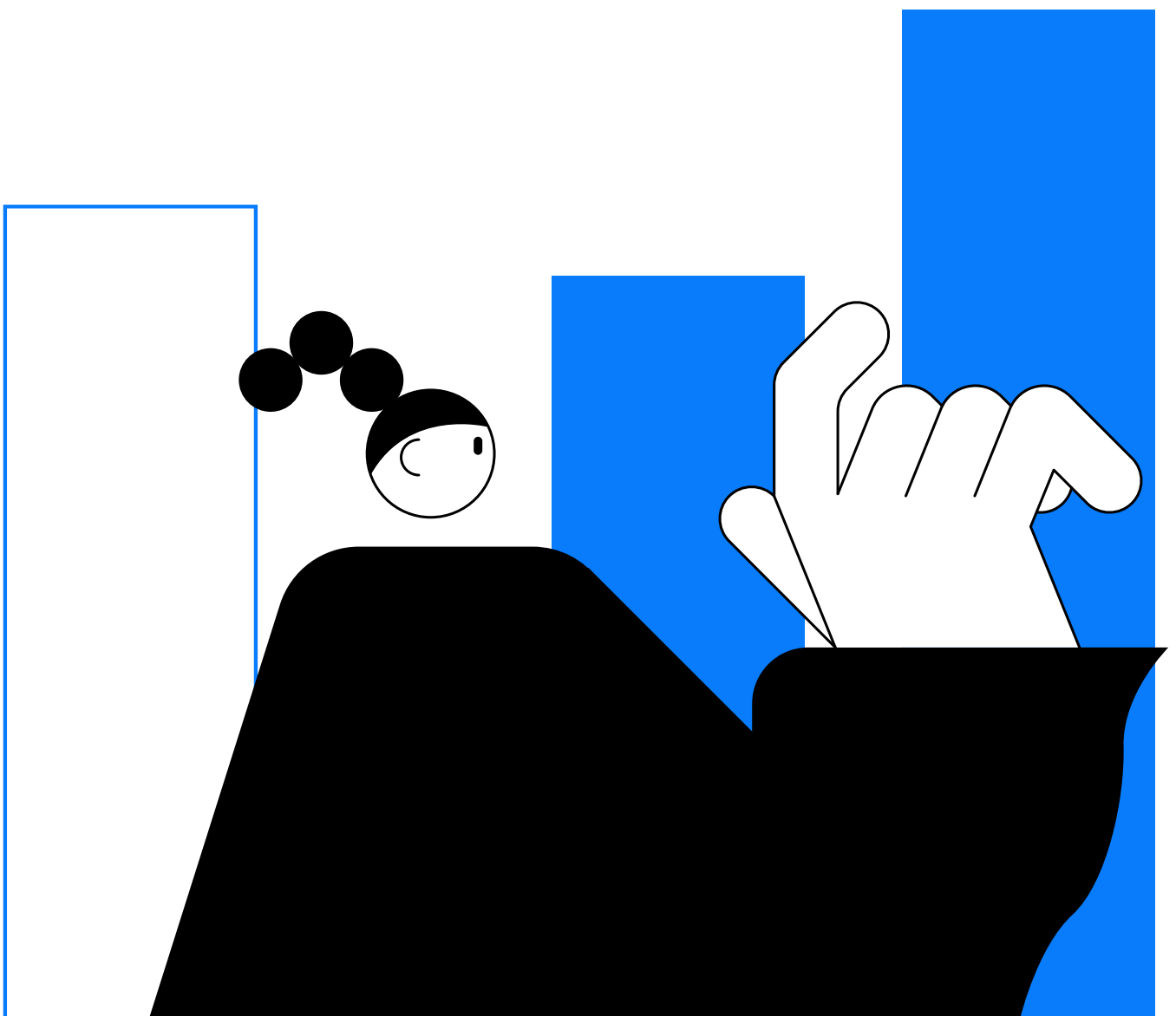


5 Ways To Make Your **Data Storytelling** More Robust With Datalore



Who is this whitepaper for?

This whitepaper could be relevant for:

- Data science and data analytics managers in small and mid-size teams who want to upskill and educate their teams in data storytelling.
- Data scientists, data analysts, BI analysts, product analysts, and sales and marketing analysts who want to improve their data storytelling skills.



Introduction

Data storytelling is the process of communicating insights, answering questions, and providing context through data to your audience in an understandable way. It's not just about telling a story; it's about creating a coherent narrative that provides data touchpoints and makes it easy for stakeholders to understand the insights and conclusions.

Historically, data storytelling relied on static dashboards or presentations, which offered limited flexibility, interactivity, and narrative capabilities, especially in complex analytics projects with huge amounts of messy data from different sources. While tools like Tableau offer interactive dashboards, they are cumbersome for presenting and understanding narratives longer than one page and require extra data preprocessing steps in other tools.

