

Chapitre 3

Tools and Technologies Used in Data Science

Presented par :
Dr. Bilal Dendani



جامعة باجي مختار - عنابة
BADJI MOKHTAR - ANNABA UNIVERSITY

Dr. DENDANI Bilal



Chapitre 3 : Tools and Technologies Used in Data Science

- Data Storage Tools
- Data Preparation Tools
- Data Visualization Tools
- IDE and Notebook Tools
- Comprehensive Data Science Platforms

Introduction

- In data science, mastering **tools** and **technologies** is essential to **transform raw data** into **actionable insights**.
- Data scientists use a wide range of tools for each stage of the data science lifecycle, from data storage to analysis and visualization.
- These tools make it easier to **handle large datasets**, automate data preparation tasks, and enable advanced analyses.

3.1 Data Storage Tools

- Data storage tools are essential for managing **large volumes of data** in a **centralized** and **secure** manner.
- They ensure durable data retention while facilitating efficient access for analysis.



3.1 Data Storage Tools

Which data storage solutions should I choose?



3.1.1 Relational Databases(SQL)

- Use a **structured model based** on tables and relationships between them.
- Widely used to store and query well-organized data.
- Data is organized in **tables (rows and columns)**, each representing an entity (e.g., *Customers, Sales*).
- The query language used is **SQL (Structured Query Language)**.
- **Examples:** MySQL, PostgreSQL, Oracle.



3.1.2 NoSQL Databases (Not Only SQL)

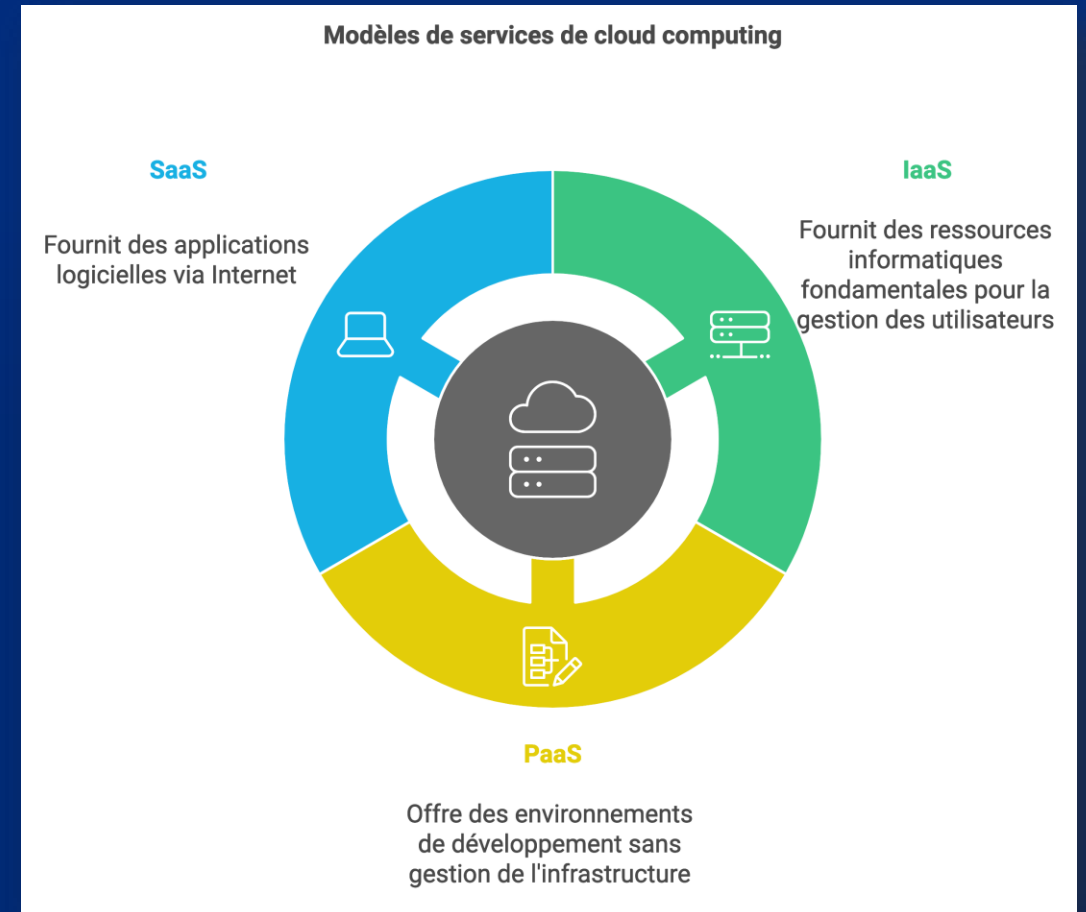
- Offer **high flexibility** for storing **unstructured or semi-structured** data such as JSON documents, social media posts, or sensor data.
- NoSQL is used for **large-scale** and **real-time** applications, emphasizing speed and scalability
- They are designed for scalability and flexibility, making them well-suited for big data applications and real-time analytics.
- **Examples: MongoDB, Redis, Cassandra.**



<https://www.datacamp.com/blog/nosql-databases-what-every-data-scientist-needs-to-know>

3.1.3 Cloud Storage

- Provides **remote access**, **high availability**, and **scalability** on demand.
- **Examples**: Amazon S3, Google Cloud Storage, Azure Blob Storage.



3.1.4 Data Lakes

- **Data lakes** are used for **storing raw data**, whether **structured or unstructured**.
They are primarily utilized for **advanced analytics** and **machine learning** applications.
- **Associated technologies:**
 - HDFS, AWS Lake Formation, Azure Data Lake.

