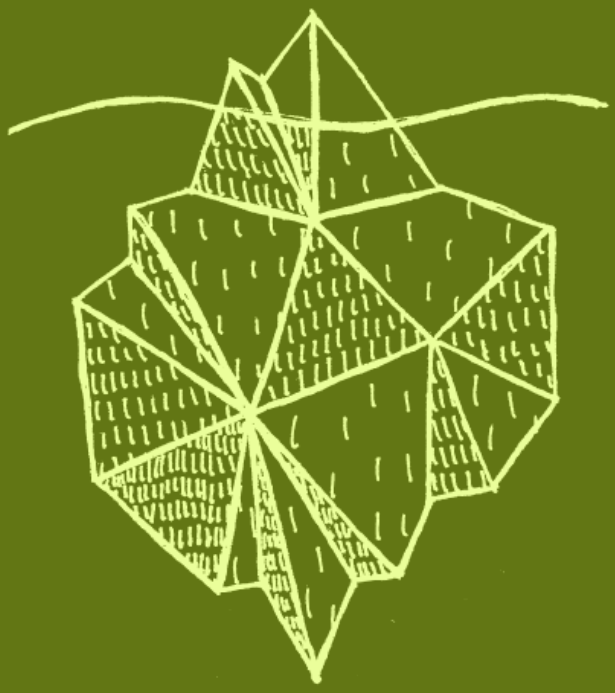
Avoid Misrepresenting Data



Avoid Misrepresenting Data

Written by: Matt David
Reviewed by: Blake Barnhill

Table of Contents

Cognitive Biases

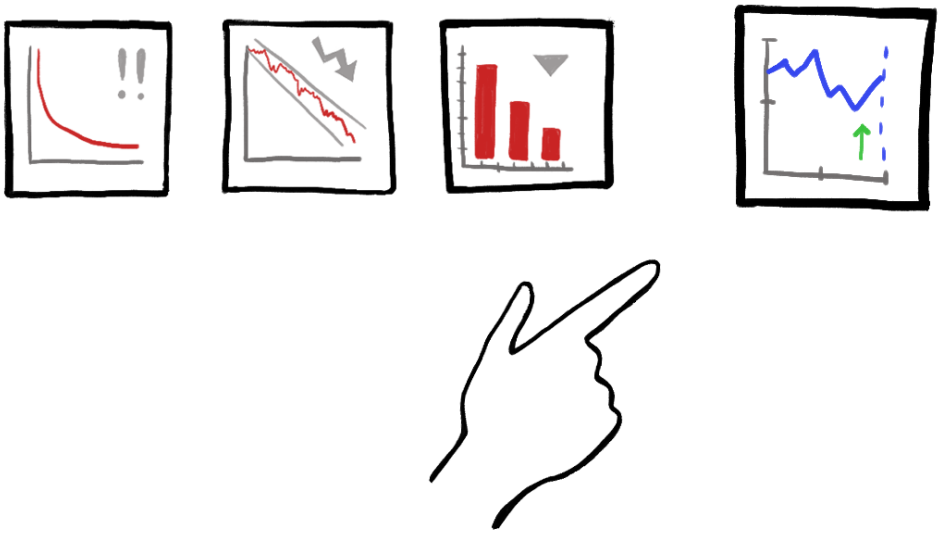
Confirmation Bias

What is Confirmation Bias?

Confirmation bias is the tendency to **seek out or interpret data to confirm beliefs you already hold**. It does this to the exclusion of contrary evidence.

In a business context, this means ignoring data that is suggesting that some aspect of your feature, product, or business is not working because you found another metric that seems to suggest that it is working.

Confirmation bias pressures you to ignore the negative signs and focus on the positive.



How to Detect Confirmation Bias?

There are 3 common signs that confirmation bias is influencing the data you are looking at.

- Only good news
- Limited metrics reported
- Obscure metrics being used

Only good news

If the only news is good news confirmation bias is probably hiding some inconvenient metrics. Any business decision, product, feature, or process change is likely to have both positive and negative consequences. What you want is for the change to have more positive consequences than negative ones.

Limited metrics reported

Usually people present a few metrics because they are trying to be brief and that is desirable in most cases. However it is unethical to not present data that contradicts their

story if they have it. If someone shows a single metric in their presentation, more often than not important contradictory data is being left out.

Obscure metrics being used

Many times we cannot measure exactly what we want so we have to settle for proxy metrics. As these metrics become more and more abstracted our confidence in them should diminish in how accurately they represent what we want to be measuring. If people are confidently sharing convoluted proxy metrics they are likely looking for ways to find positive signals in the data.

What to do about confirmation bias?

Institutionalized disconfirmation

Institutionalized disconfirmation is how academia addresses the confirmation bias problem. Anytime you want to publish a paper it typically has to be peer reviewed. So you have many other self interested scientists making sure that what you submitted has sound logic and analysis behind it.

However academia is finding that this level of review is insufficient, especially in the social sciences. Too many professional articles report data that does not hold up in subsequent trials. In fact it has been dubbed the Reproducibility Crisis. Reproducibility is a much higher bar than examining the analysis, it provides more data to confirm what the paper has laid out.

How can you apply this to your company?

The first step is to create a peer review process. Before an analysis is presented to the company have a system in place where another analyst reviews your analysis and checks to see if they reach the same conclusions. If you have the time and resources it may be worth trying to replicate a test run within your company to confirm it's findings.

3rd party audit

Hiring a 3rd party audit is how government tends to address the confirmation bias problem. They have different organizations randomly audited by a 3rd party (the Government Accountability Office) to see if records are accurate and to determine how the organization is really doing. Is money being spent properly, and are they actually achieving the results they claim?

At times, a journalist can serve as the 3rd party who investigates and exposes shortcomings of government agencies and companies.

How can you apply this practice to your company?

Use data analysts to randomly check analysis being shared. If there are any mistakes or misleading charts, have them reach out to who created the analysis and go through the issues together. Once corrected send an update to anyone who was using or viewing that data.

You can also hire consultants to come in and audit what is going on in your company. This is commonly used in the accounting and finance departments during due diligence periods for investors.

Insurance rules

This is one way business tries to address the confirmation bias problem. It does not prevent confirmation bias so much as it places rules on the company to prevent certain errors from happening. For instance it may require an inspector (3rd party audit) before u can get a policy on a house. Or it will not cover your life insurance if you choose to skydive or engage in other behavior that is unpredictable and risky. These rules can be overly burdensome but they do prevent certain silly mistakes from happening.

How can you apply this practice to your company?

Create parameters of what people can and can't claim with their data. Restrict certain types of phrases such as anything causal. If they make claims that exist outside of those parameters, it does not have approval to be shared and so claim is not "insured" by the company.

Culture

This is another way businesses tries to address the confirmation bias problem. They create a culture in which failure is not expected but is understood and does not cost the employee large consequences. This has been done most famously in Toyota, where any employee can stop the whole production line if they notice a mistake and in Google X where they celebrate a projects failure with champagne and bonuses.

This culture helps you to second guess your motives for only showing successful metrics.

How can you apply this practice to your company?

Emphasize the mission of your company. Show that the mission is more important than feeling good about the success of a product or feature. To make real progress toward the goal we are going to mess up a lot and we need to know how we messed up so we can make better decisions in the future.

Selection Bias

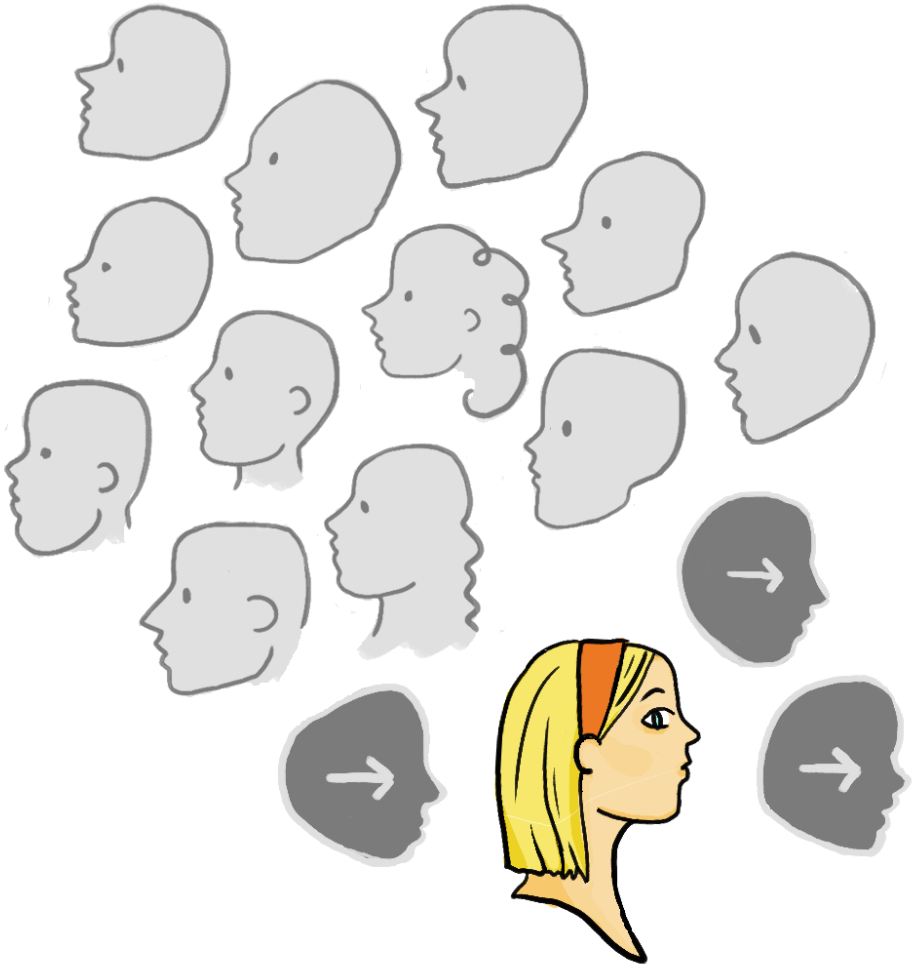
Selection Bias is when the group chosen to be analyzed is not representative of the population you are trying to draw conclusions about.

How do you get an unrepresentative group accidentally?

- Convenience
- Self Selection

Convenience

You choose a group of people to analyze in a way that is not representative of the population because they were convenient to measure.



Example:

You ask three of your friends if your new feature is valuable. While easy to ask them, are they actually representative of your customer base?

Convenient Biases Selections:

- Engaged customers
- Latest cohort of users
- People in a particular geographic region

Self-Selection

The group of people who opt in to be analyzed have characteristics that are not representative of the whole population.

Example:

You send out a survey to all of your customers to gauge their satisfaction with your product. While this seems like it would provide for good feedback, you are likely to get responses from people who are very opinionated, very angry, or people who are trying to waste time at work.

Common Biased Self Selectors:

- Very negative people
- Very positive
- Early adopters
- Power users

Selection Bias in Business

Let's say you want to introduce a premium feature in your BI Tool. You send out an email to the most active users asking them if they are interested in trying it out. Several people respond to the email and you begin giving them access to the feature.

This seems rational, they are the most engaged, they deserve a sneak peek and they might have great insight about the feature.

Why might this selection of people cause our analysis of the feature to be wrong?

- They might try every new feature regardless of if it provides value to them.
- They might see this as a way to get in touch with someone at the company.
- They might want to share their ideas for features.

While these motivations aren't necessarily bad, their feedback can be misleading.

How to fix:

Be deliberate about reaching out to a representative sample of people to test new features. Use qualifying questions to understand more about them and to give you an opportunity to select a balanced sample.

If you do send out a large email and most of the people who respond are early adopter types you may want to aggregate their feedback down to what portion of your customers they represent. Then you can weigh their feedback more evenly with the few typical customers who tried out the tool.

Summary

- Selection Bias
    - Make sure the group of people you test something on is representative of the population you want to impact. Do this by randomizing your sample.
- Self Selection
    - Make sure the people who voluntarily participate in something you are analyzing are representative of the population you want to analyze. Do this by having qualifying questions.

Survivorship Bias

What is Survivorship Bias?

Survivorship bias is the tendency to draw conclusions based on things that have survived some selection process and to ignore things that did not survive. It is a cognitive bias and is a form of selection bias.

