**Types of reports,**

Welcome to the third chapter! So far, we learned the importance of storytelling to impact the decision-making process, the steps on the data storytelling road and the elements to consider when choosing a suitable format to deliver our results.

In this chapter, we will learn how to structure a written report. We'll learn when to use several types of reports, how to achieve reproducibility in data science, and methods to write precise and clear reports.

**Written reports**

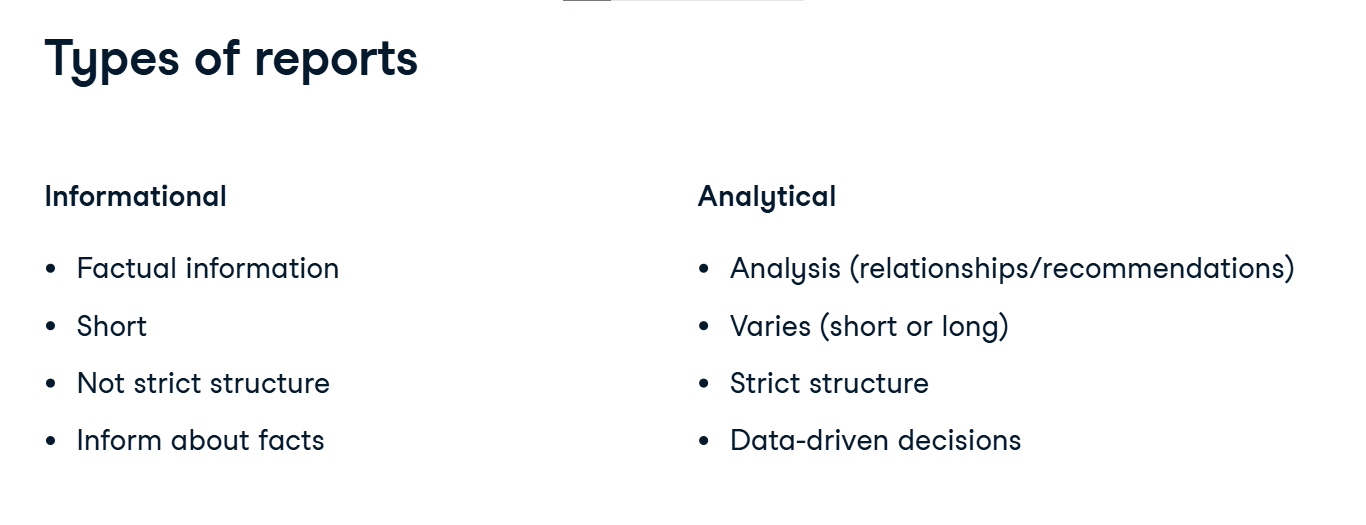
Imagine that at communicatb, we need to write a report explaining our project focused on sentiment analysis of product reviews.

**Written reports**

Our report should communicate our findings regarding the negative ratings due to delayed shippings, and the prediction accuracy our model displayed. We need to present our work according to the industry standards so our results are validated, and our recommendations drive the expected change.

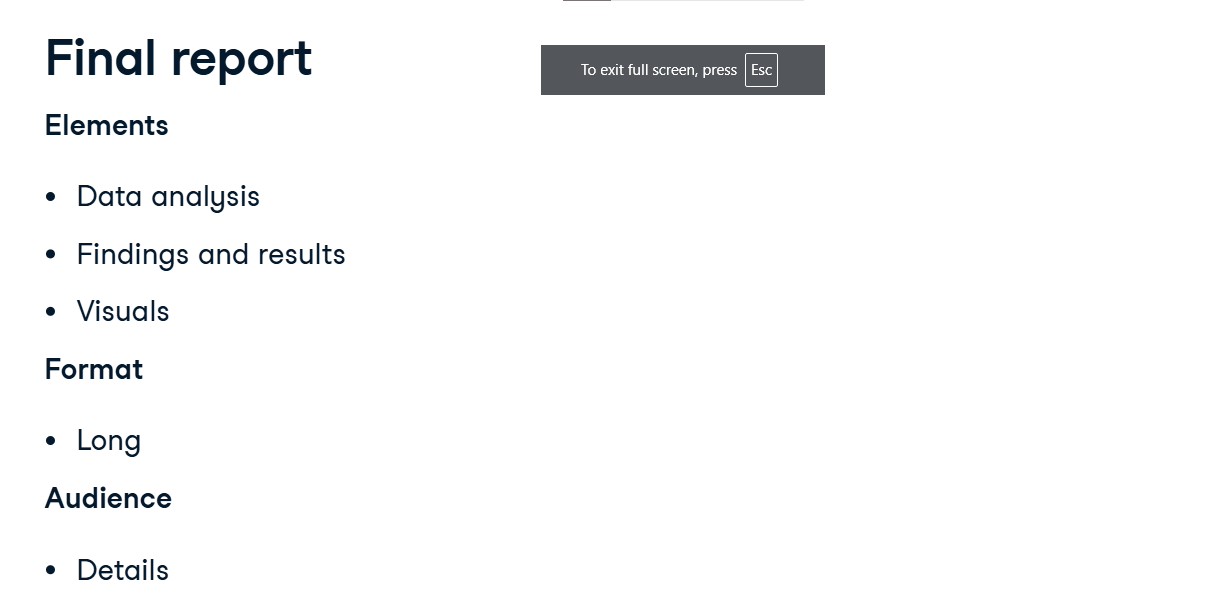
**Types of reports**

There are different types of reports. Informational reports provide factual information, are usually short, don't have a strict structure, and their purpose is to inform about facts without adding any analysis. Analytical reports provide analysis, and demonstrates relationships or recommendations. They can vary in size, but have a strict structure, and their purpose is to drive decision based on data.



**Final report**

Final reports include detailed data analyses, findings and results as well as visuals. They are usually long, because they are intended for audiences that need technical details.