Datalore, a collaborative data science platform from JetBrains, provides a new way of approaching data storytelling. Datalore allows users to combine code, data, and narrative text in a single report, making communicating insights and telling stories about data easier.

In this whitepaper, we will explore five ways to more robust your data storytelling with **Datalore**.



“Datalore allows us to do data storytelling very well, since we have one place where we pull the data, do complex manipulations with Python, create visualizations, and export the results into a format which is friendly for business consumers.”

Moreno Raimondo Vendra

Senior Machine Learning engineer at TrueLayer

Table of contents

1. Define the Storyline and Audience

2. Choose the Right Toolkit for Data Preparation and Visualization

Data retrieval

Data quality

Data visualization

3. Address Reproducibility and Ad Hoc Changes

Reproducibility

Ad hoc changes

4. Make your story stakeholder-friendly

5. Mind Automation and Building Up On Your Work

Scheduling analysis updates

Letting stakeholders deep dive into your story

Conclusions

Next steps

1. Define the Storyline and Audience

Defining your narrative and audience is the first step in creating a compelling data story. Before you start your analysis, take the time to think about what story you want to tell, who your audience is, and what insights you want them to take away from your analysis. Consider the following steps:

- **Start with a clear objective:** Identify the question you are trying to answer or the insights you want to communicate.
- **Define your audience:** Consider who you are communicating with, their background and knowledge, and what information will be most relevant to them.
- **Identify the key points:** Break down your analysis into key points or takeaways. These should be the most important insights you want your audience to remember. You can even create a template of the story structure and reuse it across multiple projects.
- **Create a table of contents for longer narratives:** Use your key points to create a clear and concise table that outlines your narrative. For short dashboards or one-page stories, this might be unnecessary.
- **Explicitly mention stakeholders:** Ensure that your stakeholders are explicitly mentioned in your narrative and that their questions or concerns are addressed in your analysis. This can help to build trust and engagement, as stakeholders are more likely to engage with data that is relevant and understandable to them.

By taking these steps, you can ensure that your **analysis is focused, relevant, and tailored to your audience**. This helps you create a narrative that is engaging and impactful, providing real value to your stakeholders.

The screenshot displays a data analysis tool interface. On the left, a 'Table of contents' sidebar lists items: 'Sales report for different buyer personas', 'Difference in buying behavior', 'Overview of price distribution', 'Sales for different persona segments', and 'Major takeaways'. The main area shows a report titled 'Monthly Sales Analysis November' with a subtitle 'Sales report for different buyer personas'. The report text states: 'We want to find out if there is a difference in buying behavior across geographic regions in different product categories. This report is designed for regional sales managers, sales operations managers and CRM marketing leads. What is inside: Difference in buying behavior, Overview of price distribution, Sales for different persona segments, Major takeaways'. Below the report is a SQL query:

```
[1] select "City", "Product line", ROUND(SUM("Total")) as "Total", ROUND(AVG("Total")) as "Average" from SCREENCAS group by "Product line", "City" HAVING "City" in ('Mandalay', 'Yangon', 'Naypyitaw') order by 'Total'
```

. At the bottom, a table shows the results of the query with columns for City, Product line, Total, and Average. The table data is as follows:

	City	Product line	Total	Average
0	Yangon	Health and beauty	75587.0	525.0
1	Naypyitaw	Electronic accessories	113814.0	677.0
2	Yangon	Home and Lifestyle	134503.0	679.0
3	Yangon	Sports and travel	116237.0	646.0
4	Yangon	Electronic accessories	109902.0	601.0
5	Naypyitaw	Home and Lifestyle	83375.0	604.0

2. Choose the Proper Toolkit for Data Preparation and Visualization

Data retrieval

Data Storytelling might require **dozens of iterative calculations**. Retrieving and analyzing your data in separate places is prone to make you lose track of the exact data used in your work, as well as making it harder to update your work with fresh data. Therefore, keeping your data and analysis in the same place is essential for reproducibility.

Datalore offers integrations with databases and S3 buckets right from the user interface, allowing you to easily access your data within your Jupyter notebook. This means that you can combine native SQL and Python code in one notebook, get access to coding assistance and seamlessly transition from querying with SQL to exploration with Python.

