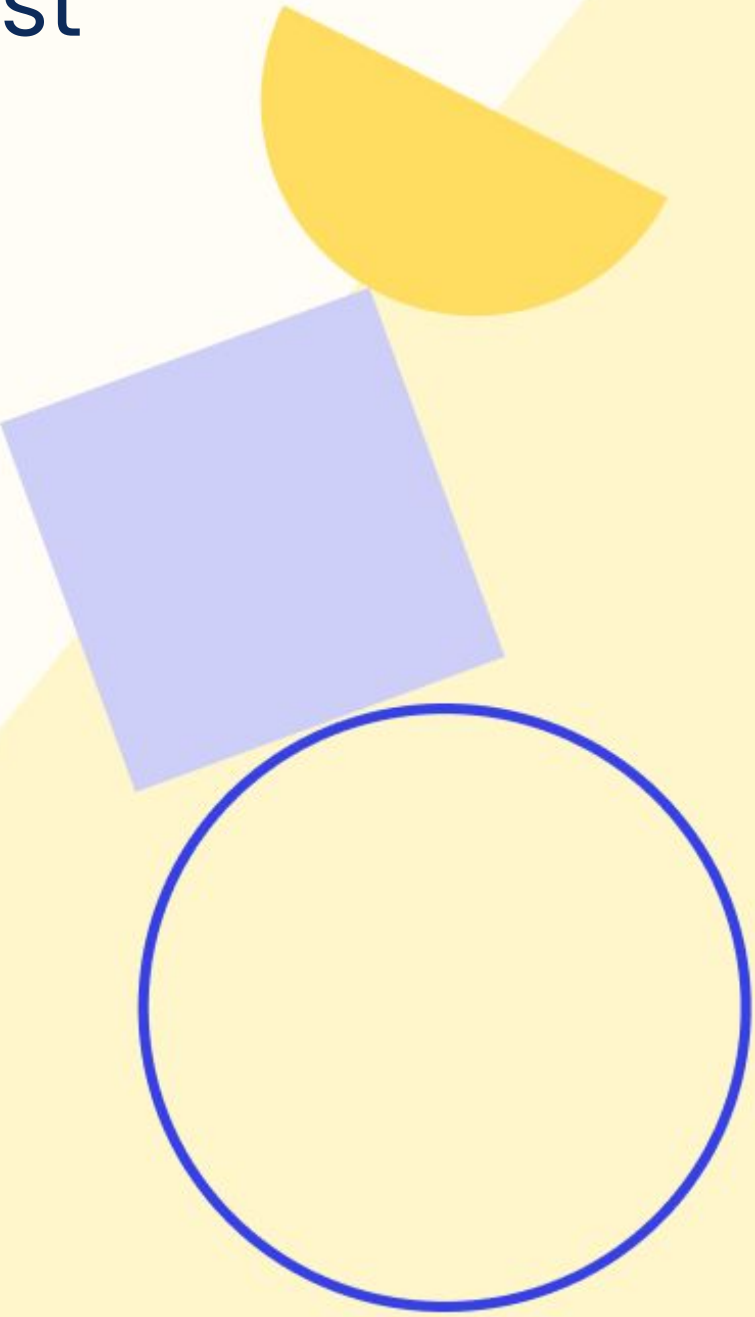
Contentsquare

A/B Testing

How to run an A/B test



What is an A/B test?

A/B testing is a way to compare two versions of something on your site/app to figure out which performs better. To start, decide what two things you want to compare. To measure the difference in impact you'll first need to define **goal metrics,** or KPIs. By using those metrics, you evaluate performance.

Why did you plan this modification?

- Increase the click rate on a CTA?
- Reduce exit rate or increase the page consumption?
- Increase the engagement rate with a specific element?

Think of a **specific objective**, which may be different from the macro objective of your website

Test on a product page will have the "Add to Cart" objective as primary KPI, and the "E-commerce conversion" as a secondary objective.

Do not set more than 2-3 objectives: if you have more, you won't be able to make any decisions at the end of the A/B Test. The more you are measuring, the more likely that you're going to see random fluctuations in your results.

*Read more about **how to set your KPIs***

Note! You always want your test ideas to be based on a solid data-driven hypothesis. Use Contentsquare to find actionable insights to base your hypothesis on.

*If you want to read more about **hypothesis testing**, read this article.*

Designing an Experiment

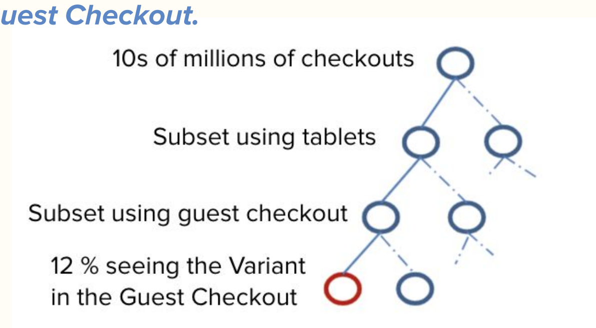
Once you have your hypothesis and the list of KPIs you want to monitor, its execution time. Before launching your test, there are a few considerations to take into account:

How large (size) does your experiment needs to be?

- Or how safe is the experiment? For large changes where you're uncertain how your users might react, you may want to consider some ramp-up tactics (e.g., exposing the Treatment to only 15% of the traffic and gradually ramp the traffic until it reaches the desired exposure level).
- Does the experiment need to share traffic with other experiments, and if so, how do you balance traffic requirements?

Example of a client running a test on their Guest Checkout.

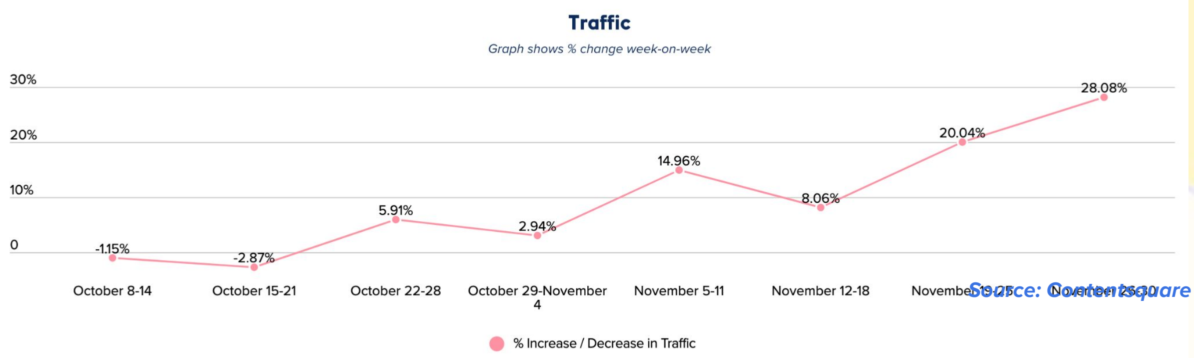
If the client wants to segment their analysis by looking only at the tablet users, they need to consider exposing the test to a sample big enough to allow for capturing differences in a population representing only 12% of the whole population (traffic).