**Summary report**

Summary reports include key findings and recommendations, as well as visuals. They are usually short, less than five pages long. Because they are a summary of a final report, they can include links to the main document. They are intended for decision-makers that do not need technical details.

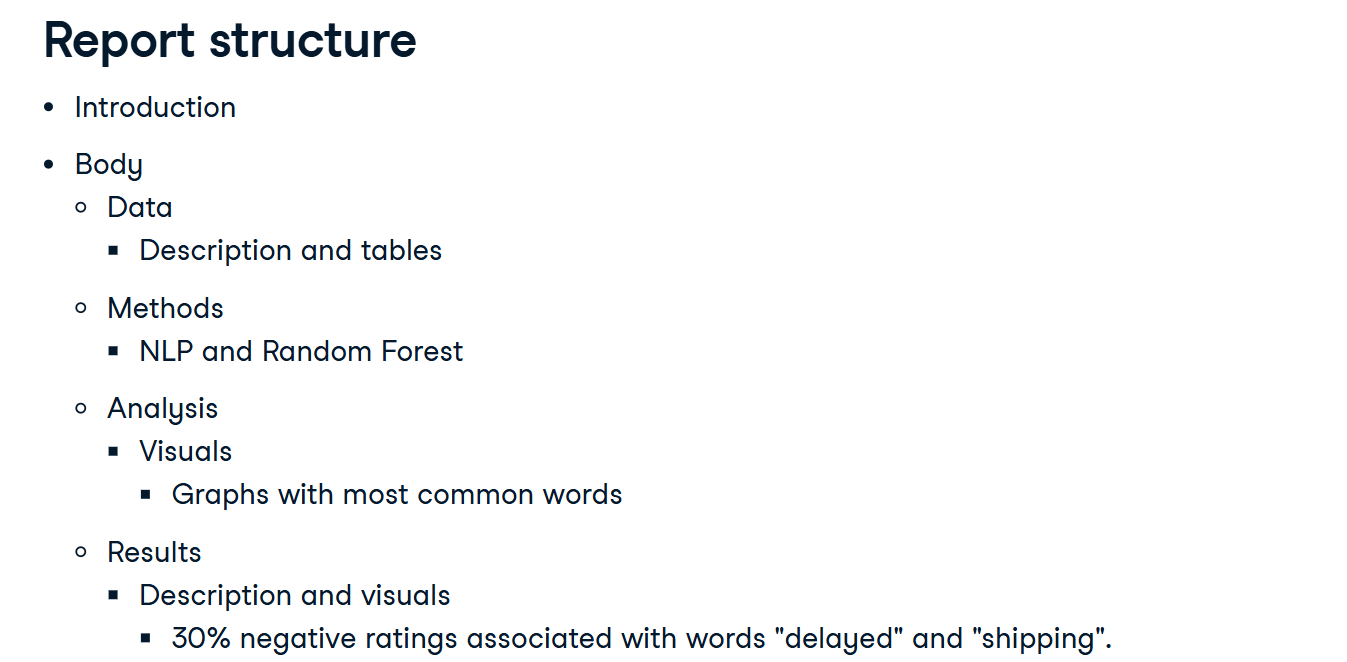


**Report structure**

As we mentioned informational reports do not follow a strict structure. But there is a straightforward structure we should follow for analytical, final and summary reports. The first section is the introduction. Here, we summarize the purpose of the report. In our example, the purpose of the report is to show the analysis on the product reviews, and the results of the rating predictions. After that, we need to include contextual information about why we performed the analysis shown. What motivated our example report is an increase in negative reviews. Finally, we should summarize our analysis questions. We ask which factors are driving the bad user experience.

**Report structure**

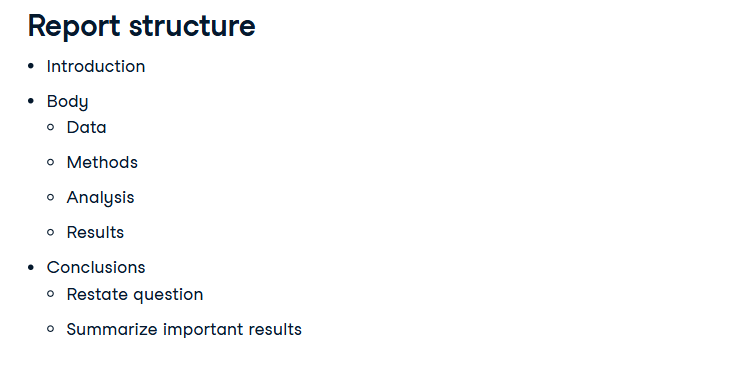
Then, we create the sections for the body of the report. In the data section, we write a description of the most relevant data used. We can include tables. In the Method section, we describe the methods used to gather and analyze the data, and build the model. In our case, we explain we used natural language processing and a random forest model. In the Analysis section, we include the selected data for the analysis and model using visuals, for example, a graph with the most common words. In the results section, we describe and explain the results of our analysis. We could also include visuals to evaluate them. For example, we state that 30% of the negative ratings were associated with the terms "delayed" and "shipping".



**Report structure**

Finally, we create a conclusion section. Here, we restate the analysis questions, as well as summarize the most important results from the analysis. This part is the appropriate location to add our recommendations for the next steps.

The structure we just saw is well-suited for data reports in journals. In a business context, your audience is different so you need to adapt. An efficient method is the 1-3-25 approach: 1 page of abstract, 3 pages max of executive summary and 25 pages max of details. The executive summary is enough for people to understand your conclusions and decide whether the report is worth reading in its entirety.