The screenshot displays a Datalore notebook titled "Monthly Sales Analysis November". The interface includes a left sidebar with "Attached data" (Notebook files, S3 Datasource, Snowflake database, Google BigQuery database) and a main workspace with a code editor and a data table.

```
select "City", "Pr", ROUND(SUM("Total")) as "Total", ROUND(AVG("Total")) as "Average" from SCREENCAST
group by "Product Line"
HAVING "City" in ( PREVIOUS_DAY(timestamp:TIMESTAMP, dow:VAR...
order by 'Total'
```

	City	Product Line	Total	Average
0				525.0
1	Nay			677.0
2	Yangon	Home and Lifestyle	134503.0	679.0
3	Yangon	Sports and travel	116237.0	646.0
4	Yangon	Electronic accessories	109992.0	601.0
5	Naypyitaw	Home and Lifestyle	83375.0	604.0
6	Mandalay	Food and beverages	91290.0	597.0
7	Mandalay	Fashion accessories	98479.0	521.0
8	Mandalay	Electronic accessories	102307.0	609.0

Below the table, there is a text block: "We want to find out if there is a difference in buying behavior across geographic regions in different product categories. This report is designed for regional sales managers, sales operations managers and CRM marketing leads."

```
select * from SCREENCAST
```

The bottom of the notebook shows a table with columns: ID, Invoice ID, Branch, City, Customer, and Gender.

2. Choose the Proper Toolkit for Data Preparation and Visualization

Data quality

Before starting any data analysis or visualization, it's important to ensure that your dataset is of high quality and ready for analysis. This includes identifying any potential data quality issues and taking steps to address them.

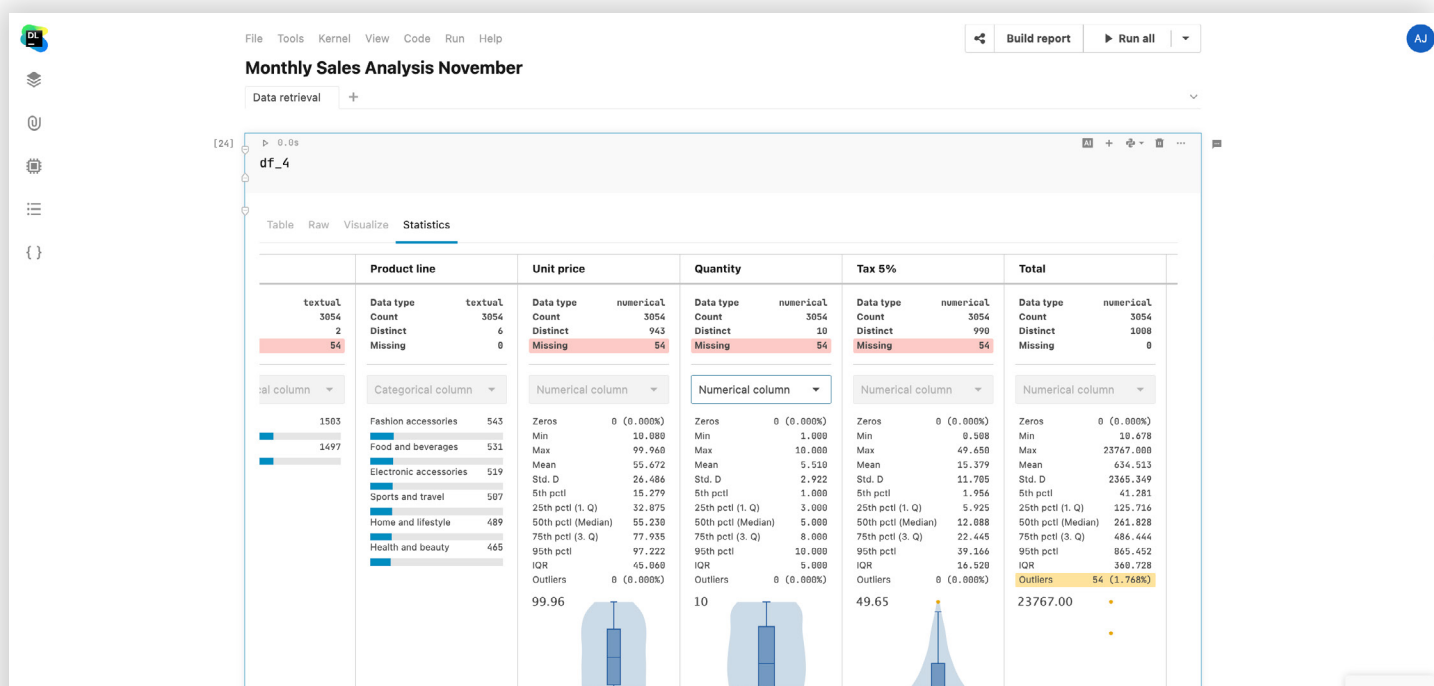
One common issue is **missing data**. It's important to understand the reason for missing data and decide on the best approach to handle it. You may choose to remove the rows with missing data, impute the missing values with the mean or median, or use a more advanced imputation method. **Datalore's statistics tab** lets you quickly see missing values in your dataset, making it easier to identify and handle missing data. **It is always a good idea to leave a comment in your data storytelling about what strategy for dealing with missing data.**

Another issue is **data inconsistency or errors**. This can include spelling mistakes, incorrect data types, or data that falls outside of expected ranges. It's important to carefully review your dataset and correct errors before starting your analysis.