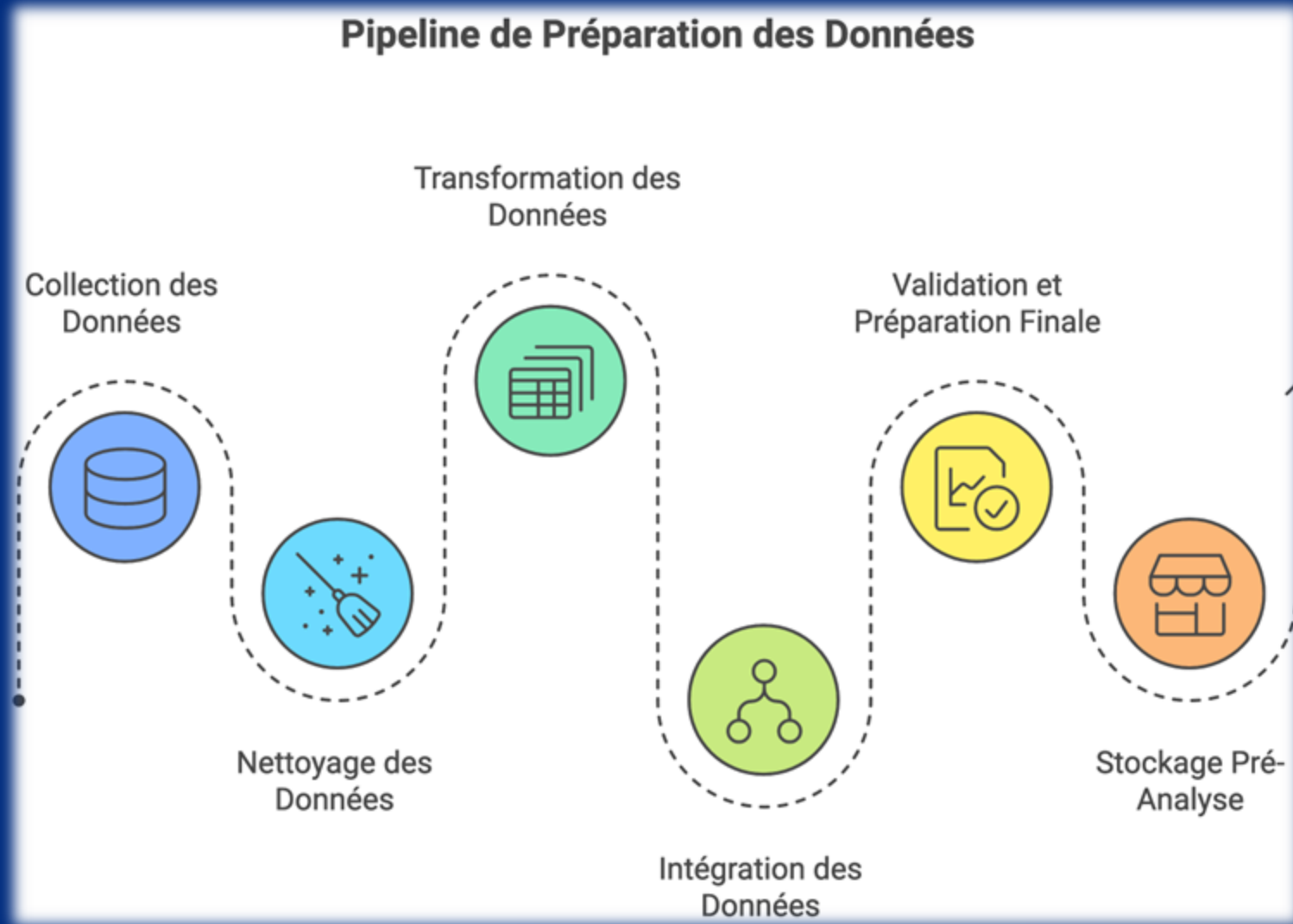


3.2 Data Preparation Tools in Data Science

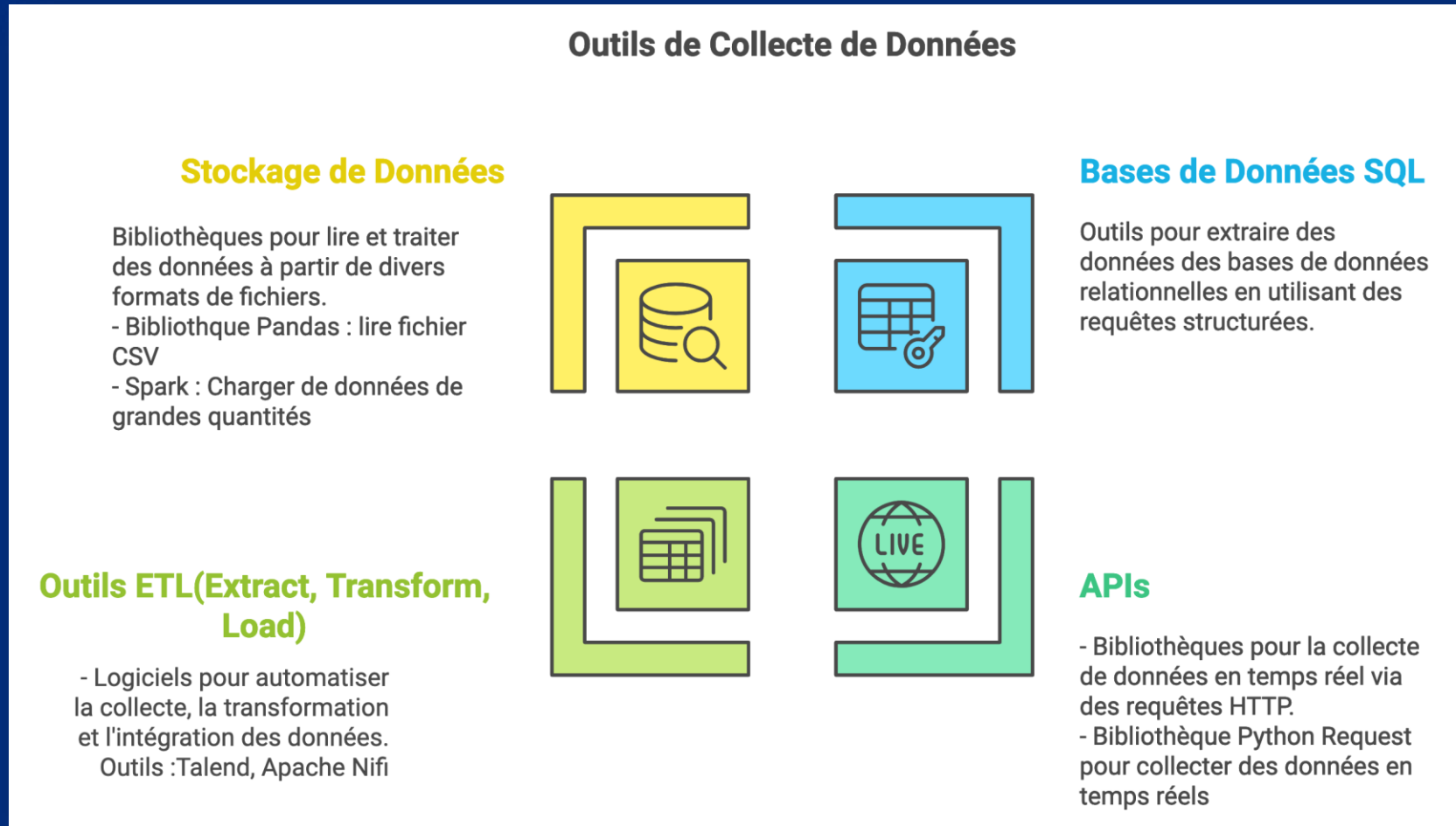
- Data preparation is a **key step** in the data science process.
- Its goal is to **clean, transform, and integrate** data to ensure its **quality and usability**.
- **Different tools** are used for each of the stages of data preparation.