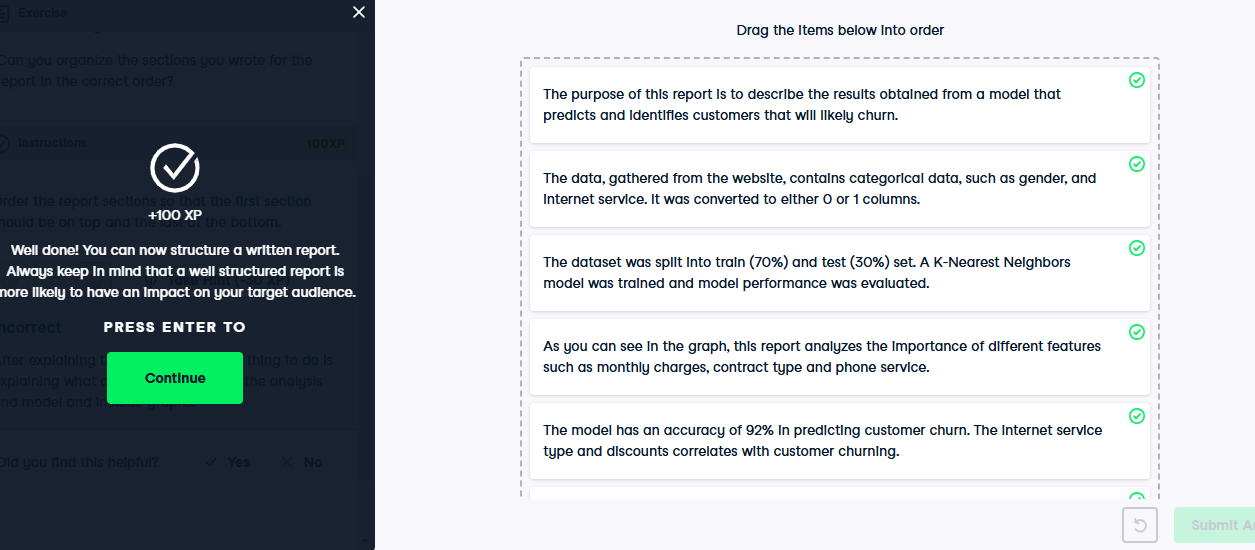


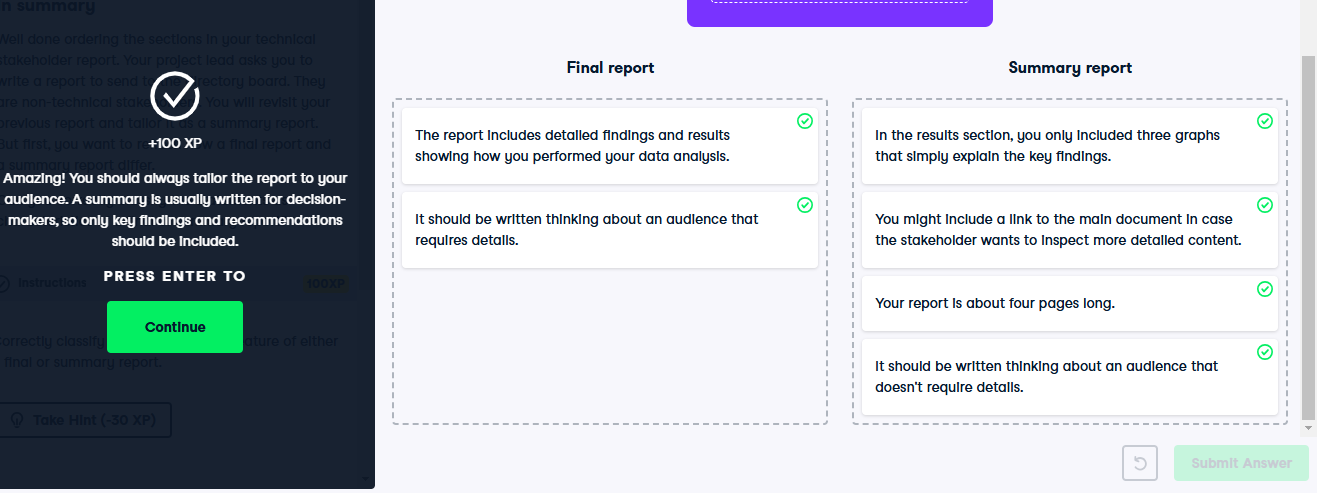
**Audience**

Again, we should keep our audience in mind. Each stakeholder is interested in different parts of our report. So we should tailor it accordingly. For example, people with little time but interested in the topic will read the introduction and conclusion and then scan the body for particular points.

The executive team will quickly read the introduction and the conclusion to find clues related to their main interest: our recommendations.

Finally, a technical stakeholder is mainly interested in the body of the report. They want to understand and validate our technical methods and analysis.





**Reproducibility and references**

Good job working on those reports! Understanding how to structure a report is tremendously important.

**Written report**

A fundamental part of communicating our findings, it making sure a report is clear and reproducible.

**Reproducibility example**

For example, say we have a friend that always cooks this amazing chocolate cake. We got at his place and follow his exact steps using his equipment: that's reproducibility. We trust he can actually bake this delicious cake. On a data project, if you run your Jupyter notebook again today and tomorrow, our results should be identical (provided the dataset hasn't changed).