To find out more about other potential data pitfalls, read [this article](#).

Once your dataset is cleaned and prepared, it's time to consider **the slice of data** you will use for your analysis. This involves deciding which variables to include in your analysis and selecting a time period, geographical region, or other relevant parameters.

By carefully selecting and preparing your data, you can ensure that your analysis is accurate and reliable. This will also make it easier to identify meaningful insights and communicate them effectively to your stakeholders.



2. Choose the Proper Toolkit for Data Preparation and Visualization

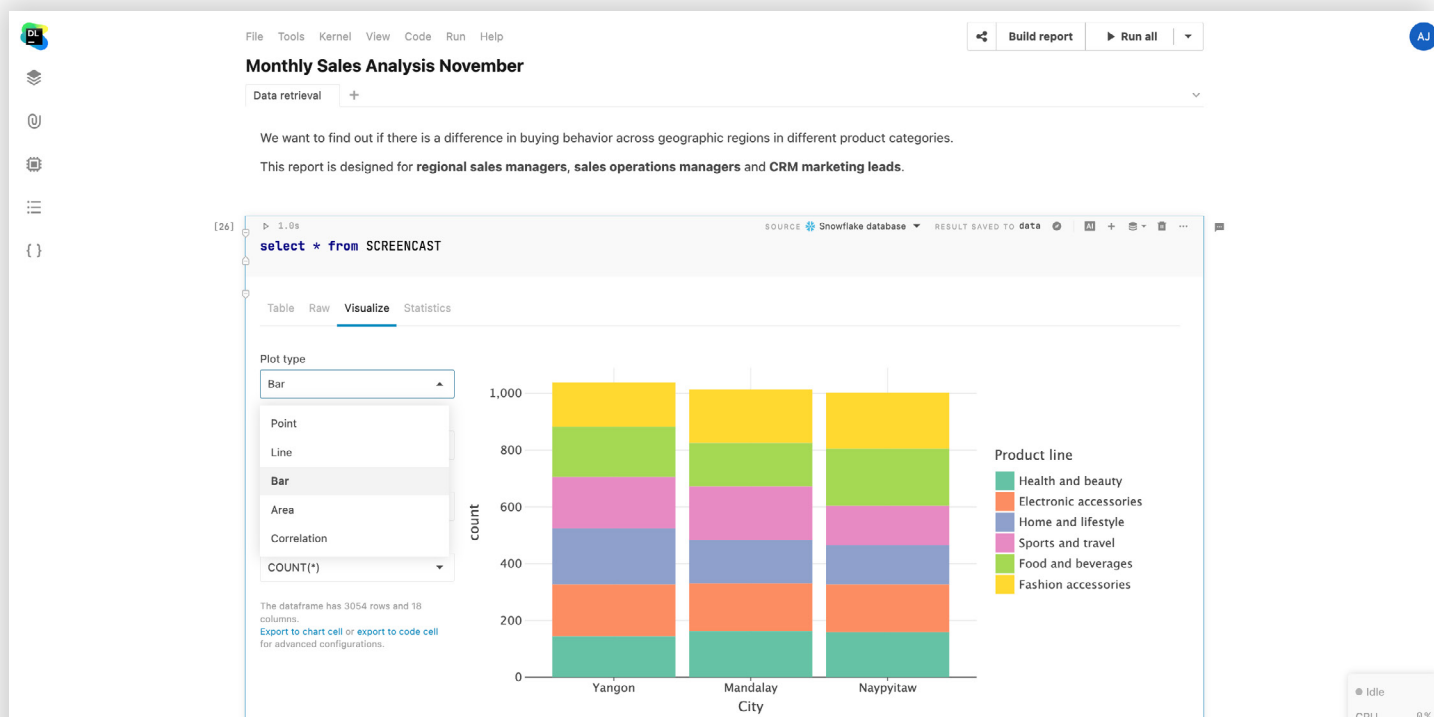
Data visualization

When it comes to data visualization, there are several types of plots that you can choose from, each with its own strengths and weaknesses. Here are some common types of plots and their use cases:

- **Bar charts** are used to compare the values of different categories or groups. They are often used to show Sales data distribution or compare different groups or categories.
- **Line charts** are used to show trends over time. They are often used to show historical changes in data or to compare trends between different groups.
- **Scatter plots** are used to show the relationship between two variables. They are often used to identify correlations between variables or to find data patterns.
- **Heatmaps** are used to show the distribution of data across a two-dimensional space. They are often used to show data density or to identify clusters or patterns.
- **Box plots** are used to show the distribution of data and to identify outliers. They are often used to compare data distribution between different groups or categories.

Once you have chosen the correct type of plot for your data, it's important to think about how to make it interactive. Interactive visualizations allow your stakeholders to explore the data and discover insights on their own.

Datalore provides several tools to make your visualizations interactive, including widgets and controls that enable your stakeholders to adjust the parameters of your visualizations.