Pipeline of data preparation

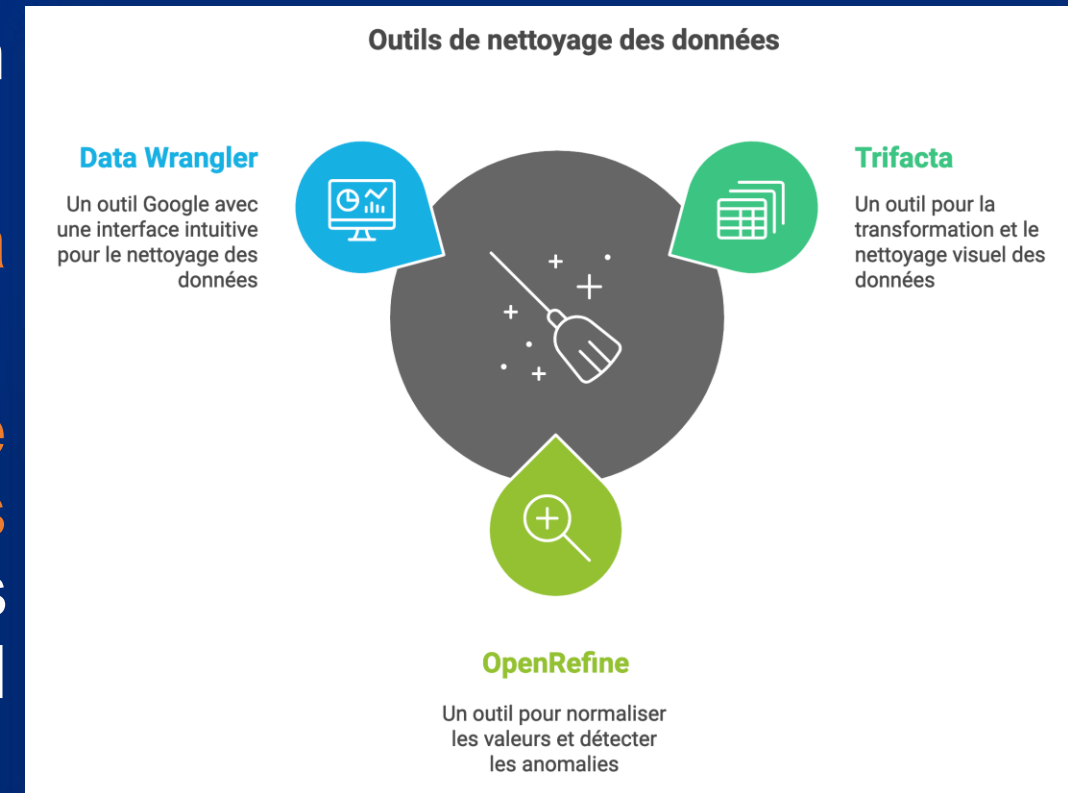


3.2.1 Data collection tools

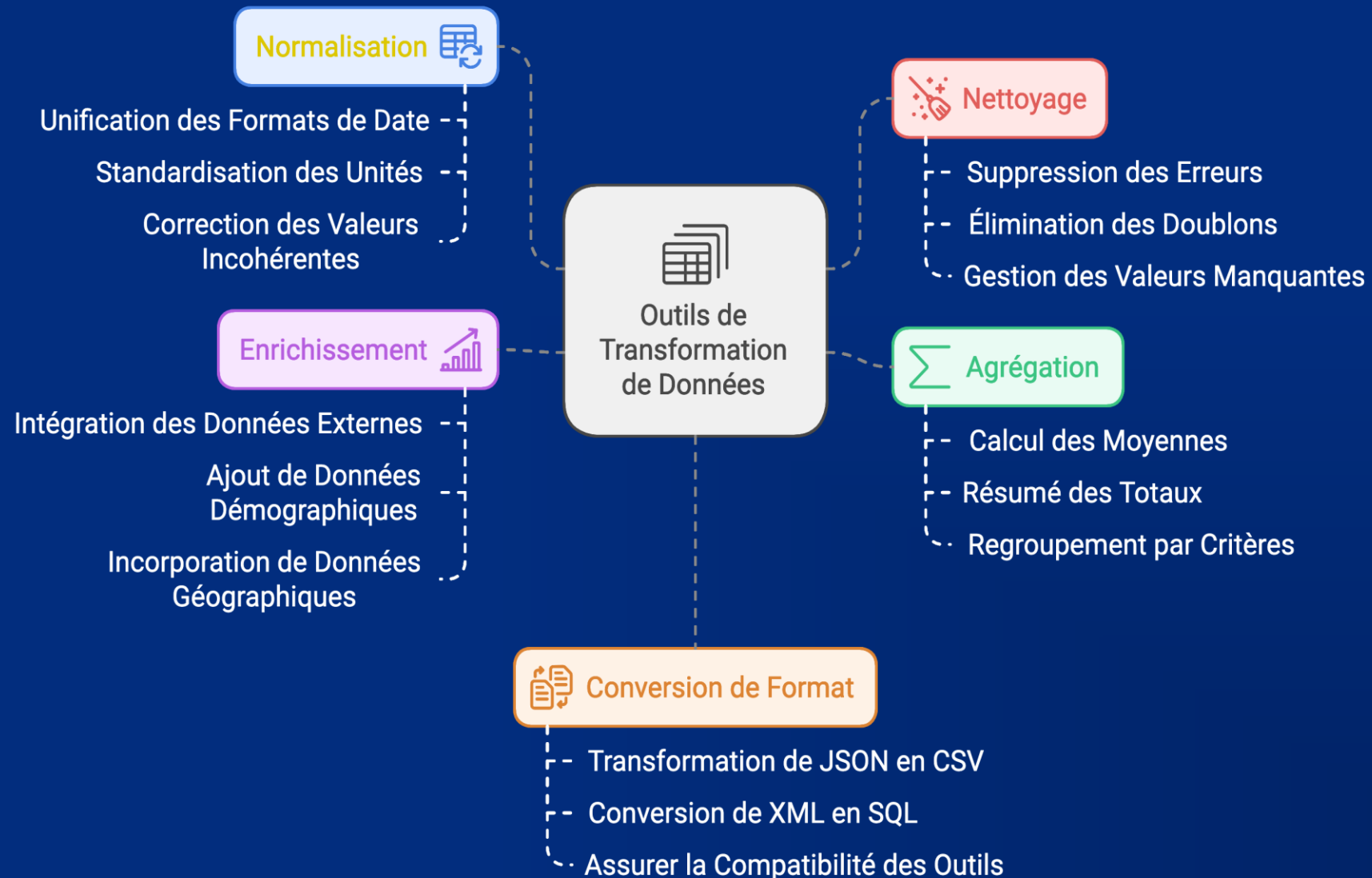


3.2.2 Data Cleaning Tools

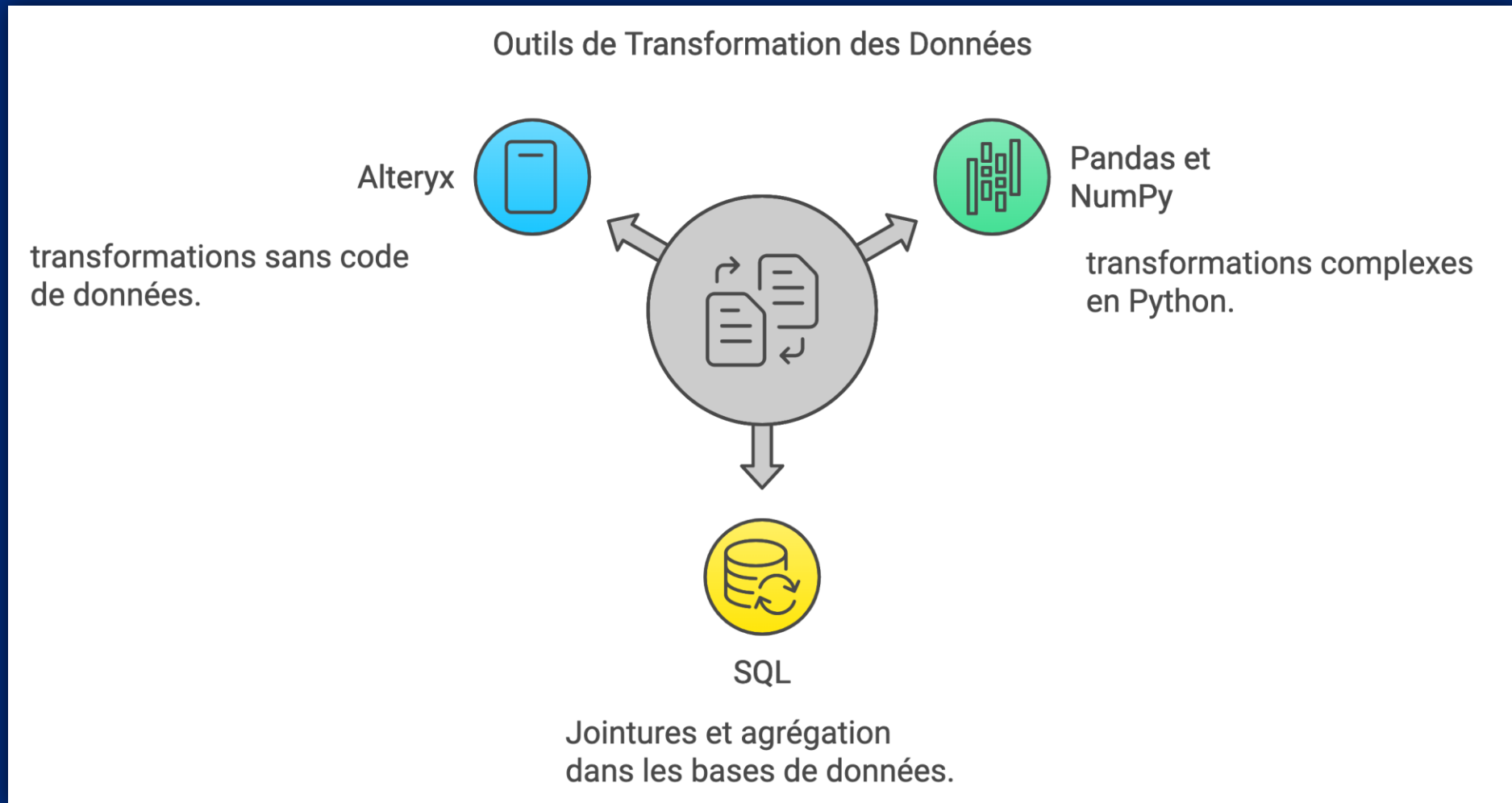
- Data cleaning is an **essential step** in the data science process.
- It plays a key role in **improving data quality**.
- The cleaning process helps **remove or correct outliers and erroneous values** that could distort the results of machine learning models and statistical analyses.



3.2.3 Data transformation tools

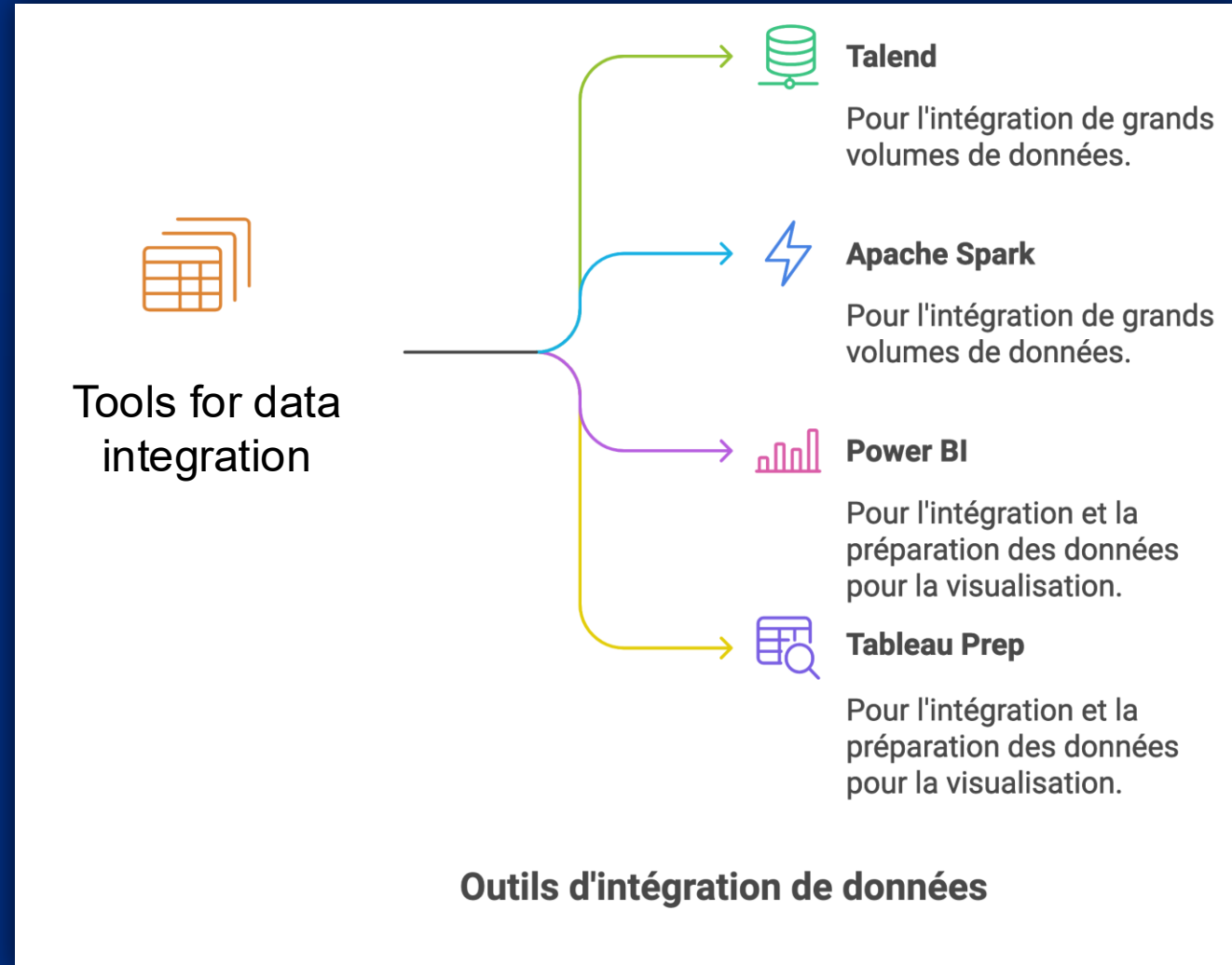


3.2.3 Data transformation tools



3.2.4 Data integration tools

- Data integration is the process of **combining** data from **multiple sources** to create a unified and coherent view.
- It involves collecting, transforming, and consolidating the data so that it can be effectively used for analysis and decision-making.



3.3 Data Visualization Tools

- In data science, data visualization tools play a crucial role in **exploring datasets**, **identifying patterns**, and **communicating results**.