**Replicability example**

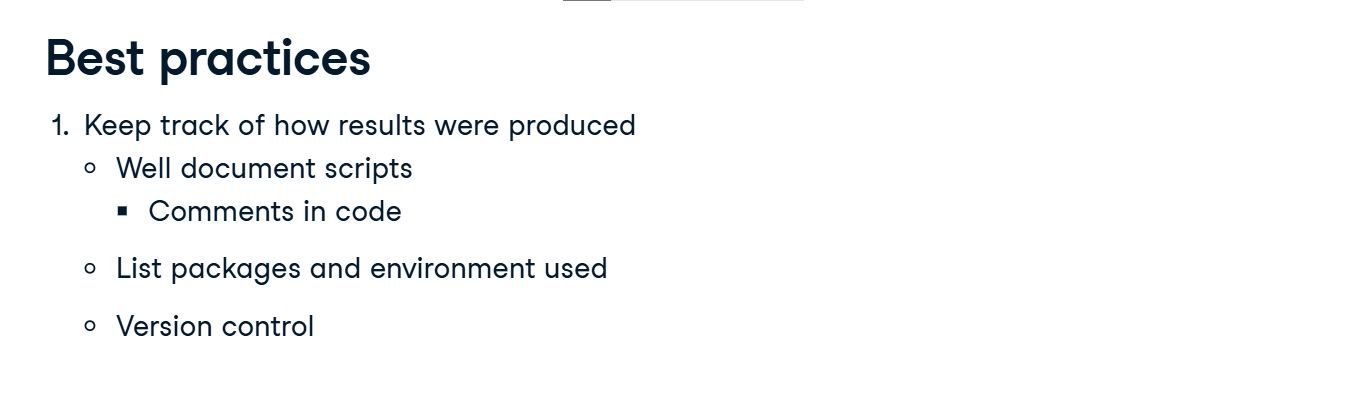
In baking replicability would be the ability to cook that cake again ourselves, following the recipe but using our own utensils and ingredients. In a data project, it would be the ability to obtain similar results using the same general approach, but a different environment (the data is processed in the same way, but using a different language on a different OS).

**Reproducibility and replicability virtues**

This scientific approach is crucial because it prevents duplication of effort and allows other data scientist to build upon preexisting work, enabling them to build upon previous work and focus on new challenges. Last but not least, it allows peer-reviewing. Although we tend to focus on coding examples in this course, you should always make sure that your code is easily reproducible and replicable, whether we use Excel spreadsheets, Tableau visualizations or Jupyter Notebooks. What matters is to keep track of how the results were produced.

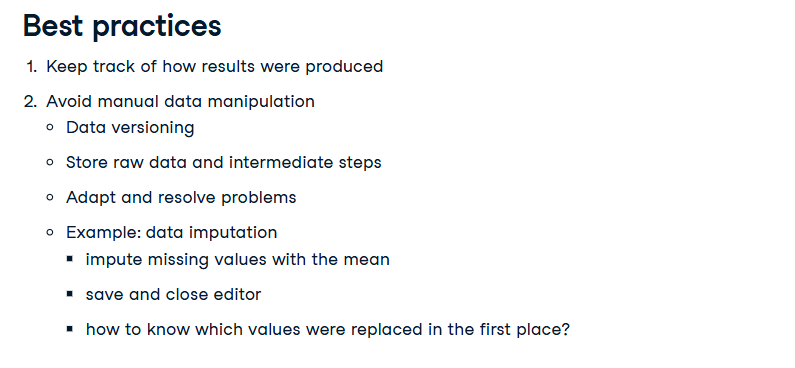
**Best practices**

We should document all the scripts we used to obtain our results using, for example, comments in our code, and list all the packages used in our environment. One helpful technique is to use a version control system (like Git), that tracks all the changes and versions of our scripts.



**Best practices**

We should avoid doing any manual manipulation of the data. We must never change the dataset directly in an editor. If possible, we should save all versions of our dataset. Also, we should save the raw data together with a script with intermediate steps. It can help us tell the story of our data transformation and create a narrative around it. Such a clear view ensures we know at any point what is happening with the data, and therefore can adapt and resolve problems. Take a data imputation example. There was a bug in the data pipeline, and you're told you can use the average values to impute missing values for a specific product. You go ahead and push the edits manually, then close your editor. Right after, your colleague informs you that actually, those products weren't available for sale on those dates, so the missing values should be zero. The data is already overwritten, so it's going to be hard to know which values were changed in the first place. Had we versioned the changes, it would be much simpler.



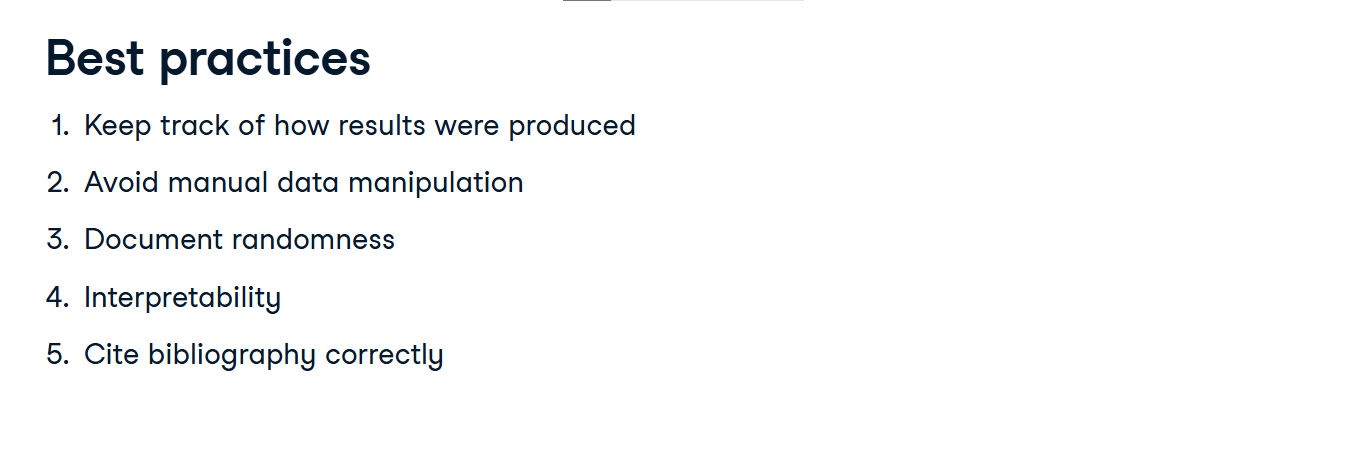
**Best practices**

We usually use machine learning algorithms or pipelines to create our workflow. Some of them might involve randomness techniques. We can usually set a random seed and introduce reproducibility into our model outputs. The random seed controls confounding variables: it lets us ensure a change was due to the model and not just randomness.