2. Choose the Proper Toolkit for Data Preparation and Visualization

Data visualization

When creating interactive plots, you can also choose from a variety of modern open-source Python packages, such as Plotly, Bokeh, Altair, Lets-plot, HoloViz, Plotnine, and more.

```
import plotly.express as px
fig = px.box(data, x="City", y="Total", color="City")
fig.show()
```

Monthly Sales Analysis November

City

- Yangon
- Naypyitaw
- Mandalay

Total Sales

Overview of price distribution

Let's take a closer look on how much revenue we got from each city and what contribution it brings to the biggest revenues.

Below, you can see Total sales distribution diagram and piechart.

[42] Choose max total order price

Idle
CPU 8%
RAM 2 GB
Calculated 20
In process 8
Errors 0
CPU S

3. Address Reproducibility and Ad Hoc Changes

Reproducibility and ad hoc changes come together since one is impossible to address without minding another.

Reproducibility

Reproducibility is essential in data storytelling. It ensures that your analysis can be repeated by others and that your insights can be validated. [Datalore](#) makes it easy to achieve reproducibility by providing a single notebook that combines code, data, and narrative text. Moreover, each notebook has an isolated environment. It means that the code and the data story itself will not be affected by errors related to different package or language versions.

First, before sharing your data story, we recommend you rerun the whole analysis from the beginning. It helps avoid situations when you've made inconsistent changes when iterating through your analysis.

Second, we recommend that you properly document your notebook with Markdown, so that you and others could quickly grasp the purpose of data ingestion, transformation, and slicing.

If you are curious to learn more about the reproducibility challenges in data science, watch [this webinar](#).



“Datalore has been really useful for reducing the friction of onboarding and documenting our workflows.”

Surya Rastogi

Senior Staff Data Scientist Chainalysis

3. Address Reproducibility and Ad Hoc Changes

Ad hoc changes

Ad hoc changes can be a challenge in data storytelling, especially if your data pipeline is scattered across multiple tools. Ad hoc changes can ruin the reproducibility and cause a lot of extra work modifying code in several places.

With Datalore, you can handle ad hoc changes by simply opening the original notebook in one click, applying the fix, rerunning your work, and republishing the story. All this can be done in mere minutes and helps you deliver the updates on the fly.

Notebooks in Datalore are easy to read, and anyone on the team can apply the required changes, especially while you are away.

The screenshot displays the Datalore interface. On the left, the 'Report builder' sidebar shows a list of cells. Cell 17 is selected and displays a chart titled 'Sales Information per City' with a bar chart showing values for 'City' and 'Yangon'. Cell 19 is an interactive cell with a slider for 'Choose max total order price' ranging from 0 to 1000. Cell 20 is a code cell with `import pandas as pd`. The main area shows a report titled 'Monthly Sales Analysis November' with a '273 Median revenue per sale' KPI card and a pie chart showing the contribution of three cities: Mandalay (40.6%), Naypyitaw (33.8%), and Yangon (25.6%). A modal dialog titled 'Update report for Monthly Sales Analysis November' is open, showing options for 'Access settings' (Public access), 'Report type' (Interactive report), and 'Full width mode' and 'Reactive mode' (both checked). The dialog also includes a 'Delete report' button.

