How to Structure your Data Science Notebook to be Easy to Follow

Clear steps to create an organized notebook, including examples

Ryan Carters Dec 22, 2021 · 5 min read ★

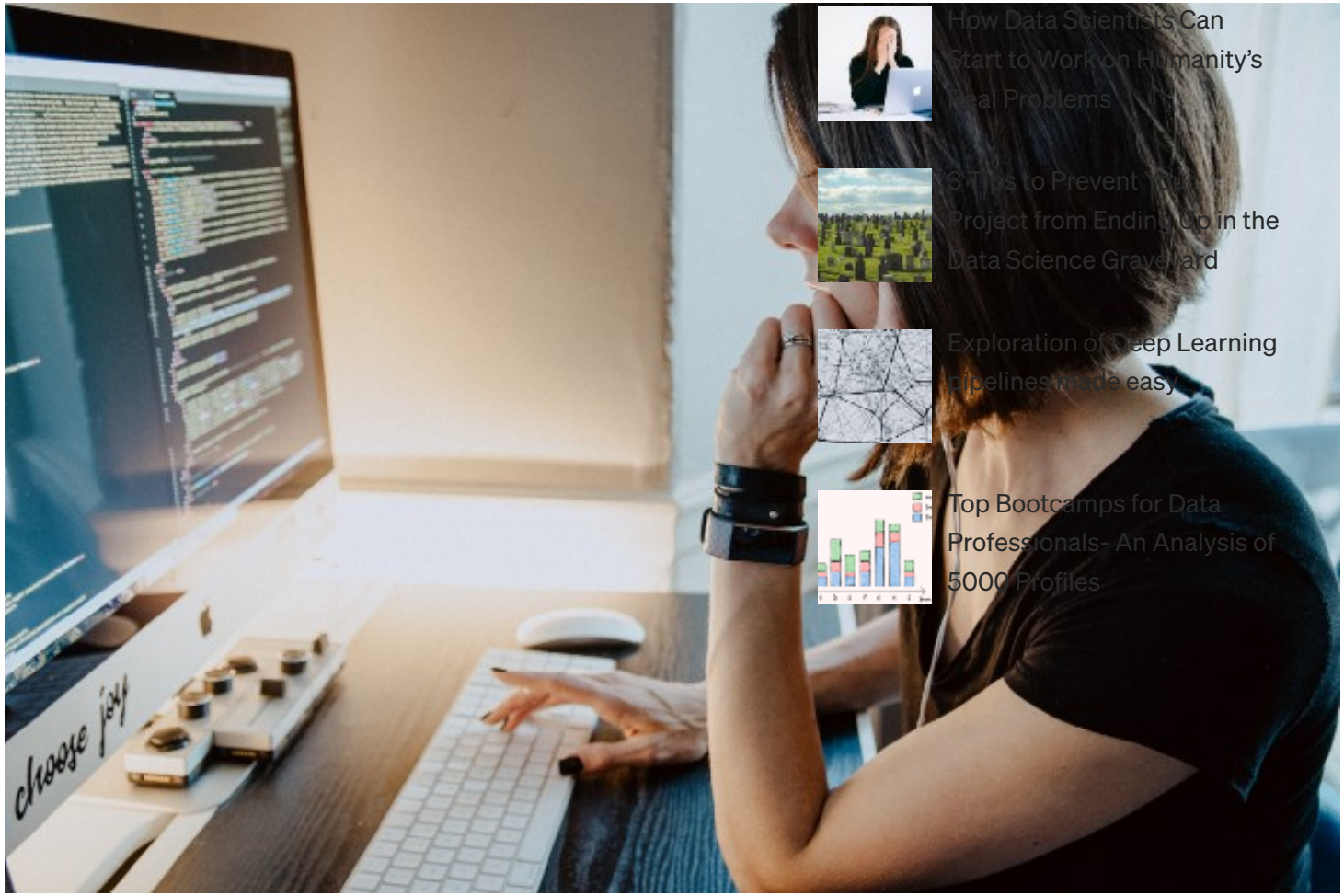

Photo by Kelly Sikkema on Unsplash

The structure of your notebook matters more than you think. The reader needs to get a clear overview of your work before understanding the details.

To achieve that, your notebook needs an organized structure with intuitive

sequential sections. Perhaps someone is only interested in a certain part of your data solution. So the main parts of your work need to be easy to identify.

But we don't need to make things unnecessarily complex. So going straight to the point, in general, we can simply use the following five sections:

1. Introduction
2. Data Wrangling
3. Exploratory Data Analysis
4. Modeling (optional)
5. Conclusion

Below we detail what must be included in each of these sections, describing all of the **necessary subsections**.

1. Introduction

1. Introduction

First, we have to describe the project in terms of **business goals**. Start by giving the **context** of your work, where it originated, and what you want to achieve: the main goals.

For example, say you're investigating a dataset containing sales data from your marketing company. The first part of the introduction would talk about where the sales data came from and state which questions you want to answer from the data — e.g. what aspects of your product (color, size, brand, etc.) is more correlated with sales.