10s of millions of checkouts
Subset using tablets
Subset using guest checkout
12 % seeing the Variant in the Guest Checkout

How long to run the experiment?

- Keep in mind the number of visits and days you'll need for this A/B Test to be significant. Use **this** online calculator to determine duration.
- Here are some other factors to consider: day-of-week effect as you may have a different population of users weekends than weekdays; seasonality, primacy and novelty effect as most experiments tend to have a larger of smaller initial effects that takes time to stabilize as users adapt to the change.

Traffic fluctuations WoW during Black Friday 2020. *Running an A/B test during sales events/campaigns can have a huge impact on both the amount of traffic and traffic decomposition being exposed to your A/B test.*



Source: Contentsquare

Tip! Have the test running for at least one business cycle.

Interpreting the Results

Before you look at your key KPIs, you also need to run some sanity checks to make sure the test was run properly and there were no bugs that could invalidate the results. A good practice is to have **invariant** metrics, or metrics that shouldn't change between Control and Treatment.

***Use cases:** Check the click recurrence of the CTA in your new form; Check the load time of the page you have the test on.*

Now it's time to look at your results.

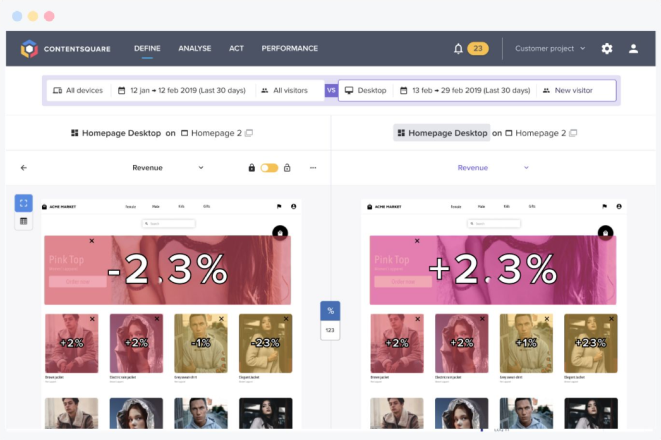
Tip! We recommend always using your traditional analytics/ AB testing tool along with Contentsquare to have a robust measure of your test's impact.

1. Use your traditional analytics tool to get an overview of how your test may have impacted your overall funnel.

- Get a high level understanding of your customer journeys
- Check whether it aligns with your testing tool data
- Compare your test performance using your basic metrics (Pageviews, Visits, Visitors)

2. Use Contentsquare to understand the how and the why behind your test's performance.

***Use case:** Understand why a Treatment won or lost utilising Contentsquare's side by side page zoning for a visual view to understand why and how visitors acted differently from the Control*



Presenting your Results

Before you put together your presentation slides, here are 3 questions to consider:

What are you trying to achieve? Your objective is to summarize how well the different treatments fared against the control, the confidence in the results and your recommendation for next steps.

Who is your audience? How technical they are? What expectations do they have? How busy they are? If you only have a couple of minutes of your C-suite's time you will plan a completely different presentation (and very 'to the point' slides) compared to an all-hands or lunch and learn session where you'll ideally present only the high level data, omitting the too technical details.

What's your medium? Here you have to decide the format to present your analysis (e.g., written report vs dashboard, a PowerPoint vs infographic). Key is to pick and present only the relevant information so you don't lose your audience's attention.

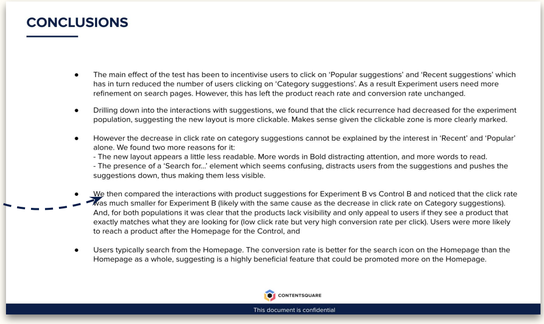
The purpose of your presentation is to convey a clear message by telling a simple story. To do so here a few tips to keep in mind:

- Have a clear understanding of the purpose of the question you're trying to answer with your presentation.
- Have a clear understanding of your audience, their needs and expectations.
- Choose your visuals carefully so the message you want to convey can shine through.
- Have one key message per visual, table or slide.
- Use signposts such as titles, axis labels, and highlight colors. Make them as easy to read as possible (e.g. avoid vertical titles that require head tilting etc.).
- Make sure your viewer can easily obtain the main message without doing any additional mental work or calculation (e.g., placing the legend of a bar chart too far from the chart where the viewer has to make an effort to read the columns and associated them with the right data).

Examples of presentational slides

The Bad

A lot of text, probably more than the rest of the deck put together but zero action.



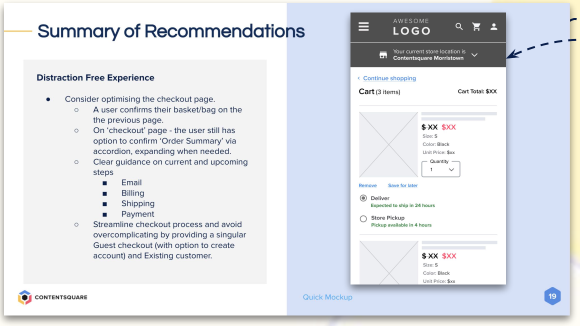
CONCLUSIONS

- The main effect of the test has been to incentivise users to click on 'Popular suggestions' and 'Recent suggestions' which has in turn reduced the number of users clicking on 'Category suggestions'. As a result Experiment users need more refinement on search pages. However, this has left the product reach rate and conversion rate unchanged.
- Drilling down into the interactions with suggestions, we found that the click recurrence had decreased for the experiment population, suggesting the new layout is more clickable. Makes sense given the clickable zone is more clearly marked.
- However the decrease in click rate on category suggestions cannot be explained by the interest in 'Recent' and 'Popular' alone. We found two more reasons for it:
  - The new layout appears a little less readable. More words in Bold distracting attention, and more words to read.
  - The presence of a 'Search for...' element which seems confusing, distracts users from the suggestions and pushes the suggestions down, thus making them less visible.
- We then compared the interactions with product suggestions for Experiment B vs Control B and noticed that the click rate was much smaller for Experiment B (likely with the same cause as the decrease in click rate on Category suggestions). And, for both populations it was clear that the products lack visibility and only appeal to users if they see a product that exactly matches what they are looking for (low click rate but very high conversion rate per click). Users were more likely to reach a product after the Homepage for the Control, and
- Users typically search from the Homepage. The conversion rate is better for the search icon on the Homepage than the Homepage as a whole, suggesting is a highly beneficial feature that could be promoted more on the Homepage.