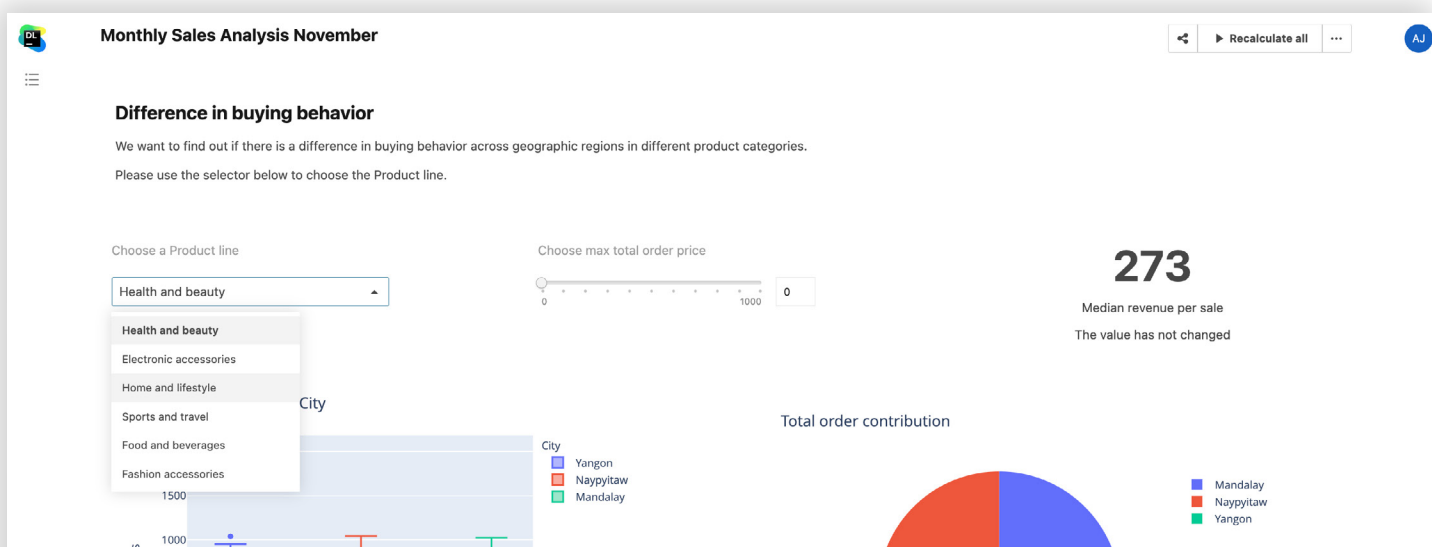
4. Make your story stakeholder-friendly

Making your data storytelling report accessible to stakeholders is crucial for its success. When creating your report, it's essential to consider your stakeholder's background and expertise. Below is a checklist for stakeholder-friendly data stories:

- **Use concise data column names** that are easy to understand and follow a consistent naming convention. For example, instead of using "col1" or "variable_x", use descriptive names like "Total Sales" or "Customer age".
- **Provide context for the data slice**, including timeframes, geographies, and relevant demographic information. For example, if you are presenting data on customer satisfaction, provide context around the survey method used, the sample size, and any significant changes in the market that may have influenced the results.
- **Use clear, concise language** that avoids technical jargon, and explain complex concepts in plain terms. For example, instead of "multicollinearity", use "when two or more predictors in a model are highly correlated".