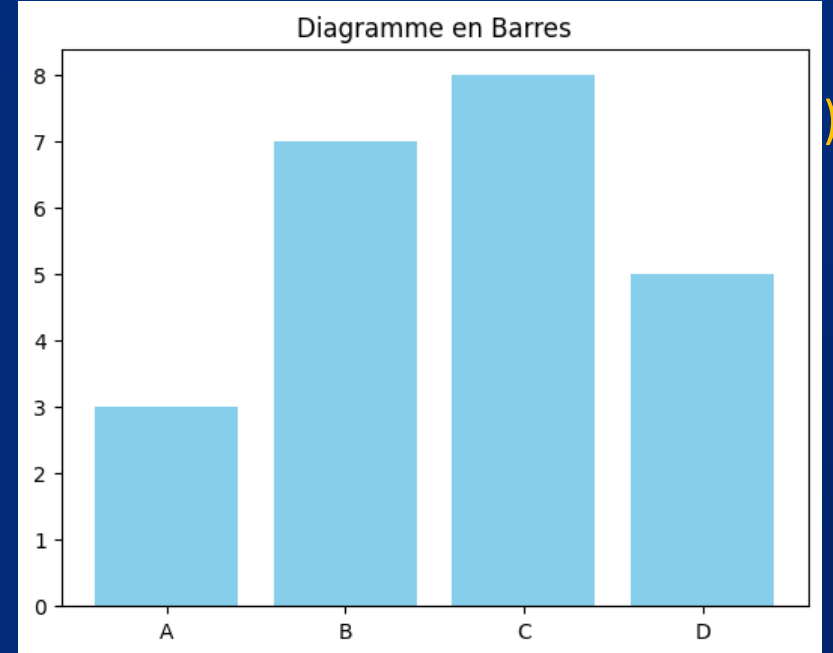
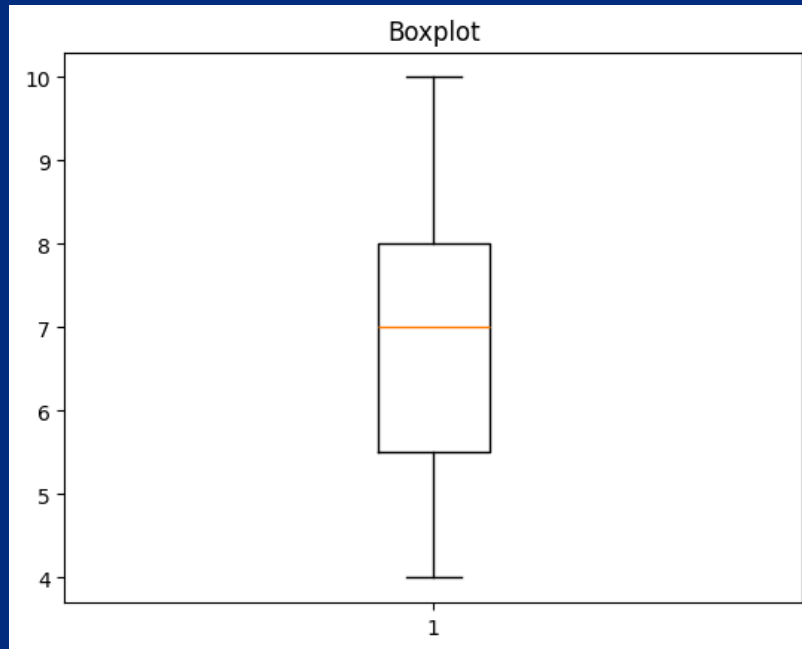
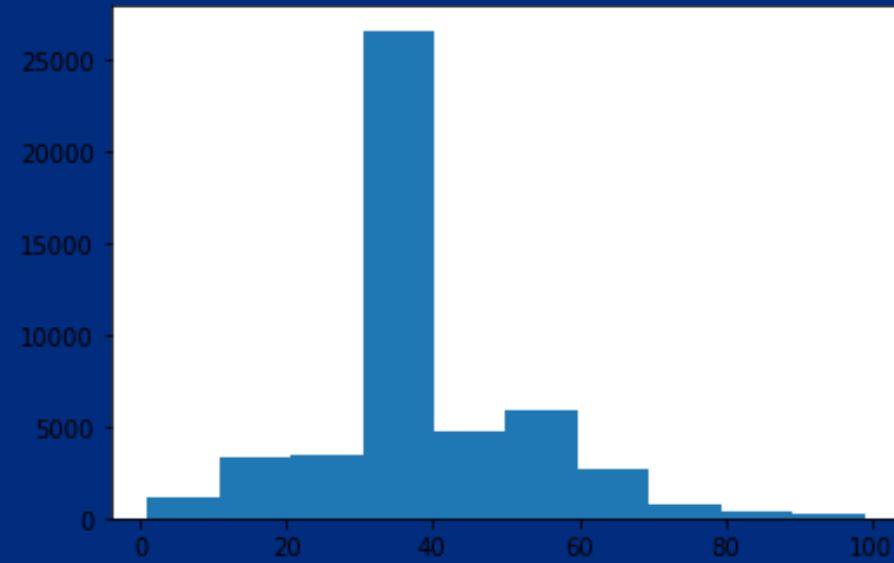
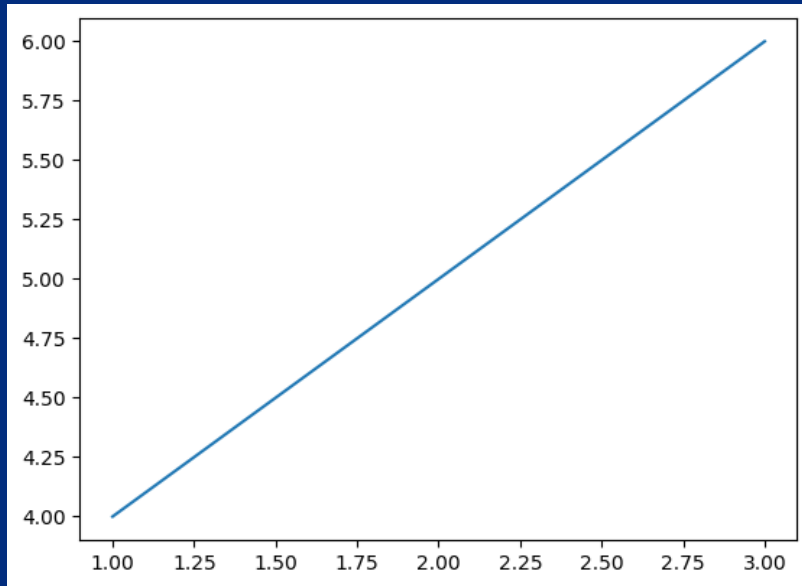
Below are the main visualization tools, categorized by type and typical use cases:

1. Python Visualization Libraries:

- Python is one of the most widely used programming languages in data science, and it offers a rich ecosystem of libraries dedicated to visualization.

3.3.1 Matplotlib

- A **foundational Python library** used to create static, animated, or interactive visualizations.
- Commonly used to build line **charts**, **histograms**, **bar charts**, and **scatter plots**.
- One of its main strengths is its high level of customization, making it highly adaptable and compatible with other libraries such as pandas and NumPy.



<https://colab.research.google.com/drive/1n4ulmR1BCV8ZJl1AJCHJvQMMcMANWuy5>

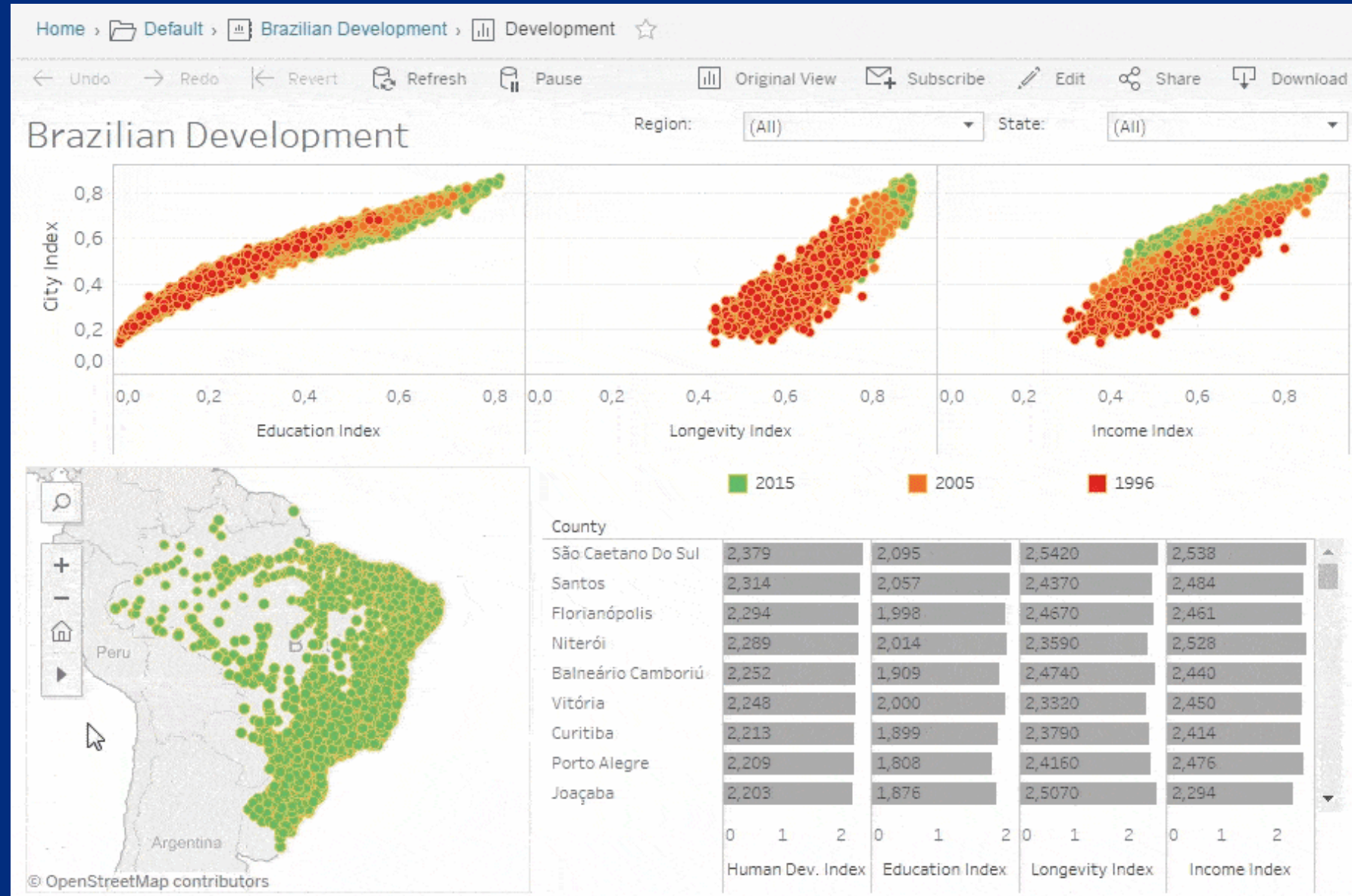
3.3.2 Seaborn

- Built on top of **Matplotlib**, **Seaborn** is designed for creating advanced statistical visualizations.
- Use cases: Visualizing distributions, heatmaps, regression plots, and multivariate relationships.
- Strengths: Offers beautiful default aesthetics and simplifies the creation of complex statistical plots.

