⋮

This is a graded discussion: 8 points possible

due Apr 3 at 4:29pm

Assignment 1.1: Explaining Potential Errors in Simplified Representations of Data [Assignment 1.1]

15 63

- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506891)**
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506893)**
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506894)**
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506895)**
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506896)**
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506897)**
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506899)**
- ✓ (https://classroom.emeritus.org/courses/9054/modules/items/1506900)
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506901)**
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506903)**
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506904)**
- ✓ **(https://classroom.emeritus.org/courses/9054/modules/items/1506905)**

There is one assignment on this page. Please scroll down to complete it.

Assignment 1.1: Explaining Potential Errors in Simplified Representations of Data [30–60 minutes]

Learning Outcome Addressed

- Identify an example of Simpson's paradox in a specific industry and how you would solve it.

This is a required discussion and will count toward course completion.

When making data-based decisions, it's crucial that you identify exactly what you want out of the data because sometimes lurking variables can arise, which alters how the data is interpreted. This is referred to as Simpson's paradox.

For the first part of this assignment, we encourage you to learn more about this phenomenon in **Simpson's Paradox and Interpreting Data: The challenge of finding the right view through data. (https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765)** Towardsdatascience.com.

Next, think of another example of Simpson's paradox, either in your own industry or one of your choosing, where the data was misinterpreted and a poor business decision was made. What is the data, what does it represent, what conclusions are drawn from it, and what is your business decision? Assume a managerial role in this part of the assignment, and explain why you are taking these actions. Your data should have some flaws as well as the method you're taking to implement a business decision. This scenario could be one you've experienced or one you could see yourself experiencing.

- Post your scenario, business decision and reasoning of at least 200 words to the discussion board.
- Then, read someone else's post and think about the flaws in their data and their interpretation of the data. Was their business decision a poor one, and what would you change about it? Was the data simplified too much and poorly understood? Respectfully tell them how their representation is incorrect and misleading.

Submission instructions:

Estimated time: 30-60 minutes.

Rubric: Assignment 1.1

Grading criteria	Exceeds expectations	Meets expectations	Below expectations
Includes a well-developed Simpson's paradox scenario	**4 pts** Submission includes a detailed and well-developed real-world Simpson's paradox scenario that addresses all	**3 pts** Submission includes a relevant real-world Simpson's paradox scenario that addresses at least	**0 pts** Submission does not include a real-world Simpson's paradox scenario that addresses the

	of the supporting questions and clearly explains what their business decisions are and the reasoning behind them. The submission is at least 200 words in length.	two of the supporting questions and explains what their business decisions are and the reasoning behind them. The submission is at least 150 words in length.	supporting questions, nor does it explain what their business decisions are and the reasoning behind them.
Response to fellow participant	**4 pts** Submission includes at least one response to a fellow participant that analyzes their business decision and data and respectfully tells them how their representation is incorrect and misleading.	**3 pts** Submission includes at least one response to a fellow participant that analyzes their business decision and data, but did not include how their representation is incorrect and misleading.	**0 pts** Submission does not include at least one response to a fellow participant that analyzes their business decision and data representation.

Search entries or author

Unread

✓ Subscribed

↶ Reply

○

https:/

Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Mar 28, 2024

"Simpson's Paradox" is a statistical phenomenon in which a trend is observed with in individual groups and undergoes a reversal when these groups are aggregated. It is a commonly occurring avoidable phenomenon in Clinical Research firm by taking precautions while designing the study protocol.

A classic and simple scenario of Simpson's Paradox I have came across in my past experience:

During Clinical trials, after receiving data from different sites to Data management system and and will undergo certain data cleaning activities and will be handover to our team to visualize the analysis by using statistical procedures in a Graphical representation or Tables. To evaluate the safety and efficacy of two drugs(Treatment A and Treatment B), while testing on selected individuals after screening process. The overall results favored Treatment A when the data is aggregated. However, when the data is stratified by gender and age category, Treatment B has shown superior results in certain categories. A sample of effectiveness of Treatment A vs Treatment B:

Example of sample data:

Note: Percentage calculations are not based on real values. Sample data is to understand table structure for categories and sub categories

	Treatment A (500 subjects)	Treatment B (135 subjects)
Effectiveness in Male(%)	(60/200) *100 =30%	(90/180) * 100 =50%
Age <18	(20/60) *100 =33.33%	(35/90) *100 = 38.9 %
Age >18	(40/60) *100 = 66.67 %	(55/90) *100 = 61.1 %
Effectiveness in Female(%)	(240/300) *100 =80%	(45/50) * 100 =90%
Age <18	(100/240) *100 =41.67%	(20/45) *100 = 44.45%
Age >18	(140/240) *100 = 58.33 %	(25/45) *100 = 55.55 %
Combined(%)	(300/500) *100 =60%	(135/230) * 100 =58.69%

Clinical studies are designed in such a way that end results shouldn't be impacted by groups or subgroups and decisions cannot be relied on which is impacted by Simpson's Paradox and we have to take certain steps to overcome this phenomenon. As part of precautionary steps, we'll have to work on data identify the complexities and confounding variables. To overcome impact of Simpson's Paradox following steps have taken:

Randomized Clinical Trials: In this, a group of people would be screened and randomly divided into different cohorts without favoring any specific data variable to achieve balanced distribution and the randomization should be kept blinded for both the participant and investigator until certain analysis have been performed on data. So the results should be unbiased irrespective of gender and other characteristics.

Blocking confounding variables: For example Based on previous analysis performed and revealed a paradox related to gender, identify gender as confounding variable in current analysis until the database lock.

Edited by **Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)** on Apr 2 at 10:59pm

Reply

○

(http

Turki Alghusoon (https://classroom.emeritus.org/courses/9054/users/229165)

Mar 30, 2024

Hi Manjari,

This is an interesting example. I am curious to know if you had considered the number of participant in each cross-section of sex and age-group?

Do you think the Simpson Paradox occurred in your scenario because the test trials for Treatment B we more concentrated in the females sub-groups (where Treatment B had a smaller marginal increase efficacy over Treatment A), compared to the male subgroup (where Treatment B's marginal efficacy increase was much higher than Treatment A)?

Reply

○

(http

Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Apr 1, 2024

Hi Turki-

Thanks for your response. Yes, I agree it is hard to extract a scenario of Simpson's paradox from my work experience as clinical studies are strongly designed to avoid these scenarios where the results impacted any type of variables like gender, race, ethnicity...I mainly focus on steps we have taken to overcome the impact of Simpson's Paradox :)

Reply

○

(http

Yossr Hammad (https://classroom.emeritus.org/courses/9054/users/229118)

Apr 2, 2024

Hello Manjari,

Interesting example. from where i understand it seems similar to Quanta and Optima example that we had.

in this case quanta is treatment B and optima A. would the size of the sample data be the reason behind the high numbers in B. I am curious, if the sample sizes are the same would the results be in treatment B still?

Thank you

Reply

○

(http

Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Apr 2, 2024

Hi Yossr-

Thanks for your response. Actually there is a slight difference between the concept of Quanta and Optima vs Simpson's Paradox.

In Quanta and Optima, there is a chance of change in decision while digging deeper into the data.

Where as in Simpson's Paradox, we'll get different results when the data is combined vs data is divided into sub groups.

Reply

○

(http

Javier Di (https://classroom.emeritus.org/courses/9054/users/226884)

Apr 4, 2024

Very interesting case and seems like the higher effectiveness comes from the majority of subjects being females, with much higher effectiveness rates Vs Males in general and Females getting higher effectiveness with Treatment B.

Out of curiosity why is the effectiveness of both treatments so different between Ma;e & Female? it's more than 40% apart? What explains this and what type of treatment were these? Thanks

Reply

○

(http

Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Apr 11, 2024

Hi Javier-

As I explained, this is just a sample data to specify the structure and the variables included and not the sample of actual data that was part of analysis:)

Reply

○

(https:/

Turki Alghusoon (https://classroom.emeritus.org/courses/9054/users/229165)

Mar 30, 2024

While I have not encountered the Simpson Paradox in my industry, one scenario I could think of where the Simpson Paradox could be encountered is in real-estate development.

The Scenario:

In this scenario, I am the Manager of Strategic planning at a construction company where the business model is to buy older properties at a low price, renovate them, then sell them at a higher price. In a given location, I can help the company increase profits by targeting for renovation properties that have the best potential for selling at higher prices.

The company is trying to choose between 2 categories of properties:

1. Row-houses;
2. Single-family homes.

In order to decide, I decided to analyze data related to prince trending for both property categories in one of the counties where the company operates.

The Data:

The available data provides information on the change in market value for 1,200 properties in a residential county for the past 5 years. For every property, the data includes the following data points:

- Street address.

- Zip code.
- Property type (single-family home / row-house)
- Change in price over the past 5 years.

The Business Conclusion:

After performing a quick analysis on the data, I learned that over the past 5 years:

- Single-family homes had an average price increase of 6%
- Row-houses had an average price increase of 17%.

Based on that information, I recommend that we focus on renovating row-houses since they have significantly outperformed single-family homes in recent years. In doing so, I believed we would increase our profit margins.

The Outcome:

After investing our capital in developing row-houses, our construction workers started to notice that single-family homes were outperforming our row-houses almost in every instance. Unbeknownst to us, we had fallen for the Simpson Paradox.