- **Balance text, numbers, and visuals** to provide a well-rounded and engaging story. Use visuals to highlight key points and text to provide context and explanations. For example, use a line chart to show trends over time and use text to explain why those trends occurred.
- **Include actionable insights and recommendations** based on the data analysis to help stakeholders understand what actions to take based on the insights. For example, if you are presenting data on customer churn, provide recommendations on how to improve customer retention rates.
- **Make the story compact**, not a lengthy report. Keep the story concise and to the point, focusing on the most important insights and information for your audience.

By following these tips, you can ensure that your data story is not only informative but also engaging and actionable for stakeholders.

After completing the analysis, you can use Datalore's report builder to create a clean narrative and hide unnecessary steps. Stakeholders will then be able to explore the data and gain a deeper understanding of your work.



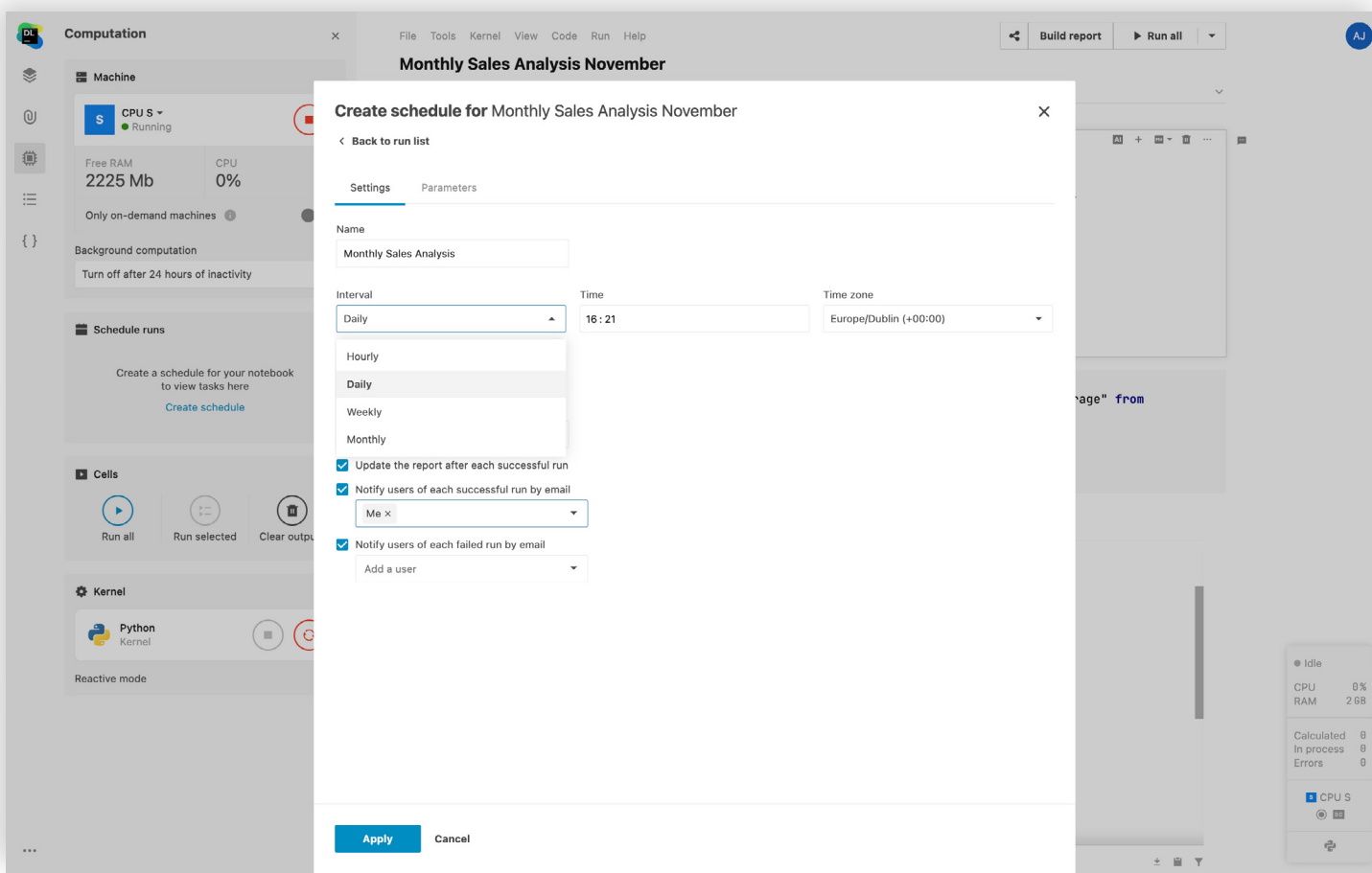
5. Mind Automation and Building Up On Your Work