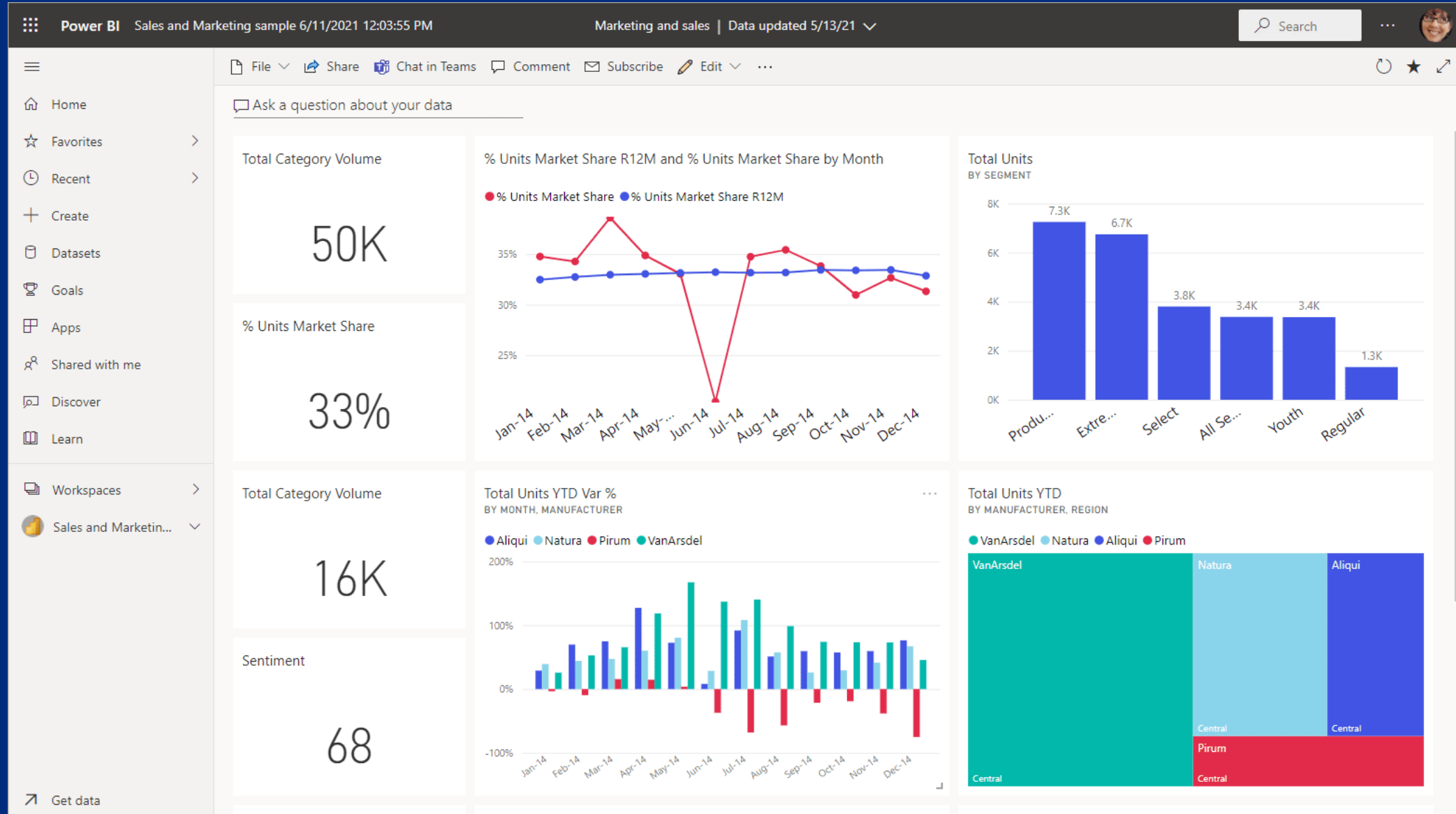
3.3.3 Plotly

- An interactive tool for creating **dynamic visualizations** and dashboards.
- Used to build interactive **maps, 3D charts**, and complex diagrams.
- One of this library's key strengths lies in its compatibility with Dash, which allows the creation of powerful and interactive dashboards.


3.3.4 Data Visualization Software: Tableau







3.3.5 Data Visualization Software: Tableau: PowerBI



3.3.6 Google Data Studio


 [Resources](#) [About](#)

  English   [My Account](#)

[< DATA JOURNALISM](#)


Lesson 11 of 13

Data Studio: Make interactive data visualizations




Data Studio: Make interactive data visualizations





Give life to your datasets by creating powerful interactive visualizations with an easy-to-use studio.

 [Download Lesson](#)