CONTENTSQUARE

This document is confidential

The Good



Summary of Recommendations

Distraction Free Experience

- Consider optimising the checkout page.
  - A user confirms their basket/bag on the the previous page.
  - On 'checkout' page - the user still has option to confirm 'Order Summary' via accordion, expanding when needed.
  - Clear guidance on current and upcoming steps
    - Email
    - Billing
    - Shipping
    - Payment
  - Streamline checkout process and avoid overcomplicating by providing a singular Guest checkout (with option to create account) and Existing customer.

CONTENTSQUARE

Quick Mockup

19

Creating a distraction-free, easy to visualize experience for the viewer

From Results to Decisions

The goal of running an A/B test is to gather data to drive decision making. Here's an example of the decision making process and the factors that need to be taken into account before making a launch/no launch decision:

- Do you need to make tradeoff between different metrics?

  If the engagement rate goes up but revenue goes down, should you launch?

- What is the cost of launching this test?

  What are the costs for developing and maintaining the new feature? Can the expected gain cover them?

- What is the downside of making a wrong decision?

  What is the opportunity cost if you forgo a change that has real impact?

Key is to establish not only statistical significance, but also to decide how big of a difference in the main KPIs actually matters from business perspective: is the difference worth the costs of making this change? In other words, what change is **practically significant**? Depending on your business, a 0.2% change of revenue-per-user might be practically significant, in other cases - this change might be too small and you are only looking for changes that improve by 10% or more.

Example of a decision-making matrix for understanding practical and statistical significance

How do the results relate to your business objective?	**What did you observe in your test results?**		
	Main KPI Negative	**Main KPI Flat**	**Main KPI Positive**
Practical Significance Negative	*Iterate, perform deeper analysis on the page or use another insight to do a test*	*Iterate, perform deeper analysis on the page or use another insight to do a test*	*Magnitude of change may not be sufficient to outweigh other factors such as costs. Do a deeper ROI-based analysis.*
Practical Significance Flat	*Iterate, perform deeper analysis on the page or use another insight to do a test*	*Iterate, perform deeper analysis on the page or use another insight to do a test*	*Magnitude of change may not be sufficient to outweigh other factors such as costs. Do a deeper ROI-based analysis.*
Practical Significance Positive	*Repeat the test with more units to gain more statistical power*	*Repeat the test with more units to gain more statistical power*	*Launch*

What if your test is flat or losses?

You analyze the treatment, the results across segments, and improve your hypothesis. Then – you test again! Be prepared to run many test rounds for each page.

Be also aware of **cannibalization** effects and **side effects** from the changes you make. Even a subtle change can have repercussions on how your visitors perceive and behave on your site.

Use case: *Client X have run a test on the visualisation of their colour swatches. They were surprised to see that the test was losing even though the design was considered an industry best practice and they saw an increase in clicks on the colour swatches. After an analysis in Contentsquare, they realised that this was due to a decrease in the visibility of reviews. Reviews acted as social proofing and had a considerable impact on conversion*

Tips:

- We recommend having an A/B Test follow-up document where you can easily record key information on context.
- One step at the time, one A/B Test = one modification. Don't change too many things on the same page. If you do, you won't be able to analyze the real impact of your modification.
- Are you sure your modification can have a relevant impact?
- Be sure to make your test when users are behaving "normally" (Ex: we don't recommend you to implement an AB Test during Black Friday or other sales periods).

Iterate

What if you build yourself a winning test? You test again!

Having "**an iterative mindset**" is key for having a successful CRO program. It's pretty rare that you're going to get everything perfectly right with only trying once. So, plan from the beginning to collect some learnings and then iterate.

Conversion optimization is a systematic, repeatable process. You test, measure the impact of your test, then - you analyze again, look into a different part of the page, build a new hypothesis and test again. Because if you don't iterate, you could move on to the next project losing out on some great opportunities.

End-to-End Case Study

In a previous analysis you've noticed that the newly released "Bestsellers" section on your Homepage has an exposure of only 30%. However, you have also noticed a very high Attractiveness rate of 64%. Your goal now is to increase the section visibility.

***Reminder*. *Attractiveness rate* shows you the % of visitors clicking on an element after being exposed to it. *Exposure rate* identifies how far down the page the average visitor is scrolling.**

A zone is considered as seen once over half of it was viewed by a visitor

Setting up your KPIs and Hypothesis

Identified problem: *" As it's pushed below the fold, 70% of the visitors don't see the 'Bestsellers' section on your Homepage. However, the Attractiveness rate is more than 60%.*

Proposed solution: *" By increasing the visibility of the 'Bestsellers' block through pushing it higher up the page, you will increase its exposure, thus increasing its usage resulting in a overall increase of browsing depth, product pages' reach and conversion rate. You will know this by looking at the following metrics:*

Primary (test-specific) KPI: *Click rate on 'Bestsellers'*

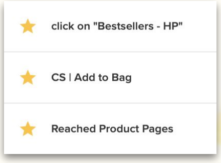
Invariant KPIs: *Click recurrence on 'Bestsellers', Homepage Bounce rate*

Secondary KPIs:. *Conversion rate per click (Goal= Reach Product pages), Click rate on 'Add to Bag' CTA*

Macro KPIs: *Conversion rate, Revenue, Average Order Value"*

How to assess the performance of your testing KPIs in Contentsquare:

1. Create and favour the relevant page goals:



★ click on "Bestsellers - HP"

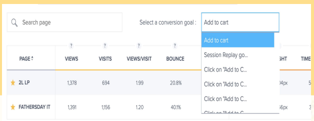
★ CS | Add to Bag

★ Reached Product Pages

2. Apply them in the different CS modules:

PAGE COMPARATOR

See if the landing page drove goal conversions across different segments



Search page

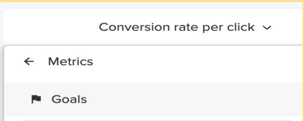
Select a conversion goal : Add to cart

Add to cart
Session Replay go...
Click on "Add to C...
Click on "Add to C...
Click on "Add to C...

```
PAGE   VIEWS  VISITS  VIEWS/VISIT  BOUNCE   ...GHT  TIME
2L LP  1,378  694     1.99         20.8%    ...4px  5
FATHERSDAY IT 1,391 1,156 1.20     40.1%    ...6px  3
```