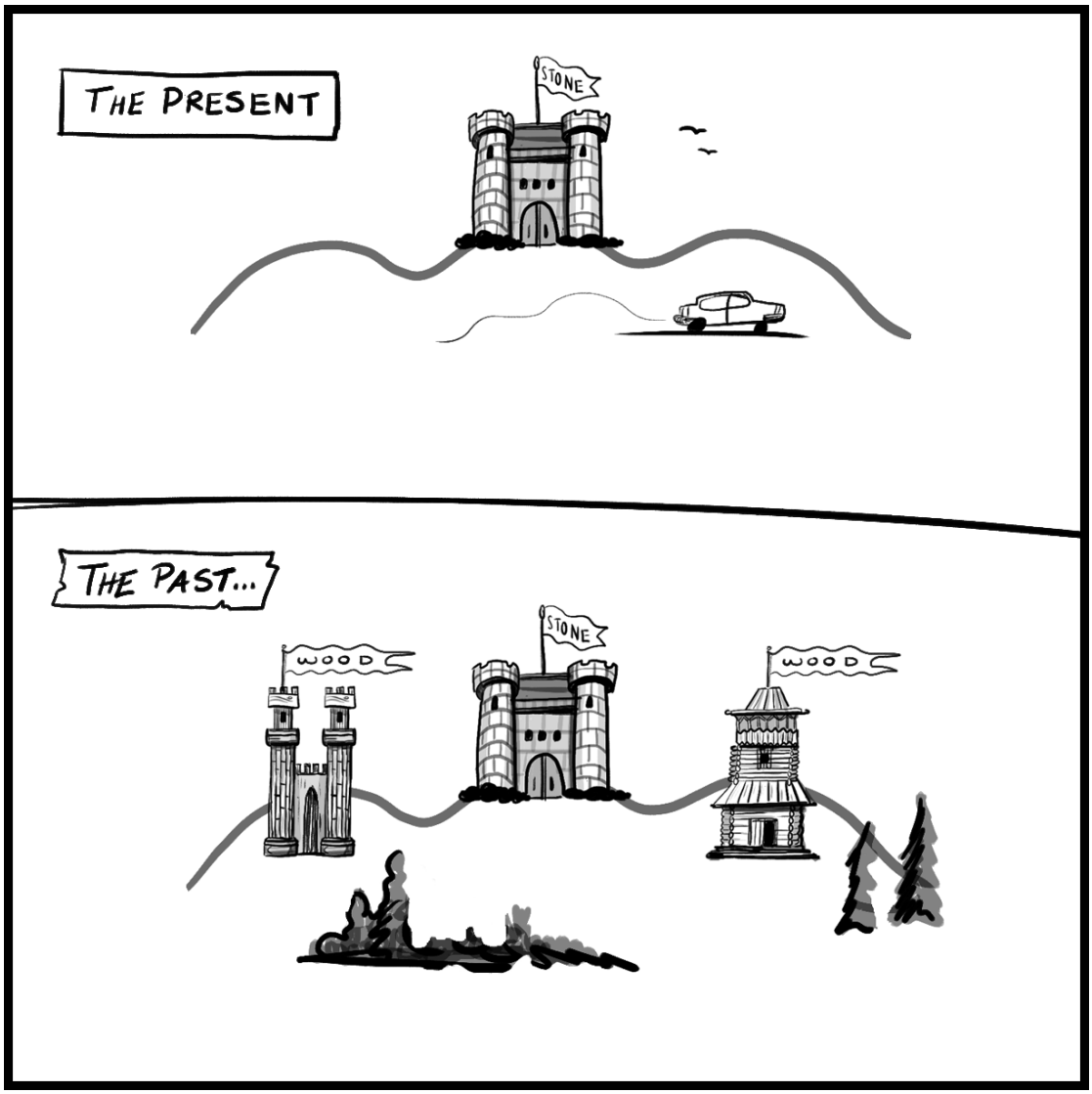
There are two main ways people reach erroneous conclusions through survivorship bias, inferring a norm and inferring causality.

Inferring a norm

The things that survived a process are the only things that ever existed

Example:

“Most castles were made of stone” vs “most castles were made of wood but were destroyed by fire or withered away over time.”



Present

All we see is the surviving stone castle

Past

In the past we see the same stone castle but also the more numerous wooden castles. The wood castles did not survive to the present.

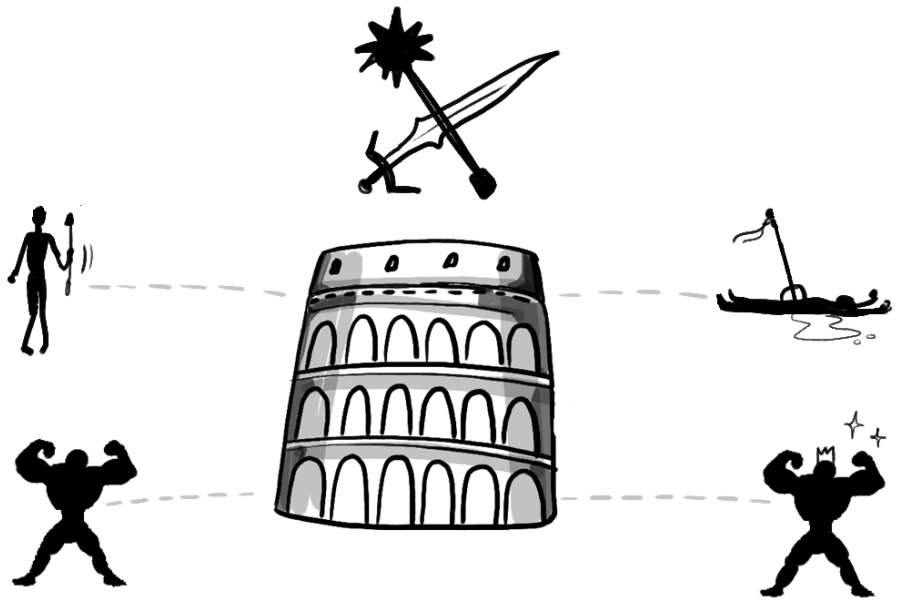
People assume that what they see or have concrete evidence of in the present are the only things that have ever existed. When in fact most of the things that have existed in the past do not exist in the present.

Inferring causality

Anything that survived a process was impacted by that process.

Example:

“Men get tough fighting in the coliseum” vs “only tough men survive the coliseum.”



People assume the competition in the coliseum caused the outcome but actually it filtered out the weak people and the strong people survived. People did not necessarily grow from their experience in the coliseum so much as it exposed who was the most fit for the coliseum

Survivorship Bias in Business

These biased ways of interpreting events and data is common in business. Let’s imagine you work at a business intelligence software company with a two week free trial period that just launched.

After one week, in the middle of the trial, you only have a few people still active. Let’s say everyone who is still engaging on your site is a data analyst. The data analyst users are creating progressively more complicated analyses in your BI tool.

What conclusions could you draw?

- My BI tool resonates with data analysts
- My BI tool empowers deep analysis

Why might these conclusions be wrong?

My BI tool resonates with data analysts (inferring a norm)

Without examining the people who stopped using the tool, we do not know if that group of people in the trial had data analysts in it as well. Let's say everyone that started the trial was a data analyst and you had more people give up than keep engaging. This would directly contradict your previous conclusion.

We need to do more investigative work to find out about those that did not survive the free trial process before drawing conclusions. Just because all of your engaged users are data analysts does not mean your product resonates with all data analysts.

How to reach a more informed conclusion:

Analyze everyone who started the trial to find patterns that truly separate cohorts. Maybe there is something unique about those who engaged and those that did not continue the trial, but we must look at the non survivors before inferring a norm.

My BI tool empowers deep analysis (inferring causality)

Right now you are not comparing your users' capabilities against a control. They may be doing the most advanced thing in your BI tool, but they might be capable of even more advanced analysis when using other BI tools. They might be very skilled analysts and succeeding in spite of your design, not because of it.

How to reach a more informed conclusion

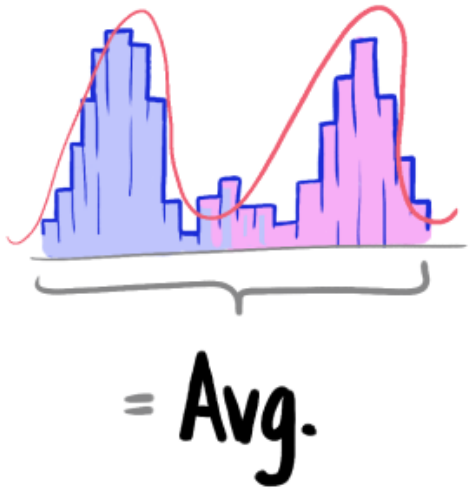
Do a pre-assessment of users skill level to see if your product/features are actually adding value. Run tests of people performing similar tasks with the tool(s) they currently use to accomplish this analysis. These tests become a great PR piece for your product or feature if you can cite a positive benefit: "Users complete BI analysis 2 times faster on our tool than our leading competitor's tool".

Summary

- Analyze the full cohort, not just the users who are still engaging after two weeks
- Compare user behavior with competitor tools to judge impact of your product
- Conduct pre-assessments to judge current skill levels

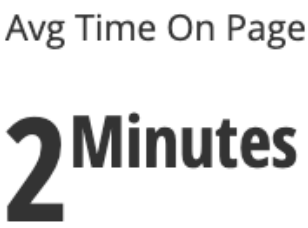
Analysis Mistakes

Statistic vs Distribution

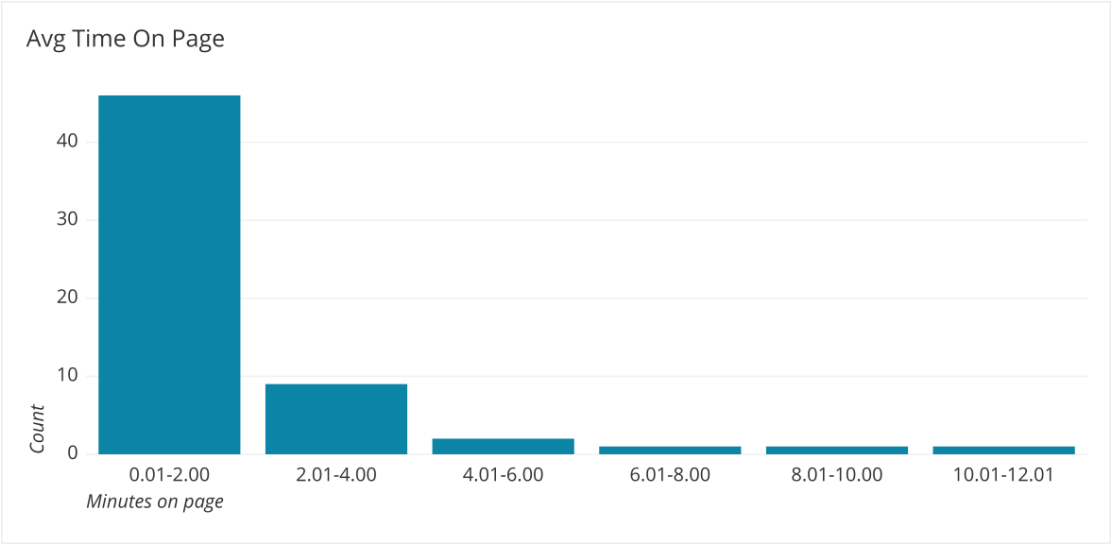


The problem with a single statistic

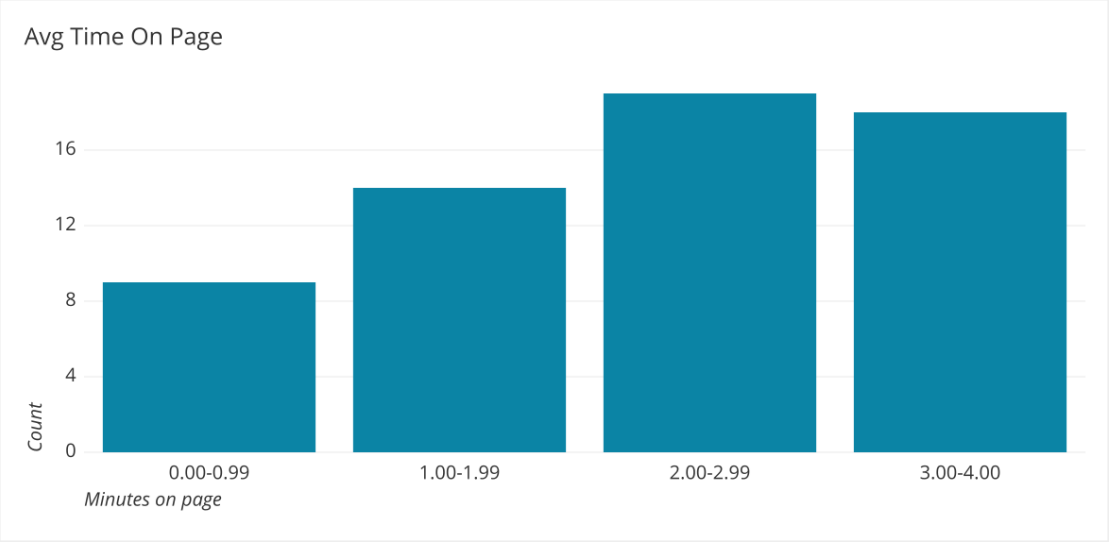
Most metrics are reported as a single statistic: Average time on page, Number of Active Users, Customer Acquisition Cost. While high-level stats can be informative, relying on them to accurately represent the underlying data can be problematic because they can hide important patterns in the underlying data.



The amount of time on page above seems respectable! Let's look at the actual underlying distribution of data points.



Here we can see most people are on the page for under 2 minutes, and we have some outliers that are affecting the average time on page. The statistic (Avg Time On Page) doesn't represent the actual data well. On the other hand, if your data is fairly normally distributed, then the average will represent the underlying data well:



Distributions help you tell a much more nuanced story than a single metric.

Create a Distribution

While you can get a stat quickly with SQL with commands such as:

```
SELECT AVG("Time On Page")
FROM Traffic;
```

Creating a distribution is a bit more complex. First you have to create buckets for the data, which means you need to organize "evenly sized" ranges for your numeric data to fit into.

If you had the numbers {1,2,3,3,6,6}, you could bucket them into two groups: 1-3 and 4-6. The first bucket 1-3 would have 4 values in it {1,2,3,3} and the second bucket 4-6 would have two values in it {6,6}. You could also bucket them into three groups which would be 1-2, 3-4, and 5-6.

Bucketing can be done using CASE WHEN. Bucket sizes should be the same with the exception that the last bucket can have an open ended upper limit if there are extreme outliers. Figuring out the correct bucket size to use takes some trial and error to capture the right amount of variation in the data.

Put the buckets into a Common Table Expression and then use a COUNT aggregation on your newly created column.