Geo map





Line



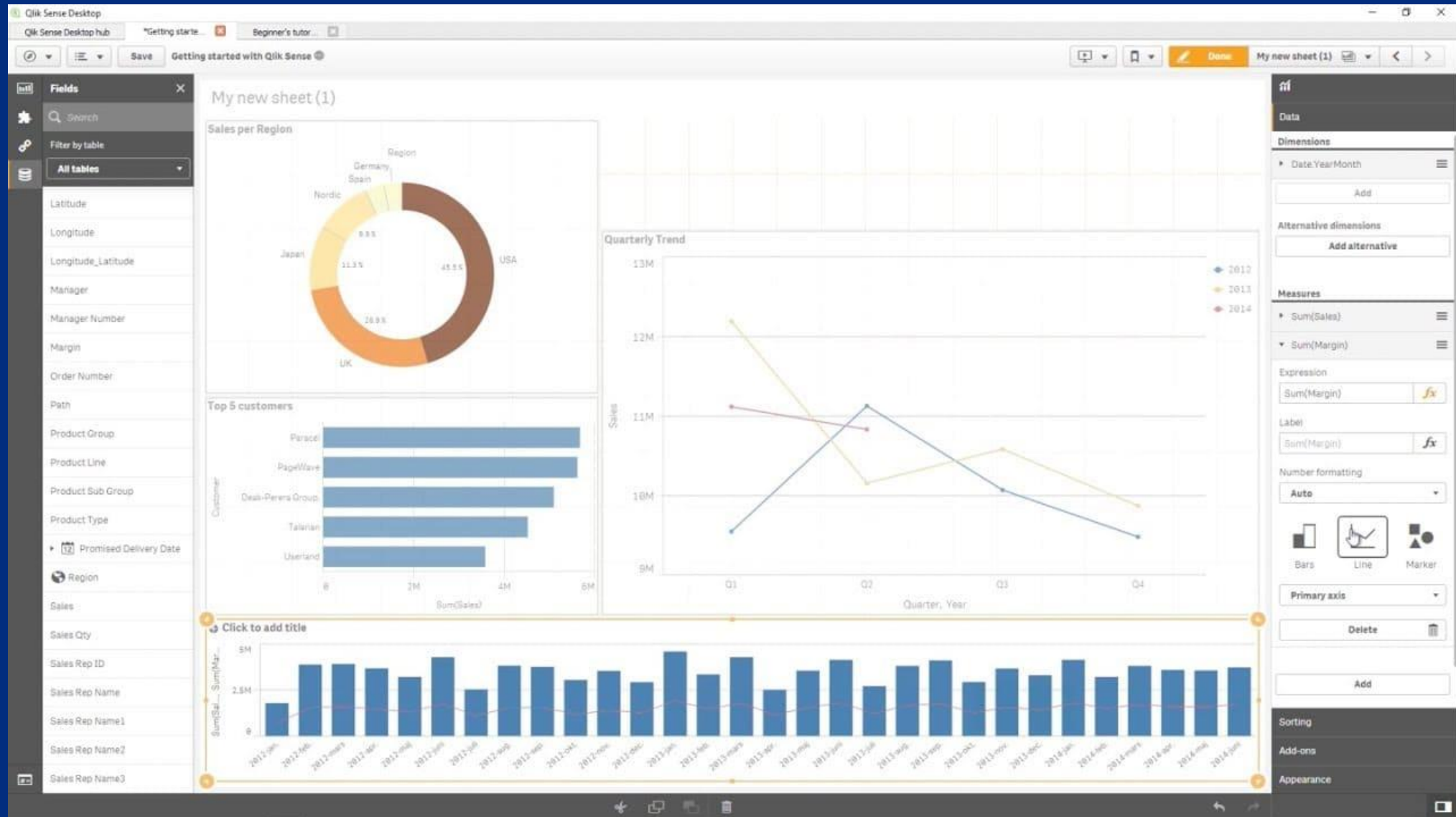
Area



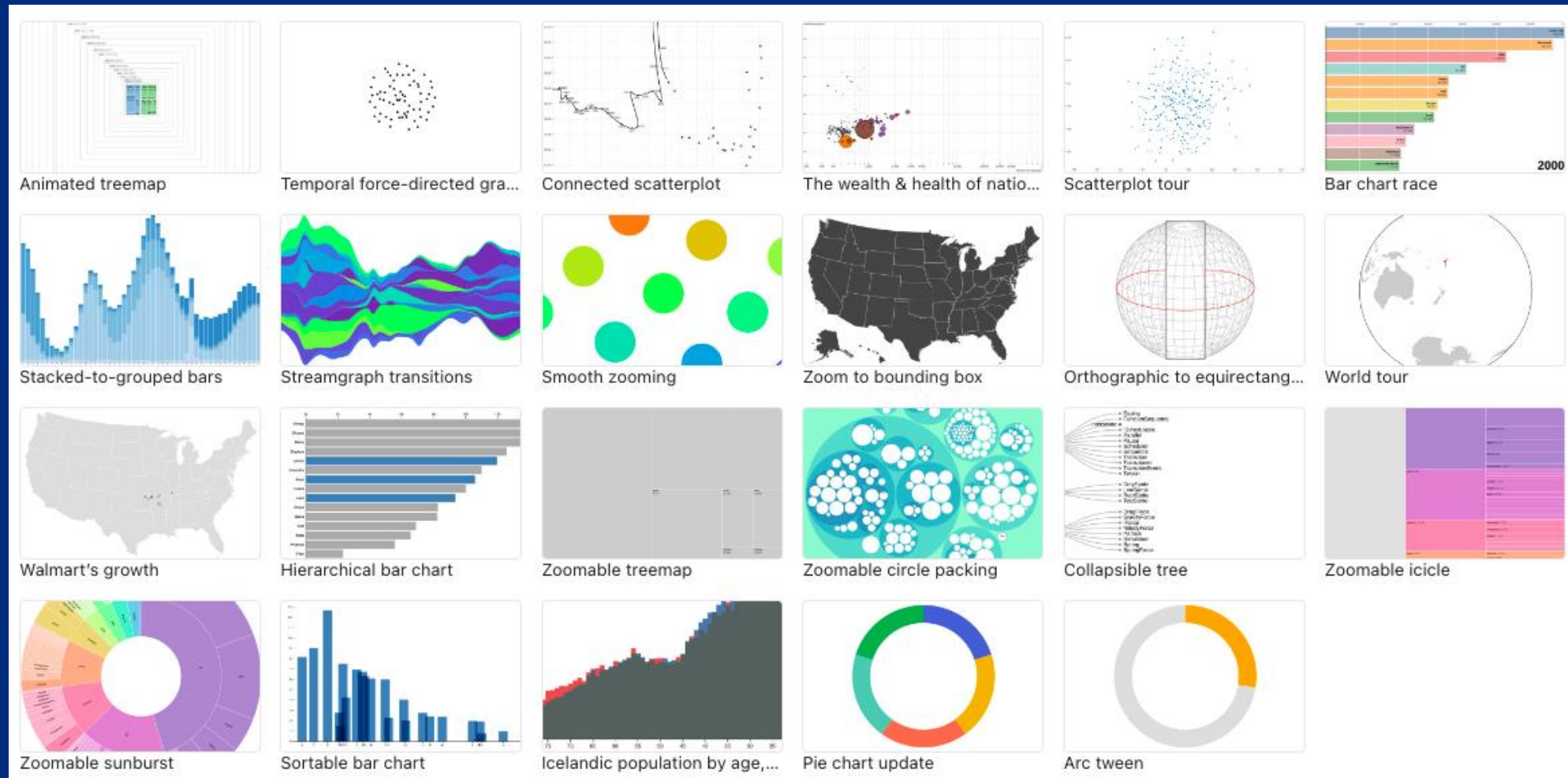
Scatter



3.3.7 Qlik Sense



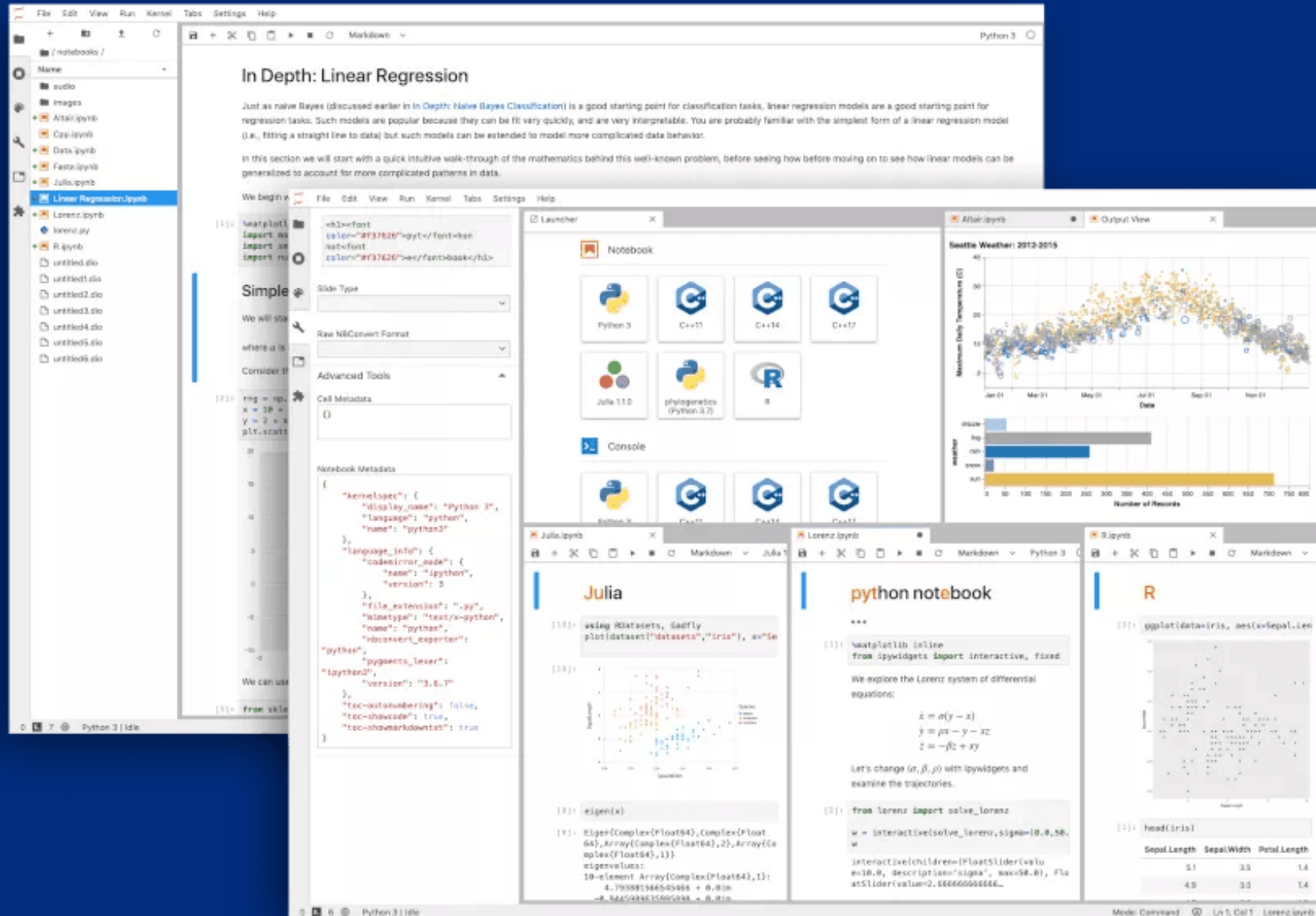
3.3.8 D3.js



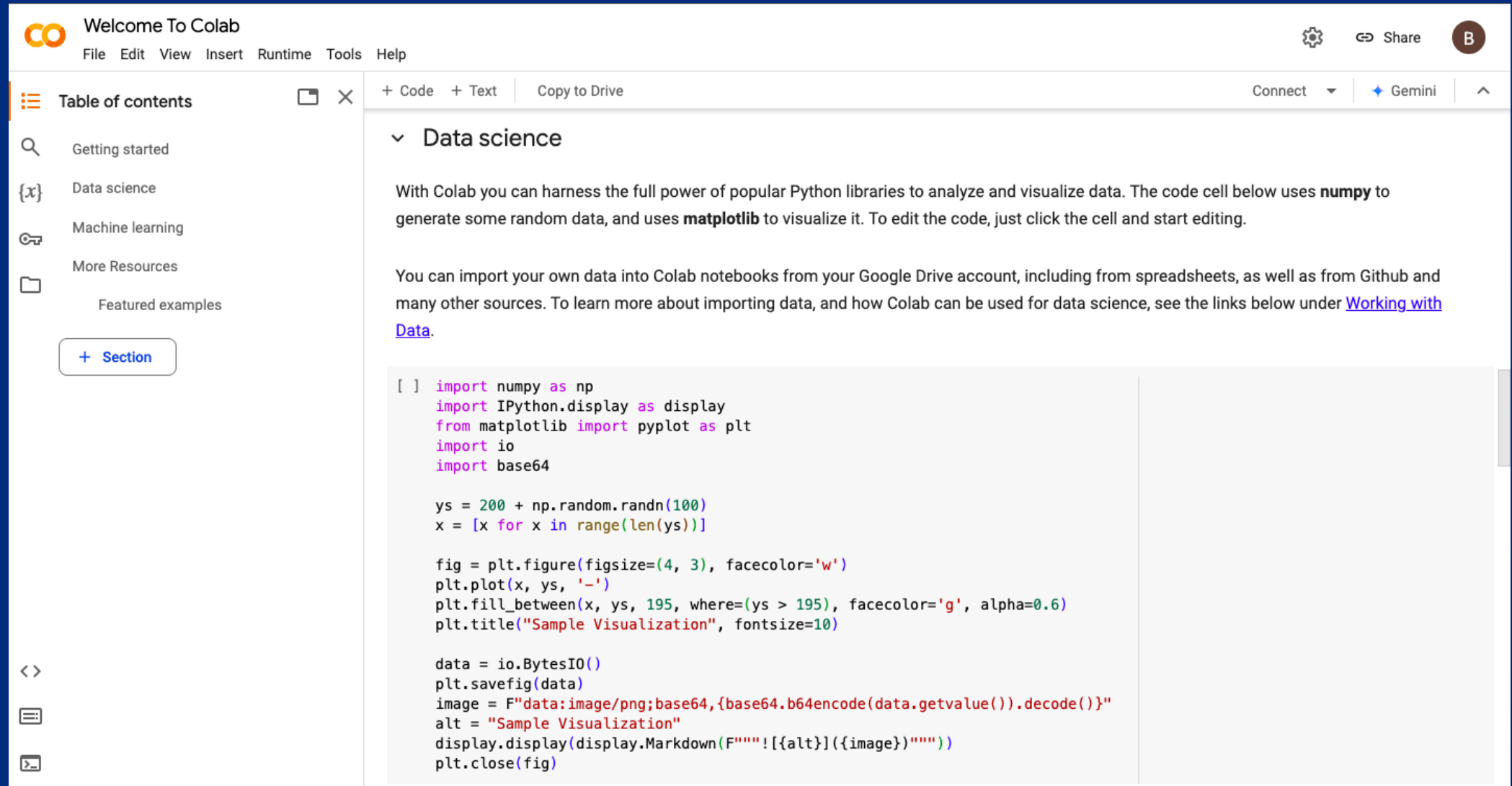
3.4 IDE Notebook Development Tools

- IDE notebooks (**Integrated Development Environment Notebooks**) are interactive environments
- that allow users to write and execute code while integrating visualizations, textual explanations, and data analysis within the same workspace.
- They are particularly useful in data science for exploring datasets, testing algorithms, and presenting results in a clear and reproducible manner.

3.4.1 IDE notebooks tools: Jupyter notebook et Jupyter Lab



3.4.2 Google Colab



The screenshot displays the Google Colab web interface. At the top, there's a 'Welcome To Colab' header with a menu bar (File, Edit, View, Insert, Runtime, Tools, Help) and a user profile icon. Below the header, a sidebar on the left contains a 'Table of contents' with links to 'Getting started', 'Data science', 'Machine learning', and 'More Resources'. A '+ Section' button is also present. The main area is titled 'Data science' and contains two paragraphs of introductory text. The first paragraph explains that Colab allows using Python libraries like **numpy** and **matplotlib** for data analysis and visualization. The second paragraph discusses importing data from Google Drive, GitHub, and other sources, with a link to 'Working with Data'. Below the text is a code cell containing Python code that generates random data, creates a plot, and displays it as a base64-encoded image.

```
[ ] import numpy as np
import IPython.display as display
from matplotlib import pyplot as plt
import io
import base64

ys = 200 + np.random.randn(100)
x = [x for x in range(len(ys))]

fig = plt.figure(figsize=(4, 3), facecolor='w')
plt.plot(x, ys, '-')
plt.fill_between(x, ys, 195, where=(ys > 195), facecolor='g', alpha=0.6)
plt.title("Sample Visualization", fontsize=10)

data = io.BytesIO()
plt.savefig(data)
image = F"data:image/png;base64,{base64.b64encode(data.getvalue()).decode()}"
alt = "Sample Visualization"
display.display(display.Markdown(F""""!{alt}({image})"""))
