**Selecting the right data**

In this second chapter, we will deepen our knowledge by learning how to prepare for communicating data. We will learn how to identify the different targeted personas to appropriately select findings and statistics, visualizations, and choose the appropriate format to make an impact with our results.

**Data storytelling road**

Let's dive right in. We learned why designing a story is important. We also learned to decide between a technical or non-technical approach.

Now, it's time to select the right data.

**The right data**

But what does it mean to select the "right data"? It implies including sufficient contextual insights in our story to support our main point without overloading our report with information. In other words, we include the minimal amount of results that will support our story.

1. 1 Dykes, Brent. Effective Data Storytelling. Wiley.

**Data storytelling**

And why is selecting insights important? Remember that data is one of the three central elements in any story. Garbage in, garbage out: if our raw data is bad and unreliable, then our analysis will be too. Same for our results But when we find the insights, we need to display our findings properly to highlight a different view of the data for each audience.

**Stakeholders**

Different stakeholders will be interested in our findings differently, so we need to tailor how we show our results to each of them. A stakeholder can be any person or group of persons that has an interest in the outcome of a project, or a decision or activity that is derived from it. As we mentioned before, stakeholders can be technical or non-technical. Among them, we may have different types of audiences.

**Identifying personas**

A helpful tool is to identify audience personas. It means describing the interests of the project outcomes and the technical knowledge of someone who represents our ideal target audience group. Defining personas helps selecting tailored findings to better convey our key insights.

**Identifying personas**

Let's go back to our communicatb food project. After finding our story, we need to define personas in our targeted audience, which will allow us to select the right findings in order to adjust our message.

**Executive team**

The first persona we might be reporting to is the executive team - most likely, the CEO, an investor, a director, or a founder. They have basic knowledge about technical aspects. And they usually need to inform their decisions using our project findings.

**Project manager**

In our example, when telling the story to our project manager, we can show the overall cost of the proposed next steps, and general metrics such as the predicted increase in number of customers, as well as the risk that our plan will fail.

**Tech team**

We can also present a report to our project collaborators or a technical supervisor with high technical knowledge. They are likely interested in replicating or continuing the project.

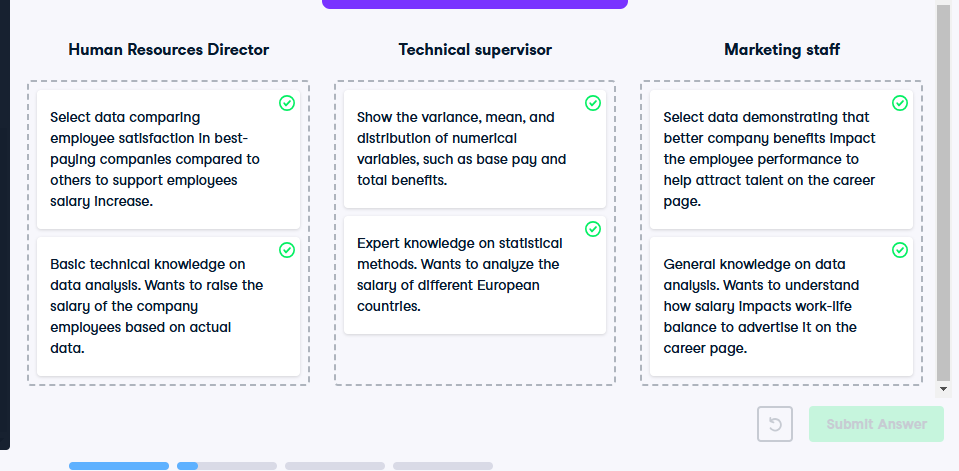
**General audience**

Lastly, we can show our story to an external customer or to other department staff that likely have basic or general technical knowledge, and want to understand the impact of the project.

In our example, we should show the historical decline in profits, and explain how the rebranding of the chocolate will increase next year earnings.

**Audience skepticism**

Finally, the audience can be more or less receptive to our results. Some will be more skeptical about our results than others. If people are easily convinced of our results, fine. But we should be prepared to bring in more data, insights and information to convince skeptics if our reasoning is put in doubt. A classic approach recommends to first convince yourself, then a friend then a skeptic.



**Showing relevant statistics**

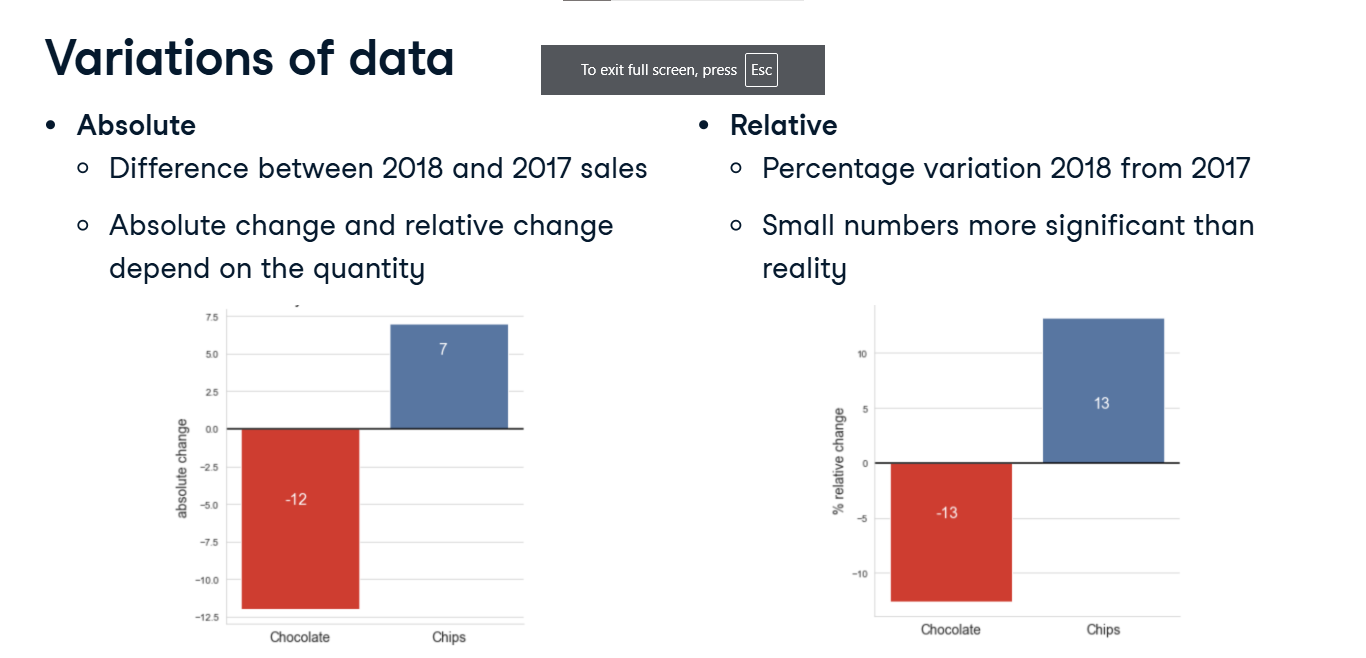
Great job identifying and profiling different types of audiences.

