divide/

Olalde, M. (2021, April 30). *Environmental groups settle with World Logistics Center*. The Desert Sun. https://www.desertsun.com/story/news/environment/2021/04/30/environmental-groups-settle-world-logistics-center/4883449001/

Povich, E. S. (2022, November 17). *States Take Key Role in Fighting Fake Online Reviews.* The Pew Charitable Trusts. https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2022/11/17/states-take-key-role-in-fighting-fake-online-reviews

WORKS CITED

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, *64* (C), 102771. https://doi.org/10.1016/j.jretconser.2021.102771

Saraiva, A., & Albright, A. (2023, April 5). *The U.S. Warehouse Capital Boomed During the Pandemic. Now It's Facing a Slowdown*. Supply Chain Brain. https://www.supplychainbrain.com/articles/36950-the-us-warehouse-capital-boomed-during-the-pandemic-now-its-facing-a-slowdown

Southern California Association of Governments. (2022, December). Regional Briefing Book. https://scag.ca.gov/sites/main/files/file-attachments/briefing_book_2022_final.pdf?1669774904

Torres, I. (2021, April 15). *ISR Research Draft Report*. Earthjustice. https://earthjustice.org/wp-content/uploads/warehouse_research_report_4.15.2021.pdf

World Economic Forum. (2021, August 10). *Fake online reviews cost $152 billion a year. Here's how e-commerce sites can stop them.* The World Economic Forum. https://www.weforum.org/agenda/2021/08/fake-online-reviews-are-a-152-billion-problem-heres-how-to-silence-them/