Interpretability is the degree to which a human can understand the cause of a decision or can consistently predict a model's result. Telling a story with a compelling narrative helps our stakeholders understand and interpret our findings. Because they are able to understand them, our conclusions can be reproduced.

1. 1 Molnar C. Interpretable Machine Learning. 2019.

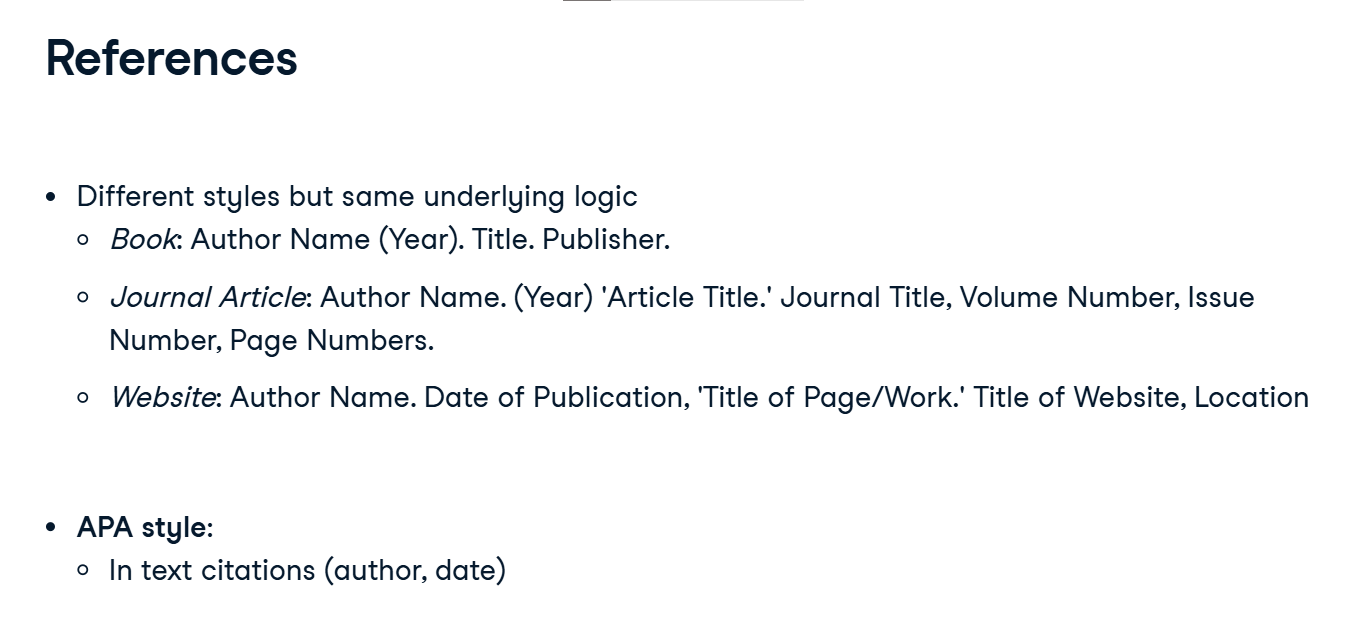
Lastly, we should always correctly cite other people's work used in our analysis.



**References**

Citations are the basic information required to identify and locate a specific publication

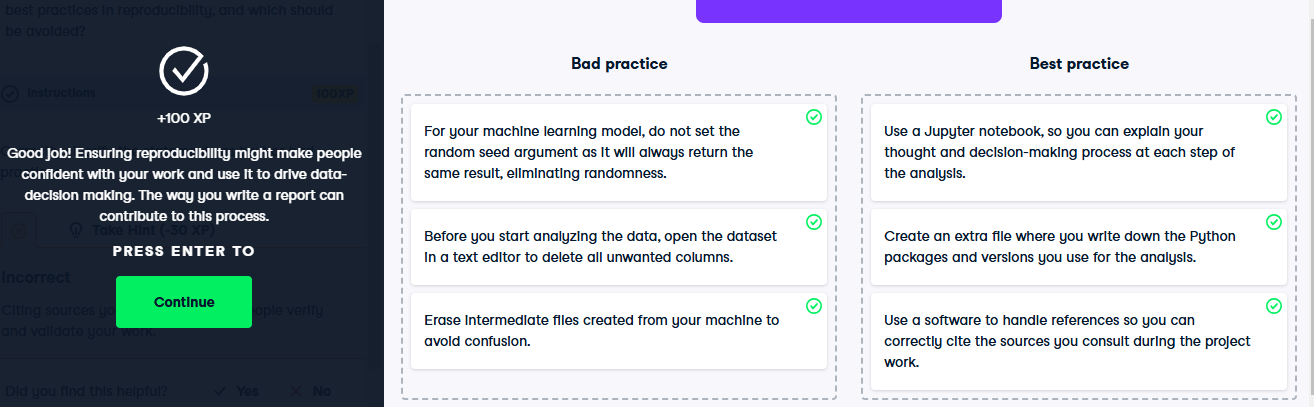
There are different styles, but they all have the same underlying logic, with slight variations. The most common style is APA, which uses in-text citations, putting the author name first and the date of publication next.



**Reference**

There are a number of reference management tools that can ease the burden of tracking all the citations. They help us automatically switch styles, and search online sources for references. EndNote, Mendeley, and RefWorks are some examples, but there are many more.

In a business context, these editions rules are much more relaxed. Most people will simply include a hyperlink to the source. What matters in the end is that the information is easy to retrieve, if the reader wants to refer to the sources material.