**Data storytelling road**

We now know what selecting the right finding means, and why it is important for storytelling. Now let's cover how specific metrics help convey different messages.

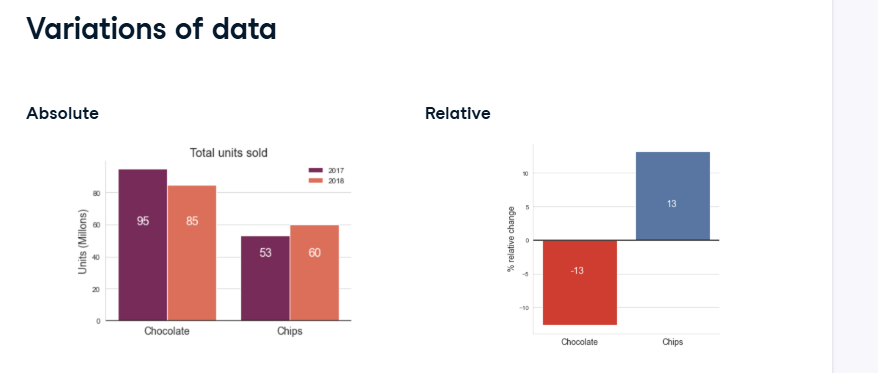
**Variations of data**

Sometimes we want to compare one or several specific variables over time. The difference can be expressed with an absolute number (3,000 more units sold) or a growth rate (10% more units sold). If we focus on one variable, an absolute number is OK, but when looking at several variables, the growth rate tends to be more informative as it allows comparing different scales (products sold in volumes of thousands and millions, for example). For example, for absolute change, we care about the difference between 2018 and 2017 sales. When we focus on the percentage change from 2017 to 2018, we use relative change. An absolute change for a small number can be small even if the relative change is large. On the opposite, absolute change for a large number can be large even if its relative change is small. Relative changes on small numbers can appear larger than they are because a small absolute change can result in a large percentage change. Which one we use depends on the question we want to answer.



**Variations of data**

Say we were asked to explore proportion changes in sales volumes between years. In the graph, we see the total units sold in 2017 and 2018. It's a good plot to show that more chocolate is sold than chips, or that chocolate sales decreased while chips increased. But since we're interested by proportion, we'd better show the percentage of change. Now we have the respective percentage changes for each product. So let's see different situations and which type of metrics we should use.

****

**Ratio**

A way to overcome this issues is to calculate a ratio. It is a comparison of two variables expressed as a quotient, such as the revenue per customer; calculated as total product revenue in dollars divided by number of customers as we can see in the graph. Ratios help normalize values, which in turn helps compare the distribution of data of originally different scales.