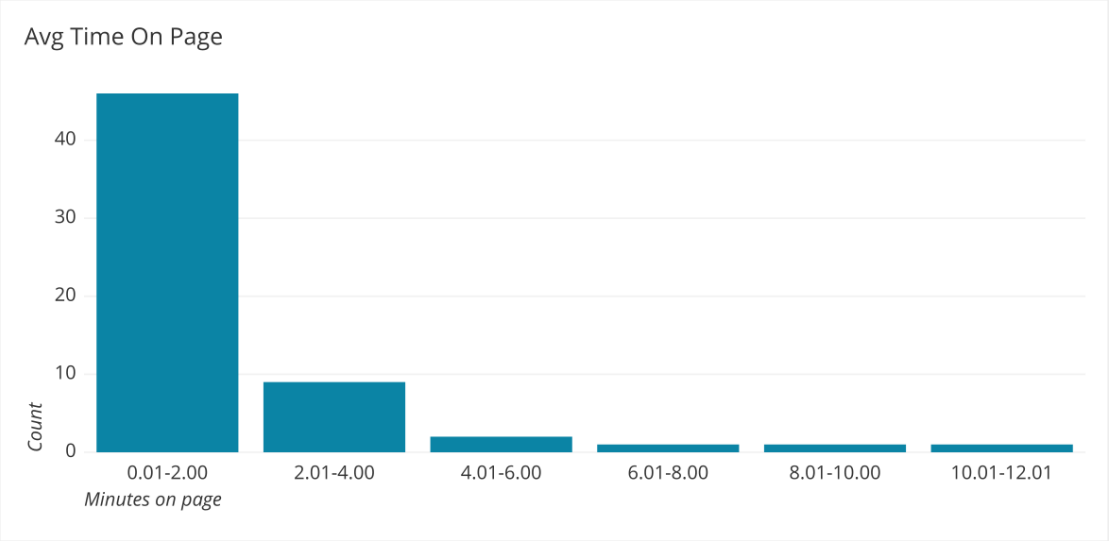
```
WITH 'Buckets' as (
  SELECT
    CASE WHEN "Time On Page" < 1 THEN '0.00-0.99'
    WHEN "Time On Page" < 2 THEN '1.00-1.99'
    WHEN "Time On Page" < 3 THEN '2.00-2.99'
    WHEN "Time On Page" < 4 THEN '3.00-3.99'
  END AS "Minutes on Page"
FROM data)

SELECT COUNT(*)
```

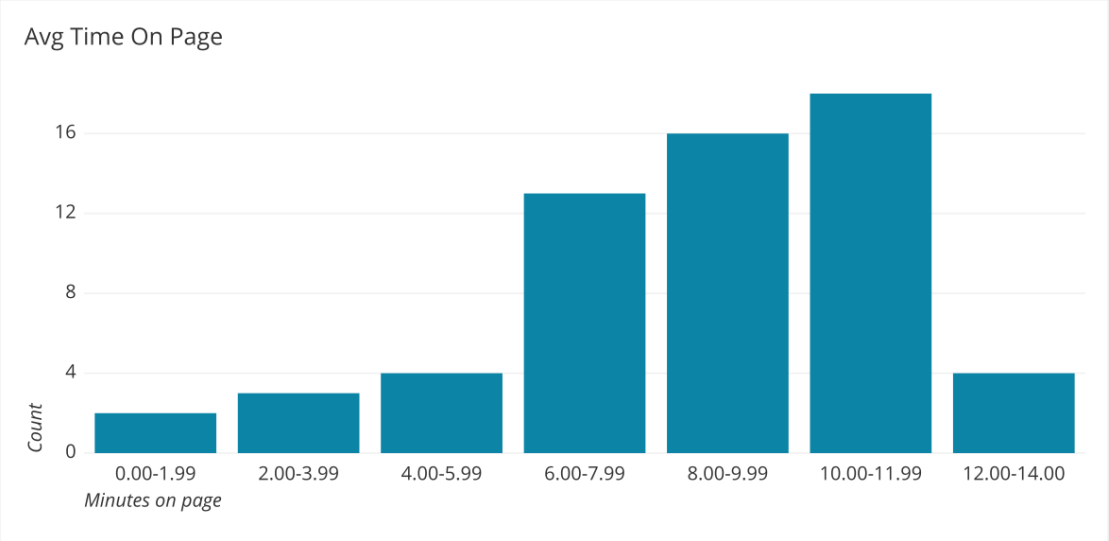
```
FROM Buckets
GROUP BY "Minutes on Page";
```

In many BI tools creating a histogram is a built-in type of chart that can take in any numeric field, bucket it, and then chart it appropriately.

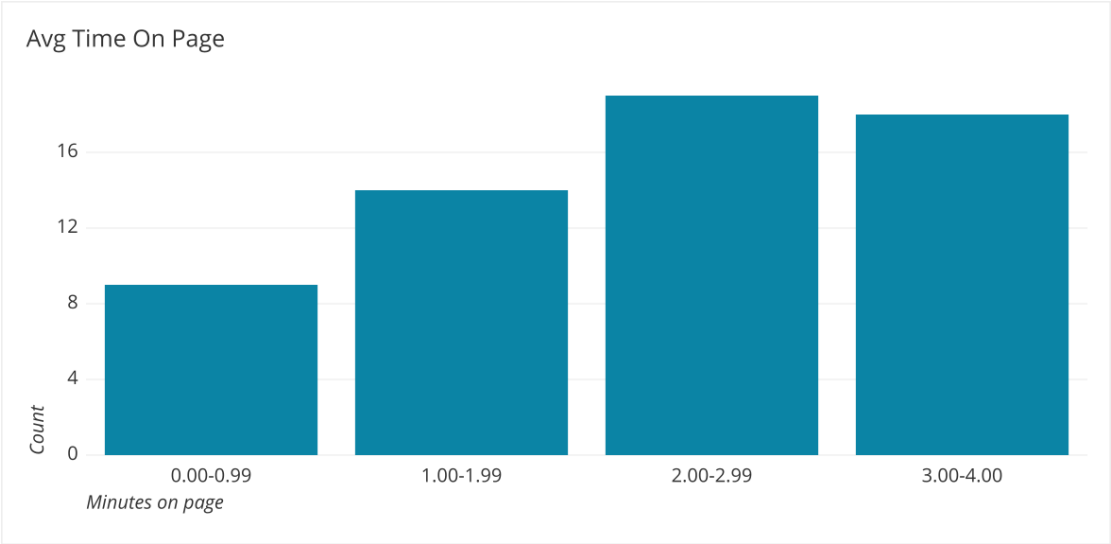
Interpret a Distribution



Right Skewed - Since most of the data is lower than the average, using a median instead of an average would be more representative of the data because it falls more in the center of the actual data. This is because it is less affected by values in the tail.- Most of the data is lower than the average, using a median instead of an average would be more representative of the data because it falls more in the center of the actual data. This is because it is less affected by values in the tail.

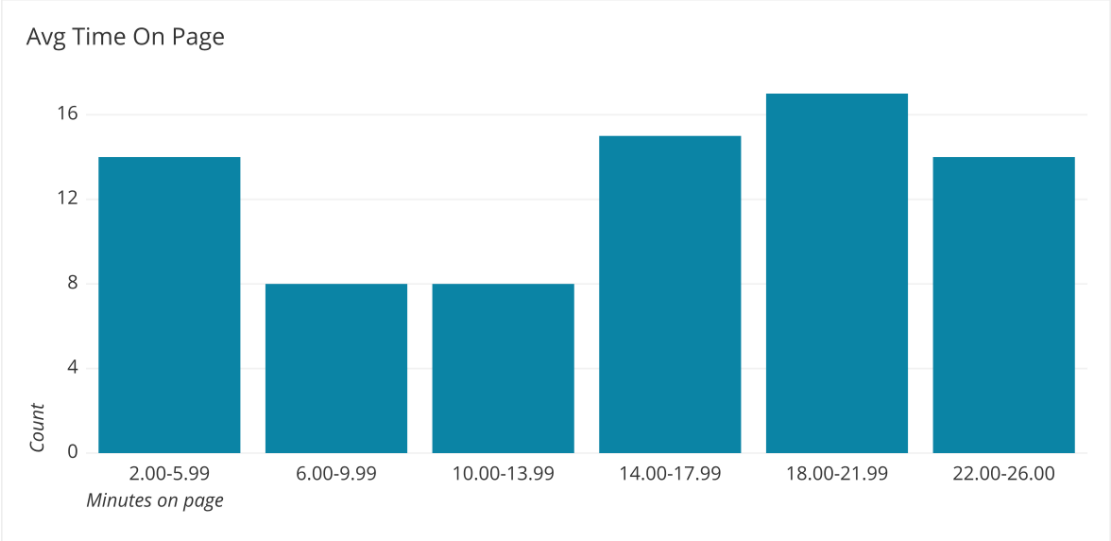


Left Skewed - Since most of the data is higher than the average, using a median instead of an average would be more representative of the data because it falls more in the center of the actual data. This is because it is less affected by values in the tail.



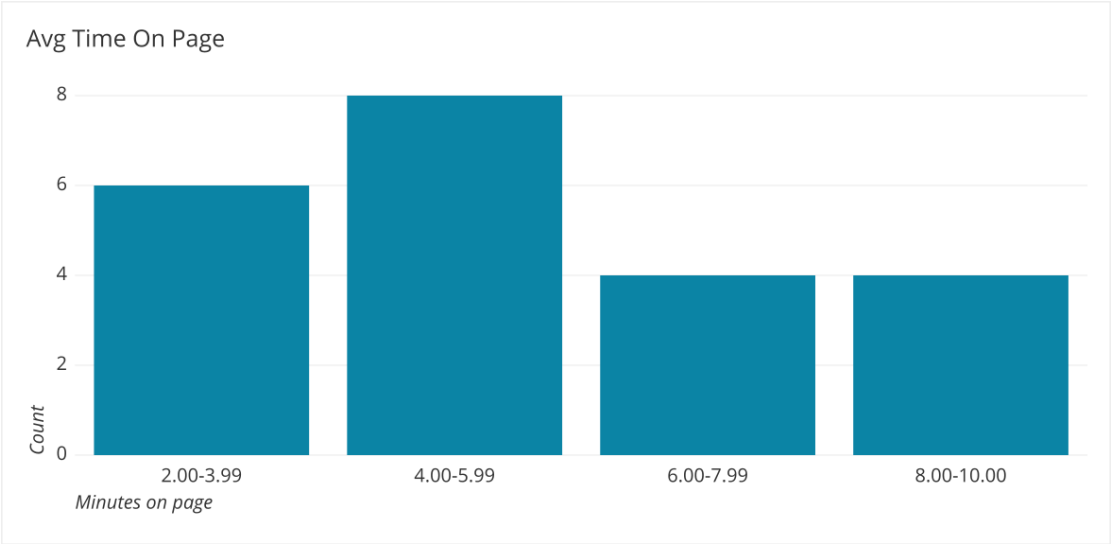
Normal - Using an average or median here is acceptable because they both fall within the middle of the data.

Note: This is technically a **unimodal** symmetrical distribution, but often people will refer to distributions that looks like this as a normal distribution. To be a real [normal distribution], it needs to have a very specific set of criteria that this distribution does not have.

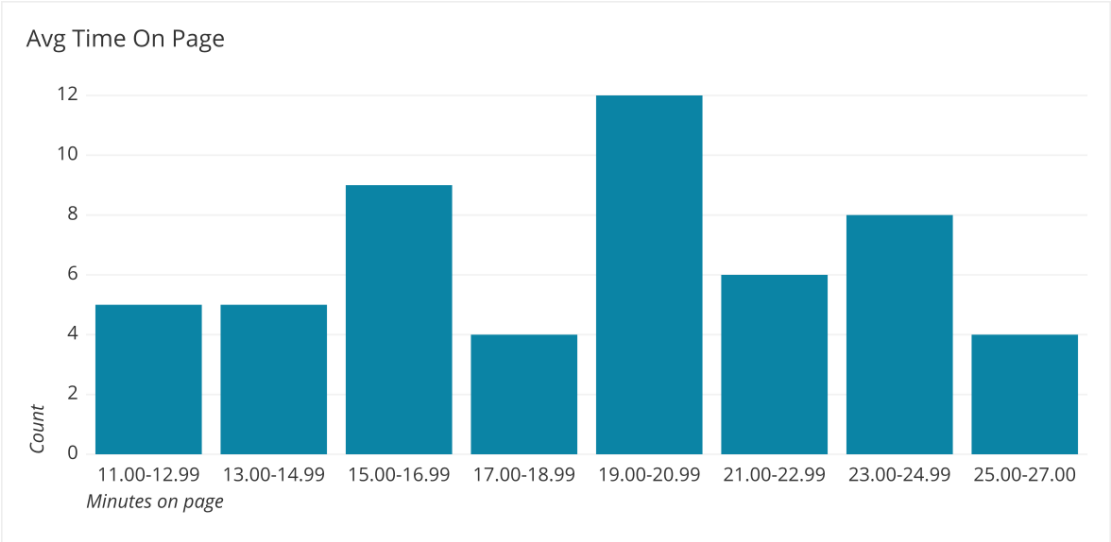


Bi-Modal - Neither an average or median is representative because there is more than one peak in the data. Split the data between the peaks and then report a summary stat on each section of the data.

We can look closer at the peak on the lower end by making the bucket size smaller and filtering the data to be less than 10 minutes on the page.



It looks to be normally distributed, now we can look at the higher end peak by making the bucket size smaller and filtering the data to be greater than 10 minutes.