ZONING ANALYSIS

See how many visitors achieved the goal after clicking a zone



Setting up your Contentsquare

In Contentsquare you can:

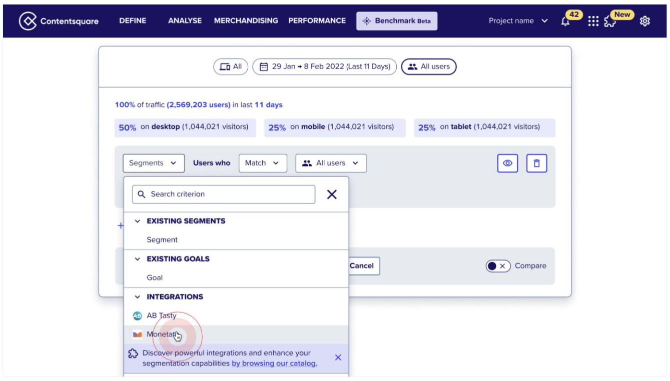
- Create your test **Segments** with Custom or Dynamic Variables and do an analysis of your A/B test in real-time
- Setup your **Zoning** and combine different metrics to really understand how your users are interacting with the content to improve the UX even further
- Build your own custom **Workspace** to monitor the performance of your key KPIs over time

How it works?

1. Setup your segment and zoning

- How to set up i**ntegration segments**?

1. If you have not yet integrated your A/B third-party tool with Contentsquare, go to the Integrations Catalog to request an integration.
2. If you have an A/B integration with Contentsquare, those segments will appear in the analysis context menu's integration segments for you to select while using any feature



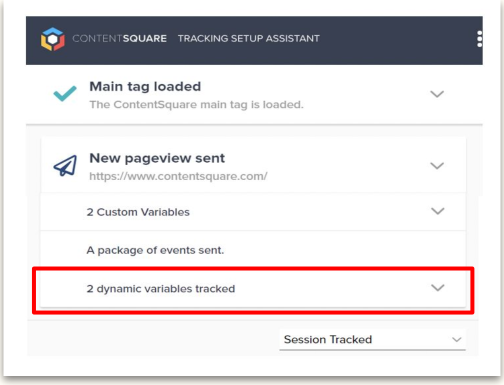
100% of traffic (2,569,203 users) in last 11 days

50% on desktop (1,044,021 visitors) 25% on mobile (1,044,021 visitors) 25% on tablet (1,044,021 visitors)

Discover powerful integrations and enhance your segmentation capabilities by browsing our catalog.

- How to set up your segment based on **Dynamic or Custom variables**?

Your A/B test segments pass through to Contentsquare as **dynamic** or **custom variables**. To set them up you can follow these steps:



Main tag loaded
The ContentSquare main tag is loaded.

New pageview sent
https://www.contentsquare.com/

2 Custom Variables

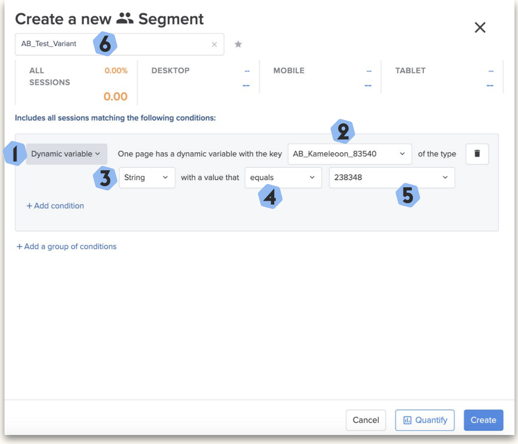
A package of events sent.

2 dynamic variables tracked

Session Tracked

1. Check if your tests are passed through as custom or dynamic variables. You can check what information is passed through to Contentsquare by using **Contentsquare Tracking Setup Assistant.** Get your extension here.
2. After its installed, open the extension
3. Then, open the dynamic variable drop-down. The drop-down will give you the information about the key, the type and the value of your dynamic variables on the opened url.

Setting up your Contentsquare

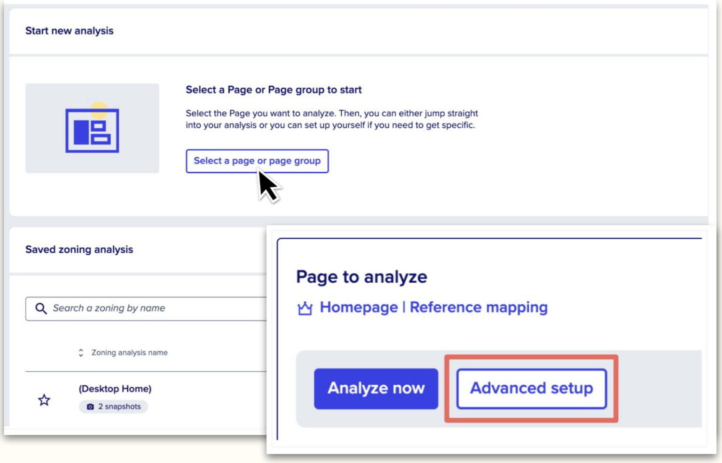


1. Navigate to the **Segments** section and choose to create a **New segment**
2. You can create the relevant segments by using the dynamic variable condition
3. Select the **key** of the variable. This is the name of the dynamic variable
4. Chose its **type** ('number' or 'string'). Number contains a numerical value, string contains a value containing letters
5. Choose its **value**
6. Give your segment a name and click on **Create** to save it

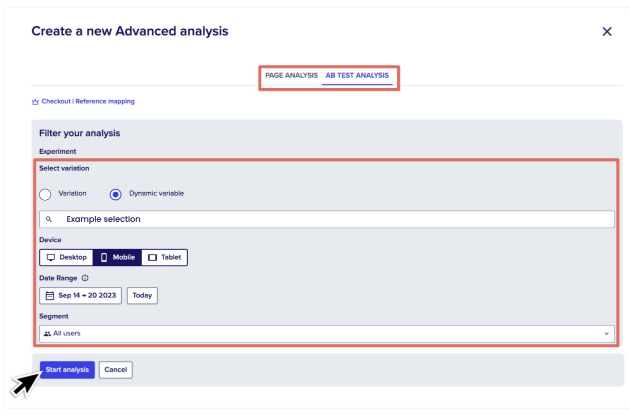
- How to set up your **Zoning**?

Before opening Contentsquare, in a separate browser tab open the test version you want to analyze.