Explanation:

While it was true that row-houses outperformed single-family homes at a high- level, “Zip code” was a lurking variable that we didn’t consider, and it made a huge difference in the interpreting the data. After deep diving into the data and stratifying according to Zip code, I noticed the following:

	Single Family Home		Row House	
	# of Properties	Price increase	# of Properties	Price increase
Top Performing Zip Code	70	30%	330	25%

Average Performing Zip Code	170	15%	230	10%
Low Performing Zip Code	360	-3%	40	-7%
Total	600	Weighted Average (6%)	600	Weighted Average (17%)

The stratification showed that while row-houses outperformed single-family homes at a high-level, single-family homes outperformed row-houses across all zip codes. The issue was that a large proportion of the single-family homes were in the low-performing Zip code which had the worst price performance in the county. In contrast, Row-houses were concentrated in top performing Zip codes which helped boost their overall performance.

Reply (3 likes)

○

(http Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Apr 2, 2024

Hi Turki-

Thanks for sharing an interesting example. I'm wondering whether the collected datapoints street address, zip code, type of house and change in price is good enough to take a decision? Also, as zip code is one of the key variable for any data related to housing, I'm curious to know the reason for not considering that variable.

Reply

○

(http Turki Alghusoon (https://classroom.emeritus.org/courses/9054/users/229165)

Apr 2, 2024

Hi Manjari,

Thank you for sharing your observations.

I think more datapoints (e.g. changes in median Income, # of police-reported incidents, average difference between listing price and actual selling prices) would have definitely resulted in a more complete insight with multiple indicators to support the decision-making process.

As for the reason for excluding the zip code: One explanation to why Zip code was not considered in the initial analysis could be a limited understanding of the business area and over-reliance on data without a business context. Such limited understanding would result in an over-simplified analysis where potentially important factors such as Zip-code are overlooked.

Edited by **Turki Alghusoon (https://classroom.emeritus.org/courses/9054/users/229165)** on Apr 2 at 7:27pm

Reply

○

(http

Todd Engle (https://classroom.emeritus.org/courses/9054/users/228910)

Apr 3, 2024

Loved your example Turki, I learned from it. I like how you showed the stratification of the data, helped me understand it better.

Reply

○

(http

Victor Flores (https://classroom.emeritus.org/courses/9054/users/197659)

Apr 3, 2024

Hi Turki,
From the available data, it is interesting to see how a low performing zip code drives a higher price increase for Row Houses and that's were the eyes of investors will be. Also, the # of properties for top performing zip codes is higher and can yield higher returns to your company. Do you know if the trend of your analysis will change if data is regrouped and other variables included?

Reply

○

(http

Swati Sharma (https://classroom.emeritus.org/courses/9054/users/236938)

Apr 3, 2024

Hello Turki, very nice to meet you, thank you for sharing! Your example of real estate development (
very different from my domain and very interesting) highlights the deceptive nature of

Simpson's Paradox. Initially favoring row-houses due to their higher average price increase, after deeper analysis revealed that single-family homes outperformed across all zip codes.

Reply

○

(http

Javier Di (https://classroom.emeritus.org/courses/9054/users/226884)

Apr 4, 2024

This is a great example and explained by the different distribution and composition and Single Family Homes having a high concentration on the low performing zip code and Raw Houses having a high concentration on the high performing zip codes

Reply

○

(https:/

Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)

Mar 31, 2024

Backdrop

In humanitarian interventions, data analysis is used to assess the impact of interventions such as

1. hygiene promotion campaigns
2. nutrition
3. in kind (e.g. providing food) vs cash transfer (giving money to buy food)

Applied data analysis

Evaluation data are often not disaggregated, analysts then apply national demographic ratios (e.g. 52% females and 48% males) to the aggregated figures. Even with this caveat analysis rarely considers gender. Unfortunately, it’s usually a single-direction analysis. Conclusions for such analysis could be misleading. This becomes apparent when the same intervention design is promoted as successful and, then, replicated or scaled up. Then in due time, the replicated interventions do not produce the same impact result as the original pilot project.

Probable confounding factors

A deeper sceptical look into the data from the scale-up may reveal that

Area of Intervention	Pilot Context	Replica probable confounding factors	Comments
hygiene promotion campaigns	Awareness of waterborne disease (current cholera outbreak) Availability of water	No perceived risk of disease Water Scarcity	Handwashing is less likely in the replicated project
nutrition	Certain cultural behaviours (female and male children are the same.	Gender inequality among children (i.e. boys are more cared for than girls) ratio of M:F children is not the same as that of the pilot	The aggregate percentage of undernutrition among children is like the pilot. However, it could still be a case of *Simpson paradox* (higher % of undernutrition among female children).
in kind (e.g. providing food) vs cash transfer (giving money to buy food)	Most of the households that got the cash were headed by a female	Certain cultural norms allowed for males to control how to spend the money	A significant amount of Cash/money was not spent on the intended items (e.g. food)

The many contextual confounding factors some highlighted in the Replica column, will need to be assessed before deciding if to replicate and where best to replicate a project. The development and humanitarian sectors continue to garble with “uncertainty” (now I know they are more *confounding factors*).

Example of Flawed Data

In 2007 (relatively early days of online social platforms and pre-smart phones) I evaluated a development project. This project was a replica of another country's very successful pilot. The logic of the project was intuitive:

1. Provide quality training for females in a certain age group on sawing machines (and provide training incentives).
2. Connect those females with relevant employment opportunities.
3. Finance purchase of sewing machines

4. Expected impact: those females would gain improved income (either through employment or saving money by sewing at home rather than buying ready-made)

The evaluation showed the percentage increase in income was insignificant (compared to the original baseline). Slicing and dicing the data showed a notable dropout among female trainees and those that continued had almost no income increase. Further assessment of the reasons for low female worker retention rates yielded, that quitting and dropping out were the result of a conservative culture. Many female trainees' motive to join the training and get employment was to interact with potential suitors (an opportunity that was not available for many of them), and when they got married, they conformed to cultural norms and stayed home.

Heightened sixth sense

More than a decade later I still think of the needed data that would have pre-empted this shortfall and possible eluding analysis dimensions!

Nevertheless, since then this experience has continued to serve me in the form of being sceptical and consistently searching for, what I now know as, the *confounding factors*.

Edited by Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864) on Apr 3 at 8:08am

Reply

○

(https:/

Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)

Mar 31, 2024

Industry: Banking and Financial Services

Scenario:

I manage the marketing team for a large bank that utilizes an aggregation platform to gather customer transaction data from customers of the bank using bank credit and debit cards and from other financial institutions who accounts and transactions have been aggregated. Given the vast amounts of accounts and transactions, there is a an opportunity to leverage this data for to provide tailored customer experiences with personalized marketing, offering other bank services and financial products.

Data and initial conclusion:

During the analysis of all customer transactions, there are observations of different segments with varying frequencies of overdraft fees. We decide that there is an opportunity to target customers with the highest constant frequencies of overdrafts as we identify higher frequencies of overdrafts as signs of financial struggles and they can be targeted for loan and credit card suggestions.

Lurking variable:

The Simpson’s paradox was revealed when we realized we overlooked income level, a crucial lurking variable. Low-income earners may be currently struggling financial and would be interested in loans and credit card applications.

However, high income earners, use the overdraft accounts as a buffer since they observe a lot of monthly transactions in their checking accounts, and do not need additional credit. Targeting these high income earners unnecessarily can lead to consequences such as leading clients to distrust the bank, result in ineffective allocation of resources and frustrations from the client and team ends.

To address Simpson’s paradox, we refined our model strategy to include net worth and rolling average of cash balances to help distinguish the high income earners from the low income earners who would be more likely interested in loans and credit cards.

Revised Business Decision:

Now we our new revised and enriched model, we continue to tailor our products to our intended target audience, the low income earners.

We will provide products such as debt consolidation with low interest loans and/or low interest credit cards with low interest balance transfer for customers with financial hardships and demonstrate a record of successful on time payments.

Reply 👍

○

(http

Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Apr 2, 2024

Hi Roy-

Thanks for detailed explanation of scenario. Can we consider this as Simpson's Paradox. As per my understanding, we should get different results when data is combined vs data is divided into sub groups. But explained scenario is based on overseeing a variable. Am I missing anything?

Reply

○

(http

Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)

Apr 6, 2024

Hi Manjari,

Thanks for your feedback. You are right "we should get different results when data is combined vs data is divided into sub groups".

And that is what I have tried to capture here. "Combined" customers with very frequent overdrafts are good targets for product offerings. The "subgroups", low earners vs high earners, yielded different results.

Without the lurking variable, income level, the interpretation of the data for the entire group was under the assumption that all these customers with frequent overdrafts are under financial struggle and can be targeted for loans and credit card offers. When the segmentation was done based on income level, the interpretations/"results" changed. In this example having frequent overdrafts do not necessarily mean there is financial struggle and that they are candidates for loans and credit card product offerings.

Edited by Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552) on Apr 6 at 5:38am

Reply

○

(http

Javier Di (https://classroom.emeritus.org/courses/9054/users/226884)

Apr 4, 2024

Great example Roy and curious why you didn't set up an yearly income treshold such as only for clients making under say $60,000 for example? As I understand customers have to disclose their incomes to have obtained those cards in the first place?