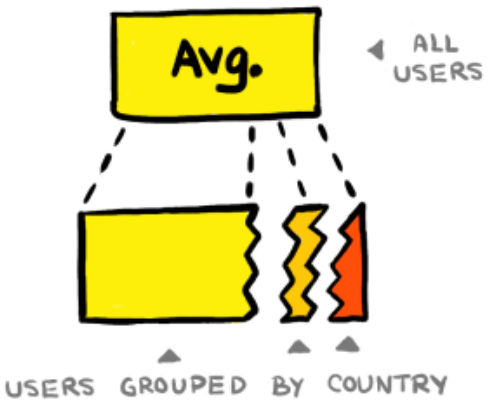
By splitting and re-bucketing we can see in greater detail what the underlying data looks like and which statistics would be a better representation of the actual data.

Summary

While statistics such as a mean or median are commonly used and easy to understand, a distribution adds more nuance and clarity to the data. Even if you do not end up displaying your distribution, you should look at it to know how well your summary stat represents it.

- Always look at the distribution of the underlying data.
- Verify that the high level statistic accurately represent the underlying data.
- There are many types of distributions:
    - Right Skewed
    - Left Skewed
    - Normal
    - Bimodal
    - And more

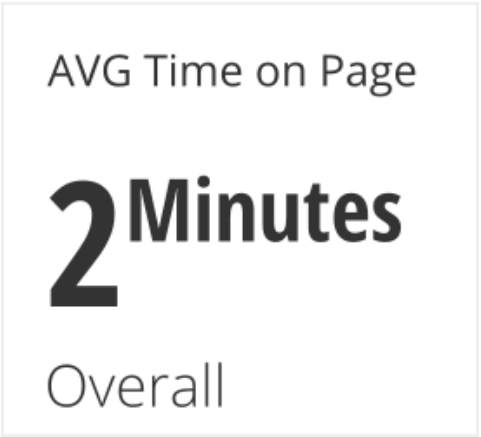
Overall vs Groups



The problem with overall statistics

Overall statistics that describe all of your users or visitors can be misleading, the overall statistic does not show you nuanced patterns about the underlying data. As you saw in statistic vs distribution you know that distributions help you see these patterns but there are multiple ways to examine the underlying data. Another method is to group the data by different categories.

Let's start with the same high-level statistic:



When you group this high level metric of AVG Time on Page by different countries you can see how it varies:

The AVG Time on Page is much higher for France and is lower for the other countries. Why is this the case? This is a tough question to answer since the data only shows you what is going on and not why it is happening. However, this is the right question to be asking and the sort of question that you only come to once you start grouping the data.

Common ways to group data about web site visitors:

- Country
- Age
- Gender
- Device
- Education
- Products purchased
- Start date

Create metrics for groups

Group by Category

To get a high level metric we can use an aggregation on a column in SQL:

```
SELECT AVG("Time on Page")
FROM Users
```

To get a high level metric broken out by group we need to add the group to the SELECT and then put it in the GROUP BY clause:

```
SELECT Country, AVG("Time on Page") as "AVG Time on Page"
FROM Users
GROUP BY Country
ORDER by 1 DESC
```

Order it by the group as well to be able to scan the data quickly for outliers.

Group by Start date

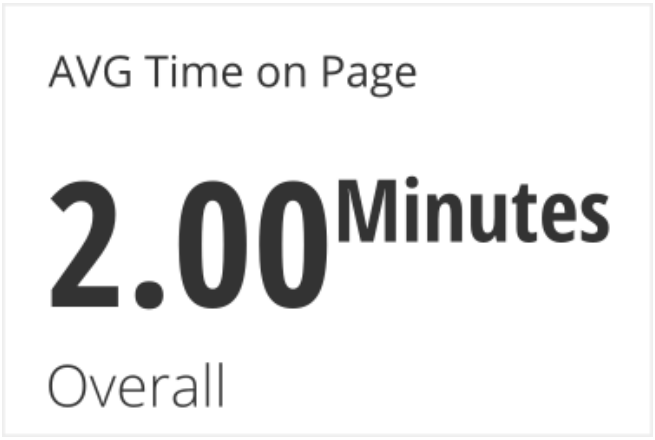
It is quite common to group data by date in analytical queries. You can turn dates into truncated strings using the TO_CHAR function. Here we will use it to truncate down to the Year and Week.

```
SELECT TO_CHAR(First_Visit, 'IYYY"-W"IW'), AVG("Time on Page") as "AVG Time on
FROM Users
GROUP BY Country
ORDER by 1 DESC
```

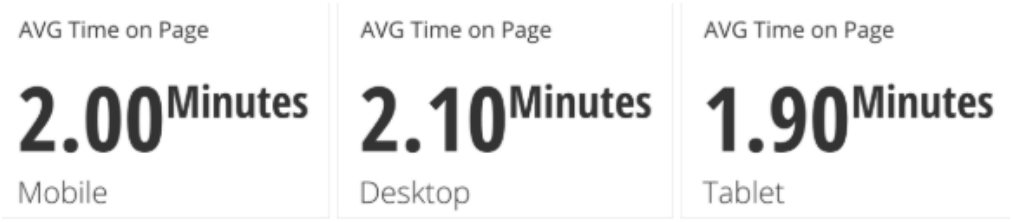
Interpret Grouped Data

Once data is grouped, the statistics for each group vary. How much they vary can give you different ways of investigating the data

Low Variance



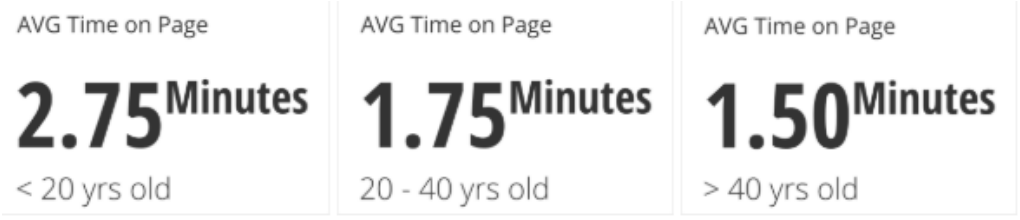
Grouped by Device



There is not a meaningful difference between these statistics and the overall statistic does a good job of representing these groups. However this might also be an indication that this type of grouping might not be the most informative. Try grouping the data in a few different ways before feeling confident in the overall statistic.

Medium Variance

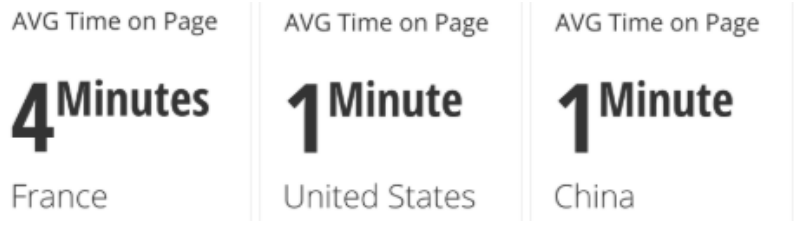
Grouped by Age



There are meaningful differences here but they still revolve around the same overall statistic. These differences might be large enough to investigate outlier groups more closely. It is a common practice to report a high level stat and provide a margin around how much it varies. Providing this extra context can help you feel more confident in the overall statistic

High Variance

Grouped by Country

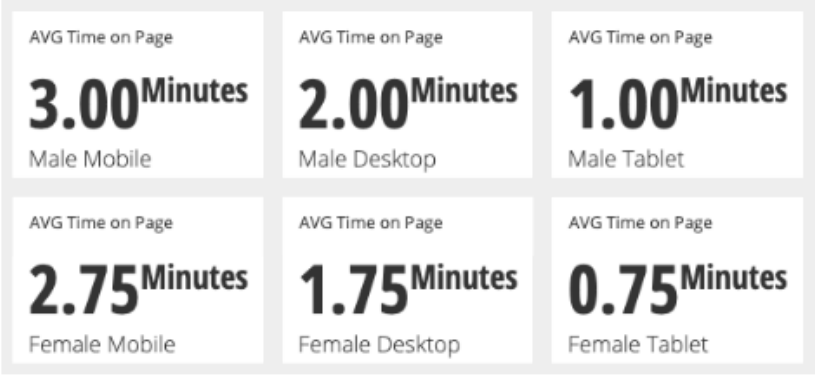


Returning to the original example, the overall statistic is not representative of this data. Do not use an overall statistic when the variance between groups is high. You should investigate the largest outliers to determine what is going on. You can choose to isolate any outlier groups or you can perform separate analyses of the data similar to how you would address a bimodal distribution.

Simpson's Paradox

Grouped by Gender

Females have an overall higher average Time on Page. However when we group by Gender and then by Device we see that in every category females have a lower average time on page than males.



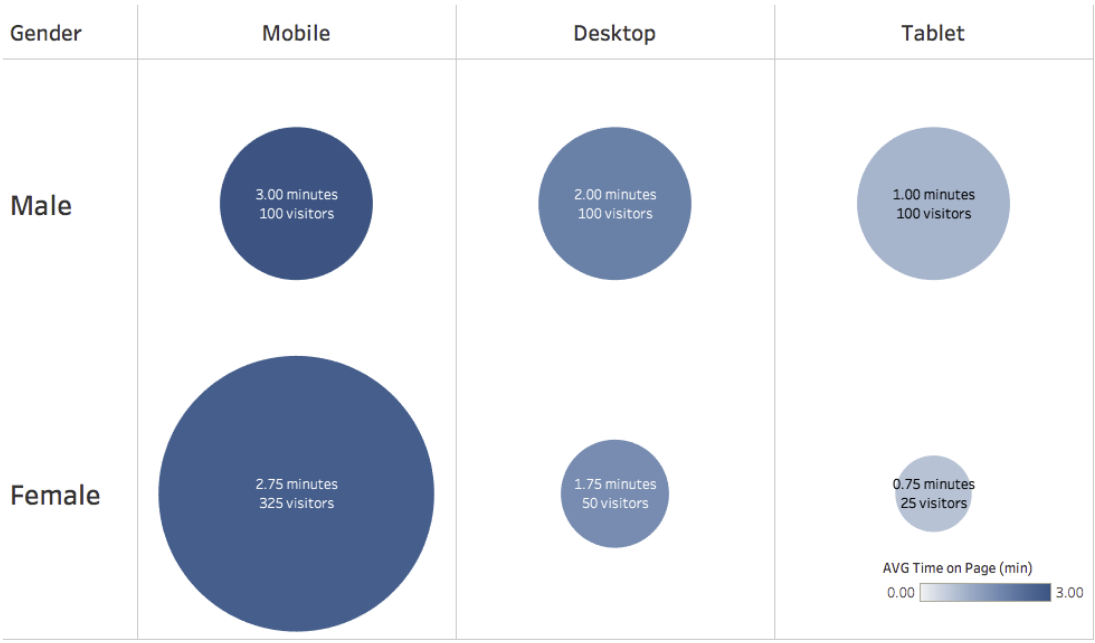
Even though females had a lower average time on page than males for every device they still have a higher overall time on page, how is this possible?

This is because the amount of people behind each one of these average time on page statistics is different:

- The amount of male visitors per device was 100 mobile, 100 desktop, and 100 tablet.
- The amount of female visitors by device was 325 mobile, 50 desktop, and 25 tablet.

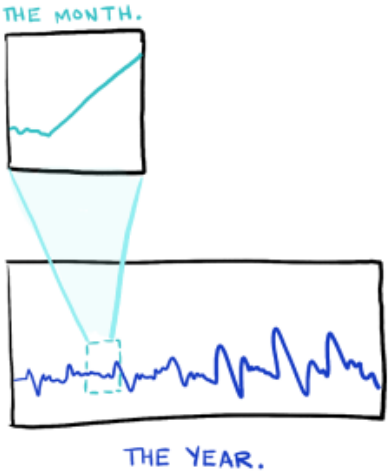
Since the female mobile group was so disproportionately large it dragged the average up with it in the overall statistic. This is an example of Simpson's Paradox.

We can visualize this more clearly by mapping number of people behind a statistic with the size of a circle and increase the saturation of a color to show the increase in the AVG time on page:



This phenomenon is important to consider when comparing groups: you should examine what the total number of observations are behind any statistic. For scenarios like this neither the overall statistic nor grouped statistics are sufficient explanations of the underlying data by themselves. You should present all of this data so that people understand the patterns at play.

Trends

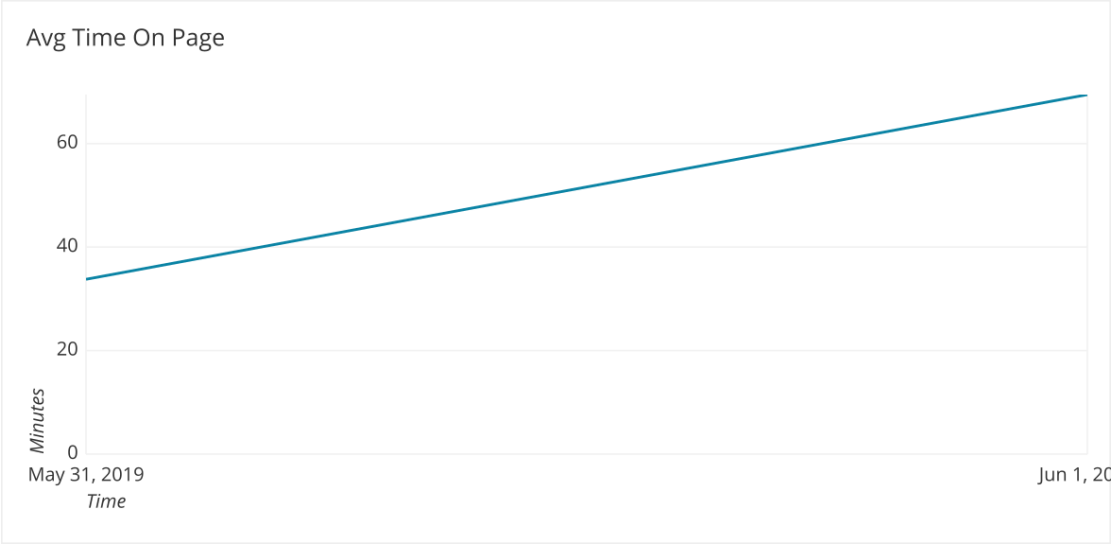


The problem with trends

When we look at a time series chart, the size of the time period we examine can drastically alter the conclusion about whether some number is trending up, down or is relatively stable.