Then, open Contentsquare and go to the **Zoning analysis.** Select **New zoning analysis** and when the **zoning creation menu** appear:


Start new analysis
Select a Page or Page group to start
Select the Page you want to analyze. Then, you can either jump straight into your analysis or you can set up yourself if you need to get specific.
Select a page or page group
Saved zoning analysis
Search a zoning by name
Zoning analysis name
(Desktop Home)
2 snapshots
Page to analyze
Homepage | Reference mapping
Analyze now
Advanced setup

1. Click **Select a page or page group**
2. Define which page you want to analyze, click **Apply** and select **Advanced set up** in the next window.
3. In the window that opens, navigate to the **AB test analysis tab.**



4. Select either **Variation** or **Dynamic variable**

5. Choose the variation or dynamic variable you want to analyze from the drop-down list.

6. Set appropriate device, date range and segment for the test analysis and click **Start analysis**

Note: *If you don't have a live AB test integration, you can setup your zoning by going through the default setting of creating a zoning by capturing a snapshot from your live website or by capturing a retroactive snapshot from Session Replay.*

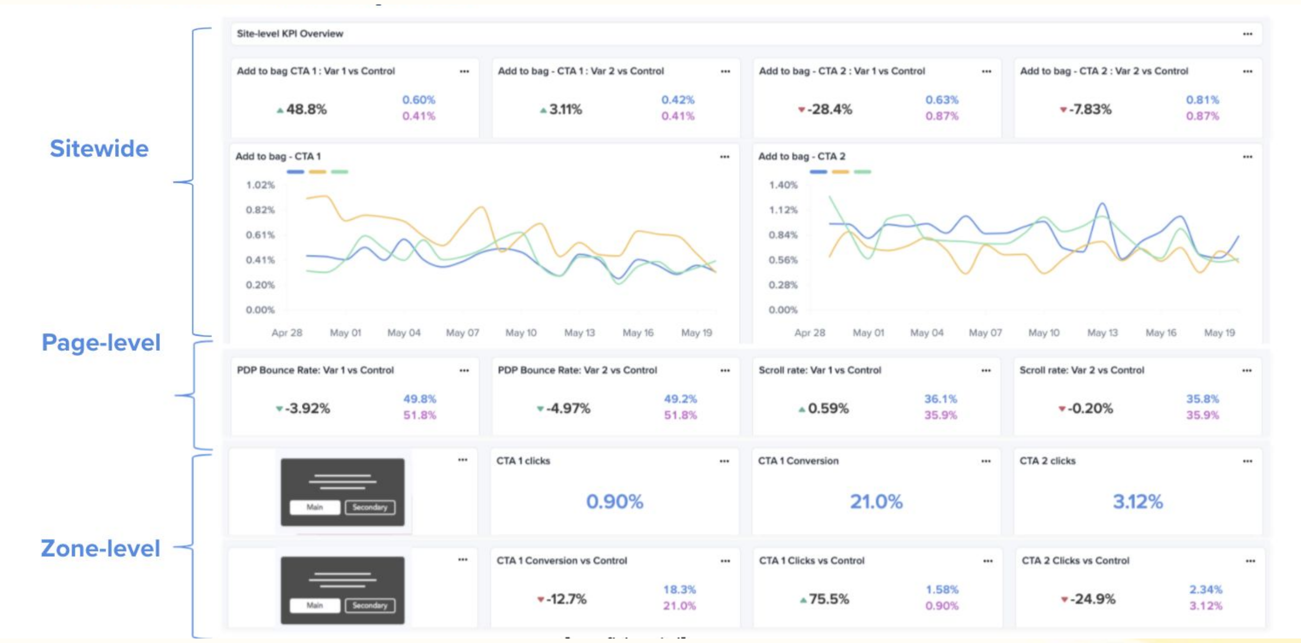
2. Setup your Workspace

To monitor the performance of your A/B test you can create a **Workspace** to view important metrics at a glance.

- **Site-level performance -** create widgets with site-level metrics like conversion rate, average cart and bounce rate to understand how the test impacted the the KPIs of each population
- **Page performance** - create widgets with page level metrics such as bounce rate, load time and percentage of visits achieving the test micro goals (e.g., add to bag, reach Checkout)
- **Zone performance** - create widgets for measuring the usage of the tested content (e.g., click rate and attractiveness rate of a new Homepage banner)

Example of Workspace created by a client running an AB test with a Control and two Variants for primary and secondary CTAs on the PDP.

- *The top part of the workspace uses single value widgets to look at the site-level add to bag conversion rate for each CTA. This is using the conversion rate metric for the goal of Add to bag and layering on the relevant segments for the population falling into each variant of the test.*
- *The middle part contains page-level metrics for monitoring the impact the test is having on the user's engagement with the PDP: bounce rate and scroll rate are compared to the Control for each of the two variants.*
- *The bottom part of the workspace is designed to look at a zone-level data: the click rate and conversion rate per click driven by each CTA within the variants.*



Note: *Use these widgets to create Alerts, ensuring you find out quickly whenever something unusual happens. See how to create an alert and a widget.*

Analysing the Test Results

1. ***Check the Test Performance in your AB Test Solution Tool***

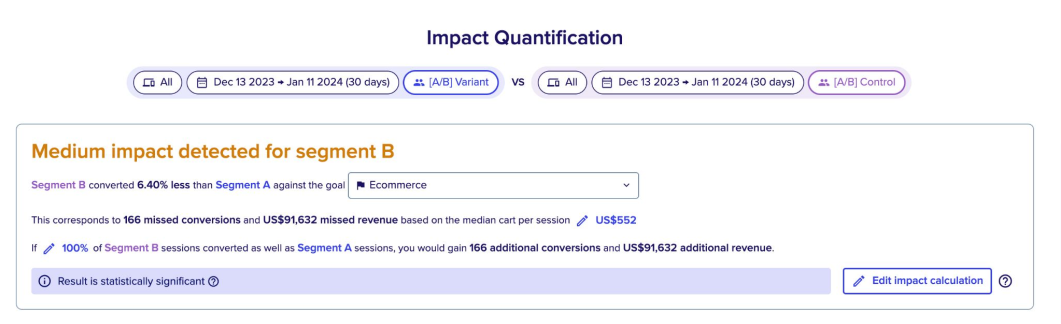
The first step is to take a critical look at the results of the AB Test from your AB test solution tool. Note down the answers to the following questions:

- Are the results significant?
- Is the 'Bestsellers' block in the Variant getting more clicks? How is the Homepage bounce rate compare between Control and Variant?
- What about its effect on conversion? Is it improving in some indicators but not translating to an overall improvement in conversion? If so, we need to start analysing *why* this is so.
- Is the variant significantly changing the way customers use the Homepage? Now is the time to start really drilling down and make a deep dive analysis in Contentsquare.

2. ***Analyze the Test Performance in Contentsquare***

Impact Quantification

The first step is to take a critical look at the results of the AB Test. Using the Impact Quantification to start, you can get a view of the session level metrics for each group.