Thanks

Reply

○

(https:/

MATT DEFREITAS (https://classroom.emeritus.org/courses/9054/users/220100)

Apr 1, 2024

Being a fan of American football and the NFL draft is in the next few weeks, I will use an example of a Quarterback's (QB) completion rate. Typically, a franchise will be heavily invested in the QB position as not only is it typically the face of the franchise but also the person who has their hands on the ball the most throughout a game. With so many metrics available within the sport, completion rate is one that can be significantly impacted by Simpon's paradox. In this scenario, I will be the individual responsible for selecting the best available player for my team's success for years to come.

Let's use QB A and QB B as an example and examine how the Simpson's paradox can be evident when dealing with completion rates.

QB A has an overall completion rate of 80% whereas QB B has an overall completion rate of 70%. On paper, QB A appears to be more appealing to franchises as from that view alone he appears to be the most accurate player. However, as the person responsible I need to continue to ask questions about the data to ensure I am providing the best recommendation to the organization. One area we quickly start to dive into is the length of their throws. This is important as in the game you want the QB to be able to throw short, medium, and long passes to confuse the defense. Once we started looking into these different levels we started to see a different story.

When we looked at the mix of throws, 90% of QB A's throws only go less than 5 yards and his completion rate is 85%. While that is a positive completion rate, he lacks the ability to drive the ball downfield which can limit the offense. He rarely distributes the ball past 5 yards but when he does, he is far less accurate. QB B on the other hand has a great mix of throws ranging from 5 yards to 25+ and he consistently outperforms in all other categories except for less than 5 yards.

As an organization, the decision to select a QB that will be able to get the ball down field often and accurately is critical so in this situation while the aggregate rate is better for QB A the better choice is QB B.

The flaw identified in the example above is like the reading's example with the flavors of soda. The aggregated data did not factor in other variables. Even though the survey was straightforward, it's results were not. I don't know if I would have concluded that the results were inconclusive as if both men and women enjoy peach and the aggregate data said others that leads me to believe the design of the survey may have been flawed as if the intention was to gather the sex of the respondents why not display that information originally?

Reply 👍

○

(http **Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)**

Apr 2, 2024

Hi Matt,

Thanks for posting this! The fact that QB B outperforms on all other categories except for less than 5 yards is interesting. Could it be the weight and or size? Meaning if QB B is smaller than many of the players he can be intimidated. The psychological pressure due to his physiology could be affecting their confidence. Wonder if those psychological stats are collected?

Reply

○

(http **Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)**

Apr 3, 2024

Another very interesting business case application, Thanks for sharing, Matt.

Careen stories of Tom Bradly (TB) and Micheal Jordan (MJ) come to mind. TB (I believe) was a late draft and not the top start-up choice for his team, while MJ was only successful with the Bulls.

Given that, a good portion of top-rated athlete stats are available, If the disaggregated completion rate approach was run for two (or x number of drafted) QBs. Both (or most) of those QBs had statistically insignificant differences when they were drafted. However, one far excelled (TB) over the other(s). Would the applied detailed completion rate approach have anticipated this? This also echo's Roy Nunez point on players' ability to handle pressure and how they mentally mature over time. There are some examples in soccer of player who became stars after being let go by their clubs, by getting mentally stronger.

Do you think the completion rate approach would sufficiently benefit if coupled with a prescriptive model (e.g. the athlete needs to play an average of so many minutes in x number of games in one season in a top-flight competition)?

Edited by **Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)** on Apr 3 at 6:51am

Reply

○

https:/ **Dawn Prewett (https://classroom.emeritus.org/courses/9054/users/233112)**

Apr 1, 2024

Years ago I held a position within the waste management industry in which I oversaw reporting for the tonnage collected by our garbage, recycling, and food waste trucks. This data was critical for cities to direct their recycling efforts, spot-check for contamination, and secure funding through grants. Due to trucks servicing multiple cities, tonnage data was aggregated based on the percentage of stops made in each city—a method in use for a decade before I assumed the position.

Upon reviewing the data, I discovered discrepancies: the total reported tonnage was less than what was actually collected. This discrepancy arose from several factors, including variable customer container sizes, unaccounted for service stops, incomplete driver reports, and averaged truck tonnages for shared routes. This flawed aggregation method masked the true distribution of waste collection across cities, a classic example of Simpson's Paradox.

Around the time that I started to uncover these issues, the company was working on the introduction of new truck technology, which automatically recorded each stop and provided the perfect opportunity to refine our data collection process. Despite initial resistance due to legal concerns about past inaccuracies, I advocated for leveraging this technology to correct and improve our data accuracy. By transparently focusing on these improvements and directly addressing cities' concerns regarding potential impacts on grant funding, I facilitated a transition to a more accurate reporting system.

The revised algorithm calculated the tonnage per load, allocating it across stops weighted by container size, and aggregating these figures by city. This adjustment resulted in a match between collected and reported tonnages, though it significantly altered some cities' data. Despite challenges, these changes were ultimately recognized as beneficial improvements, showcasing the importance of accurate data collection and the willingness to address and correct longstanding issues.

Reply (1 like)

○

(http

Priscilla Annor-Gyamfi (https://classroom.emeritus.org/courses/9054/users/226376)

Apr 3, 2024

Great post Dawn.

I would add that, to improve the situation, a more thorough understanding of the data collection process and its limitations should be studied and implementing new algorithm is just one aspect of improving data accuracy. It's also important to ensure that the data interpretation methods are sound and that any potential biases or inaccuracies are addressed.

Reply

○

(http

Dawn Prewett (https://classroom.emeritus.org/courses/9054/users/233112)

Apr 3, 2024

I did exactly that, which was how I discovered that the trucks were getting the new technology that enabled most of the changes I implemented. My timing was pretty lucky as well as I was able to influence how that data was collected and reported to better align with our objectives. The new algorithm was built for my local area, but ended up being adopted enterprise-wide from what I've been told - the completion of the project actually resulted in a new job offer. It was one of the more impactful activities I've led.

Reply

○

(http

Lee Lanzafame (https://classroom.emeritus.org/courses/9054/users/231975)

Apr 12, 2024

Great example, depending on your teams budget, you could implement advanced sensors that could identify the volume of waste by load sensing during pickup or upon arriving at a garbage disposal organisation, these places usually weigh the trucks and they charge per tonne. This weight could be recorded by the driver and centralised, a model could then be built based on actuals to improve the algorithm.

Reply

○

(https:/

Lee Lanzafame (https://classroom.emeritus.org/courses/9054/users/231975)

Apr 2, 2024

I work at a telecommunications company. We measure customer satisfaction by using a metric called NPS (Net Promotor Score) which produces a value between 1 to 10 and is called an LTR (Likelihood to Recommend).

In order to capture the NPS score, the customer gets a survey after having an episode (completing a certain task like buying a new plan) or interaction (calling a call centre with a query) with the company asking them how their experience was. In store employees are incentivised on a multiplier bonus basis when their episode and interaction scores are 9 or 10. We had an instance in a store where the employees were swapping a customers email address with their own, this meant that the employee would receive the customer satisfaction survey instead of the customer and hence cheating by giving themselves a 10 out of 10. After filling out the survey they would change the email address back. These results ultimately get feed into the employees commission.

When the area managers looked at the combined scores for the stores in the area there wasn’t anything that stood out but when they drilled into a store by store comparison this store stood out as having very different satisfaction scores compared to others in the area. It became obvious that something was up.

Conclusion that I can draw from this example is that, things may seem straight forward in the data and you could easily assume that you have a team of rockstars (and you might!) i.e., if you want to improve customer satisfaction then send customers a survey but in reality there are many intricate pieces and decisions and processes need to be constantly reviewed and monitored, this is where sophisticated modelling techniques come into it, they can help uncover hidden patterns.

If I was in a managerial role the actions to resolve this are:

- Check for sudden persist changes in satisfaction scores. i.e., if a store that was previously receiving all 5’s or 6’s suddenly has a jump in performance with no other explanations (like change in staff, major store refurbishments, or radical coaching changes)
- Rather than having a geographic specific focus to grouping stores, group on a feature basis across the geography to have better comparative data, i.e. based on store size, number of staff or even customer demographics
- Bring more 1 on 1 coaching conversations with store managers to understand how incentives are motivating staff and how key measures of performance can be improved over time to provide the right incentive

The first two points above can be solved by building a model, the third point can add a more human element. A combination of both is always best.

Reply

○

(http

Dawn Prewett (https://classroom.emeritus.org/courses/9054/users/233112)

Apr 2, 2024

This is an excellent example of Simpson's paradox, though I daresay, you could go even deeper.

These surveys, although designed to measure service quality, often fail to capture the nuances of customer-agent interactions. Customers frequently base their ratings on their overall feelings towards the company rather than the specific service received during the interaction. This discrepancy, even when instructions ask to rate the agent's performance exclusively, muddies the data and unfairly impacts agents who are judged by these skewed metrics. Such systems, where one's job performance is tied to what can feel like arbitrary evaluations, incentivize finding ways to "beat the system."