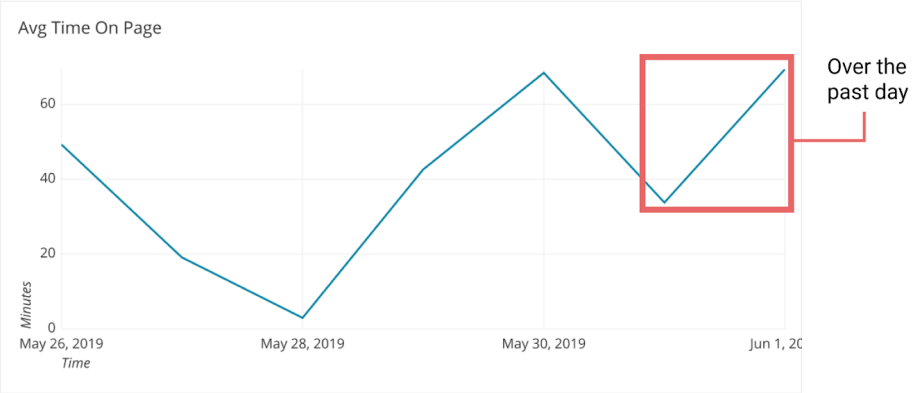
All three of the following graphs come from the same data:

Avg Time on Page over the past **day**



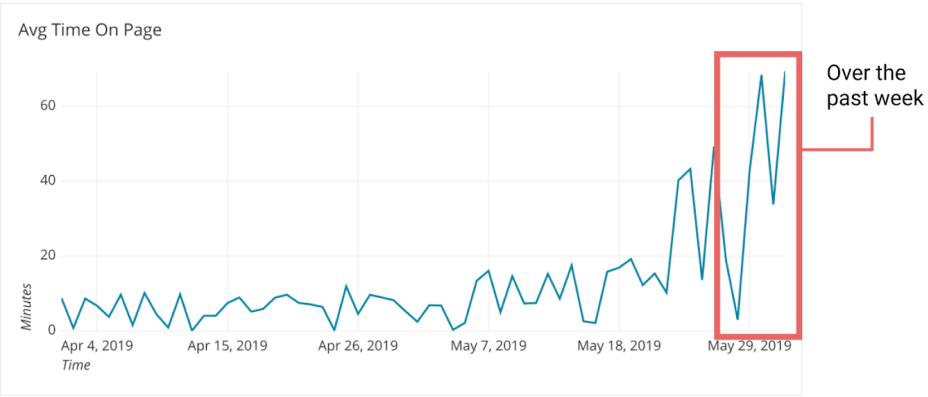
The data is moving up and to the right – things are trending up!

Avg Time on Page over the past **week**



The data has been up and down the past week, we can't be sure if the past day is part of a trend or not.

Avg Time on Page over the past **2 months**



The data has been trending up recently but the variance is much higher than it has ever been.

We can tell three different stories based on how much time we include. People gravitate to the timeline that accommodates the story they believe to be true (an example of confirmation bias). Therefore it is a best practice to show as much data as possible first, try to contextualize any large variations in the data and them zoom in to specific date ranges.

Creating the right timeline

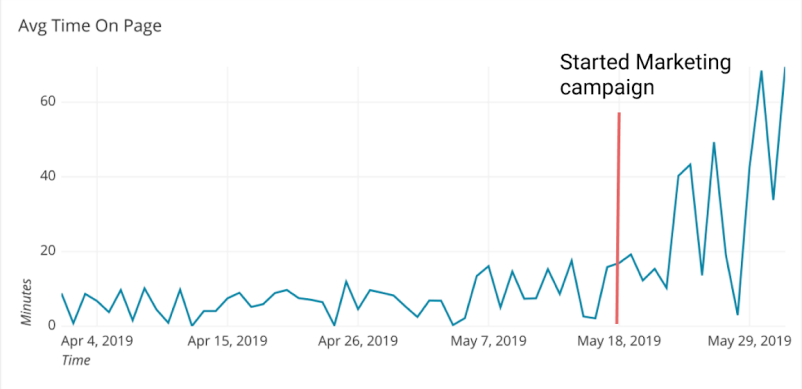
Deciding what data to show has a large impact on how people will read the chart. It is important for the reader to factor in context and variance to make their decision.

Context

Seeing what this data was in the past helps us judge the trend in the present. We can determine if this is new or something we have seen before. In the case of the example above, there is a clear trend that things are going up but variance is also increasing.

Context best practices:

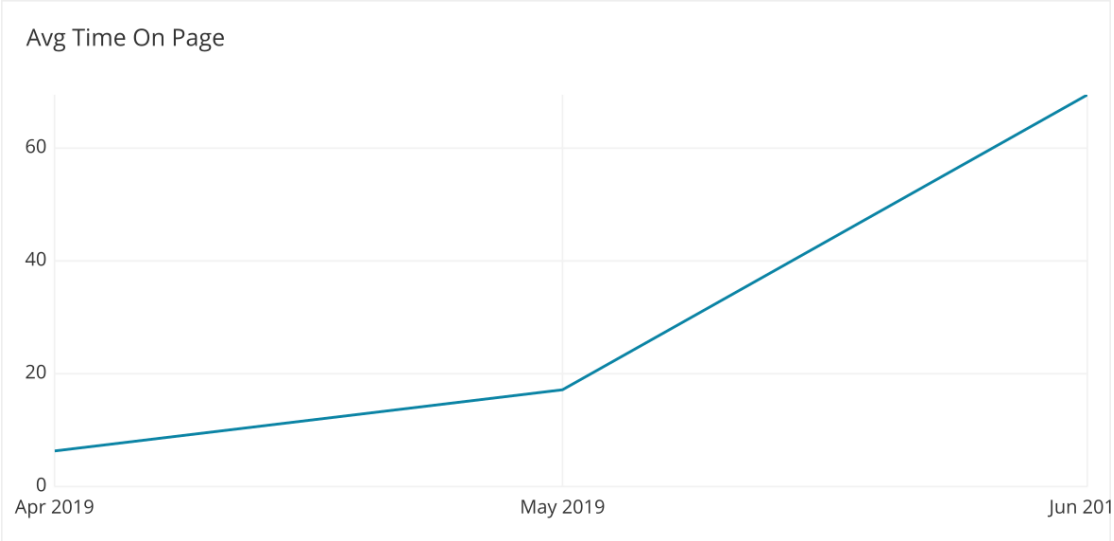
- Check the data on the longest timeline you have and then move the filter up as long as you feel the data you are excluding is not relevant or could be easily summed up.
- Include access to the full timeline to add credibility to your analysis.
- Annotate the timeline to point out why large variations happened.



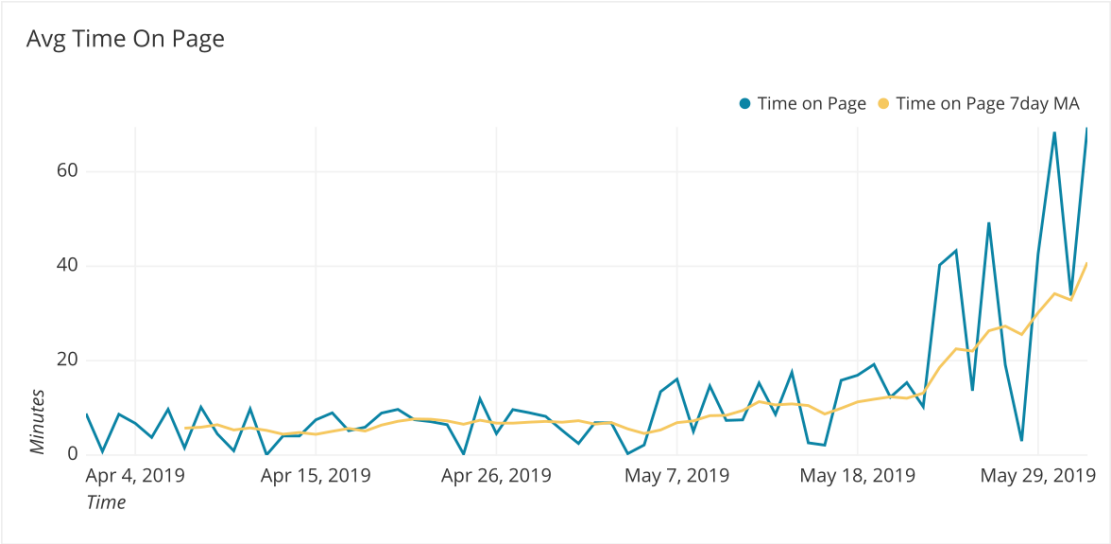
With this annotation on the chart it is more clear why this spike in time on page might have happened.

Variance

Variance in a line graph can be distracting or informative about the trend we are analyzing. If we take the daily avg time on page chart from above and look at it by month it tells a much more general story, time on page is trending up:



The longer the timeframe you are aggregating to the less variance you will see in the data. The shorter the more variance. You can also reduce variance using moving averages.



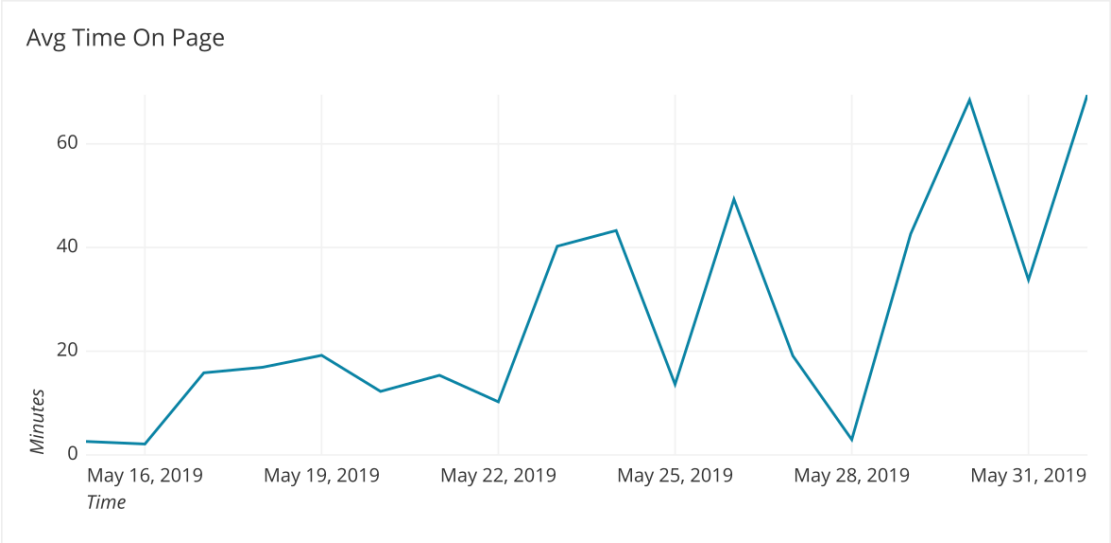
Here we can see the yellow line captures the overall trend and smooths out (hides) the variance of the daily figures.

Interpreting timelines

Whenever you are presented with time series data, take note of the axes:

X-axis: could there be more data they aren't showing? Why?

Having seen the graph above we know there was a long period where time on page was low. If we cut off that previous data it no longer feels like a new phenomenon it seems like since it was first created it is trending up. The choice of cutoff point in this graph below removes that original context.

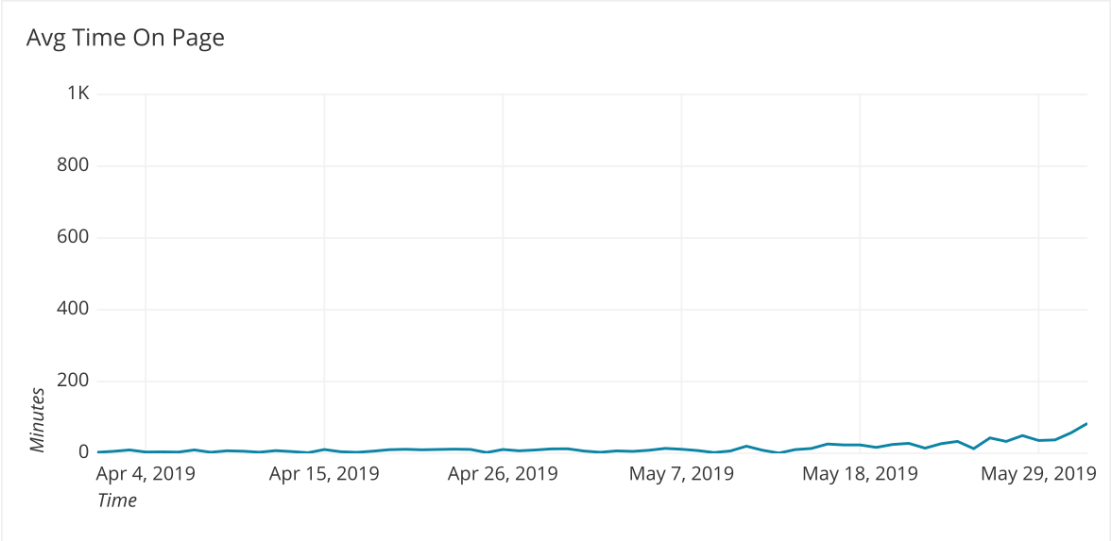


This is not necessarily bad, if you explain why the cutoff point was chosen but it can be misleading to only show recent trends in data.