Impact Quantification

All | Dec 13 2023 → Jan 11 2024 (30 days) | [A/B] Variant vs All | Dec 13 2023 → Jan 11 2024 (30 days) | [A/B] Control

Medium impact detected for segment B

Segment B converted **6.40% less** than Segment A against the goal Ecommerce

This corresponds to **166 missed conversions** and **US$91,632 missed revenue** based on the median cart per session US$552

If 100% of Segment B sessions converted as well as Segment A sessions, you would gain **166 additional conversions** and **US$91,632 additional revenue**.

Result is statistically significant

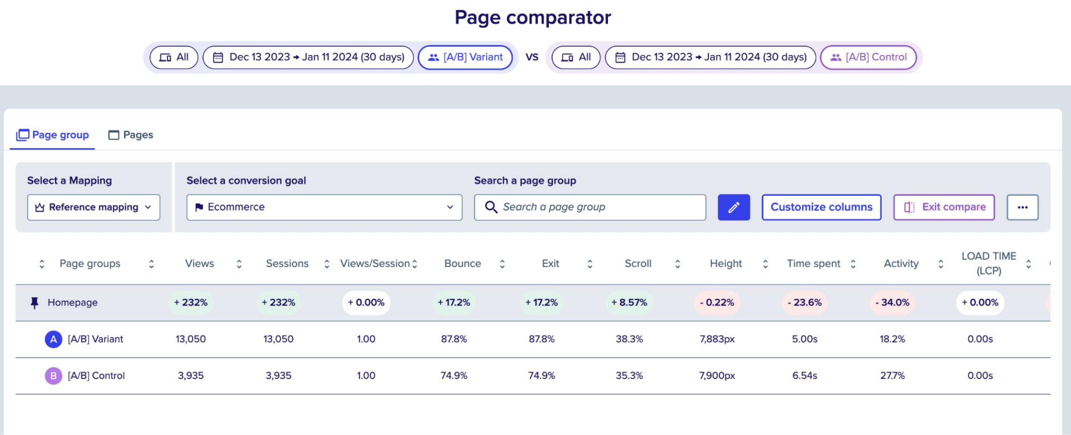
Here, by applying the different primary and secondary goals you can start to see how your AB test is impacting your visitors. You can begin to uncover questions like:

- Are the results significant?
- What about its effect on conversion and revenue?
- Are the visitors in the Variant more likely to stay on the site? Do they have deeper journeys and spending more time on site?
- How likely they are to achieve the page goal? (Are they more likely to add something to their Bags or view enter the booking funnel?)
- Is it improving in some indicators but not translating to an overall improvement in conversion? If so, we need to start analyzing why this is so.

*Read more about **Impact Quantification** here.*

Page Comparator

Page comparator is useful for getting an overview of different segments' activity on a given page: Did it impact any of the key UX metrics on the page?



Page comparator

Select a Mapping: Reference mapping

Select a conversion goal: Ecommerce

Search a page group: Search a page group

Page groups	Views	Sessions	Views/Session	Bounce	Exit	Scroll	Height	Time spent	Activity	LOAD TIME (LCP)
Homepage	+ 232%	+ 232%	+ 0.00%	+ 17.2%	+ 17.2%	+ 8.57%	- 0.22%	- 23.6%	- 34.0%	+ 0.00%
A [A/B] Variant	13,050	13,050	1.00	87.8%	87.8%	38.3%	7,883px	5.00s	18.2%	0.00s
B [A/B] Control	3,935	3,935	1.00	74.9%	74.9%	35.3%	7,900px	6.54s	27.7%	0.00s

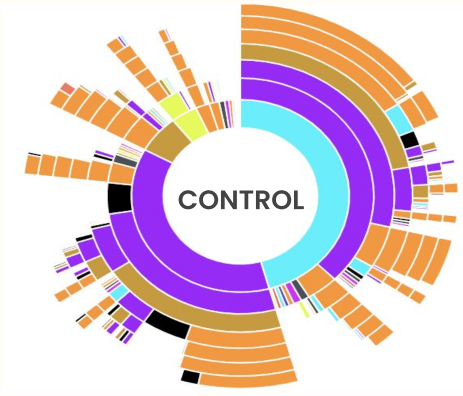
You can further drill down into your in-page performance. You can quickly see:

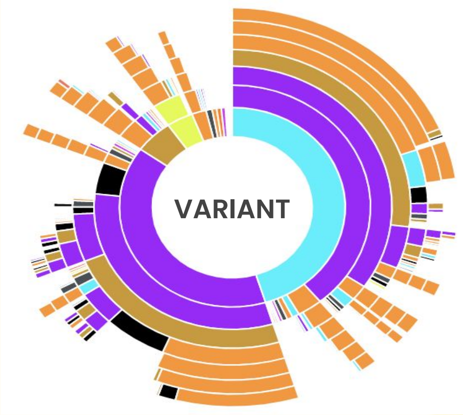
- If your test is helping in retaining your visitors on the page
- If the change had an impact on how much your visitors scroll through the page
- And if they're now more active on average and more likely to engage with the page content
- How the test is impacting the micro conversions of your page: are your visitors now more likely to achieve the goal you've set? Are they seeing more of your Product pages, adding items to their Carts?

*Read more about **Page comparator** here.*

Journey Analysis

Use the Journey analysis to understand if test has affected the key journeys on the site: Do the journeys differ between the two segments?



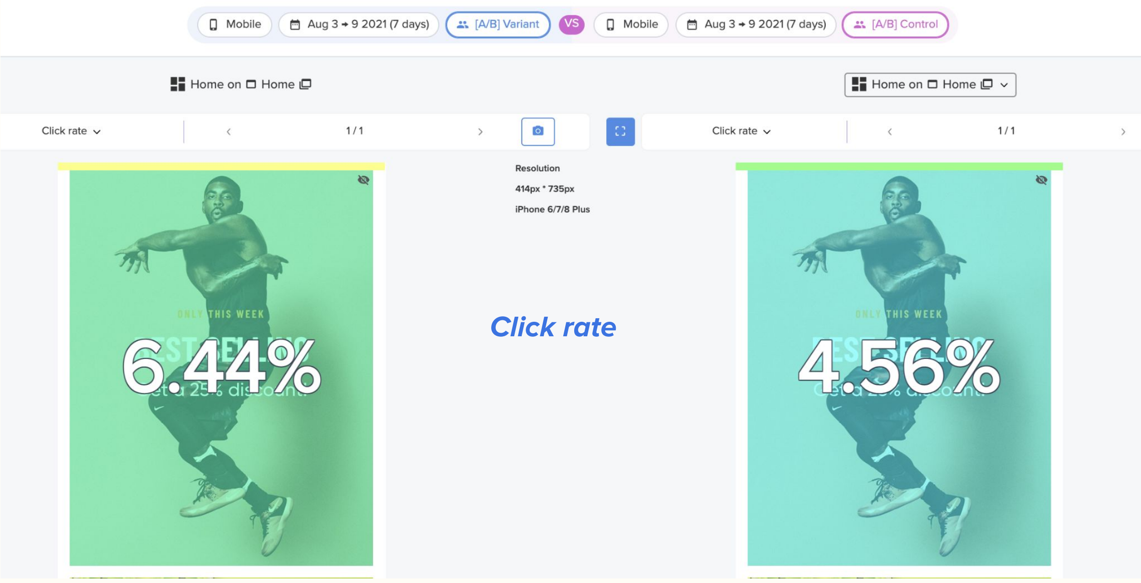


By looking at the journeys after the page where the test is running, you can really understand:

- Where your visitors go after seeing yout test?
- Is the change aiding their overall navigation or disrupting their journeys by creating stumbling blocks? Are they seeing any unexpected pages (e.g., Error pages)?
- Is there any looping behaviors: are they more likely to loop back between your Homepage and Product pages?

Zoning Analysis

How does interaction on elements on the page differ between the two versions: Is the 'Bestseller' block getting more engagement ?


Mobile | Aug 3 → 9 2021 (7 days) | [A/B] Variant VS Mobile | Aug 3 → 9 2021 (7 days) | [A/B] Control

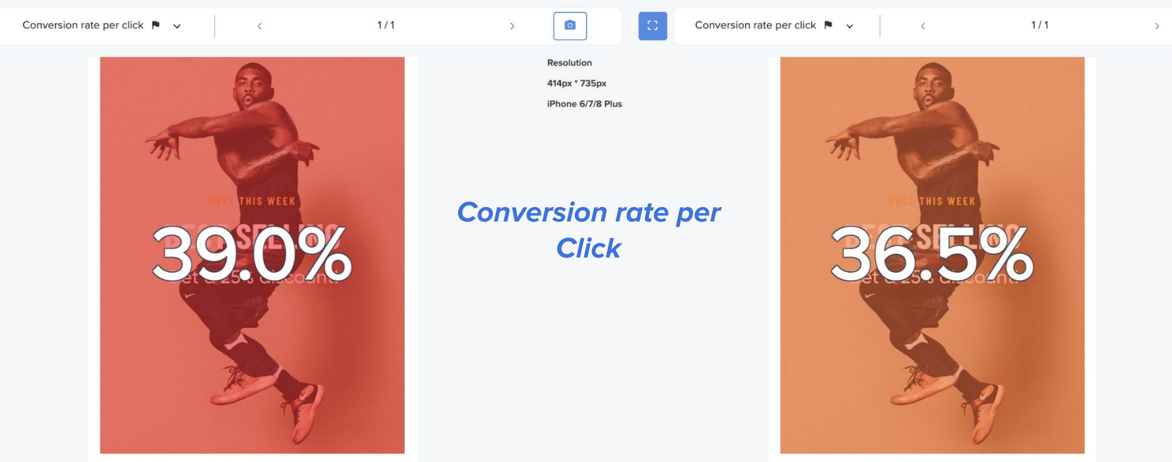
Home on Home | Home on Home

Click rate 1 / 1 | Click rate 1 / 1

Resolution
414px * 735px
iPhone 6/7/8 Plus

Click rate

6.44% | 4.56%


Conversion rate per click 1 / 1 | Conversion rate per click 1 / 1

Resolution
414px * 735px
iPhone 6/7/8 Plus

Conversion rate per Click

39.0% | 36.5%

With the zoning you can get an in-depth understanding of the user behavior on the page. You can answer question like:

- How did the test impact the attractiveness of the key element? Is the 'Bestellers' block getting more or less clicks?
- Did it have an impact on interactions in any other way? How are the Hover rate, Time before first click, Click recurrence, Conversion rate per click trending?
- Did your test impact the interactions with other key elements on the page? Did engagement with the other product pushes or the search page dropped?

*Read more about **Zoning analysis** here.*

Impact Analysis: Summary of Key Insights and Recommendations

After doing all of the analysis, it's good practice to make a list of the generated key insights, to put together all of the conclusions about how the test has impacted the user behavior.

Mapping out the key takeaways will then be used as a basis for making a decision 1) of what to do next: deploy or iterate and 2) to generate ideas for future iterations.

Example of a table summarizing key insights and recommendations for iterations

INSIGHT	RECOMMENDATION FOR ITERATION
Bestsellers is **still only exposed to around 40% of mobile visitors,** % even lower on Desktop despite being the **second highest revenue generating module** on Homepage	Shift Bestsellers higher up on the page
AOV slightly drops when recently viewed replaces Bestsellers as users don't discover as many items when they have only viewed 1-3 items. However, **PDP reach rate increases, especially for returning users**	Split out Bestsellers and Recently viewed modules, Display recently viewed items above the fold when more than 2 items have been viewed already

Appendix

The Statistics Behind Online Testing

How to do Hypothesis Testing

Statistics are fundamental to designing and analyzing tests the right way. And every good test is based on a hypothesis. Here you'll learn what is a hypothesis, why do you need one, how to set it up, and what to do after.

Hypothesis Testing

An A/B test is an example of the classic statistical hypothesis testing, or the use of statistics to determine the probability that a given hypothesis is true - you define a problem, you propose a solution and you predict the outcome.

1. Setting up a Hypothesis

The **null hypothesis** refers to something that is assumed to be true if there is no real difference between the Treatment and the Control, that any observed difference is just noise.

The **alternative hypothesis** refers to something that is being tested against the null hypothesis and is basically a hypothesis that you believe to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.

Here's an example of a good format for writing your hypothesis:

We believe that doing [A] for people [B] will make outcome [C] happen. We'll know this when we see data [D]

With a hypothesis, we're matching identified problems with identified solutions while stipulating the desired outcomes.

***Identified problem**: " The Add to Cart rate on your Product page is 11% . By doing some previous research (e.g., surveys, heuristic evaluation), a problem you identified is that you do not have product reviews on the page."*

***Proposed solution:** " By adding reviews on the product pages, you will increase social proof, trust and confidence in the product, thus increase the number of users adding items to their Carts. You will know this by looking at the Add to Cart CTA.*

The **null hypothesis** here would be: ***no reviews generates an Add to Cart rate equal to 11% (the status quo)***

The **alternative hypothesis** here would be**: *adding reviews will cause Add to Cart rate to be more than 11% (challenging the status quo).***

After setting up your test hypothesis, it's time to pick the test statistic that will help you determine if your hypothesis can be validated or not.