plt.close(fig)
```

3.4.3 Spyder

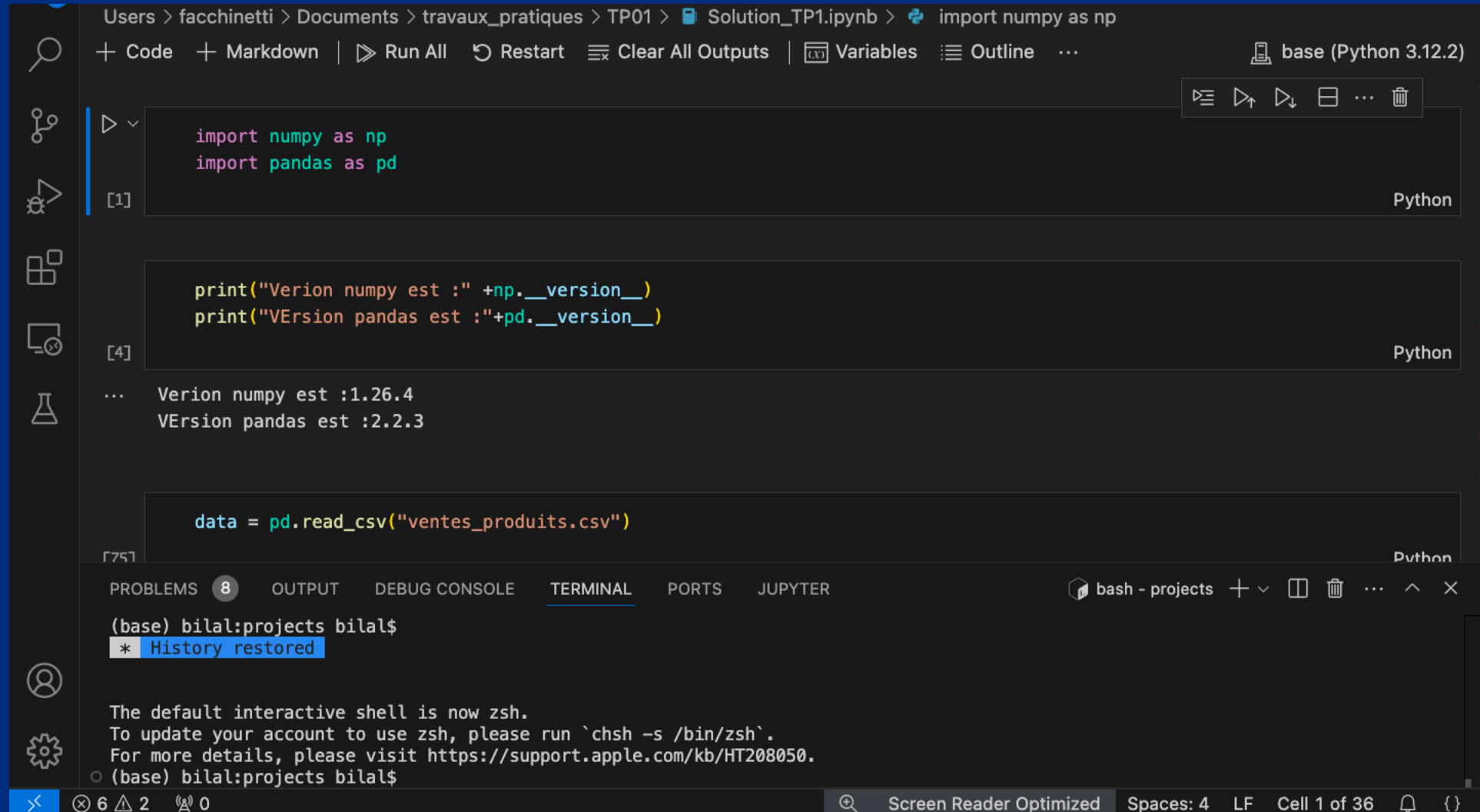
The screenshot displays the Spyder IDE interface with the following components:

- Left Panel (File Explorer):** Shows the project structure with folders for `plugin.py`, `plot_example.py`, and `plugin.py - IPythonConsole`.
- Central Panel (Code Editor):** Displays the `plugin.py` file, which defines the `Plots` plugin class. The code includes imports for `qtgui.QtCore`, `spyder.api.plugins`, `spyder.api.translations`, and `spyder.plugins.plots.widgets`. It also includes a `__main__` block and a `Plots` class that inherits from `SpyderDockablePlugin`.
- Right Panel (Variable Explorer):** Shows a table of variables with their types and values.
- Bottom Panel (Plots):** Displays a 3D surface plot of a function, with axes labeled from -84.41 to 36.73.

Name	Type	Size	Value
a	foo	1	foo object of __main__ module
filename	str	53	/Users/Documents/spyder/spyder/tests/test_dont_use.py
i	bool	1	True
my_set	set	3	{1, 2, 3}
r	float	1	6.46567886443
t	tuple	5	('abcd', 745, 2.23, 'efgh', 78.2)
thisdict	dict	3	{'brand': 'Ford', 'model': 'Mustang', 'year': 1964}
tinylist	list	2	[123, 'efgh']
x	Array of int64	(2,)	[1 2]
y	timedelta	1	2 days, 0:00:00

Bottom status bar: LSP Python: ready | conda: spyder-dev (Python 3.7.10) | master | Line 1, Col 1 | UTF-8 | LF | RW | Mem 57%

3.4.4 VS Code with Jupyter Extension



The screenshot displays the VS Code Jupyter Notebook interface. The top toolbar includes buttons for '+ Code', '+ Markdown', 'Run All', 'Restart', 'Clear All Outputs', 'Variables', 'Outline', and a dropdown for the current kernel 'base (Python 3.12.2)'. The notebook contains three code cells:

- Cell [1]:

```
import numpy as np
import pandas as pd
```
- Cell [4]:

```
print("Version numpy est :"+np.__version__)
print("Version pandas est :"+pd.__version__)
```
- Cell [75]:

```
data = pd.read_csv("ventes_produits.csv")
```

The output of cell [4] shows the versions of numpy and pandas. The bottom panel features tabs for 'PROBLEMS' (8), 'OUTPUT', 'DEBUG CONSOLE', 'TERMINAL', 'PORTS', and 'JUPYTER'. The 'TERMINAL' tab is active, showing a bash shell prompt and system messages about the default interactive shell being zsh.