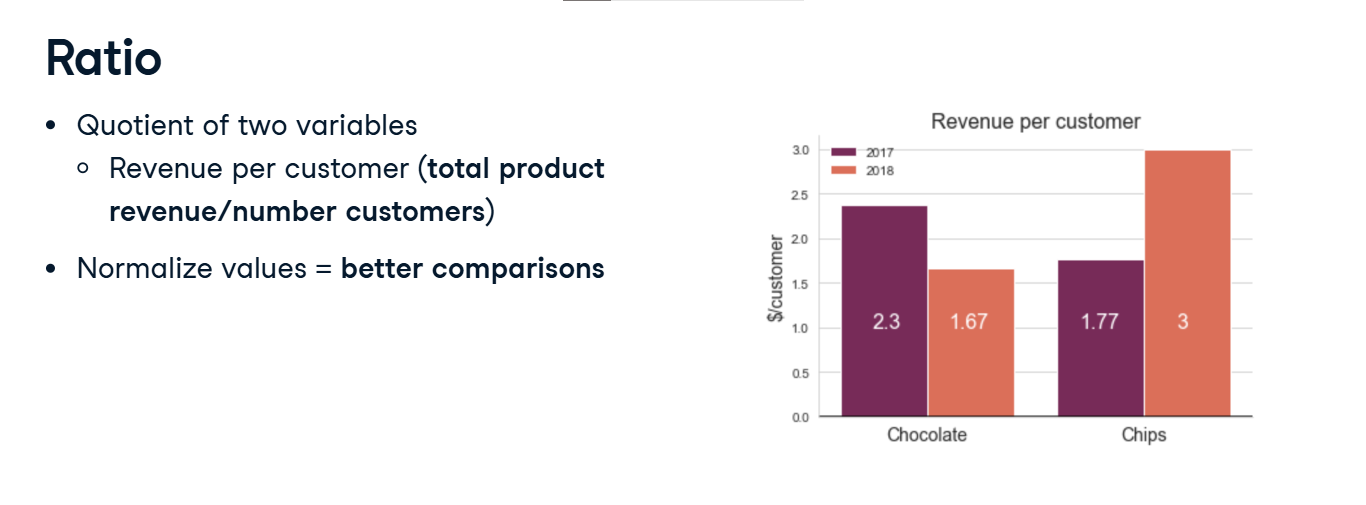
**Aggregates**

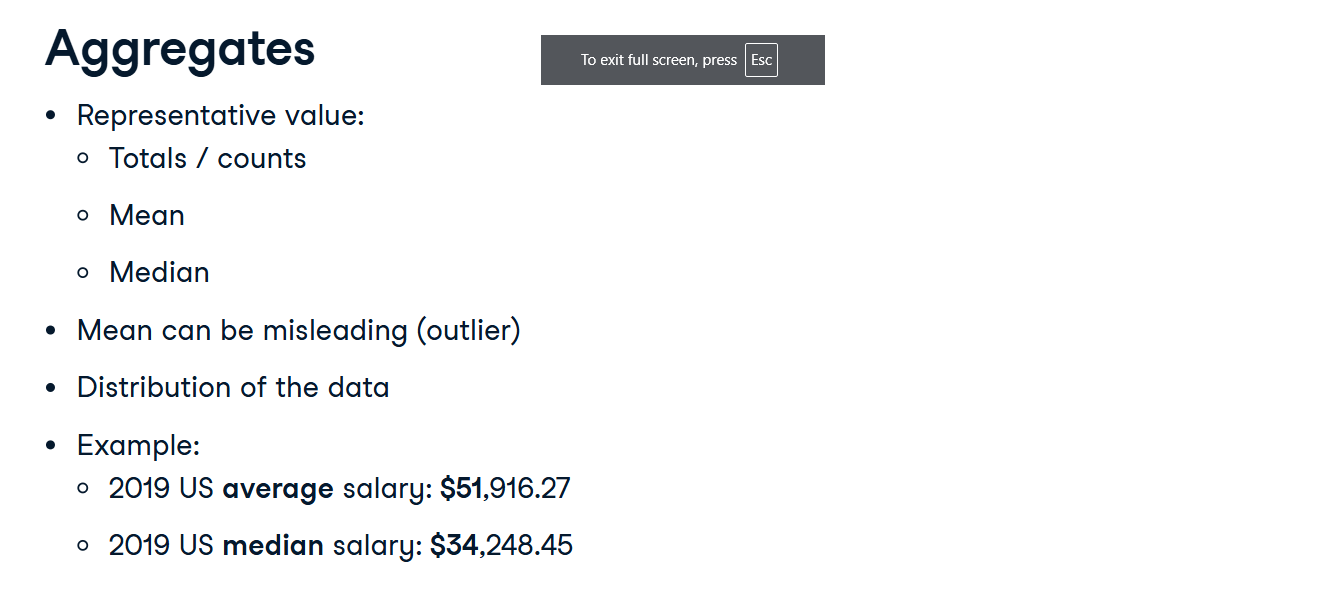
Sometimes we need to summarize numerical data into an aggregate: a number that gives an idea of an overall or representative value It can be a simple total or count, like total sales or how long a a campaign will last.

Another common aggregate is the mean. the average number of chocolates or chips sold per year, as we see in the graph.or the median price for each product.

**Aggregates**

The mean can be misleading, particularly if there are outliers (extreme values, the data is not normally distributed). In these cases, the median is a more robust metric. For example, in the US, in 2019 the average annual wage in 2019 in the US was $51,916.27, and the median annual wage was $34,248.45. Using the mean, we'd think the common tendency is a $52K wage, when actually half the population is below $34K.