In light of these observations, it's clear that while seeking to quantify customer satisfaction, we must tread carefully to ensure fairness and accuracy. Alternatives or complementary measures should be considered, ones that can more faithfully represent the quality of customer service delivered. A blend of direct feedback, peer reviews, and self-assessment, for instance, might offer a more balanced and comprehensive view. This multifaceted approach could mitigate the pressures to game the system, fostering a more genuine and constructive environment for service improvement.

Reply (1 like)

○

(https:/

Chris Cosmas (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/226607)

Apr 2, 2024

During the last phase of my Bachelor's Degree, I was tasked with writing a thesis research paper on the company for which I was interning. After discussing with my thesis supervisor I chose to focus on Operational Efficiency. As explained in the first discussion the firm is a well-established firm with robust processes which cover all operational needs. The firm was heavily process-oriented with internal wiki pages documenting every process step by step. I chose to focus my research on the ticketing process as it was the widest-reaching process involving all business units of the firm. I had noticed even though we were a small office in the firm we were receiving a large amount of tickets. Ticket resolution was one of the major tasks as they required us to fulfill customer requests, update certain data points, and other varying tasks. This was done on a ticketing system where all information was kept and allowed different teams to work on tickets at the same time. This system also allowed me to extract metadata on the requests such as time elapsed since opening, different request phases, different teams who owned the requests, and other data points. I focused my research on ticket volume and time resolution and tried to identify different operational stressors or bottlenecks employees were meeting. When examining the data at an aggregate level it seemed the that ticket resolution time was correct as teams resolved tickets in the amount of time the firm promised to its customers. But this didn't explain the abnormal level of tickets we were receiving at the office compared to other offices. The process had different levels, tickets with higher complexity were relayed to different teams until reaching us, and if it proved too difficult we sent it to the "last line of defense" which was a small team of experts. I looked at different stressors such as KPIs teams were tracked with, SOP documents to see if there were any differences between teams, organizational cultures in different offices, and so on. On the aggregate level, everything seemed to be sound, we were all instructed to perform the same steps across the organization. I started looking at data on a team level by performing standard distribution statistics such as average time elapsed, quartiles, and so on. It came to my attention that teams at the first levels of the process were offloading tickets to the next team after way shorter periods than the instructed period, lower level teams held tickets for a quarter of the time which latter teams usually took. After noticing the time discrepancy between the different teams I shifted my focus once again towards organizational stressors. Among a few new insights that came to light through interviews and surveys, I became aware of extra KPIs the first level were tracked off which we were not, the teams had a range of KPIs on their tickets such as the number of requests closed or sent to another team in a day, timeliness, amount of tickets outstanding, etc... This seemed to me a huge organizational stressor, these KPIs being directly linked to their bonuses, created a significant incentive for lower teams to offload tickets onto other teams in the organization.

Yes, requests were being resolved but this could be done way more efficiently, enabling lower levels to spend time on tickets would enable them to get more exposure to the different cases that we were getting. This could lead to higher job satisfaction which translates to further employee retention,

Reply

○

(http

Diego Milanes (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/228518)

Apr 2, 2024

Hi Chris,

I appreciate a lot the experience you have just shared. I have not much experienced beyond the academic and research environment, and your comment shed some light on how these hidden variables appear at several levels in a company. I wonder about the measures that the company adopt once you show them these bad practices. Did they react in some way?

Reply

○

(http

Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)

Apr 3, 2024

Interesting case Chris.

The lack of clear SOP, roles and responsibilities and empowerment, were the undermining issues that came out when we conducted a capacity assessment of all the organizations' country office's performance against KPI. To a certain degree,. Your case summarizes a key success factor where performance is linked to benefits (bounces) in tandem with effectively addressing the undermining issues. The leadership commitment in your company must have played a pivotal role by seeing the application of the recommendations all the way through.

Reply

○

(http

Chris Cosmas (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/226607)

Apr 3, 2024

Hello Diego, Haitham,

Thank you both for your replies.

I'd like to first start by saying I am no expert whatsoever, I am very junior in the field as well as in the professional environment, the deductions I made were based on my interpretations of the data, there must have been some bias as well to fit the purpose of my thesis as It was a requirement for me to find an issue in the company. Having said that I did formulate many recommendations in the end regarding other stressors I had found which I will not bore you with, such as actions to improve organizational asymmetry between different departments, involving other units in our meetings to increase information sharing, carry more mirroring sessions on how we complete our tasks, I had found other issues in regards to job title mismatch between the tasks we were actually performing and the labels our positions had.

Later in my internship they actually did try to change titles to reflect positions better to manage expectations better, some of the team members around me did perform more mirroring sessions and started a weekly meeting with other units. As for the KPIs that is way out of my pay grade :D my findings were confirmed by my team leader as well as the president of the unit, and I did present my findings to the managing director, but I was told there are many politics to be taken into account. Again I was a young intern with not much influence trying to finish my thesis, and the company is performing quite well, there is no need for any drastic changes.

Reply

○

(http

Lee Lanzafame (https://classroom.emeritus.org/courses/9054/users/231975)

Apr 12, 2024

I work at a telco where we have had the same problem, transparent reporting and revising KPI's is a good way to prevent this. If possible job swaps between levels could also balance any biases. I wonder if machine learning could be used to triage tickets based on complexity?

Reply

○

(https:/

Yossr Hammad (https://classroom.emeritus.org/courses/9054/users/229118)

Apr 2, 2024

Scenario:
Graduation Project : Fruit Fiesta

When i was back home my graduation project was Fruit fiesta which is now known as edible arrangements but covered with chocolates, cinnamon , nuts and etc..

we wanted to rent a booth in a specific location (a mall) so we ran a survey in that location lets call it X location. we distributed around 700 questionnaires and got back about 650. The data from the surveys assured that about 80% of the answers would buy the serving for $7 twice a week and 15% would buy it once a week and 5% would buy it once a month.

We wanted to open another booth in a different location. Let's call it Y location. We increased the order weigh and ingredients and offered to sell for $9 at Y location. we collected about 800 responses. Only 40% agreed to pay that price twice a week and 60% to buy it once a month.

We assumed the price difference is the problem but when our team dived deeper, we found that in location X most people responded to the survey make $4k+ monthly while responses we got from location Y 50% make only $3k+ and the rest make less than $3K.

We decided to change the pricing strategy based on this hidden variable. We reduced the price for location Y to $5 hence the portion of the serving and we increased the price in location X to be $9 per serving.

We ran the survey for the second time in both location. Responses from location Y was satisfying and aligning with our pricing strategy but the responses collected from location X was confusing.

60% responded to purchase the product once a month and 30% to buy it twice a week and 10% once a week.

We took another look into the data and we figured out what was wrong. most of the visitor of the mall that day was non working females. it was a clearance day for all the stores in the mall from 9 am to 3pm and most of the responses make no money monthly.

Decision made:
For location Y we keep the price $5 as based on the surveys most people around this location are relatively low income.
For location X we decided to price it as it was first attempt $7 as we thought that the location wont be full of working individuals only , we assumed that some days ,like the day we ran the second survey , the mall will have unemployed people that are not willing to spend more than $7 to buy our product.

Reply

○

(https:/

Jignesh Dalal (https://classroom.emeritus.org/courses/9054/users/229173)

Apr 2, 2024

I work in the quality domain in the telecommunications industry and we have analyzed customer satisfaction data in our firm, looking at two services: traditional landline phones and modern mobile services. Initially, the data shows that customer satisfaction rates are higher for landline services than for mobile services when viewed separately. Based on this, one might conclude that investing more in landline infrastructure would be the best business decision. But that is not necessarily true

Data Interpretation and Flaws:

Landline Service Satisfaction: 90% satisfaction among a smaller, older demographic that values reliability over features.
Mobile Service Satisfaction: 85% satisfaction among a much larger, diverse demographic valuing features, flexibility, and innovation.

The data flaw here is not considering the demographic differences and the evolving market needs. The higher satisfaction rate in the landline service is driven by a less demanding, smaller customer base, whereas the slightly lower satisfaction in mobile services spans a larger, more diverse, and innovation-driven customer base.

Business Decision and Reasoning:

Despite the initial data suggesting higher satisfaction with landline services, the decision here is to invest more in mobile services. This decision is made by recognizing the flaw in interpreting the satisfaction rates without considering the market dynamics and customer demographics. Furthermore, the landline services for telephone are declining.

Investing in mobile services aligns with the broader market trend towards mobile usage, technological advancements, and the diverse needs of a larger customer base. Additionally, mobile services offer more opportunities for innovation, new service offerings, and reaching a younger demographic, which is crucial for long-term growth in the telecommunications industry.

The rationale behind this decision is to not only address the current needs but also to anticipate future market trends and customer preferences. By acknowledging the potential misinterpretation of the data due to Simpson's Paradox, we can make a more informed decision that aligns with strategic growth and customer satisfaction goals.

This scenario highlights the importance of looking beyond the surface of data trends and considering the broader context to avoid making flawed business decisions based on misleading interpretations.