Y-axis: Is the range compressing or expanding the variation inappropriately?

By increasing the y-axis range we can compress any trend to look basically flat. This is a subtle technique that can be applied to hide the size and variance of a trend.

We can look at the same chart, but increase the y-axis limit from 70 to 1000.

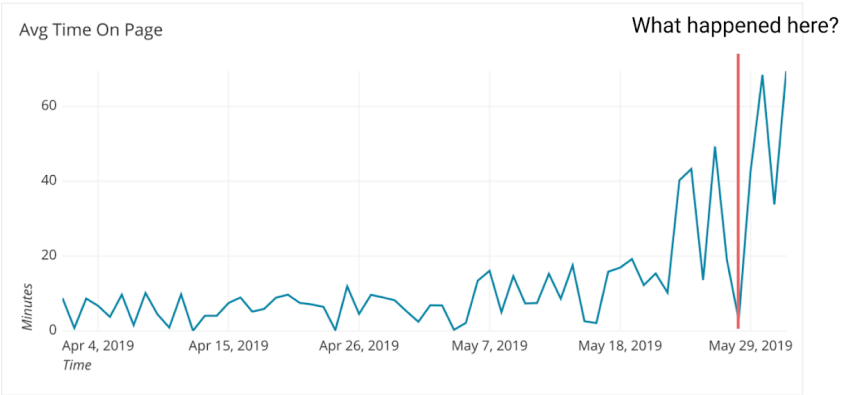


This radically changes how someone would evaluate the trend at first glance.

Annotations: Do they explain the various parts of the chart that stand out?

When a metric varies greatly it is an indication that there might have been a problem in how the data was collected or there was an outlier among very few users that dramatically affected the statistic.

Sometimes it is a true effect, but either way pointing out and addressing these points on the graph helps everyone understand it more clearly. For instance did anything happen before the Avg Time on Page started to increase? Why did the Avg Time on Page plummet after going up?



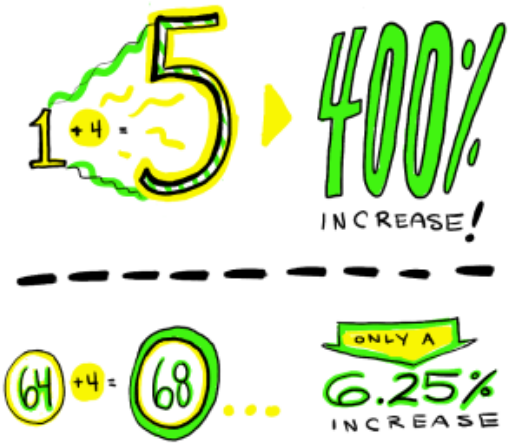
Think through what people would be skeptical of and address it. Providing clear data is an ethical responsibility of every analyst and data presenter.

Summary:

- Visualizing trends can be misleading
- It is important to get context for your data
    - Try to explain any variances from the norm
- Monitor the x and y axes:

- X- Does the time period provide enough information to draw accurate conclusions ?
- Y- Is the data effectively visualized using this scale?

Relative vs Absolute Change



The problem with Relative vs Absolute changes

When numbers change people can report how big that change was in relative or absolute terms.

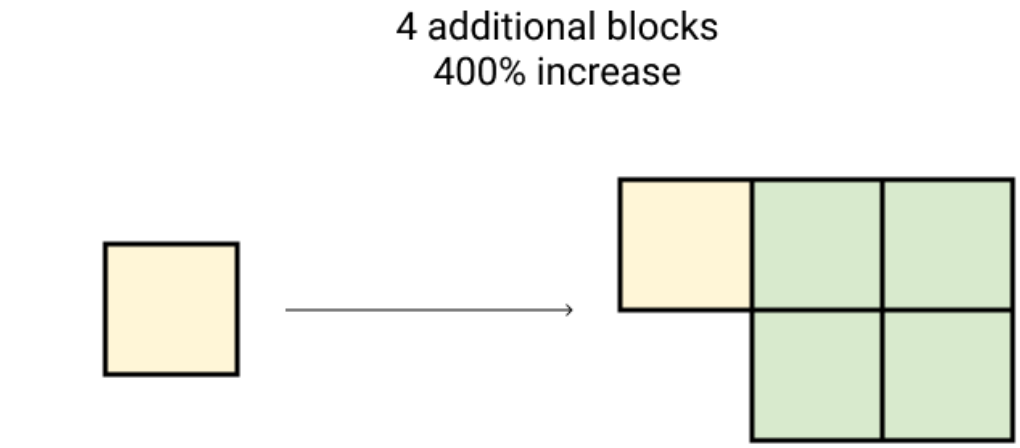
- Relative change - What percentage larger or smaller did the number get from the original number
- Absolute change - how many more or less is it than the original number

While these two statements do not sound that different, let's explore how they can each be misleading.

Relative changes

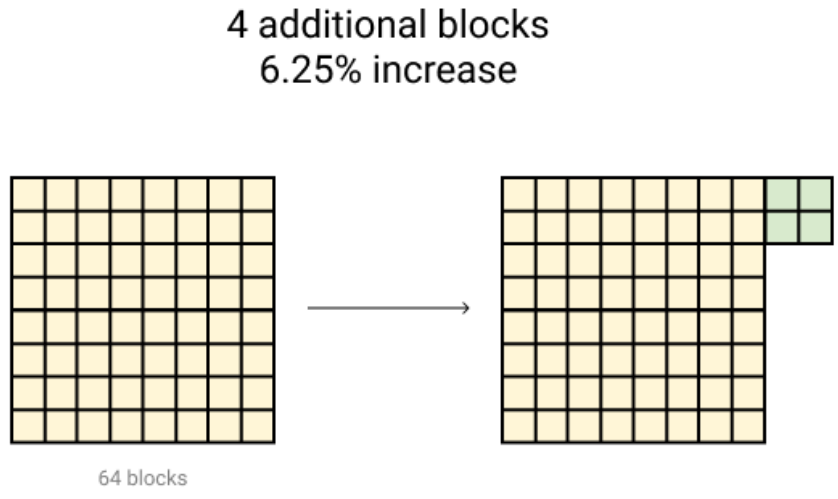
Relative changes on small numbers can appear to be more significant than they are. This is because a small absolute change in the number can result in a large percentage change.

So if I got a $50 dollar return on my $10 investment, my relative change was a 400% increase.



There is also some variation in how relative changes are reported. Here we have a 400% increase which is also the same as saying 5 times as much. This becomes more confusing when a relative change is negative. Something that is 5 times as small is an 80% decrease from the original amount.

Relative changes on big numbers can appear less significant. This is because any absolute change in the number needs to be large to show a large relative change. Even when the absolute change is large, if it is a change on a larger number the relative change can be small. Let's say that the national deficit increased by 5%. This may seem small, but the actual increase or absolute change to the $20,000,000,000,000 budget is 1 trillion dollars.



Absolute changes

Absolute changes work the other way:

- Absolute changes on small numbers can look small even if their relative changes are large. I earned $40 on my investment.

- Absolute changes on big numbers can look big even if their relative changes are small. The deficit has increased by 1 trillion dollars.

When to use Relative vs Absolute change

Use Both

When choosing between reporting a relative change or absolute change, take a second and think if you choosing the type of change that best represents what is actually happening or are you selecting the more sensational number. The best practice is to provide both numbers. Personally, I recommend putting the less sensational number first so people have context and are not distrusting when you reveal the less significant number.

Context Matters

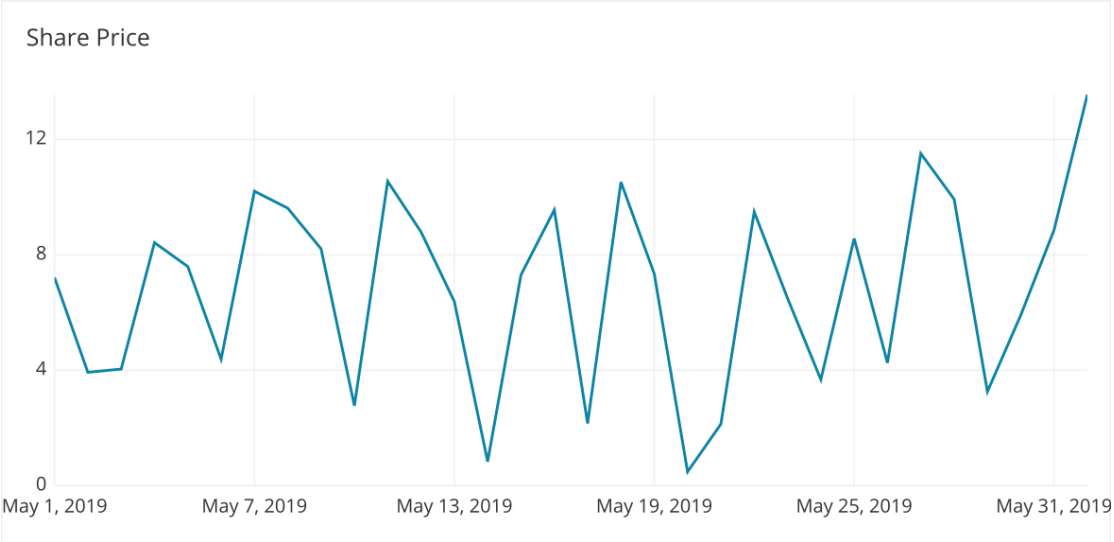
When numbers change, we want to know the reason. Sometimes the context completely changes the story. If home prices increased by 20% over the past ten years we might be concerned about this trend. However if you factor in inflation which was also increased by 20% over the past ten years, then the relative value of homes have stayed constant. Even if the change in home price was put in absolute terms, there was no actual change in the value of the home. For any chart focused on tracking a monetary value over a time period, you will always need to adjust for inflation.

Compare

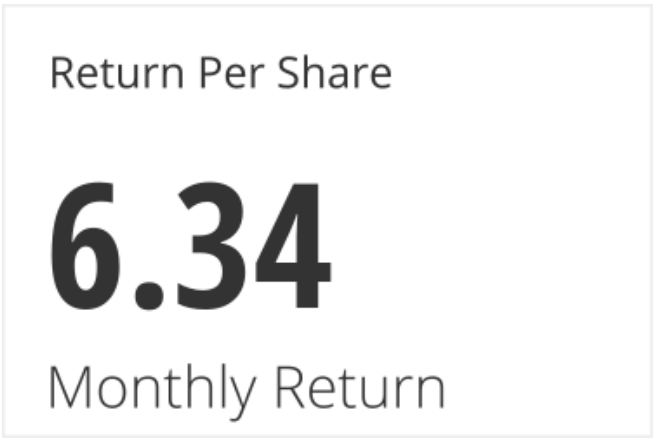
To get a clearer idea if an absolute or relative change is significant compare it to other changes that are related to it. For example, if sharks kill 16 people per year and this year they killed 20, that is a 25% increase of sharks killing people. If you compare this to heart disease killing approximately 600,000 people per year, the total number and the relative change of shark related deaths don't sound so significant. Even if heart disease declined 25% in the same year, it should still helps us better understand the lack of significance to shark related deaths.

Interpreting Relative vs Absolute change

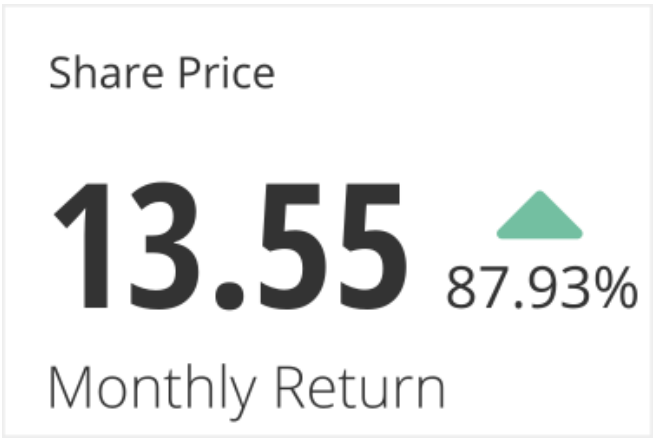
Let's look at an example of changes in share price to demonstrate how the change in price could be represented in different ways.



Here we can see the absolute change in price everyday. There is a lot of variance which can distract from the relative and absolute returns of the investment. If we bought the stock on May 1st we could look at the return per share in absolute terms.



Now this may not look like that much money but when we look at the relative change from our starting positions we can see we got a high relative return on our money.



Next time we would want to put more money in so that the absolute return would be bigger.

Summary:

- Relative changes on small numbers often look big.
- Relative changes on big numbers often look small.
- Absolute changes on small numbers often look small.
- Absolute changes on big numbers often look big.
- Explore both types of changes when looking at data

Experiment Design

Predicting Outcomes

How many jellybeans are in the jar?

The answer: 103 Jelly Beans

If you guessed before looking at the answer, great: you are acting like a scientist! You made a hypothesis and then looked to see if you were right. If you did not guess, then you might be saying to yourself: "103 makes sense." or "I would've guessed something like that around 100."