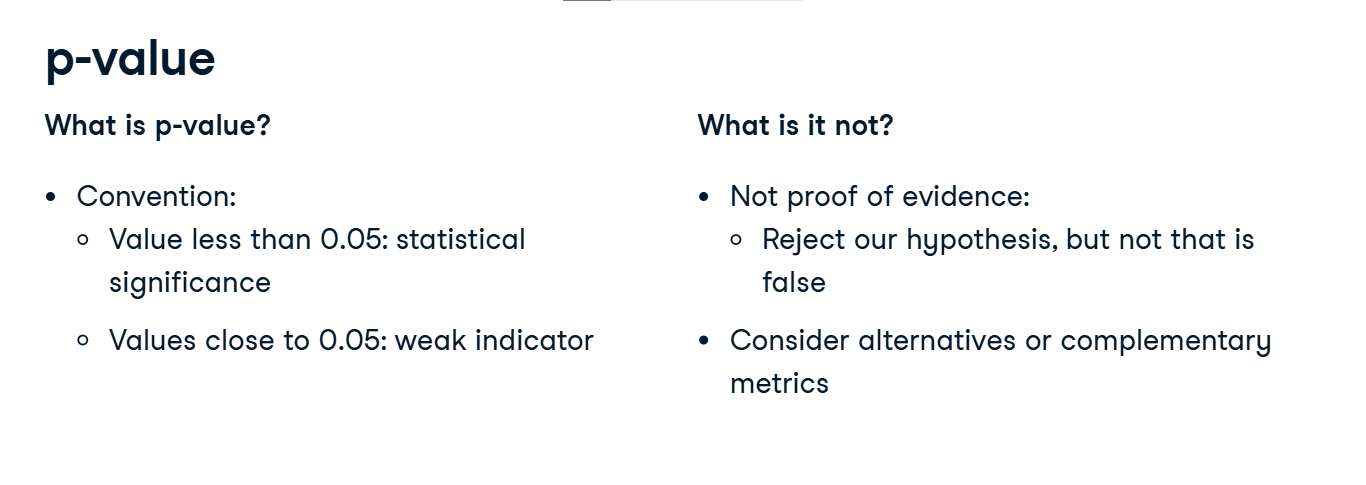


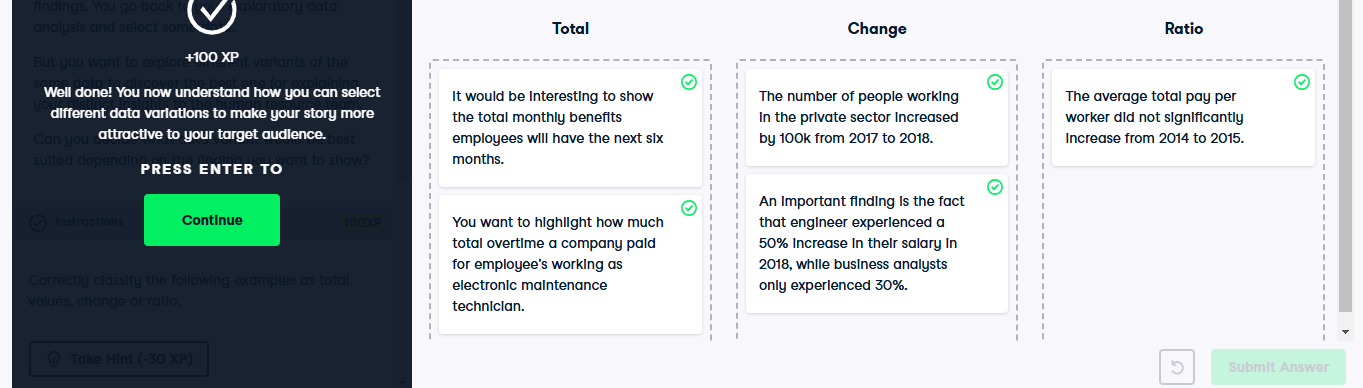


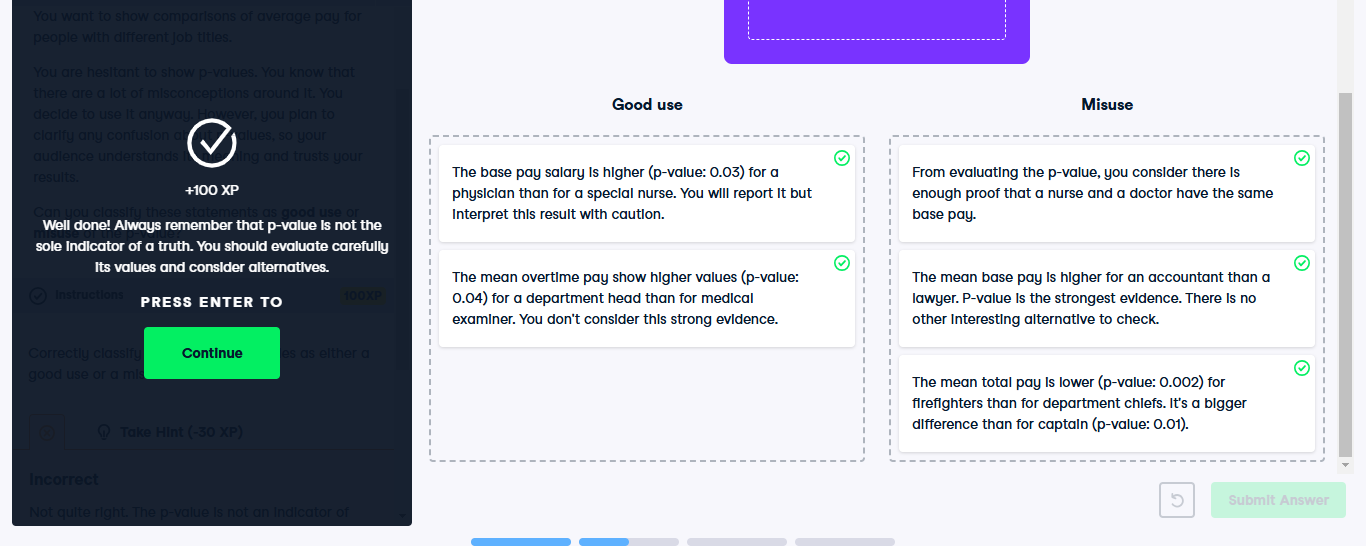
**p-value**

When communicating our results, we often need to provide proof that they are statistically significant (i.e., that they can't be due to randomness). A p-value lower than 0-point-05 is considered an indicator of significance by convention. The lower it gets below 0.5, the stronger the indicator. However, it is not a proof of evidence: it only rejects or accepts a hypothesis without saying anything about the truth of it. Because of how often p-value metrics are misunderstood or confusing the audience, consider alternative metrics or add some more in support.

1. 1 P-Values: Misunderstood and Misused. Front. Phys. 2016.







**Visualizations for different audiences**

Great job on these exercises! Let's learn about how to include visualizations in our story.

**Communication strategy**

We know how to craft a story for a technical or non-technical audience.

Now, we are ready to choose a data visualization intended for that specific audience.

**Data storytelling**

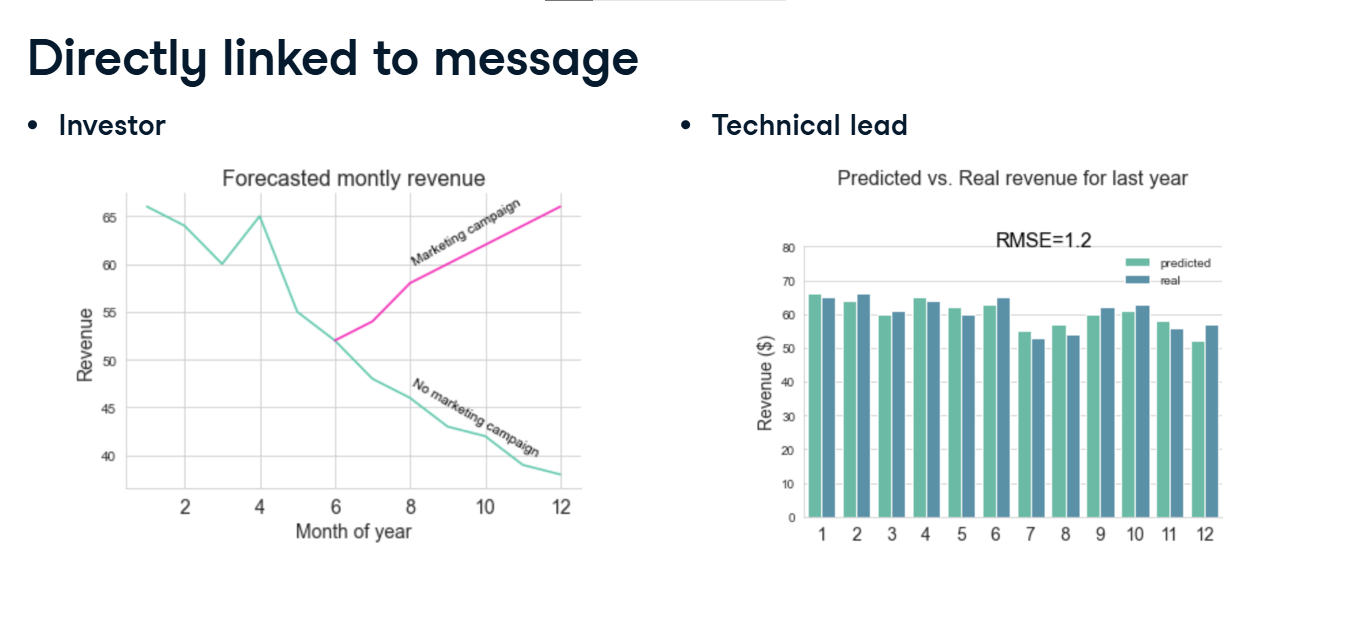
We have seen that visualizations are key elements for a data story. When we find the insights and have the right data, we still have to choose and adjust a visualization to the message we want to convey. Again, we should consider our audience expertise in the topic and familiarity with the concepts, to select what graphs they can easily interpret. For example, a density plot would be OK for a technical audience, but a histogram would be more appropriate for non-technical people.

**Tailored message**

Let's get back to our chocolate project to illustrate some best practices for visualizations. Imagine we need to communicate two different messages: the first one for an investor, and the second one for our technical lead.