Reply (1 like)

○

(http

Victor Flores (https://classroom.emeritus.org/courses/9054/users/197659)

Apr 3, 2024

Hi Jignesh,

The customer satisfaction study performed across landline and mobile services is a great example of how data can be looked from more than one lenses and change the objectives of an organization. As you mentioned although customer satisfaction was leading investors towards investing on landline services, the idea of investing on a broader market with much more opportunities for growth changed the trend and set your company's radar's screen on mobiles services instead. I envision all industries go through the same scenarios when exploring investment options in diverse and dynamic environments. Your example helped me to better understand the "Potential Errors in Simplified Representations of Data" that is address on this module and further illustrated with the Simpson's phenomenon.

What other variable besides customer satisfaction does your company study when reasoning on future investments?

Reply

○

(http

[Jignesh Dalal (https://classroom.emeritus.org/courses/9054/users/229173)](https://classroom.emeritus.org/courses/9054/users/229173)

Apr 3, 2024

Hi Victor

Thank for your message, There are multiple areas of focus that aligns with company mission other than customer satisfaction are as below.

- Market trends and demographics
- Technological advancement

- Competative Landscape

- Regulatory Environment
- Financial Performance

With above variables in conjunction company can draw a holistic understanding that drive long-term value creation for our company and stakeholders.

Reply

○

(https:/

[Koffi Henri Charles Koffi (https://classroom.emeritus.org/courses/9054/users/208039)](https://classroom.emeritus.org/courses/9054/users/208039)

Apr 2, 2024

Covid-19 Simpson’s Paradox

From one the example provided by Nassim Nicholas Teb with respect to covid-19 pandemic .

A study appears which involves country A, , citizens aged 10 to 60 plus and concludes that for that country the unvaccinated live longer than the vaccinated with conspiracy theorists rushing to claim that the vaccine is therefore negative for society .

The exact opposite tends to be true when taking a close look at data by age brackets.

- 10 to 20 year old vaccinated people tend to live longer than 10 to 20 year old unvaccinated

With the same being true for 20 to 30 year olds , 30 to 40 year olds and other subgroups.

Then why did the study show that the entire population , its unvaccinated people that tends to leave longer?

Simply because in Country A like most country prioritized vaccine distribution toward those at risk thus a lot of elderly individuals received the vaccine as a percentage of the vaccinated population therefore the vaccination group contains a larger percentage of older individuals than the unvaccinated one , who are obviously more likely to die as such the vaccine actually improve overall survival regardless of what age you are .

With the result of the study being a consequence of the fact that older individuals were over represented in the vaccinated population .

The lesson learned here is to not jump to conclusion after just looking into the data but a deep analysis is required if we want to pursue the truth .

Reply

○

(http

Timothy Andrew Ramkissoon (https://classroom.emeritus.org/courses/9054/users/226697)

Apr 3, 2024

Koffi,

This is a very good case study; it shows how interpreting data can affect the opinions of the populace. However, it's unclear what decisions would you make as a manager presented with this data.

I would consider factors such as what percentage of vaccinated persons died after receiving the vaccine, what underlying medical condition did persons who died have, what percentage of elderly patients received the vaccine and how much did that skewer the data to lead to "unvaccinated live longer than the vaccinated". I might present the data in various forms, one with young adults, one with middle aged adults and one with seniors to check the relationships between lifespan of vaccinated and unvaccinated among age groups.

This scenario demonstrated the importance of looking beyond surface-level data to make informed decisions.

Reply 👍

○

(https:/ **Diego Milanes (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/228518)**

Apr 2, 2024

I think I haven't experienced a situation like that in my professional environment. Let me create a hypothetical situation.

Let's go back to Bob's Corner (from discussion 1.3), the local beer and wine retailer with several locations across the city. Let's imagine that there are only 2 locations to facilitate the example: locations A and B, which are placed in similar neighbourhoods and offer the same type of products.
The sales analysis shows that location A outperforms B in terms of yearly overall revenue. For this analysis, the data comprises transactional information on each sale, focusing on gathering information about the customer to develop some marketing strategy afterwards. This solely information leads Bob to think on taking some action. Bob is convinced that the main real difference between the locations is the crew team and, mistakenly, he decides to change the crew of location B, and to contract a more performing team that reaches the revenue shown at location A.
Fortunately, before making the final decision, Bob asks a friend who is enrolled in a Data Science course to have a detailed look at the data. The first thing observed is that even though the locations are in similar neighbourhoods and the products offered are the same, some important differences exist in the products sold and deeply in the profile of the customer.
Both locations have sold about the same amount of beers, but A has sold more wines from a market segment that represent a higher profit margin. This implies that the performance of the crews is similar, but Bob needs to develop a strategy to engage consumers from location B with wines from the other market segment. This also implies that there are hidden differences between customers, which suggest that another positive action can be to apply surveys to the clients to increase the number of features for a better profiling in further marketing campaigns.

Reply 👍 (1 like)

○

(http **Jignesh Dalal (https://classroom.emeritus.org/courses/9054/users/229173)**

Apr 3, 2024

Thank you Diego for sharing ideas around Bob's business scenario.

Few other areas could be looked as aggregated data to make better decision making:

1. Competitor analysis: Identify main competitors, their offerings, and pricing strategies.
2. Seasonal trends: Highlight any seasonal fluctuations in sales and factors driving them.
3. Product mix optimization: Outline best-selling products and opportunities for diversification.
4. Customer feedback: Summarize key insights from customer feedback and reviews.
5. Data quality: Assess the accuracy and reliability of available data sources.
6. Financial analysis: Provide a brief overview of costs, profit margins, and potential ROI for proposed strategies.

This information will help Bob make data-driven decisions to enhance sales and customer satisfaction at Bob's Corner.

Reply

○

https:/

Todd Engle (https://classroom.emeritus.org/courses/9054/users/228910)

Apr 3, 2024

I worked for a large bank that had a high number of consultants working on any given project. Management wanted me to do a cost analysis to better understand the impact of having a high consultancy ratio. On the surface, it looked pretty simple. If an employee worked over 40 hours a week, the cost was still the same due to having a salary. Consultants, on average, made more than employees, and often worked by the hour. Therefore, if the consultants worked over 40 hours, many received overtime pay. On many of the projects, it was about 60% consultants vs. 40% FTEs.

Initial Observation: A simple calculation of labor costs by employee type and by project, you would see that consultants made up more than 60% of the labor cost. Therefore, it was assumed that, yes, consultants were more expensive and increased the cost of projects.

Finding the Confounding Variable: I am not a data analyst, but I was a consultant, and my observation was that consultants worked harder at the bank and didn’t take as many breaks as the full-time employees did. Therefore, I considered the amount of effort completed by each employee type using an earned value calculation for each of the project activities and timesheet data.

The result was that consultants consistently brought much more value to the project than FTEs. They simply accomplished more, in less time. It was evident that decreasing the number of consultants would only prolong the effort and the project timeline, thus increasing the overall cost.

This was many years ago when I was first starting out as a consultant, and I felt proud of my findings. Reading about the Simpson's Paradox I realized that I intuitively felt there was a confounding variable simply through observation of the environment I was working in did not jibe with the data.

If I was to do this again, I would include a lot more information, such as project complexity, project phases, and if I could get it, level of expertise. I'm wondering what other secrets I could pull out of the data.

Reply

○

(http

Chris Cosmas (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/226607)

Apr 3, 2024

Hello Todd,

I appreciate your example as there seems to be a never-ending debate on social media on the benefit of outsourcing consultants.

They do offer invaluable information but tend to leave a pricey bill for contracting firms.

Have you considered the possibility that consultants might have possessed much more relevant skills needed for these projects, or might have had access to resources that were not available to FTEs? Another consideration could be experience, more senior consultants might have more exposure and knowledge than the FTEs. The time frame of the work might also have an impact Full-time employees do tend to get more comfortable due to the reason that they are full-time employees and might not feel as pressured as consultants who have a much more limited time to perform.

Reply

○

(http

Todd Engle (https://classroom.emeritus.org/courses/9054/users/228910)

Apr 6, 2024

Hey Chris,

All good points. Brain drain was a problem at the firm. As you said, some of the consultants were brought in as SME's but there was little opportunity to have that knowledge bleed into the overall knowledge base of the full-time community. Knowledge expertise would have been a good data point to capture.

Reply

○

(https:/

Ricardo Anaya (https://classroom.emeritus.org/courses/9054/users/228915)

Apr 3, 2024

I work in a LATAM region where the wiereless networks are behind techologically

Source: 5G Americas

There are 314 5G networks live worldwide

there are only 39 in the LATAM region,

about 11 percent of Global Deployments

However there are 714 4G deployments Globally Vs 314 5G 43%

in the LATAM region There are 135 4G deployments vs 39 30%

So the assumption is that 5G in LATAM is behind, this has many other points of view

1) more potential to bring new 5G devices as the market is still to come

2) continue selling 4G devices as the network is not ready yet

The paradox is that number of subscribers in 4G and 5G

2024 number of Subscribers data:

World

4G LTE 58.4% 5017 Million of 8594

5G 25.2% 2165 Million of 8594

LATAM

4G LTE 73.7% 541 Million of 734

5G 7.4% 54 Million of 734

The numbers would tell that 5G is a Choice for the future as the market share is yet to be developed

However Im seeing still a lot of 4G devices, that keeps growing, and 5G is growing also but at a lesss expected rate

The Key here is that Network investments, spectrum costs, device costs (Aaverage selling price), Average Data used by month, price per MB/GB of Data, and Time to renew devices (24 to 36 months) is delaying the 5G launch, and keeping up the 4G Networks as a main opportunity and the oportunity for 5G, is there, just not yet... when, that is the question that would be the prediction and model I would like to work out.