As your data storytelling projects grow in complexity, you may find yourself repeating certain tasks that could benefit from automation. These could include data cleaning, visualization creation, and report generation. Automating these tasks not only saves time but also ensures consistency and reproducibility. In this chapter, we will explore how **Datalore** can help you automate aspects of your data storytelling projects, allowing you to focus on insights and storytelling.

Scheduling analysis updates

By scheduling report updates, you can ensure that your data stories are always accurate and reliable. For example, you could schedule a weekly report that provides stakeholders with a summary of key metrics or insights. This could include charts and visualizations that help stakeholders quickly understand trends and patterns in the data.

Additionally, **Datalore** allows you to automate the process of running your data analysis and cleaning scripts. You can **trigger Datalore notebooks** to run on demand via the API calls or use the same Scheduling feature to rerun the notebook on a schedule. By automating these tasks, you can focus on the more creative aspects of data storytelling, such as developing new insights and identifying key trends.

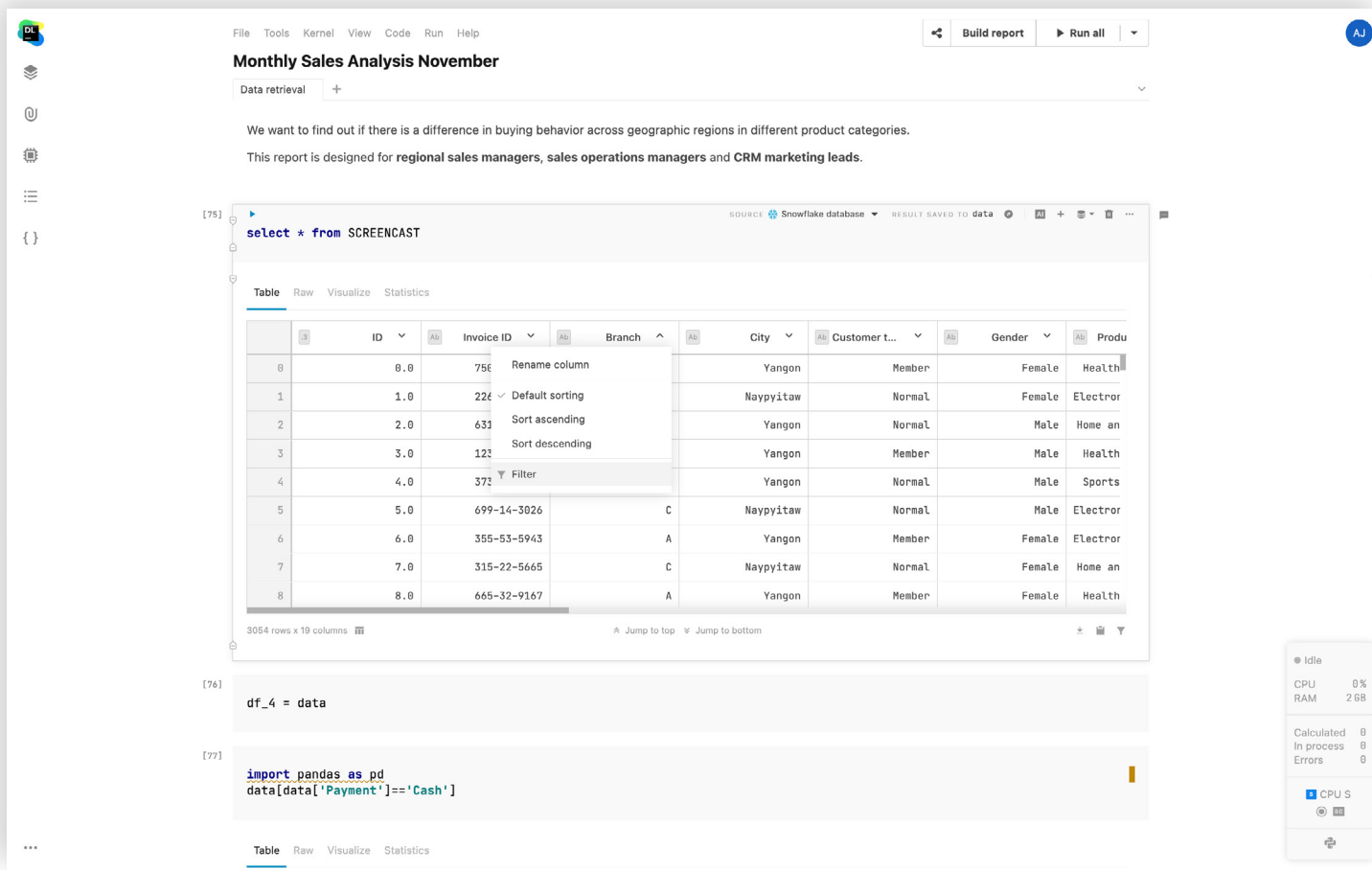


5. Mind Automation and Building Up On Your Work

Letting stakeholders deep dive into your story

One of the key benefits of using a collaborative data science platform like [Datalore](#) is that it allows stakeholders to take a deep dive into your data story. With Datalore, stakeholders can edit copies of your interactive reports, use SQL to get more data, and work with no-code statistics tabs and chart cells. They don't need to know Python to understand and analyze the data.

By providing access to these features, stakeholders can explore the data more deeply and gain a better understanding of the insights and recommendations that are being presented to them. This can lead to more informed decision-making and better outcomes for the organization.



The screenshot displays the Datalore interface for a project titled "Monthly Sales Analysis November". The interface includes a menu bar (File, Tools, Kernel, View, Code, Run, Help), a "Build report" button, and a "Run all" button. Below the title, there is a "Data retrieval" section with a plus sign. A descriptive text block states: "We want to find out if there is a difference in buying behavior across geographic regions in different product categories. This report is designed for regional sales managers, sales operations managers and CRM marketing leads." The main area shows a SQL query: `select * from SCREENCAST`. Below the query, a table view displays the results. The table has columns: ID, Invoice ID, Branch, City, Customer t..., Gender, and Produ. The data rows are as follows:

ID	Invoice ID	Branch	City	Customer t...	Gender	Produ
0	756		Yangon	Member	Female	Health
1	222		Naypyitaw	Normal	Female	Electror
2	633		Yangon	Normal	Male	Home an
3	121		Yangon	Member	Male	Health
4	371		Yangon	Normal	Male	Sports
5	699-14-3026	C	Naypyitaw	Normal	Male	Electror
6	355-53-5943	A	Yangon	Member	Female	Electror
7	315-22-5665	C	Naypyitaw	Normal	Female	Home an
8	665-32-9167	A	Yangon	Member	Female	Health

Below the table, there is a status bar indicating "3054 rows x 10 columns". Below the table view, there are two code cells. The first cell contains `df_4 = data`. The second cell contains `import pandas as pd` and `data[data['Payment']=='Cash']`. On the right side of the interface, there is a system status panel showing: Idle, CPU 0%, RAM 2GB, Calculated 0, In process 0, Errors 0, and CPU 5.

Conclusions

In conclusion, efficient data storytelling is critical to the success of any data analysis project.

To make your data stories more robust, it's essential to:

- Have a clear storyline.
- Choose an appropriate data retrieval, data quality, and visualization toolkit.
- Act with reproducibility in mind and be ready for ad hoc changes.
- Make your story stakeholder-friendly.
- Automate as many things as you can.

By following these best practices, data analysts can ensure that their findings are impactful and accessible to stakeholders, ultimately driving informed decision-making and positive outcomes.

Next steps

To learn more about how **Datalore** can help your data team excel at data storytelling, visit jetbrains.com/datalore/ or schedule a **personalized demo** for your team.

If you have any questions or feedback regarding the recommendations outlined in this whitepaper, please get in touch with us via email at datalore-enterprise@jetbrains.com.