**Write precise and clear reports**

We now know how to make our work reproducible.

**Written report**

In order to make an impact at the decision level, other stakeholders should validate our work. So our report must not only be reproducible but also clear and intelligible.

**Write precise and clear reports**

In data science, writing should be concise and precise to avoid misleadings and confusion. Stakeholders should understand our message. In our reports, there is no place for sentences that add no information or introduce potential misunderstandings. Let's explore some options to help us with this task.

1. 1 Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

**Empty phrases**

Empty phrases basically don't add meaning or information. Phrases such as It is interesting to note that, The fact that, It should be pointed out that, it is well known that, or it is obvious that.

1. 1 Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

**Empty phrases**

Empty phrases are distracting and should be avoided: be straightforward and to the point.

1. 1 Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

**Empty phrases**

Here, both sentences convey the same message, but the first one is straight to the point, while the second one is harder to follow. Basically, if it doesn't add information, remove it.

1. 1 Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

**Concrete nouns**

Good technical writing is concise and direct. So we should write concrete nouns, and avoid the overuse of the pronouns this or that or it, as it can be unclear what exactly they refer to. Also, keeping track of what the pronouns mean adds some cognitive load and distracts the readers from the actual insights

1. 1 Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

**Concrete nouns**

In the example sentence 'this shows an accuracy of 80 percent when predicting customer churn', the word this should be replaced by a concrete noun and state that the model shows an accuracy of 80% when predicting customer churn.

1. 1 Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

**More pronouns**

The active voice is criticized for placing the emphasis on the author instead of the facts. The passive voice is criticized for being too stuffy and harder to read. If the passive voice is preferred in an academic setting, in a business context your message will probably land better using the active voice.

1. 1 Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

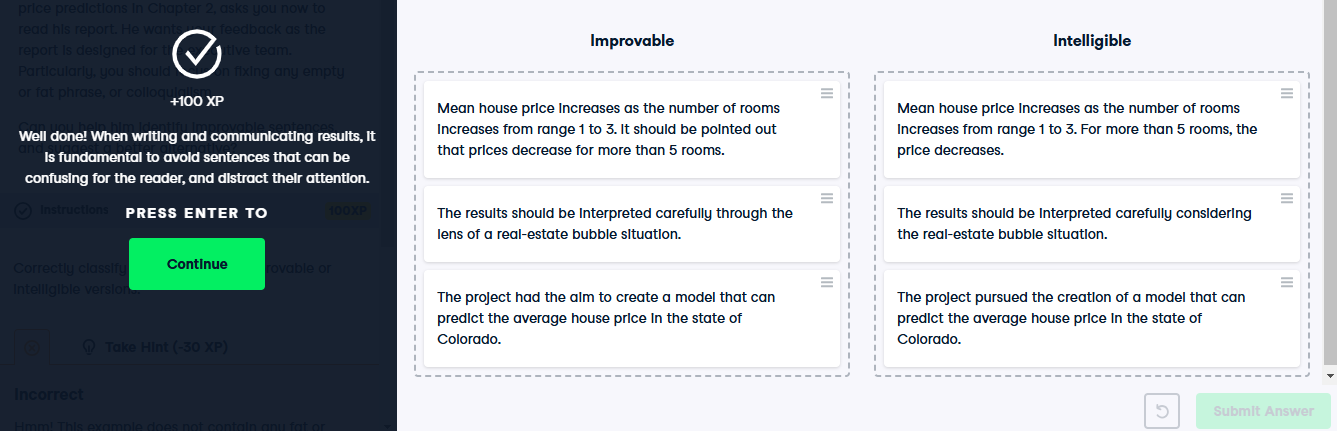
**Redundant adjectives and adverbs**

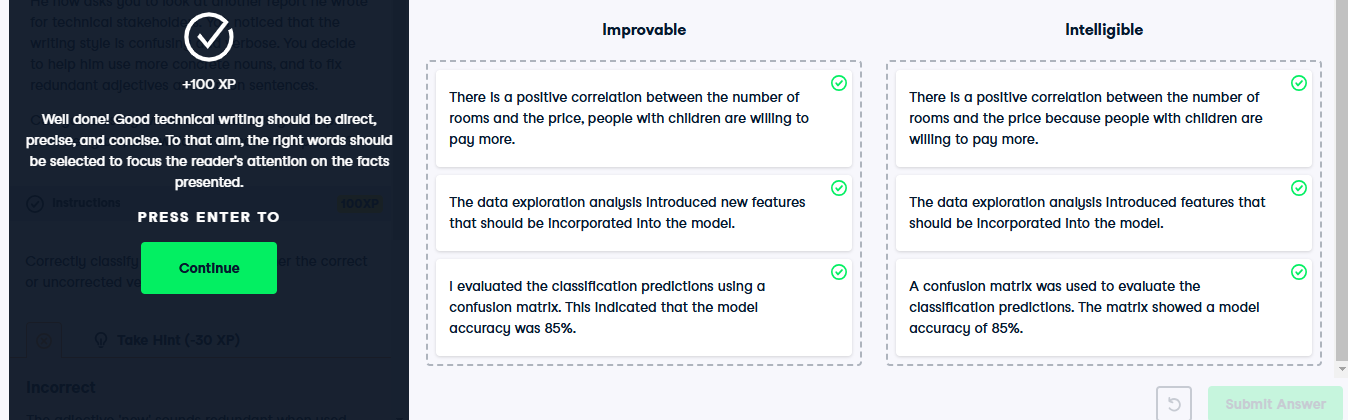
When trying to emphasize an argument, it is easy to end up using redundant adjectives and adverbs. Some examples are introducing the new, or done previously. In order to keep the report concise, we should eliminate these redundant adjectives or adverbs.

1. 1 Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

**Run-on sentences**

Lastly, run-on sentences are two or more independent sentences or clauses that are connected or punctuated incorrectly. For example, the sentence in the slide is run-on. It is composed of two full sentences but connected by a comma. To fix these sentences, we should either make two separate clauses or use dependent clauses by introducing words like because or so.