Reply

○

(https:/

STEPHEN HUTSON (https://classroom.emeritus.org/courses/9054/users/233645)

Apr 3, 2024

You are the working with a large music record label and are tasked with determining which of two Artists your label manages should headline a new music venue that is opening to generate the most ticket and merchandise sales possible. The label wants this to be a big success with the local population in hopes that they will fill this large venue and continue to return for future shows. This new venue is located in Asheville, North Carolina in the United States.

Artist A:

A popular artist who had several recent hit singles that were popular on the music charts, which quickly generated a lot of publicity and popularity. This artist began posting songs on YouTube from their home in London, and quickly became an international sensation with their songs being frequently featured on social media.

Artist B:

An older more established artist who's been playing in the music industry for over 30 years. They grew up 20 miles outside of Asheville and have written many songs about their hometown, and have a cult following of fans who are passionate about their music.

In order to make a decision on which of these two to book, you decide to look at streaming data and records sold over the last 2 years in order to determine which would be the better artist to book:

Artist A:
Streams:

- 2023: 250,000,000, $80M in revenue
- 2024: 500,000,000, $160M in revenue

Records sold:

- 2023: 800K units, $12M in revenue
- 2024: 750K units, $11.25M in revenue

Artist B:

Steams:

- 2023: 45,000,000, $15M in revenue
- 2024: 50,000,000, $16M in revenue

Records sold:

- 2023: 100K units, $1.5M in revenue
- 2024: 110K units, $1.65M in revenue

Based on the financial data you examined, you decide that Artist A is clearly more popular both in streams and in records sold, and decide that they will be the better option for the venue in terms of a higher chance of selling out the venue and generating merchandise sale, and make this recommendation to the label.

Reply

○

(http

MATT DEFREITAS (https://classroom.emeritus.org/courses/9054/users/220100)

Apr 3, 2024

Hi Stephen,

I love your example with the music industry. This is data that I've been interested in for a while now so it's great to see it in action.

As this is a localized event, would you be able to geofence the streams and/or records sold within Asheville and the surrounding areas? I think of this from a strictly marketing perspective and we would want to understand what the demand is specific to this area to see what the people purchase or stream. This way we are personalizing our artist selection to the audience rather than the larger market. This approach should appeal to the concert goers as it would reflect their own choices vs being subject to another city, state, or country's choice.

Based on your post alone, I would assume that given Artist B was from Asheville and with their longevity in the industry, they might be able to fill the large venue and give the audience a reason to come back for future shows.

The larger question is whether or not that is data that is accessible to you in order to make an informed decision. Sometimes we think data is always available when in reality it all depends on the process to capture the information.

Reply

○

(http **Shahrod Hemassi (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/224267)**

Apr 3, 2024

Hi Stephen. This is a great post. Unfortunately, I respectfully disagree with your assessment of which artist will be the better option. There are a couple things that you have not taken into account:

- The new venue is in Asheville and Artist B is from the nearby area with many songs about the area. He is likely to have a large following in the area. If you broke down the revenue numbers to the local area, you may find that Artist B generated higher revenue in the Asheville area. Artist A is from London and while he has blown up globally, his fan base in Asheville is likely to be much smaller than Artist B's.
- Artist A is a new artist who has maybe been generating revenue for 2-3 years. He is hot at the moment and that is why his revenue figures the past couple years are much higher. But Artist B has been generating revenue for over 30 years. Most of his fans likely already own his records.

Even if you averaged the number of records that Artist B sold the last 2 years, you would have 105k units and $1.575M per year. If you multiply this by 30 years, you have 3,150 units and $47.25M in revenue which greatly exceeds the records revenue that Artist A has generated. It is also more likely that Artist B used to sell a higher number of records per year in the peak of his career so the number is likely much higher than this.

You could do a similar calculation of the streams. Although streaming has not been happening for all of Artist B's career, we could estimate that he has been getting streaming revenue for the last 20 years. In the last 2 years, Artist B has averaged 47,500,000 streams and $15.5M in revenue. If you multiply that by 15 years, you would have 950,000,000 streams and $310M in revenue. Again, these numbers exceed artist A's career revenue numbers if Artist A has only been earning for the last 2 years. And again, it is likely that Artist B's career revenue from streams is much higher as he is in the latter part of a long career.

Reply

○

(https:/

Timothy Andrew Ramkissoon (https://classroom.emeritus.org/courses/9054/users/226697)

Apr 3, 2024

This is a combination of a previous experience and a scenario as I'm unable to recall specifics in the data.

I was a supervisor for a sales team that sold office products, in this case, printers.

Data:

- Team A sold printers earning $10,000 out of $15,000 in January and $40,000 out of $45,000 in February.
- Team B sold printers earning $12,500 out of $15,000 in January and $37,000 out of $45,000 in February.

Representation:
The data represents the revenue generated from sales out of the total expected revenue for each team across two months.

Conclusion:

At first glance, the two teams did comparable with Team B performing better because they have a higher success rate in January (83.33% vs 66.67% for Team A) and a comparable rate in February (82.22% & 88.89% respectively). However, when comparing the data from both months, Team A earned 83.33% of their expected revenue and Team B earned 82.50% of theirs.

Simpson's Paradox:
The paradox occurs because Team B were more consistent with their sales monthly, however, Team A were able to make up for their shortcoming in January with additional sales in February.

Business Decision:

As a manager, I would investigate further before making a decision solely on percentages. It's important to consider other factors such as the assigned regions for sales teams, customer satisfaction, competitor presence and team dynamics. If all other factors are equal, I might conclude that Team B is slightly more consistent and could be a better model for scaling up sales efforts. However, I would also recognize Team A's strong performance in February and seek to understand and replicate their success across the company.

One point to note was that the preference of the models of printers sold across the teams varied. Team A preferred to sell a model that had a higher cost-value while Team B preferred to sell a model that had more features; therefore, Team A sold a higher quantity of printers than Team B. Since our company also sold printer cartridges and services for repairs and maintenance, we were able to generate more revenue from Team A overall.

The data presented here may not have matched actual data, but the scenario remains the same for my experience with Simpson's paradox.

Reply

(https:/

Roman Jazmin (https://classroom.emeritus.org/courses/9054/users/225803)

Apr 3, 2024

A good example that comes into mind is manufacturing a traditional public gas combustion operated vehicle compared to an alternative battery-operated electric vehicle. The question arises as to which one is a better option?

Initially one will assume that the battery-operated electric vehicle would be a good choice because an electric vehicle's emissions, or lack of any, would be better for the environment when many of those same types of vehicles are used daily.

Ok let us do a deep dive and consider the factor(s) on the actual cost for acquiring the required materials, parts, and the manufacturing process to build a gas operated, combustion engine with a battery operated, electric vehicle's engines.

It is a recorded fact that the cost of extracting the material minerals to manufacture the individual parts to assemble an electric vehicle's battery is more costly compared to extracting the material parts to assemble a gas powered, combustion engine.

Consider the fact that lithium minerals, greatly needed in an electric vehicle's batteries, are costly to the environment, in terms of the amount of the extraction pollution generated, since one needs to dig deeper to extract it from the earth and the location of many natural deposits of that same mineral may be hard to reach, transport from and the supply very rare. We might end up with a low supply of it to construct enough batteries to meet customer demands.

Another hidden factor to consider is the actual cost to maintain an electric vehicle as compared to a gas operated vehicle. The fact is that when our cell phone's batteries reach their limits depending on how many times that it can be recharged, then we naturally must take out the better and replace it with a new one.

Now take that same cell phone battery and increase it s size to power a car sized vehicle, we can't just take it out and replace it with another one. The reality is that it is a costly process to take out an electric vehicle's battery, replace it, and depose of it when it is no longer usable.

The environmental impact of it not being properly disposed of or not being reusable will make having an electric vehicle a very unwise proposition. Compared to the cost of maintaining a gas-powered vehicle, we know from experience that it is much cheaper to have than having an electric vehicle.

As a customer looking for my next ride, these are the things I must take into consideration when I am calculating and determining which type of vehicle, I should buy next and be happy that it's manufacturing impact to the environment won't be as detrimental as a superficial glance of the data looks to me.

Reply

○

(http

Ahmad Abu Baker (https://classroom.emeritus.org/courses/9054/users/234460)

Apr 3, 2024

Your analysis raises important points about the complexities of choosing between gas combustion vehicles and electric vehicles (EVs). You've highlighted the need to consider not only the direct emissions of these vehicles but also the broader environmental and economic impacts of their production and maintenance.

The issues around lithium mining and the cost and feasibility of battery replacement in EVs are significant. It's true that the extraction and production of materials like lithium for batteries can be environmentally damaging and expensive. The end-of-life disposal and recycling of these batteries also present challenges that need to be addressed to fully understand the environmental impact of EVs.

However, it's important to balance these considerations with the long-term benefits of transitioning to electric vehicles. EVs offer the potential for significant reductions in greenhouse gas emissions, especially as renewable energy sources become more prevalent in the electricity grid. Moreover, technological advancements are continually improving battery efficiency, life span, and recycling processes, potentially mitigating some of the issues you mentioned.