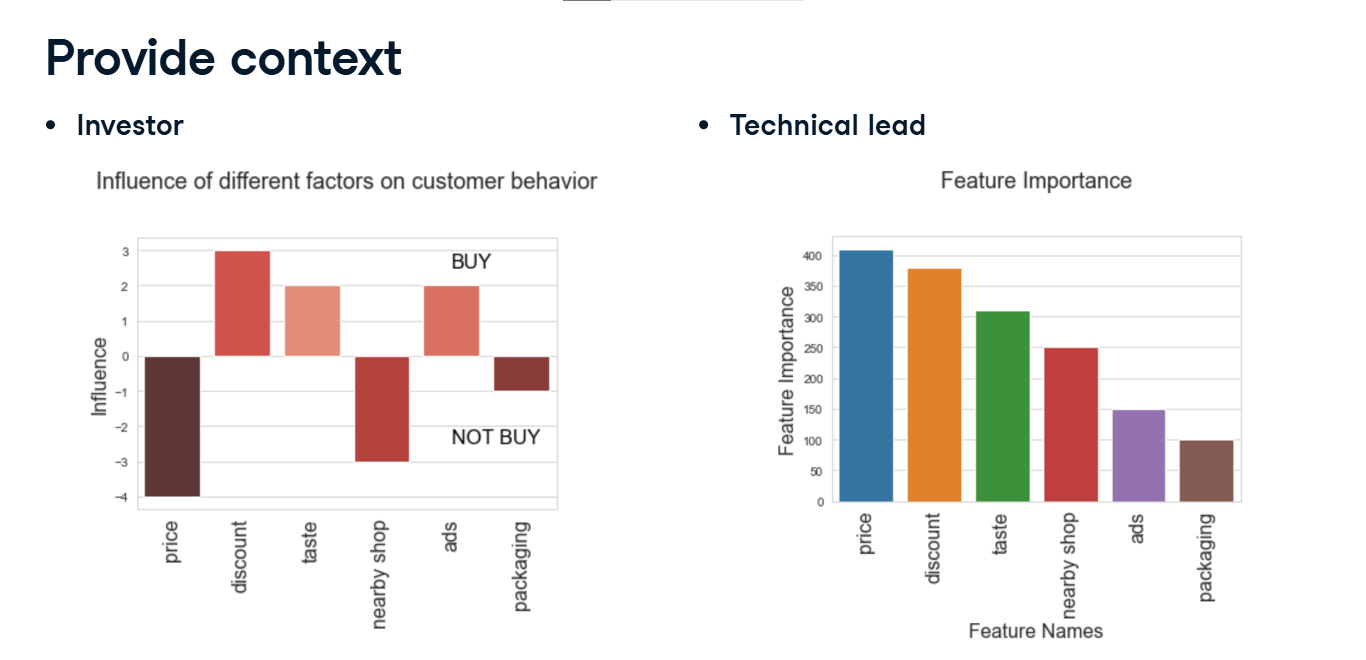
**Directly linked to message**

To include relevant visualizations, we should determine which data elements are required to help us convey our message, and only include those visualizations that are directly linked to our message. For our investor, we could show a graph showing the forecast monthly revenue for both scenarios: launching and not launching a marketing campaign. For our technical lead, we need to include a graph that shows how well our model performed on historical data.



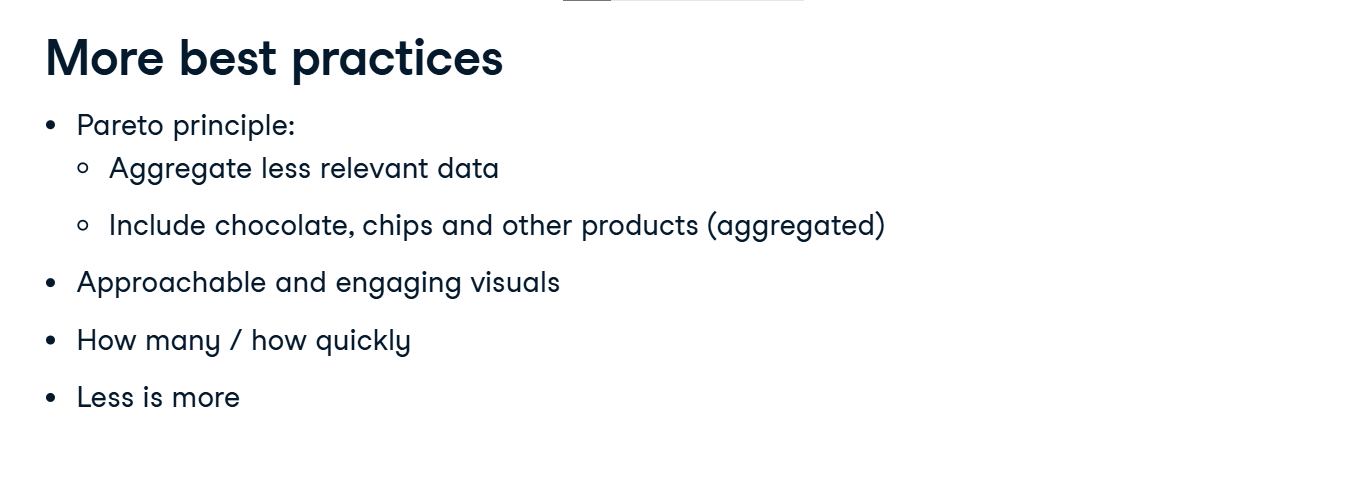
**Provide context**

Visualizations should also provide context to support our message. For the investor, we can include a graph showing the most influential factors for a customer to buy our chocolates. So for example, the higher the price the less likely the customer is to buy, and vice versa. The technical lead will be interested in this, but also more details on the analysis itself, such as the detailed feature importance.



**More best practices**

The Pareto principle states that the majority of outputs come from a minority of inputs. So outside of the data contributing to our story, there will be some data that is less relevant but that we want to include. We should aggregate it to reduce noise, such as presenting the sales for chocolate, chips and other products instead of each product separately. Our visuals can be made approachable and engaging by considering our audience background, and simplifying the visuals to the audience knowledge level. Basically, you care about how many insights your audience gets form your visualization, and how quickly they get it. A complex plot gives many insights but takes time to understand; a simple plot gives few insights but is quick to understand. In general, less is more. Instead of showing a lot of detailed visuals, we should just focus on the most simple and relevant ones that support our message.