**.Case study: report on credit risk**

Good job! Let's go through a small case study now.

**Credit risk**

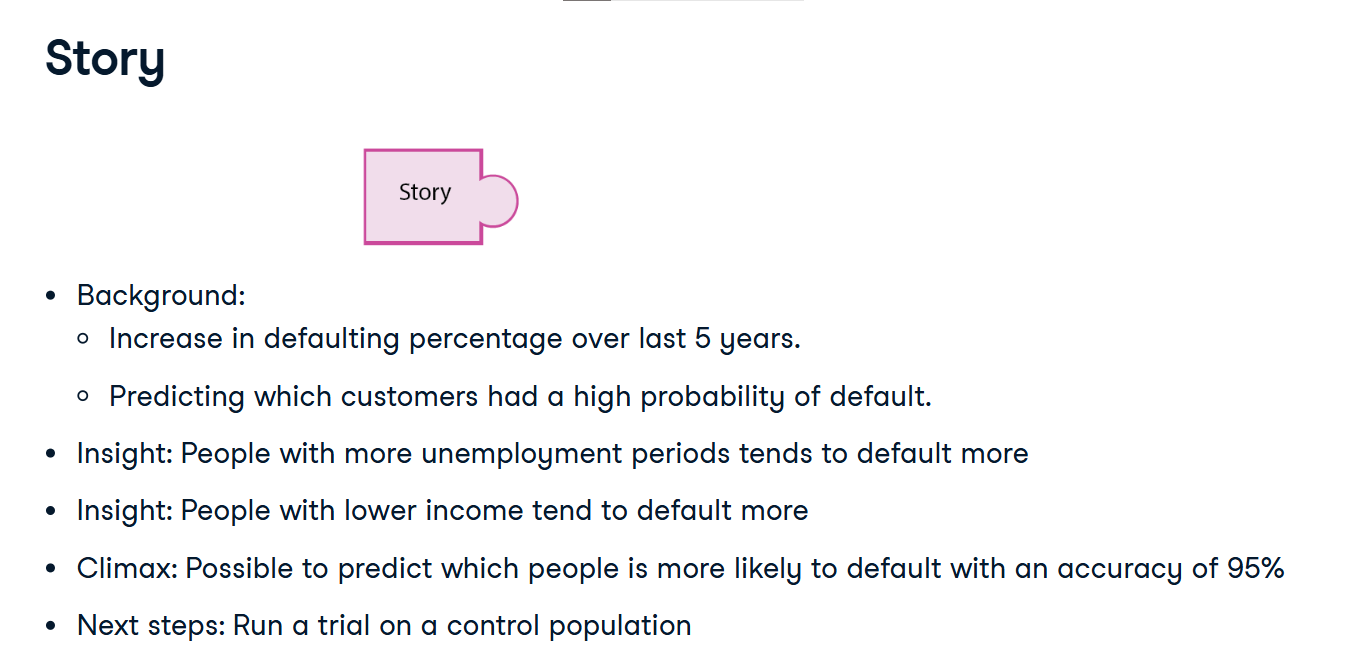
We're going to decide step by step how to write a report on a credit risk case. So let's present the case. Credit risk quantifies the probability that a customer fails to meet contractual obligations, such as credit card debts, and other loans. The finance team of a fictional Bank, Loanme, is interested in understanding and predicting which customer taking on a loan is more likely to default. They send us data available on loan details and borrowers, including age, income, loan amount, that we use to perform data exploration, and to train and evaluate a predictive model. Now it's time to show our results.

**Audience**

We are going to communicate our findings to non-technical stakeholders, mostly to decision-makers.

**Story**

So first, let's find our story, and especially our narrative. There was an increase in the percentage of defaulting customers over the last 5 years. So the bank became interested in predicting which customers had a high probability of default. We analyzed the data and saw that people with more unemployment periods tend to default more. Also, our analysis showed that younger people with less income tend to default more. After training our model, we saw that it is possible to predict which people are more likely to default with an accuracy of 95%. To confirm these results, the next step should be to run a trial on a controlled population.