In deciding between a gas-powered vehicle and an electric one, consumers like you are right to consider the full lifecycle impacts of each option. This decision-making process is a great example of the complexities that can arise in interpreting data and making sustainable choices. It underscores the importance of comprehensive analysis that takes into account not just the immediate costs and benefits but also the long-term environmental and economic impacts. Your thoughtful approach to evaluating these factors is commendable and necessary for making informed, sustainable decisions.

Reply

○

(https:/

Ahmad Abu Baker (https://classroom.emeritus.org/courses/9054/users/234460)

Apr 3, 2024

Simpson's paradox is a statistical phenomenon in which a trend appears in different groups of data but disappears or reverses when these groups are combined. This paradox can lead to misinterpretation of data and result in poor decision-making.

Example from the healthcare industry:

In a hospital, two treatments, A and B, are used for a certain disease. Looking at the overall hospital data, Treatment A shows a higher recovery rate than Treatment B. Based on this data alone, the hospital management might conclude that Treatment A is more effective and decide to use it more frequently.

However, when the data is segregated by the severity of the cases (mild and severe), a different picture emerges. For mild cases, Treatment A and B have similar recovery rates, but for severe cases, Treatment B significantly outperforms Treatment A. This reversal of trends is a classic example of Simpson’s paradox.

The initial decision to favor Treatment A for all cases would be flawed. This decision was based on aggregated data that masked underlying variations in treatment effectiveness related to case severity.

To solve this, I would analyze the data separately for different severity levels of the disease before making a decision on treatment protocols. This approach ensures that the treatment’s effectiveness is evaluated in the right context, preventing misinterpretation due to aggregated data. As a manager, I would advocate for data analysis that considers relevant subgroups and factors that could influence the outcome. This strategy would lead to more informed, effective, and patient-specific healthcare decisions, enhancing treatment success rates and optimizing resource utilization.

Reply

○

(http

STEPHEN HUTSON (https://classroom.emeritus.org/courses/9054/users/233645)

Apr 3, 2024

Great scenario Ahmad!

Agree that analyzing the data depending on severity levels would be an important factor here in ensuring that appropriate options for treatments are selected based on each patient's circumstance. It would also make sense for the hospital to look into patient's genetic factors, patient demographics like age groups/gender, and seeing if there may be underlying health conditions that also have an effect on the treatments in order for the hospital to see if these affect the success rates of the different treatments available. By continuing to factor in these different data points in the context of these treatments, the hospital will be able to make better informed decisions around which treatment will be more appropriate for particular patients.

Reply

○

(http

Mhelissa Yayalar (https://classroom.emeritus.org/courses/9054/users/233590)

Apr 11, 2024

Agreed Ahmad!

Similar to my post about gender-based weight loss marketing, each individual react to treatments differently. For instance, quantitative age factors and weight factors. You can even apply to new medicine, like booster shots for covid. Boosters shot were developed based on age. Therefore, digging deeper into the data, as well as understanding the different factors are important to consider in the analysis.

Cheers,

-my

Reply

○

(https:/

Priscilla Annor-Gyamfi (https://classroom.emeritus.org/courses/9054/users/226376)

Apr 3, 2024

Scenario:
Nsapa Collections, my beaded accessories business, is evaluating the sales performance of its primary products, Bracelets, and Earrings, throughout the previous month. The objective is to determine which category among these two has recorded a higher overall sales volume, helping to determine their respective production rates.

The Data:

- Bracelets: Sold 500 items in total.
- Earrings: Sold 400 items in total.

Taking a first look at the above data, it appeared that Bracelets have a higher overall sales volume compared to Earrings. As the manager, without any further dive into the data, I went ahead to instruct the purchase of more raw materials to start production for Bracelets to prepare for the next month after concluding that customers prefer Bracelets over Earrings.

Deep Dive:

After realizing my initial conclusion on previous analysis was not entirely correct, my team and I decided to collect additional information on the sales performance by categorizing product based on materials used, such as stones, pearls, and recycled beads. After segmenting the data by product category, it revealed the following:

- Bracelets:
  - Stones: Sold 300 items
  - Pearls: Sold 150 items
  - Recycled beads: Sold 50 items.

- Earrings:
  - Stones: Sold 150 items
  - Pearls: Sold 180 items
  - Recycled beads: Sold 70 items.

New Findings Revealed:

Although initially, Bracelets appeared to have a higher overall sales volume, when the data is segmented by product category, it shows that Earrings actually sold more items in two out of three categories: Pearls and Recycled beads whereas Bracelets made of Stones sold more.

Business Decision:

As the manager at Nsapa Collections, I recognized the presence of Simpson's paradox in the sales data and understood that the overall conclusion can be misleading without considering additional factors.

Rather than relying solely on the aggregate sales volume, I would delve deeper into the analysis of the sales performance within each product category across the three primary materials utilized. This approach would enable me to identify the essential raw materials needed and worth investing in, thereby reducing production costs through bulk procurement. Subsequently, I could tailor marketing strategies and optimize inventory management practices accordingly.

Reply

(https:/

Victor Flores (https://classroom.emeritus.org/courses/9054/users/197659)

Apr 3, 2024

Just like in medical circumstances, the Simpson's Paradox phenomenon can also be encountered by reservoir engineers and geologists when conducting characterization exercises and modeling work for specific reservoirs across the World. The work published by Y. Zee in the paper named Simpson's Paradox in Reservoir Modelling (2008) provides scientific evidence of how the Simpson's Paradox can be observed when characterizing specific downhole conditions unique to some reservoirs in North America and how this natural phenomena can lead to misinterpretations.

Through characterization studies carried out at specific stratigraphic units across North America, it has been identified that the aggregation of two geological formations can lead to the observances of the Simpson's Reversal or Paradox. For instance, if we study the porosity data for two geological formation and two facies in a carbonate reservoir, we can face the scenario which is depicted in table 2 hereunder. From this table, it can be inferred "The beach facies have higher porosities in each of the two geological formations, A and B in respect tot he shoal facies. However, the average porosity of the shoal (9.41%) is higher compared to the beach (9.13%). In other words, the ensemble of the two larger entities in the beach is less that of two smaller entities in the shoal (source: Simpson's Paradox in Reservoir Modeling and Evaluation by Y. Zee Ma, 2008)."

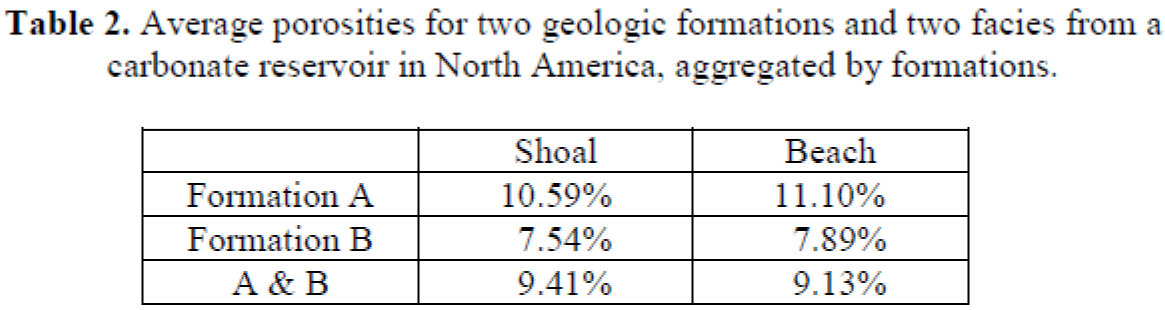

Table 2. Average porosities for two geologic formations and two facies from a carbonate reservoir in North America, aggregated by formations.

	Shoal	Beach
Formation A	10.59%	11.10%
Formation B	7.54%	7.89%
A & B	9.41%	9.13%

Source: Simpson's Paradox in Reservoir Modeling and Evaluation by Y. Zee Ma, 2008

Reservoir modeling examples should consider all variables and implications such that critical decisions can be taken in regards to the target reservoir. Many investments could be at risk if mapping is missed and reservoir properties not well determined. As a consequence, hydrocarbon presence could be in incorrectly estimated in the presence of the Simpson's Paradox phenomenon is ignored.

Note: Simpson's Paradox in Reservoir Modeling and Evaluation by Y. Zee Ma has been attached a scientific evidence used to elaborate on the Simpson's Paradox phenomenon.

SimpsonParadox_in_Reservoir_Modeling.pdf (https://classroom.emeritus.org/files/2420768/download?download_frd=1&verifier=EhKEpfGzN7ED0ncRxdEepKAH2CunB0ahVW9yGox2)

Reply

(https:/

Shahrod Hemassi (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/224267)

Apr 3, 2024

I was managing the engineering department for an EPC. It was my responsibility to report progress on each project that we were working on.

My engineering team was tasked with producing engineering designs. Here is where we stood with our progress on the engineering documents for a project:

Progress on Engineering Design Documents

Doc #	Planned Senior Engineer Hours	Planned Junior Engineer Hours	Hours Completed	Progress Percentage
1	15	35	50	100%
2	20	30	50	100%
3	40	60	100	100%
4	50	50	100	100%
5	35	65	100	100%
6	70	30	50	50%
7	65	35	50	50%
8	70	30	40	40%
9	90	60	45	30%
10	100	50	45	30%
TOTALS	555	445	630	Average: 70%

We had 10 design documents and 5 of them were completed. We also had made progress on the others. I reported that we were 70% complete.