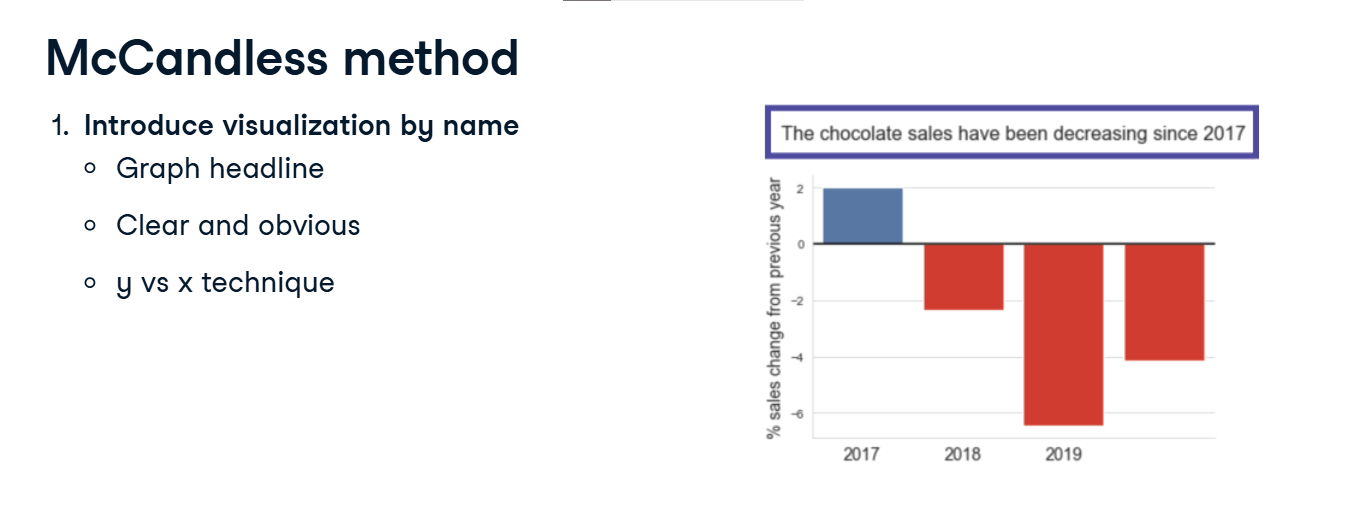


**McCandless method**

There are some steps that we can follow to include and present visualization effectively to our audience. They were established by David McCandless, a British data-journalist. First, we should give our graph a headline, which we then use to introduce the visualization and focus the audience's attention on it. The headline should be short, clear and obvious: it supports our story and clarifies the visualization. In doubt, we can use the y axis vs x axis technique, for example: "chocolate sales by month".

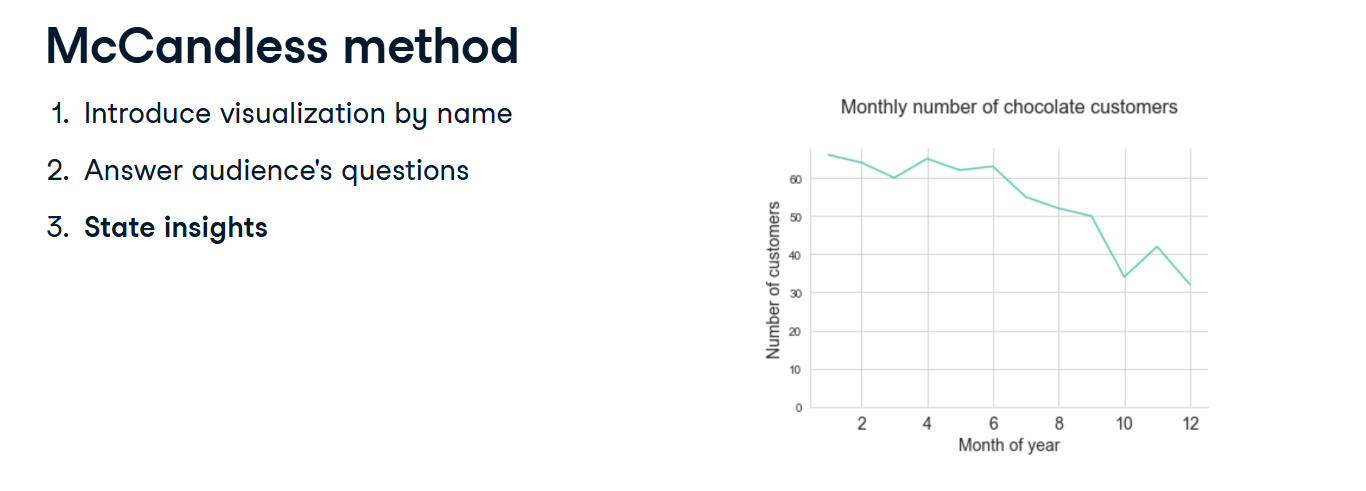
1. 1 https://artscience.blog/home/the-mccandless-method-of-data-presentation

Then, we should anticipate common questions from our audience, such as "Where does this data come from?" or "Why did you focus on this characteristic specifically"? With these questions answered before being even asked, attention is focused on our graph and our story.