> A good rule of thumb is to explicitly create a list of questions to address.

Still related to the goal of your project, the **product** from this notebook alone needs to be specified. For example, the end goal of the analysis is to create a monthly report with insights? or just a one-time paragraph with a plot to send via e-mail to the CEO to prove a point? This should be clear in the introduction.

Finally, it is also important to briefly talk about any **prior knowledge**. For example, if the sales data is only from one specific store, that should be mentioned. If at a certain period of time the company had problems with some of the products, like distribution issues, directly affecting sales, this should be stated too. Basically, we have to describe anything that helps **understand the context** of the data sources and important details.

2. Data Wrangling

This section starts with a brief description of the dataset we will analyze. Basically, that means we have to **list the column names** in each table, together with **their meaning**. If we are dealing with multiple tables, we should also describe how the tables relate to each other.

After that, we start coding, with the first step being to load the data. It is also important here to check for **cleanliness**, and completely clean your dataset for the analysis.

> Any modification in the dataset should be in this section.

The following sections will simply **look** at the data for exploration. But the data wrangling section is where you adjust the data for the dimensions you want to explore.

It is also important to document the steps taken to clean the data, **justifying every decision**. For example, if you removed observations with missing data on the column “age”, you need to clearly explain why.

Since datasets usually have many wrangling tasks, a good practice is to **list all of the modifications** performed in this section in a markdown cell.

3. Exploratory Data Analysis

This section should be guided by the questions you want to answer. In other words, you will **create one subsection for each question raised** in the introduction.

So with the data already cleaned, here you will basically **compute statistics** and **create visualizations** with the goal of answering each question.

Remember to always add comments throughout the exploration. After every plot or analysis decision, you need to **explain what was the outcome**, and the observations you got from the result.

> Lead the reader through your thought process!

Make sure to investigate each question from multiple angles, discussing aspects that can make the results more interesting. For example, instead of just adding distribution plots and describing them, think about **creating plots that support your answers**.

4. Modeling (optional)

Not every data science work includes modeling. Many business questions can be answered with just exploratory data analysis. For example, well-designed hypothesis tests can give you many interesting insights into your data.

> *More importantly, it is **not** recommended to perform modeling in a notebook.*

Instead of using notebooks, modeling should be done with modular code, directly using python scripts, with collaborative work, testing suites, etc. Nonetheless, notebooks might be used for quick modeling experimentation, such as creating simple baseline models.

So if you do include modeling, make sure to first and foremost separate your dataset into **"train" and "test" samples**. Then, work with the "train" part until a given metric is satisfied, say 85% accuracy using cross-validation. Once you're done **comparing models and performing tuning**,

you assess your final model on the “test” sample.

So the main parts to document here are: how you split the data into train and test, which algorithms and hyperparameters were used, and **how you assess that your model is indeed useful**. For example, sometimes “precision” makes more sense to tune compared to “recall”. This should be stated in the introduction and tuned here.

5. Conclusion

In the conclusion, we **summarize our findings** and the results that have been obtained regarding the questions raised in the introduction.

Moreover, we should point out where **additional research can be done** or where **additional information** could be useful. For example, the analysis could point out that interesting data for discount offers would be how long the user interacted with a product’s page. So the analyst would suggest the company add such a feature to every product’s sales page.

Finally, another very important point is to make sure that you are clear with regard to the **limitations of your project**. No project is perfect, so at least one limitation should be described in a subsection here.

> Force yourself to list at least one limitation of your work.

For example: are there any missing observations in the data — i.e., other information that is not in this dataset? Can we fully trust the statistical tests performed to verify the hypothesis mentioned in the project? Why?

I hope you enjoyed this high-level view of how to structure your data science notebook. Following the aspects mentioned above, here is a concrete example of subsections that you can include in the five sections described:

1. **Introduction**: Context, Business goals, List of questions

2. **Data Wrangling**: List of cleaning steps, Step 1, Step 2, …
3. **Exploratory Data Analysis**: Question 1, Question 2, …
4. **Modeling (optional)**: Splitting data, Training, Testing
5. **Conclusion**: Limitations, Future Work

Remember to fully use markdown cells and all of the styling resources available for markdown while describing the work and you'll do great!

Finally, creating internal links and setting up a table of contents is also useful. Here's an example:

```
# Table of Contents
<ul>
<li><a href="#introduction">Introduction</a></li>
<li><a href="#data_wrangling">Data Wrangling</a></li>
<li><a href="#exploratory">Exploratory Data Analysis</a></li>
<li><a href="#modeling">Modeling (Optional)</a></li>
<li><a href="#conclusion">Conclusion</a></li>
</ul>
```

```
# Introduction <a id='introduction'></a>

...
```

Table of contents using Markdown cells and HTML in a Notebook. Image by the author.

If you enjoy reading stories like these and want to support me as a writer, consider signing up to become a Medium member. It's $5 a month, giving you unlimited access to stories on Medium. If you sign up using my link, I'll earn a small commission.

Data Science Machine Learning Artificial Intelligence Jupyter Notebook Reproducibility