2. Picking a Test Statistic

This is the method and value that will be used to reject or validate your hypothesis. **Two-sampled T-tests** are the most common statistical significance tests to determine if there is a real difference between your Treatment and Control, and hence, rejecting or validating your hypothesis.

At its most basic, the two-sample t-tests look at the size of the difference between the two means of your two samples (your users in the Control vs the Treatment) relative to the **variance**, or the measure of variability of your data. The variance tells you the degree of spread in your data set.

The significance of the difference is represented by the **p-value,** or the probability that the difference would be at least this extreme if there is really no difference between Treatment and Control**.** The lower the p-value, the stronger the evidence that there is a real difference between Treatment and Control. By convention, any difference with a p-value lower than '0.05' is deemed 'statistically significant'.

Note: The p-value doesn't tell you if the Null hypothesis is really true or not, it shows the probability of observing the delta of the Null hypothesis is true.

3. Control for Type I and Type II Errors

As with any tests, your A/B tests are also prone to errors. In hypothesis testing, you care about **Type I** and **Type II** errors. A Type I error, better known as alpha, is to conclude that there is a difference between the Control and the Treatment when there is none. And Type II error is when you conclude there is no difference when there is actually one.

You control for the **Type I** error by inferring statistical significance only when the p-value is < 0.05 (or alpha = 5%). With alpha at 5%, it means that there is 95% confidence placed in the results. When comparing the p-value to alpha, the null hypothesis is ruled out once the p-value is less than or equal to alpha.

Type II error on the other hand, or as it's better know as power, is the ability to detect differences between the variants when there is one. Industry standard is to strive for at least 80% power in your tests. The common way is to conduct a power analysis before starting your tests to decide how many samples (or users) are needed in each condition in order to have enough power.

The hypothesis we start with is either true or false and we either accept it or reject it, which leads to the following four distinct cases.

	True	False
Accepted	*You hypothesized that adding reviews will increase the page conversion. Your visitors are indeed adding more items to the bag and purchase more. Your test analytics tells you that your test is winning.* ***You are right!***	*You hypothesized that adding reviews will increase the page conversion. Your visitors are not noticing the reviews and are not doing anything different on the page. Your test analytics tells you that your test is winning.* ***You are wrong!***
Rejected	*You hypothesized that adding reviews will increase the page conversion. Your visitors are indeed adding more items to the bag and purchase more. Your test analytics tells you that your test is losing.* ***You are wrong!***	*You hypothesized that adding reviews will increase the page conversion. Your visitors aren't notice the reviews and ate not do anything different on the page. Your test analytics tells you that your test is losing.* ***You are right!***

4. Checking for Threats to the External Validity

External validity refers to the extent to which assumptions made in control experiments can be generalized to axes such as different populations (e.g., would this trend be applicable in other locations, or other websites and domains?) and across time (e.g., will the 2% increase in CR continue to grow, diminish over time, stay stable?).

Generalizations across populations is often dubious; a feature that works on one site may not work on another. A typical solution is to rerun the experiment for different markets.

Generalization across time is even harder. You want the observed results to stand the test of time, after all you want long-term improvements. Here you have to consider two types of threats:

- **Primacy effects.** When you introduce a change to your visitors, they may need some time to really adapt to it and adopt it. They're still primed to old functionality and need time to get used to the change.

- **Novelty effects**. The novelty effect, on the other hand, is a fleeting effect with a short shelf life. When introduced to the new feature, your visitors get excited initially and are excited to try it. If they don't find it useful enough, however, they will stop using it altogether. Your Treatment will appear to be doing very well at the beginning, but will quickly decline over time.

A common way of checking for primacy and novelty effects is to plot usage over time and see of its trending up or down. You want the Treatment effect to be constant over time.

Tip: Run your test for at least a full business cycle, which is weekly in 95% of cases. Even if you reach your Minimum Sample Size in 3 days, you should not stop your test until it has run for 7 full days, or whatever duration your business cycle is.

Defining your Testing KPIs
Principles and Techniques

The success of your testing program depends on defining not only the right things to test but also the right things to measure. Is it clicks, is it transactions or revenue? How to decide? And how to align them with the overall business objectives?

Key Principles when Developing your KPIs

Ensure that your metrics are:

- **Simple**: Easily understood by all stakeholders
- **Measurable**: Even in the online environment, not all effects are easy to measure. Post-purchase satisfaction or generating cross-channel leads (e.g., online to brick-and-mortar stores' sales) can be hard to measure.
- **Attributable**: To measure the success of your test, you must be able to assign actual values to your testing segments. For example, if you want to analyze if the Treatment is causing higher app crashes or higher loading times, you must be able to attribute the issue to its variant.
- **Sensitive and timely**: Your metric should be able to capture changes that happen in a timely fashion. An example of an insensitive metric is trying to make a change to the stock price of your company by running a test with improved product placements. To impact the company stock by changing individual product for a limited amount of time (or the duration of the test) is practically impossible.

Tip: We always recommend in your analyses to combine different metrics to account for such side effects.

Here are some helpful techniques and considerations for developing metrics

- Use **hypotheses** to generate ideas and validate them through a robust data analyses. (Read more about How to set up your testing hypothesis)
- Remember that metrics are **proxies** - always come back to the reason you created your test and link this back to the objective of your page. Is your page's objective to lead more people into the conversion funnel, add to cart, give specific information or be used as a navigation tool in your sire? And what do you want to achieve with your test?

__Use case__: You want to improve the navigation capabilities of your Homepage. To do so you are thinking of rearranging your menu or adding banners and links to other key parts of your site. Success metric here, hence, could be click rate on the new element but also reach rate of the next step of the funnel.

- Think of **edge/niche cases.** Each metric has a corresponding set of failure cases.

__Use case__: As a travel site. you are heavily reliant on your search engine. You want to increase the engagement and your success metric is click-through rate. By only looking at CTR, however, might miss usability issues your visitors are encountering (e.g., seeing irrelevant results, reaching error pages). Hence, you must look into additional metrics to account for such edge cases (e.g., click recurrence, the rate of reaching a designated follow-up page).

- Sometimes is easier to **measure what you don't want to achieve,** such as frustration or disappointment, than what you really want to accomplish. It's hard to create a cut-off point for time spend on a site for someone to be classified as a 'satisfied' user. On travel sites, a short visit might mean that a user is easily finding what they're looking for, doing some preliminary research and returning to book on a later date. On others, like media sites, you want a very long visit duration as an indication that your visitors are finding the right shows and spending time actually watching them. In that sense, a decrease of a negative metric (e.g.,bounce rate) can be used as a measure of user satisfaction.