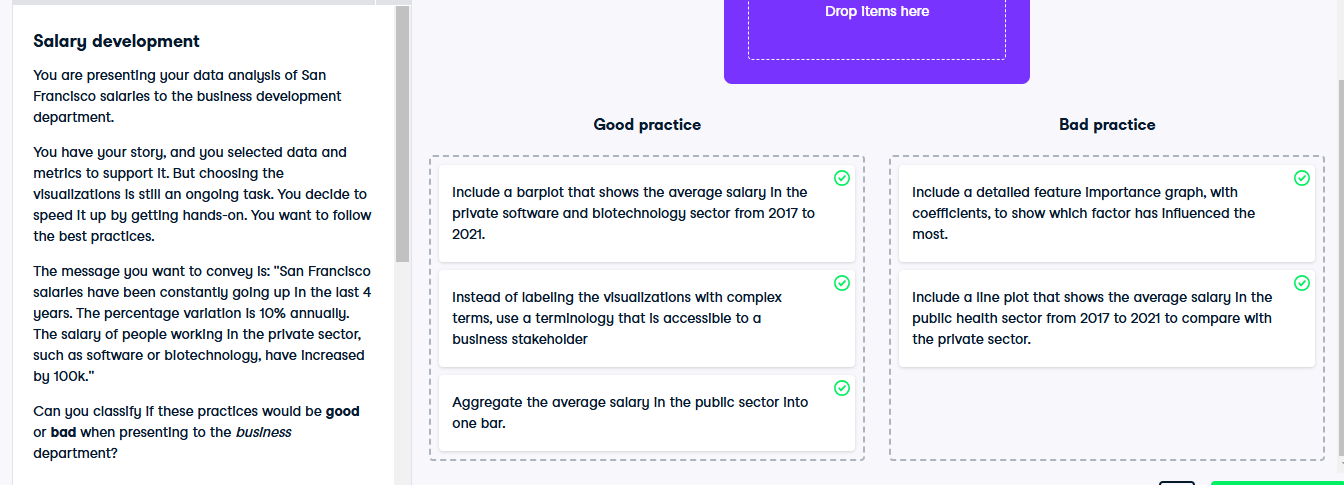
**McCandless method**

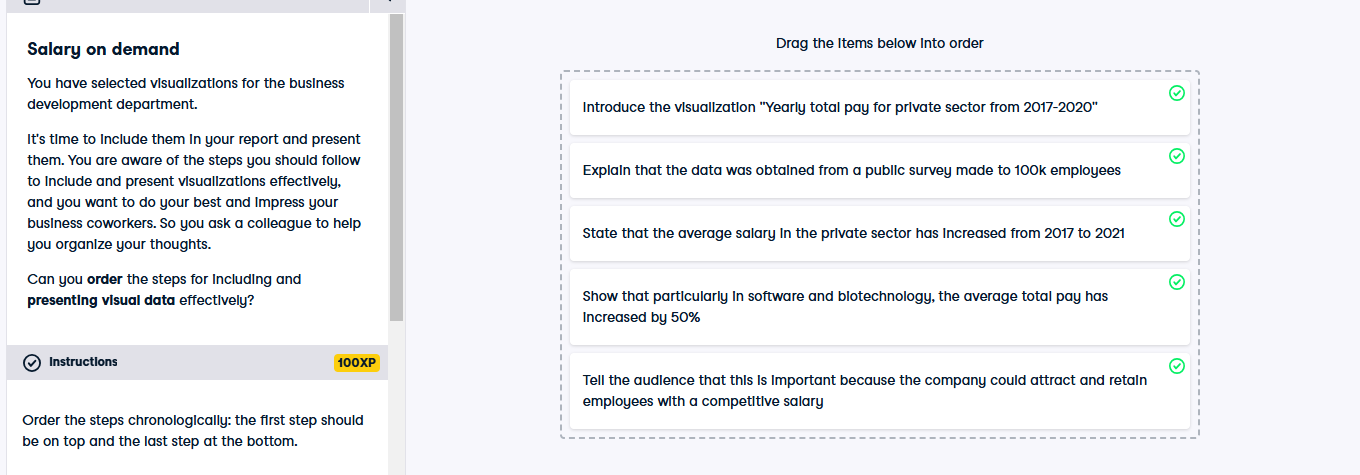
After that, we should clarify what insights the visual is showing. We should explain to the audience what they are seeing, and not assume they will ask later or understand by themselves. For example, chocolate sales have decreased with time.



**McCandless method**

Lastly, we need to help the audience relate to the graph and its insights. making sure that they understand why the insight matters, how it will support further insights in the presentation, or how this insight can be acted upon. For example, the chocolate sales are relevant to the finance department, as they impact the revenue, or to the marketing department, as advertising might help.





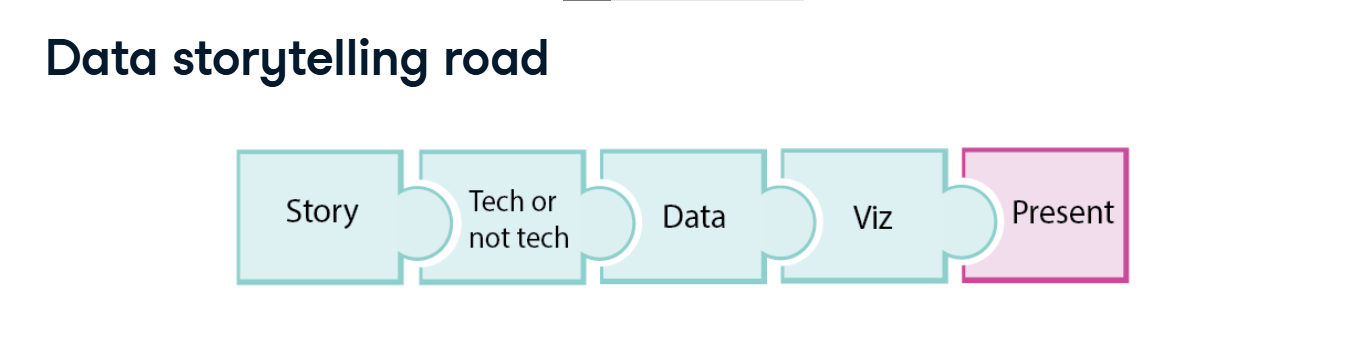
**Choosing the appropriate format**

We have just seen how to include visualizations in our story. Now, we'll learn when to use several presentation formats.

**Data storytelling road**

This is the final piece in our communication strategy. We've designed our story, we've defined our approach, we've selected the right data, and we've chosen an impactful visualization.

Now, it's time to present it to our targeted audience.



**Which format is more effective?**

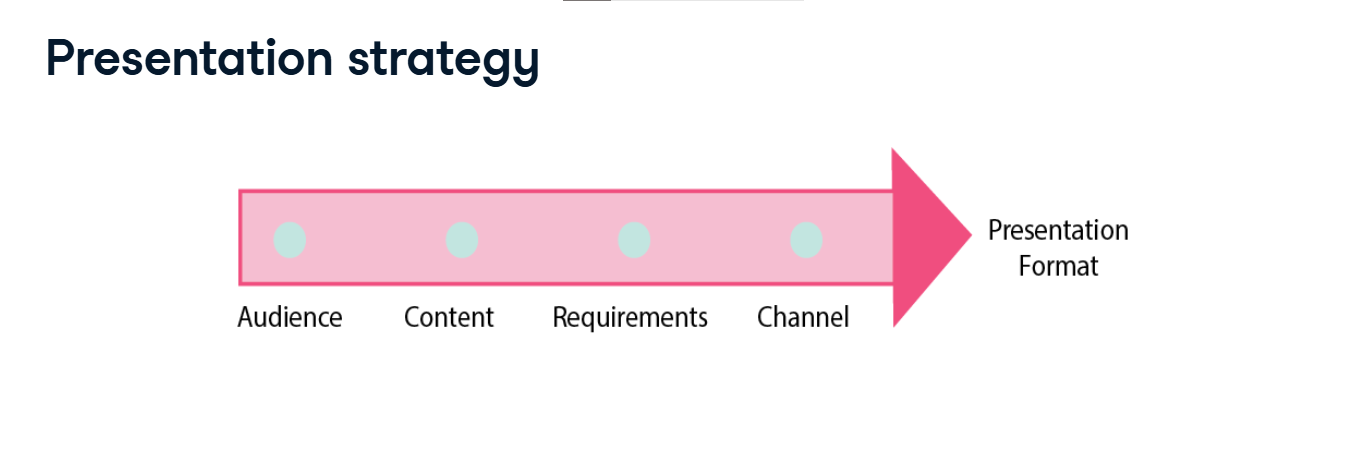
A good communication format shows key information from our project in a way that is engaging, and easy to understand.

There are two main formats that we'll discuss here: written reports and oral presentations. Most data science projects will require a written report and an oral presentation, but in the end it depends on the situation and project at hand.

**Presentation strategy**

There are several things to consider when sharing findings: Our audience, the content to include,special requirements to take into account, and which channel to use.

All these elements help us define the best format to communicate our results.



**Stakeholders**

The first thing to consider is who we are presenting to. This helps figure out Why they need to know about the findings: For accountability? To understand the methodology? How are they going to use our findings? To make a decision? To start a new project? What information do they need? Our results? The impact our findings have?