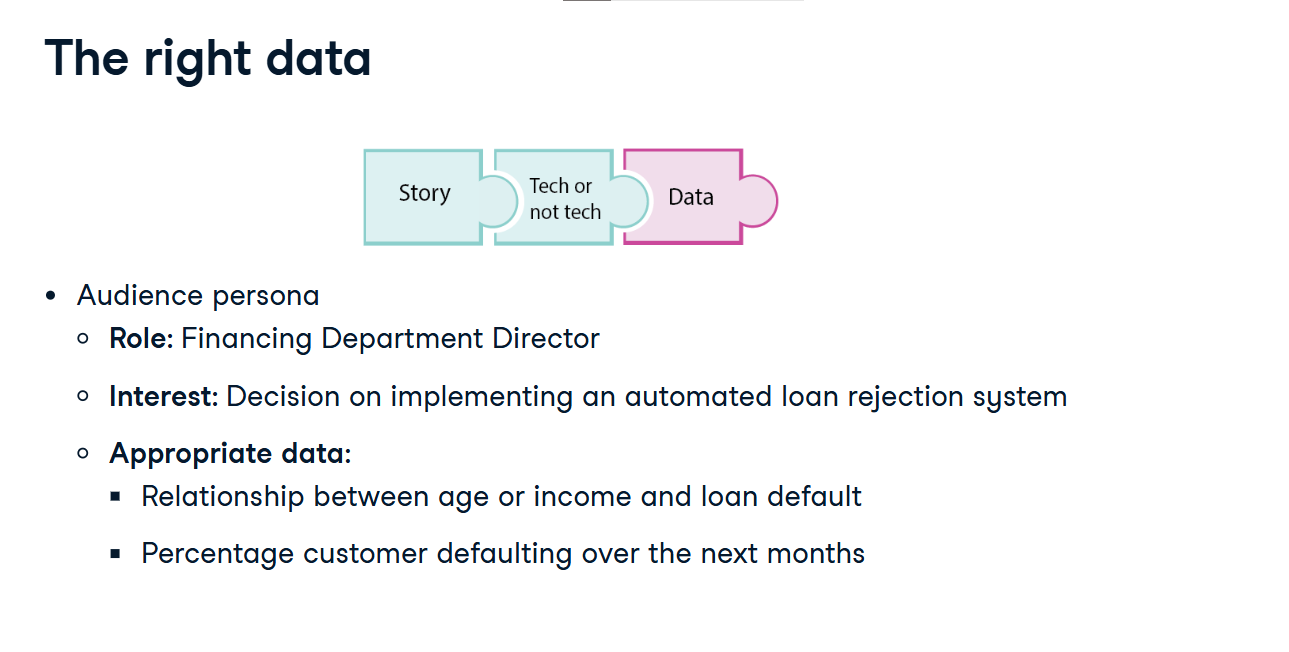


**Tech or non-tech**

We have our story. Now we need to translate technical results for non-technical stakeholders.

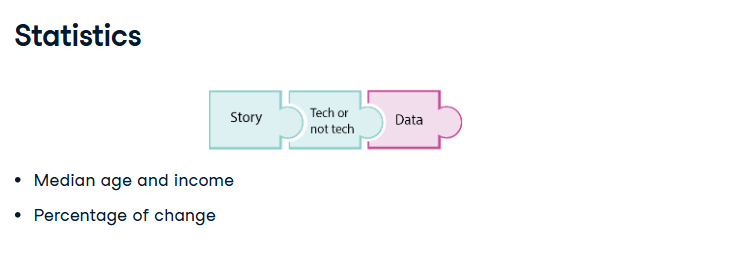
**The right data**

Let's define an audience persona. We report to the finance department director, who needs to decide on implementing an automated loan rejection system using our project findings. So we report the relationship between age or income and loan default and forecast percentage of customer defaulting over the next months.



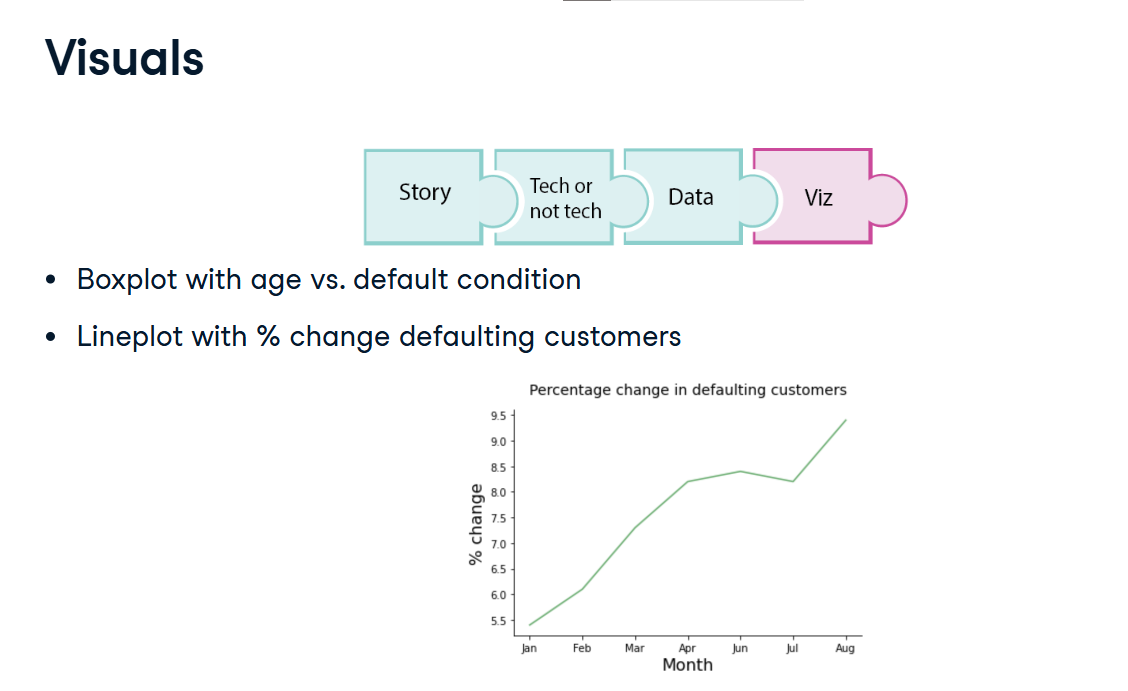
**Statistics**

Because we need to summarize numerical data into an aggregate, we show the median age and income for default versus non default customers. Also, we need to show how the number of customer defaulting changes over time, so we show the percentage of change.



**Visuals**

Accordingly, we include a boxplot showing age or income vs. default condition,and a lineplot with the percentage of change in defaulting customers over the next months.



**Correct format**

We now have all our pieces together. Let's recap. We are reporting to the Financial Department director, because she has an important decision ahead. She is interested in our key findings and recommendations, and we should communicate the findings through email before our meeting.

**Report**

All of these elements, lead us to select a written report to deliver our results. But should we select a summary or a final report?

It's a non-technical stakeholder, so a summary report should be better. Should it be an informational or an analytical report?

**Report**

We are presenting not only facts but also our analysis. So the most suitable is the analytical report.

**Summary report structure**

So how are structuring the summary report? Let's see. In the introduction, we first summarize the purpose of our report, we add contextual information about the reason of our project, and state our analysis question. Then, in the body, we are going to describe the data, and include only key findings in the result section. Lastly, we restate the questions linking it to the central insight and we add our recommendations.