The problem with not making predictions

When we do not predict outcomes, we are more likely to encounter confirmation bias. Our brain finds a way to confirm that we are good at guessing and confirms thoughts about how we would have been close if we did make a prediction (This is very similar to hindsight bias). This is problematic for many reasons but we will focus on two:

- Lack of scientific rigor
- Not improving our decision making over time

Lack of scientific rigor

The scientific method can be boiled down to 2 steps.

1. Make a prediction.
2. Test to see if that prediction is correct.

Your prediction should be based on past knowledge about the field you are in and the test should be designed to confirm or disconfirm that prediction. When you do not make a prediction, you cannot design a test to confirm or disconfirm it. Conclusions that are reached without a scientific process are less reliable. The interpretation of results are based on post hoc reasoning and therefore our decisions may not be valid. In addition, we are not able to learn clear lessons from our results.

If you do not make predictions you can gather data and make observations, but you can only report what you have found and not draw any conclusions (i.e. correlation ≠

causation). However, that information you gather, can help inform predictions in the future for experiments and tests that can be subject to statistical techniques.

Not improving our decision making over time

We have to request and commit resources for projects before they commence. So we need to make predictions about what the impact of these projects will be in order to justify the resource allocation. Saying “this project is going to be good for the company” gives you a lot of ways to “find success”. This mindset does not give you a good feedback loop on if your prediction was correct. This vague prediction provides vague information for the company

If we do not make clear predictions, we will find reasons why the data we gathered was “predictable”. We will not learn from our planning mistakes.

We need to make firm predictions and then review the outcomes. We can refine vague predictions e.g. “This project is going to be good for the company” by asking “how?”. As we closely examine our predictions and the actual results we will improve our ability to make better predictions in the future.

Tips for Making Predictions

While making an initial prediction is not too challenging, sticking to your prediction even when your prediction is inaccurate is difficult. Naturally, we want to make predictions that are correct. When we predict the outcome of some product or feature we are working on, we want the results to show a positive impact. The impulse to report inaccurate results is dangerous but can be overcome.

Things aren’t exact and can be negative

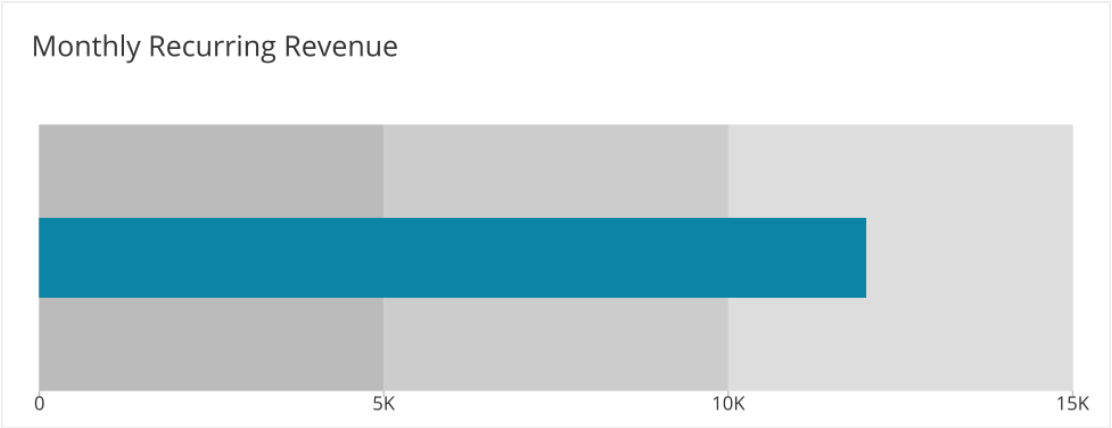
Make an upper bound and lower bound prediction and then add a confidence measure to it. “I am 90% confident that this feature will improve engagement 5-50%. Thinking in bounds and confidence levels allows you to think more broadly about your prediction and removes the stigma of needing to be exactly right.

A common misconception for people is that the range of impact of their feature is greater than 0 on whatever the metric is. The real range of the impact can be negative or positive. Your feature can have a negative impact as they often do.

Think through how feature could go wrong and whether it affects your confidence level in your range. What are you basing your prediction on that would prevent the feature from having a negative impact? Have some guidelines and actions set in place if the feature you launch has a negative impact on the metrics you predicted it to positively impact. Having this plan and following it will reduce squabbling from your team and make it easier to try another experiment. Otherwise your team will be frustrated that they spent time on something that didn’t work.

Make public facing predictions

Create a dashboard ahead of the product or feature where you have your predictions laid out. A bullet chart can be great for many metrics since there are multiple zones allowing you to define a range of outcomes.



Here we can see the gray areas defining the range of outcomes we predicted with the blue bar representing the actual value of the metric.

Ask others to hold you accountable. Confirmation bias plagues us all, so we need people to constantly check us on this. Make your dashboard public and share it with your team and your manager. This also provides you a way to update the team on progress and to review findings This will be covered in another chapter.

Summary:

- If you do not predict outcomes you will not get better at making predictions and decisions based on the data collected
- Predict in ranges and confidence levels instead of exact numbers
- Make predictions public to hold yourself accountable

Define Experiment Parameters

Even if a metric was chosen for a new feature or product, important parameters about how that metric will be measured and evaluated might not be set.

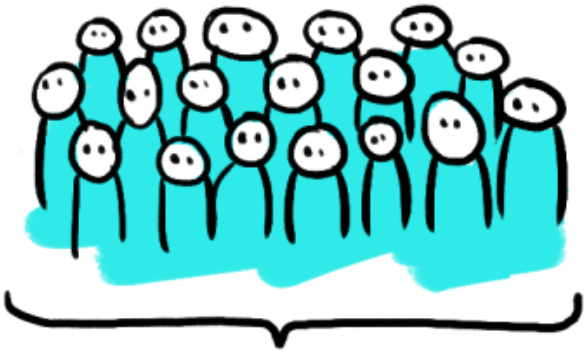
One of the things that makes analysis hard is that there are so many things happening at once that might affect the metric you are monitoring. Determining what caused the metric to move is complicated. A good experiment tries to control for as many of the other things that may be influencing the outcomes of your experiment as possible so that we can determine if our new product/feature impacted it.

In a typical business context there are three parameters that you should spend time defining:

- Cohort
- Timeline
- Controls

Not setting these parameters in the beginning allows people to move these around at the analysis stage in order to find data that makes their product or feature look more successful than it actually is. And that is bad, very bad. We want to limit the ability for people to let their confirmation bias flare up and allow them to interpret data incorrectly. Being strict on these parameters prevents people from changing how they interpret the data after the results are in.

Let's talk through each of these pieces and some important considerations.



Cohort

If you are testing out a feature or product with a subset of your user base, make sure they are representative of the user base you are expecting to use the product or feature. Otherwise the results might be biased and the feature or product will not perform how you expect when you roll it out to everyone.

Questions to ask yourself:

- Is the cohort unique in any way?
- Am I picking this cohort out of convenience?

If the answer is yes to either you need to find a different cohort.

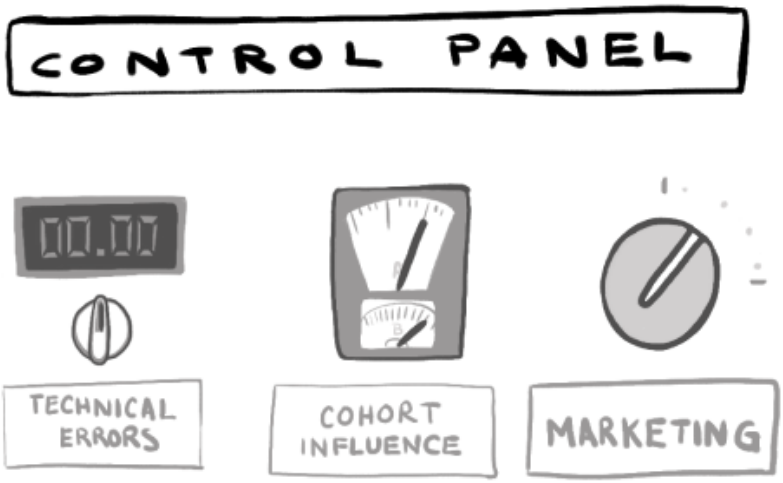
Timeline

People report good news too early and bad news too late. If day one after you launch there is a big spike in the metric you are tracking it is natural to feel excited and send out a message to the company. Yet, this can be dangerous, not only to your reputation of coming to conclusions prematurely but to what the company learns about your new feature. If the next day, the spike drops back down, you have to communicate all over again to try and undo the harm.

Before the product or feature launches ask and answer yourself two questions:

- “When will we know if this is successful?”
- “When will we know if this is not successful?”

Most launches inevitably have a bit of marketing surrounding them so the first day or first week is not usually reliable data. You do not want a spike and then a return to normal, you want a sustained increase. Pick a date before launch to review and share the impact of the data.



Controls

You need to account for a lot of different things when creating an experiment to test a feature or product within a company. Think through what other initiatives are going on at your company and what world events are coming up that may affect your metrics.

Common reasons your results were higher than expected:

- Marketing Campaign recently launched
- Internal usage is being factored into the data

Common reasons your results were lower than expected:

- Data isn't being tracked correctly
- Bug in the code
- Broken links
- Weekends and National Holidays

Look at it from other people's shoes

The unfortunate truth is that most products or features will not have a big impact, and you should be prepared for a modification or revision to have no or negative effect! People in your organization will almost always be skeptical of large positive results. You should prepare for this because the goal of your product or feature is not only to provide value to your customers but to create knowledge your company can build off of.

Ask yourself:

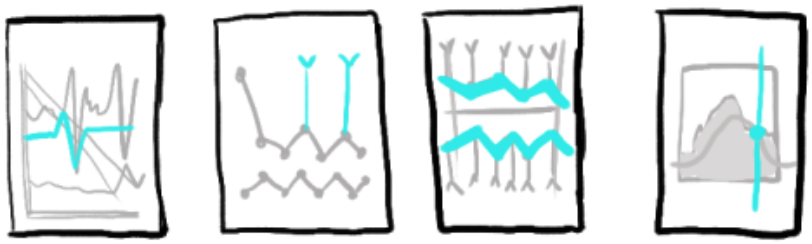
- What would make me skeptical of the results?

Take your answer to that question seriously and do your best to address that in your experiment design.

Summary:

- Define the cohort who will be involved in your experiment, note any characteristics that make them unrepresentative
- Set a timeline for when you will evaluate the success of your experiment
- Do your due diligence when conducting an experiment to make sure your results won't be affected by other peoples' actions.

Review Outcomes



We just launched our new feature and our metrics look like they have improved. How should be feel about the following graph?:

We should feel skeptical! Start asking questions!

- What caused this large positive change?
- Is this the right time window to examine these metrics?
- Are we sure the data is accurate?

Remember all the analysis mistakes and mental biases covered in this book. It is easy to misrepresent or misunderstand data. We need to take a closer look.

How to investigate changes in data

When data changes dramatically we want to take credit if it was a positive change and deny responsibility if it was a negative change. Ultimately we should want to know *why* it changed. To figure this out we need to investigate a few common ways data changes.

Marketing

A lot of times we are focused on the work we are doing and do not see what other parts of the company are doing to influence buyers and users. One of the most common reasons for any spike is there being a marketing push that day or that week. Reach out to that team to confirm no additional spend/effort has been going on to remove this potential reason

Technical Reasons

Did the site go down? Did a data point stop being tracked?Is there a mistake in the tracking? (e.g. double-counting of hits, etc.) Simple technical problems can have huge impacts on data especially in the negative direction. It is important to reach out to the development team to establish if they might have had an impact.

Cohort Influence

When rolling out new functionality or products, the people who buy or engage with them are usually not representative of the whole customer/user base. It is important to try and group the data by different demographic and engagement metrics to see if there is anything unique about those whose activity has influenced the metrics you are tracking.

Sharing results tips

Avoid sharing data (positive or negative) too soon with people outside your direct team. People will naturally make judgements and assumptions off of this data. And it can be hard to undo these learnings. Give your feature or product some time before declaring it a victory or defeat. Time is the ultimate judge!

When you feel confident enough time has elapsed it is important to communicate clearly the results of your work. Let people know what they should take away from this experiment and what they should not. When the results are inconclusive say so. It is much more dangerous for people to learn the wrong lessons than to learn nothing.

Rarely do lines move smoothly up and to the right. You should distinguish trends from variance (signal from noise). And point out the amount of certainty you have – or lack thereof – about the trends in the data.

Simple questions to stay in check:

- Did our change cause this?
- How should we interpret a spike?