Also, we used many engineers who all had differing rates. We calculated a blended rate of $85/hour for the engineers, so I calculated that the following costs were needed to complete the work:

Cost to Complete based on Hours = $85/hour x 1,000 total hours x 30% remaining = $25,500

So I reported that we were 70% complete and would need $25,500 to complete the remaining work.

Reply

○

(https:/

Swati Sharma (https://classroom.emeritus.org/courses/9054/users/236938)

Apr 3, 2024

In my current organization, we identified Simpson's paradox as well. We were working on identifying terminations based on hourly and salaried employees and by the different locations. We had mentioned that Location A was doing better than Location based on the turnovers.

However, when looking at the data closely, we found that Location A had significantly less employees and therefore their turnover percentages were very low compared to Location B. When this was weighted based on the number of employees, we identified that the percentages were vastly different.

For weighting, we divided the number of employees terminated from Location A to the total number of employees represented for both locations. This weight allowed to bring the percentage down for Location B and identified that Location B was performing better. This helped my organization to focus on the right location and implement strategies and retention programs for Location B

Reply 👍 (1 like)

○

(http

Mariana Flores (https://classroom.emeritus.org/courses/9054/users/237198) ⋮

Apr 3, 2024

Hi Swati, so nice to meet you. Great post - people analytics is truly fascinating. Let’s say, for example, Location A has 150 employees and Location B 850 employees this brings the total number of employees to 1,000 in a given time-period. Location A is doing better than Location B in terms of turnover, by having 35 and 210 employees correspondingly this brings the total number to 245 employees. The total turnover percentage across both locations is 24.5%, 23.3% for Location A and 24.7% for Location B. Did turnover vary by month and/or across salary and hourly employees?

P(Turnover Employees for Location A) = P(Turnover Employees for Location A | Salary)P(Salary) + P(Turnover Employees for Location A | Hourly)P(Hourly)

35/150 = (28/125)x(125/150) + (7/25)x(25/150)

P(Turnover Employees for Location B) = P(Turnover Employees for Location B | Salary)P(Salary) + P(Turnover Employees for Location B | Hourly)P(Hourly)

210/850 = (52/250)x(250/850) + (158/600)x(600/850)

For example, let’s keep time constant and say, that when we add compensation type, 22.4% of salary and 28.0% of hourly employees turnover for Location A while 20.8% of salary and 26.3% of hourly employees turn over for Location B. Turnover for salary and hourly employees is lower for Location B compared to Location A. It is my understanding that this trend would be an example of Simpson’s Paradox because the lower percent turnover for Location A when the data is combined would disappear when we group employees into salary and hourly.

I agree with you - Location B performs better when we group employees by compensation type, salary and hourly and focusing on Location B with 85% of total employees. Thank you for sharing.

Reply (1 like)

○

(https:/

Mariana Flores (https://classroom.emeritus.org/courses/9054/users/237198)

Apr 3, 2024

Models are a simplified version of reality and thus it is of essence to understand the real world we are modeling after to ensure that all appropriate variables are accounted for when deriving data-driven recommendations. I recently worked with a small business owner who was interested in measuring performance of various campaigns. Campaign data included channel, tactic, audience segment, timestamp, browser, device type, impressions, clicks, views, purchases, etc.

This data represents raw historical campaign data on performance across two years. Conclusions were based on historical findings analyzed at a high level to understand overall performance by channel and campaign objective. When evaluating campaign performance on a high-level view and controlling for campaign objective there were specific channels with strong performance across specific stages in the conversion funnel. Decisions had been made in terms of where to allocate funding for each channel to strategically meet or exceed business objectives. Yet return on investment across specific products was not exactly what was expected so I was brought in to delve deeper into the data and provide recommendations.

After further analysis into the campaign performance, I included lurking or confounding variables including product and audience segment. I noticed performance varied across both and that this trend was observed across channel and campaign objective. Thus, to reach the target business objectives it was recommended to adjust investment to incorporate for these findings. Measuring campaign performance a few months later, we saw improved efficiency across key performance indicators (KPIs), although without hypothesis testing and a test and control group results cannot be directly attributable in a scientific manner. We were moving in the right direction and continued to iterate from our learnings to make intelligent data-driven decisions.

Reply

○

(https://

Javier Di (https://classroom.emeritus.org/courses/9054/users/226884)

Apr 4, 2024

Assigment 1.1 Sympsons Paradox:

A business situation I was involved with in the investing field in which I work where the data was misinterpreted and a poor business decision was made relates to deciding to not buy small caps on an up market in 2024 based on the 2023 data and behavior, even though small caps have historically outperformed large cap stocks (Figure 1) and even more so in up markets.
The data the team had looked at was the 2023 US stock market returns. This data showed the following relevant returns for the indices S&P 500 (top 500 Companies Market Weighted) and Russell 2000 which contains Small Caps:

2023 S&P 500 Return= +26%
Russell 2000 Return= +16.9%

Based on this data, the team made a rushed decision that Small Caps were not working properly, market conditions may had changed and now didn't make sense to buy them anymore to participate in the market upside as they didn't outperform. At an aggregate high level, this is what the data indicated, but breaking down the data provides a different picture in line with the Simpson's Paradox.
The data was simplified too much by not looking at its components and a poor decision made. What would have been the right thing to do? The right thing to do was to understand what was behind these numbers, where they come from?

Disaggregating and Understanding the Data:
1) First understand how each index return is measured. The first insight is that the S&P500 is a market weighted Index. This means that the weightings are considering based on the Market Caps of the components. So very large companies get a disproportionate weighting in the Index.
What has happened over the last decade is that the larger companies in the index such as

Microsoft, Tesla, Apple, Amazon, Nvdia have produced outstanding returns and their weighting as % of total S&P 500 weight has grown from ~14% in 2015 to ~29% currently. These companies returns have greatly exceeded all others and their weighting jus kept getting larger.

2) Knowing this fact as illustrated in Figure 2, we can then adjust the data to consider what the S&P 500 Returns in 2023 would have been if done equal weighted and without the Magnificent 7 and the compare that with small caps to see if Small Caps still outperformed after adjusting for the distortion of the large returns and market weighting of the magnificent 7. Adjusting this data would look like:

2023 S&P 500 Return= +26%
2023 S&P 500 Return Equal Weighted= +12%
S&P 500 Return Without Magnificent 7= +8%
Russell 2000 Returns= +16.9%

Cleaning up and reviewing the data in this view would indicate that the Small Caps Russell 2000 Index still outperformed the Large Caps in the S&P500 Index either Equal Weighted or eliminating the Magnificent 7 and their Distorstions as:

Russell 2000 Returns= +16.9% >> 2023 S&P 500 Return Equal Weighted= +12%

And thus drives to better decisions, demonstrating that it is a good decision to still buy Small Cap Stocks

Figures:

1. **Small Caps have historically outperformed large caps:**

2. **Seven largest companies share of S&P 500 (%):**

Reply

○

(https:/

Gustavo Santana (https://classroom.emeritus.org/courses/9054/users/120927)

Apr 9, 2024

When working as a Data Analyst for a physiotherapy digital company, I gathered data about their participants and how well they did during their treatment.

At first glance, when gathered all the information, we could see that only 12% of all our participants had progress from the beginning of the year, which seems odd since the main feeling is that the app generally works well in improving the well-being of the ones eager to use it as recommended.

Looking further we could see that we should not calculate all the participants, but only the ones with at least 2 assessments, being possible to track if they have progressed, separating from 70% of our users that only had 1 assessment or none at all.

Now we could see that 92% of our users with 2+ assessments had improvement in their mobility and reduced pain. We imagined what that would be if we separated this information by age categories, and the best way to be seen, that our older participants were the most benefited ones, having higher scores restoring their movements.

We used this data in marketing and sales presentations, showing not only the quality of our service but also the quality of our data management.

Reply

(https:/

Mhelissa Yayalar (https://classroom.emeritus.org/courses/9054/users/233590)

Apr 11, 2024

An example of when Simpson's paradox may apply is through data analysis of whether men vs woman have better chance of losing weight. If so, develop target marketing campaign to women.

1. Data/Signal:

- Weight loss total

- Gender: Categorized as male or female.

2. Aggregated data result:

- All participants, women tend to lose more weight than men on average

- The average weight loss for women is 18 lbs, while for men, it's 10 lbs.

- Within each weight loss program, men outperform women.

3. **Issues**:

- The data does not include other factors that influence losing weigh, such as preexisting conditions or the types of weight loss program

4. **Simpson's Paradox**

- If data is combined, it would appear that women lose more weight on average than men. If the data is analyze further, the men's average weight loss data is higher for some of the weight loss programs.

5. **Adjustments and results:**

- Instead of gender-based marketing, the analysis should dig deeper into data related to the types of weight loss program that are generating actual weigh loss. Then, include gender to focus the campaign.

- Assuming that all women will lose weight because of the aggregated data does not yield the best outcome because, not all women are created equal. Likewise, with men.

Reply