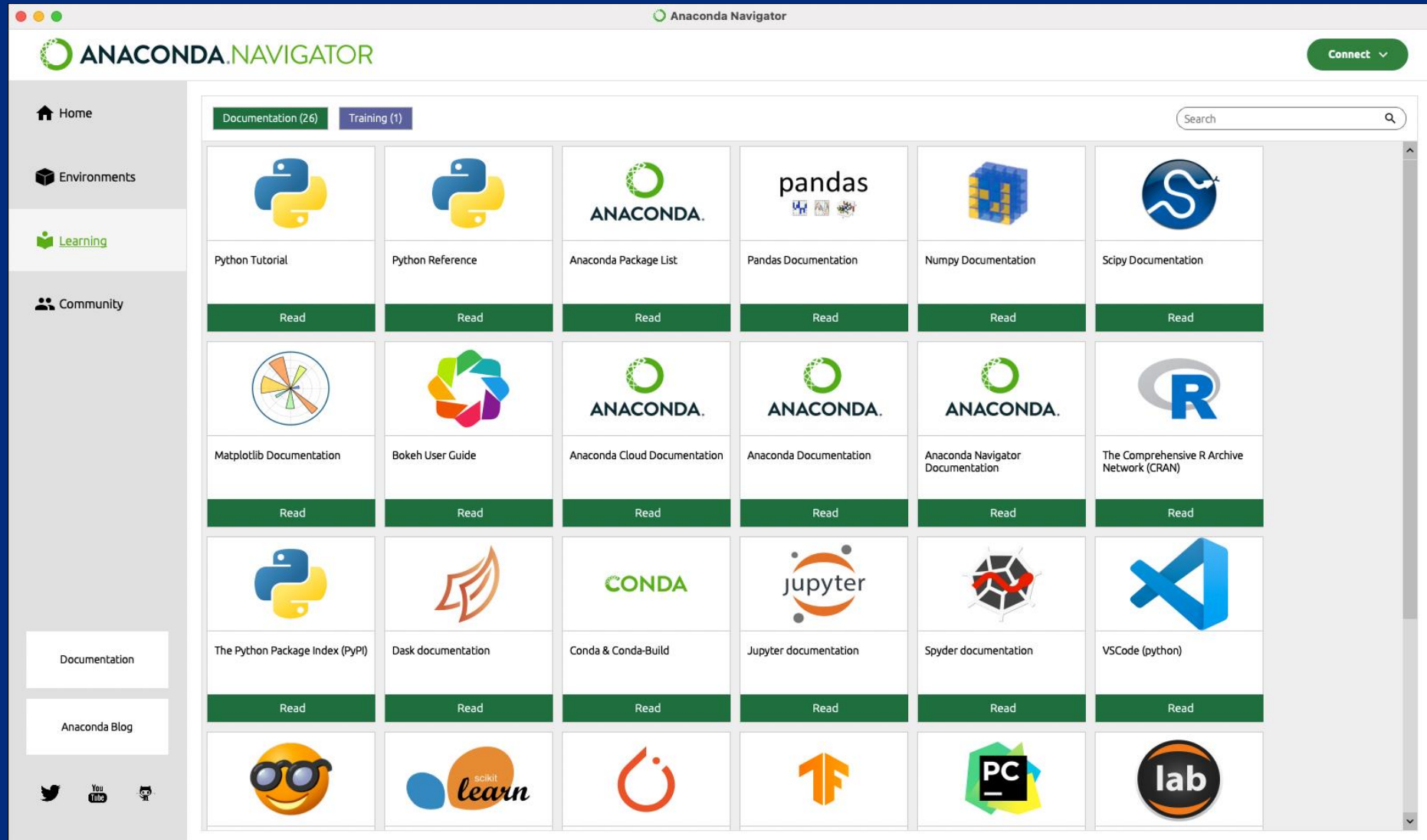
```
(base) bilal:projects bilal$
* History restored

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) bilal:projects bilal$
```

3.5 Comprehensive Data Science Platforms

- Comprehensive Data Science platforms are integrated environments that bring together all the tools required to carry out the entire data science workflow — from data collection to visualization and model deployment.
- These platforms enable data scientists to collaborate efficiently, automate workflows, and manage complex projects in a unified workspace.

3.5.1 Anaconda



3.5.2 Microsoft Azure

Platform Services

Security and Management

- Portal
- Active Directory
- Multi-Factor Authentication
- Automation
- Key Vault
- Store/Marketplace
- VM Image Gallery and VM Depot

Compute

- Cloud Services
- Service Fabric
- Batch
- Remote App

Web and mobile

- Web Apps
- API Apps
- API Management
- Mobile Apps
- Logic Apps
- Notification Hubs

Developer services

- Visual Studio
- Azure SDK
- Team Project
- Application Insights

Hybrid Operations

- Azure AD Connect Health
- AD Privileged Identity Management
- Backup
- Operational Insights
- Import/Export
- Site Recovery
- StorSimple

Integration

- Storage Queues
- BizTalk Services
- Hybrid Connections
- Service Bus

Analytics and IoT

- HDInsight
- Machine Learning
- Data Factory
- Event Hubs
- Stream Analytics
- Mobile Engagement

Data

- SQL Database
- SQL Data Warehouse
- Redis Cache
- Search
- Cosmos DB
- Tables

Media and CDN

- Media Services
- Content Delivery Network (CDN)

Infrastructure Services

Compute

- Virtual Machine
- Containers

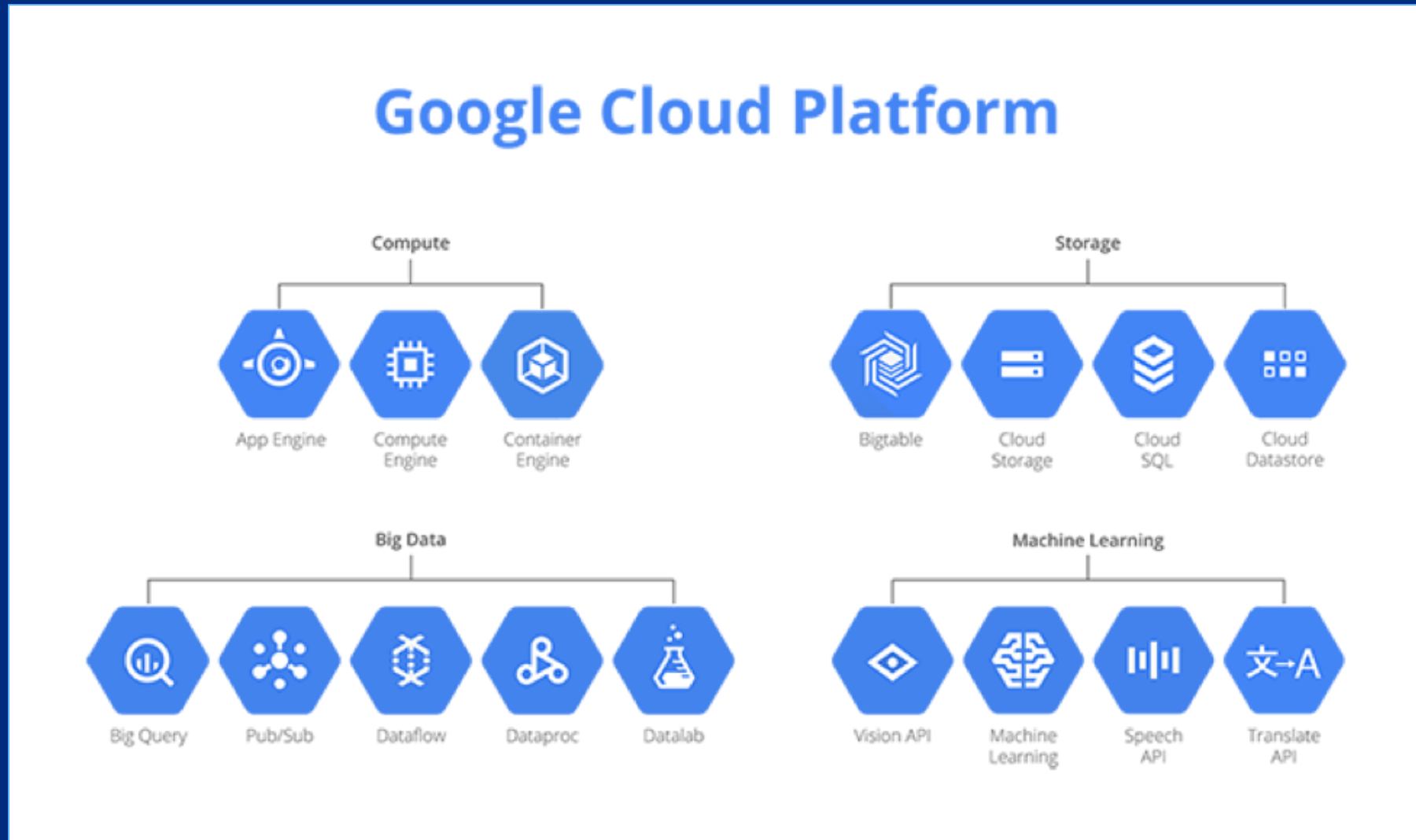
Storage

- BLOB Storage
- Azure Files
- Premium Storage

Networking

- Virtual Network
- Load Balancer
- DNS
- Express Route
- Traffic Manager
- VPN Gateway
- Application Gateway

3.5.3 Google cloud platform



3.5.6 Data science on AWS