Summary:

- Your metric may move for reasons other than your feature or product
- Marketing can give a metric a temporary boost
- Bugs in the code can drastically under or over report some metric
- Your cohort might not have been representative of the whole population
- Share your results when you feel confident they will hold, do not share at the first sign of positivity

Extras

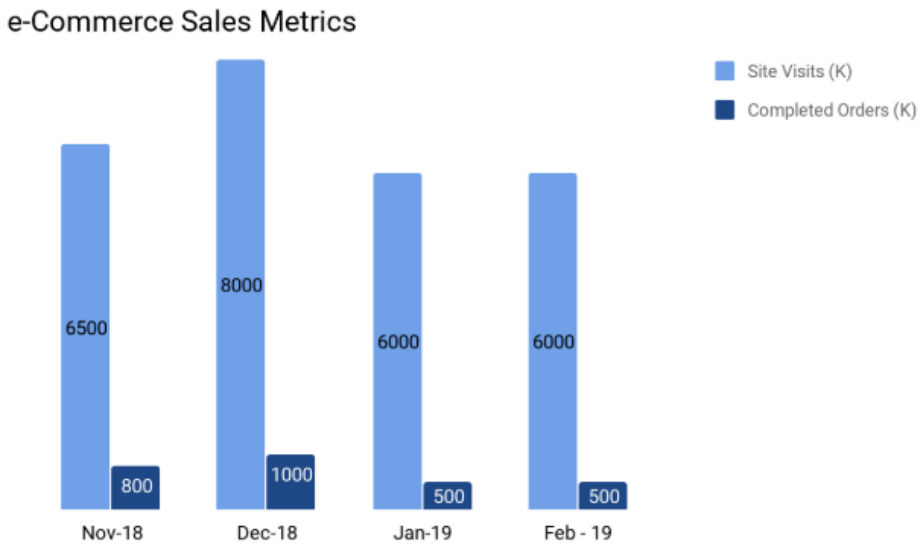
Increase Ecommerce Sales with Metrics

The problem with a single metric

I worked with an e-commerce client who was looking to get a grasp of his site's conversion rate so he can find ways to increase it. The sporting good site had a 25% conversion rate and he wanted to develop a strategy on how to raise it to between 28% to 30% by the end of year. My client was fairly new to e-commerce, he had a successful career in retail sales and now he assumed responsibility for digital sales for his organization.

There are different nuances between shopping online versus in-store that he needed to learn. For example, customers interact with sales people when in the store. They're able to ask questions and see and touch the product live versus browsing online. If the sales person sees that the customer is about to buy a product, they will throw in a discount to close the deal. When a customer is shopping online we only get to see metrics and data about their behavior which is harder to interpret and intervene.

During our first meeting we reviewed the reports given from the marketing department. The marketing report showed 6M visits to the e-commerce site on average per month for the past two quarters. He wanted to know why with so many visits there were so few sales. In retail he could rely on that single metric to explain and predict sales. Fewer people in the stores fewer sales. And 6 million visits to the physical store would have meant a lot more sales than what he is seeing online



At 500K completed orders per month the e-commerce site, online sales were making up 20% of total orders. Seeing the total visits broken out by month was not helping my client see a way to increase conversion. He needed multiple metrics to help understand how to influence increased online sales conversions.

Use Multiple Metrics

We need to explore more metrics to figure out what is going on here.

- Is visits the right metric?
- What other metrics would help contextualize it better?

I met with the analytics director to get a better understanding of how data was being tracked. She explained that the marketing team is tasked with “bringing customers to the top of the funnel”. My client is measured on sales. Visits alone is not enough to help him develop a sales strategy.

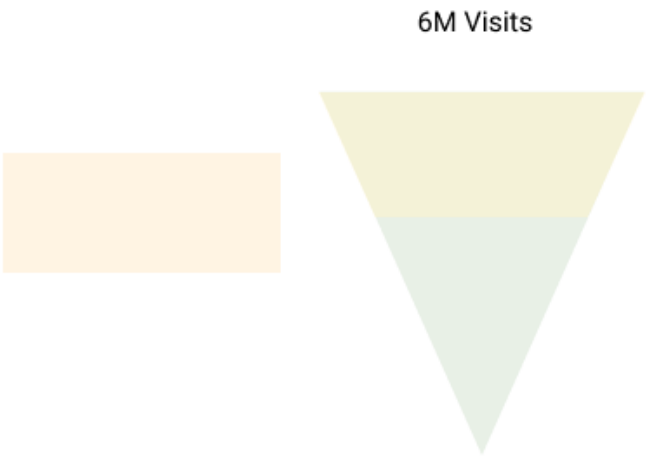
The analytics director gave me insight into more metrics:

1. Unique visitors - 2M on average per month for the 1st half of 2019
2. Abandoned shopping carts - 100K on average per month with values of $100 or greater
3. Number of orders entering sales flow - 600K on average per month
4. Completed orders - 500K on average per month

When I regrouped with my client, I used brick and mortar analogies to walk him through high level customer flows.

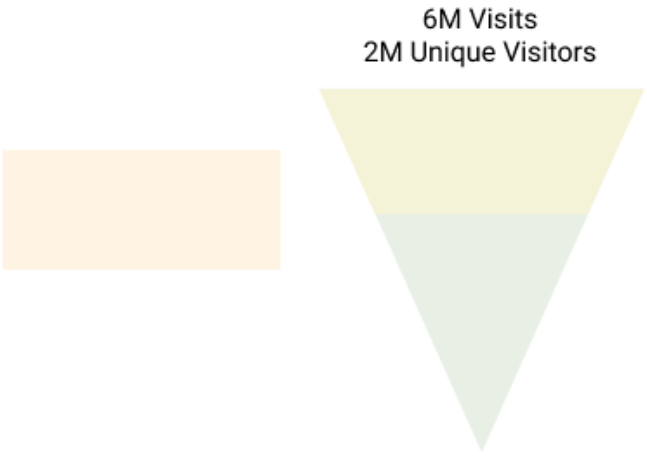
Visits

I explained that marketing uses this metric to measure how many customers they are bringing to the “store”. This equates to 6M visits on average per month coming to the e-commerce site’s homepage.



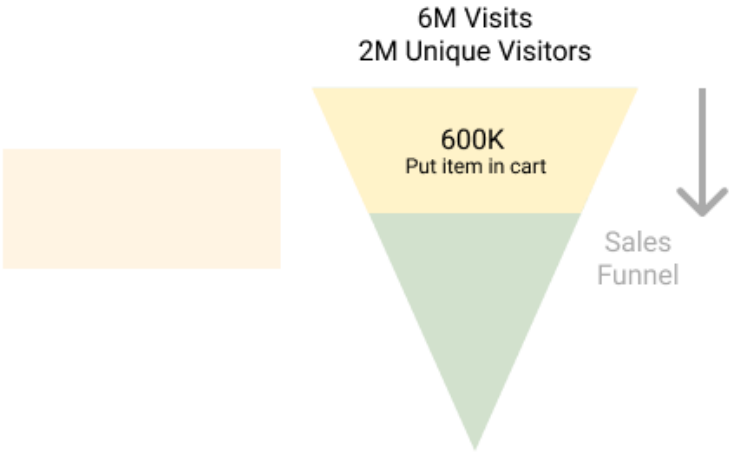
Unique Visitors

While the site has 6M visits a month only 2M represent unique visitors. We have 2M individual customers on average per month coming to the e-commerce homepage. So a customer can make multiple visits but they are one customer coming to the “store”. When we divide visits by unique visitors, we get 6/2 = 3 visits/unique visitors on average per month. So the average online customer is making 3 visits to the site per month.



Orders Entering Sales Flow

This metric represents the number of customers that have filled up their shopping cart and ready to check out. Understanding the sales flow can be very important in analyzing online sales. My client wants to make sure that online customers have the most optimal experience and nothing is hindering or obstructing the sales process. Currently there are 600K orders per month on average coming to the checkout flow.



Completed Orders

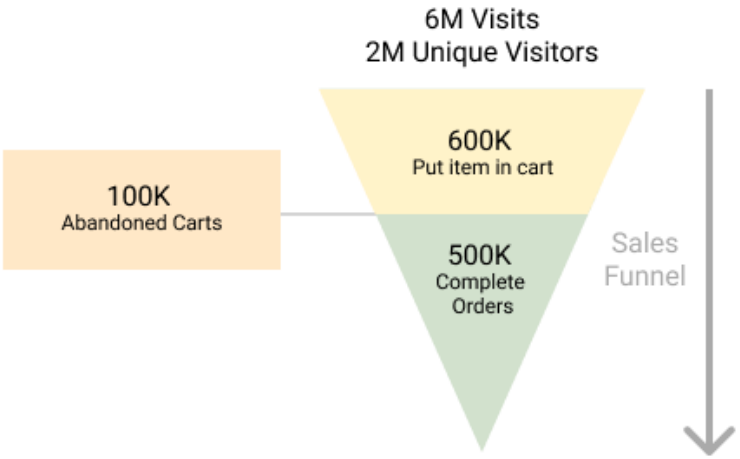
This metric lets us know how many customers made purchases on the e-commerce site. This is the most important metric for the sales VP as it gives him the bottomline number on performance. On average, 500K orders are being completed per month. One item to explore with my client is why there are 600K - 500K = 100K orders falling out of the sales

flow.

Abandoned Carts

Another metric that can help with sales strategy is abandoned carts. When talking with the analytics director, we agreed that there are customers who are “window shopping” and/or doing price comparisons against other competitors. However when a customer leaves $100 or greater worth of product, we might want to reach out to them via an email or retargeted online ad.This way we are reaching out to the customer to see if there is still interest in a particular product they were interested in and inform them on a possible sale.

Currently there are 100K abandoned cart orders of $100 or greater on average per month. I will discuss with my client how to work with marketing to see if we can do a targeted campaign to these customers.



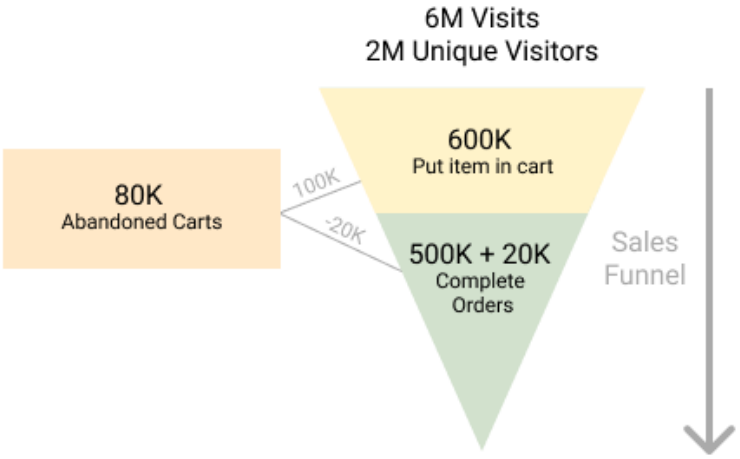
Interpret Multiple Metrics

My client now had multiple metrics to help him brainstorm a sales strategy. With 500K orders in monthly sales we know how many orders are being processes. With the average order being $1000 we can estimate that the site is generating 500K x $1000 = $5M in sales. So there are two clear ways to increase sales, completed orders and average order value.If either is increased, sales can increase. Let’s focus on completed orders.

We can optimize the top of the funnel which would be visits and unique visitors or we can focus on the bottom of the funnel which would be cart conversion.

Bottom of the funnel

We have 100K abandoned cart orders per month on average. I discussed with my client the possibility of working with marketing to target customers who have abandoned carts with orders of $100 or more.

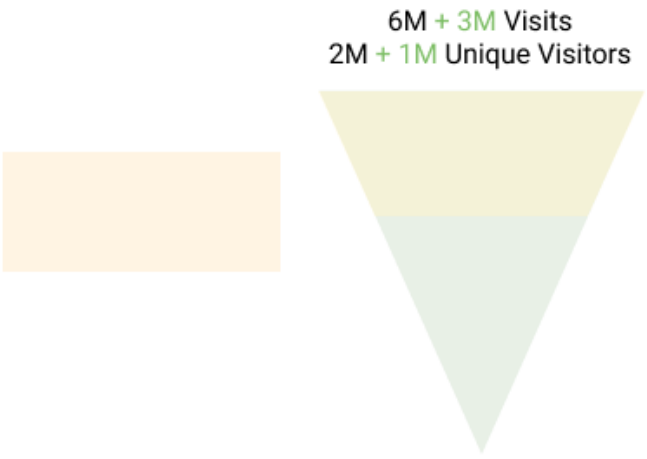


The idea here is that with this segment, there are a good amount of customers who might convert when given an email reminder and/or incentive. They are low hanging fruit in our goal to increase potential sales. Assuming we captured 20% of the abandoned cart segment, that's 20% of 100K = 20K additional converted customers. Given us a total of 520K completed orders on average per month.

Top of the funnel

There are 2M unique visitors (potential customers) that visit the e-commerce site each month. This equates to a customer conversion rate of 500K/2M = 25%. These two variables number of sales and number of unique visitors can be analyzed to speculate how many sales or customers are needed to increase conversion. We can use this ratio to predict sales more accurately as we can easily relate orders to an individual customer.

For example, if we sent out a series of email and ad campaigns, we might be able to attract an additional 1M unique visitors to the homepage. Based on our current ratio, we have 3 visits for every unique customer. Therefore 1M unique visitors would equate to 3M additional visits to the e-commerce site on average per month. We can leverage visits as customer touch points knowing we have on average 3 times per month to reach out to a customer.



While it was beneficial to understand how many visits come to the e-commerce site, this alone doesn't give insight on the sales impact. Visits tell us the level of engagement a customer is having with the site not their likelihood to purchase.

Using a brick and mortar example, customer like visiting Macy's Herald Square as it's a must-see for tourists visiting NYC, a nice escape for others on their lunch break, and a place many people window shop. The total number of visits people make might not correlate to sales. We want to know how many of those people are unique visitors so we can spend time marketing to those likely to convert and not those who just want a photo or to browse. Total visits and unique visitors help tell a bigger story than alone.

Conclusion

To understand the overall sales flow and how my client could begin to impact the conversion rate, we had to look at various metrics. My client and I worked with marketing and analytics to understand what metrics in addition to visits is available to analyze sales performance. We sketched out a funnel to map out the customer journey and sales flow. Using multiple metrics we were able to identify areas that we could improve to drive up sales.

Using various metrics helps form a holistic story on what is going on with the e-commerce site from different perspectives like sales performance and engagement. Multiple metrics helped my client gain more insight, better communicate his needs, and set reasonable goals for sales.