**Content**

The answer to these questions will help us decide which content to include. Should we focus on results, conclusions, include recommendations, or just explain the methods in detail?

1. 1 Photo from Unsplash

**Requirements**

The next step is to identify whether any stakeholder has special requirements. Do they have enough time to read a detailed report, or is a short meeting better? Do they report to someone else, and need a document to back up a claim? Are they in a different timezone, so a written communication is preferred?

**Consumption**

We also need to think about how our presentation will be consumed. We could write a document such as a Word doc, a Jupyter notebook, a blog post or an article. We could also build a slide deck. Then, how will it be delivered? Will we present directly to stakeholders, being able to answer comments or questions, or will we share it by email or Slack and get questions later? Finally, how big would the audience be? You wouldn't present things the same way in a conference room of six people or a ballroom of two-hundred.

1. 1 Photo from Unsplash

**Oral communication**

Selecting the correct format depends not only on the audience, but also on the strengths and weaknesses of the different formats. Oral communication allows for building a relationship with the audience, and for immediate feedback or actions. It delivers a rich message because body language and voice add meaning. However, unless you provide a recording, the message cannot be revised because there is no permanent record of communication, and it is not suitable for lengthy messages, as the longer the presentation is, the more chance there is that the audience will lose focus at some point.

**Written communication**

The written format provides records of communication so the message can be analyzed on the longer term. It is easy to share with large audiences, and it's less prone to emotional reactions. It is also suitable to share code with any technical stakeholder for review or replication. However, not seeing the audience reaction makes it harder to adapt, as the feedback is not immediate but will come later in the form of comments.

**Appropriate format**

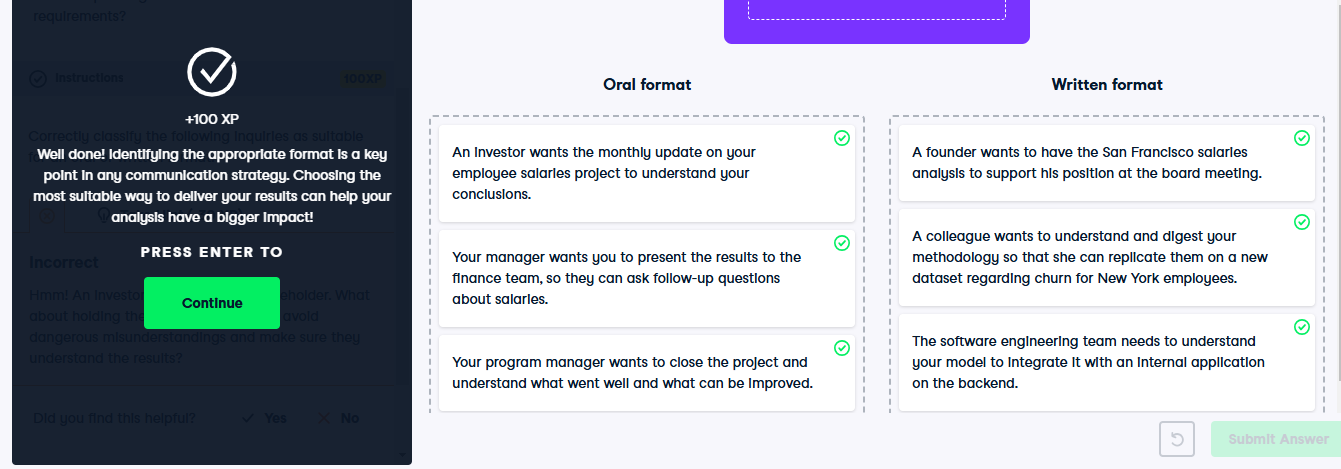
Let's look at an example. We are presenting our findings to the communicatb company CEO. She is interested in our conclusions and wants quick answers to her questions. She's rather busy, but happens to be visiting our office at the end of the week. This is a perfect occasion to set up a quick meeting and deliver our findings using an oral presentation.

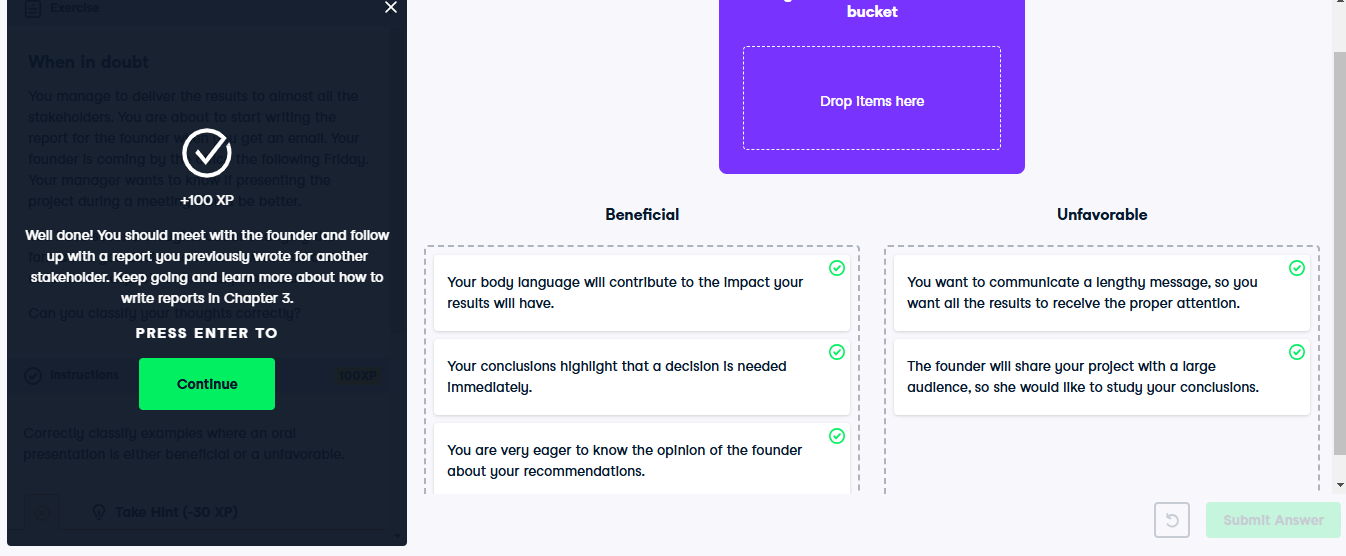
1. 1 Photo from Unsplash

**Appropriate format**

After the meeting, the CEO tells us important decisions will be based on our conclusions and reported to investors. Now, it's time to structure them in a written report, taking into account the questions and feedback we got from her.

1. 1 Photo from Unsplash



****