



IMT Atlantique

Bretagne-Pays de la Loire

École Mines-Télécom

Data Understanding

Data Science Toolkit & Applications

Lina Fahed & Laurent Brisson

Data Understanding: The Backbone of Data Science

How to **Explore**, **Clean**, and **Validate Data** for Insights

The Critical Role of Data Understanding

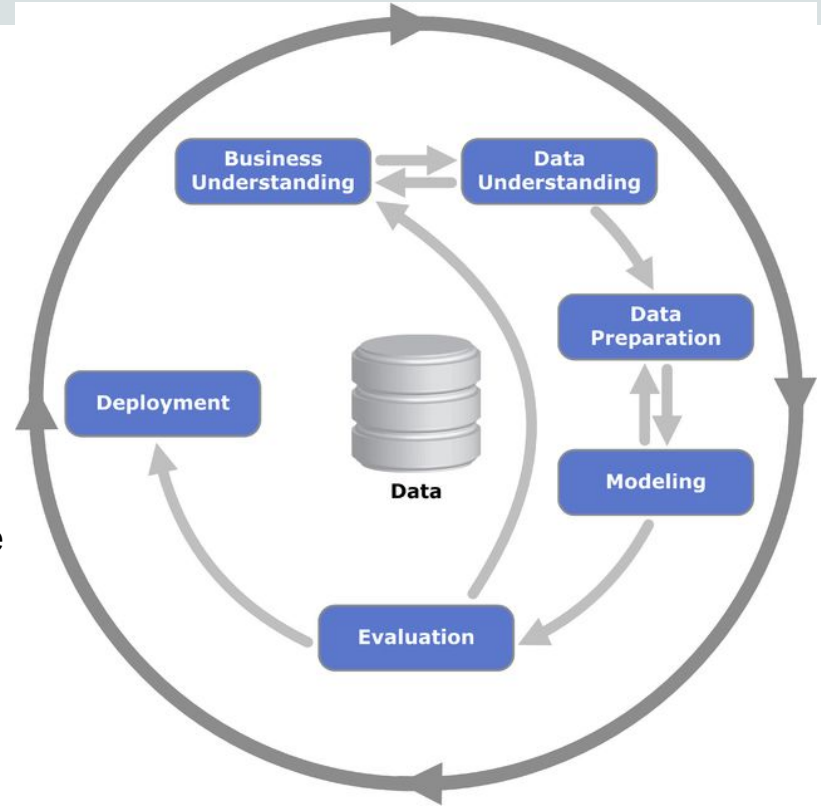
- 60% of data science **time** is spent on data exploration and cleaning
- **Garbage In, Garbage Out** (GIGO): Poor data = poor insights.
- Example: A model trained on **biased data** will produce **biased predictions**.

Where Does Data Understanding Fit?

CRISP-DM Steps:

1. Business Understanding ✓
2. Data Understanding (Current focus)
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Data Understanding **Goal**: Collect, explore, and validate data to ensure it aligns with the business problem.



What Does Data Understanding Involve?



1. **Data Collection:** Gather all **relevant** datasets (internal/external).



2. **Data Exploration:** **Understand** structure, variables, and relationships.



3. **Data Quality Check:** Identify **missing** values, **outliers**, inconsistencies.



4. **Data Validation:** Ensure data matches the **business problem**.

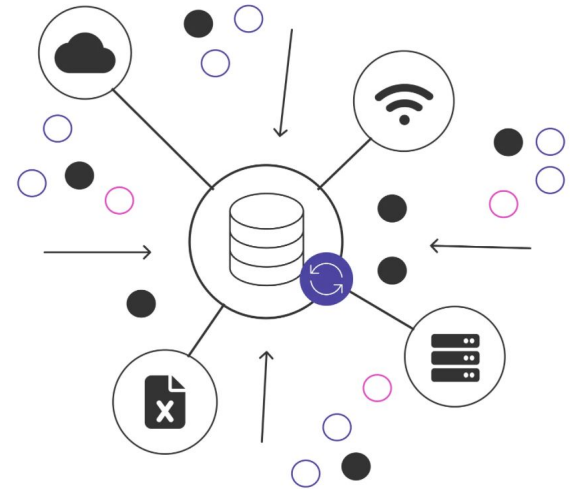
Where Does Data Come From?

Data can come from:

- **Internal Sources:** Databases, CRM, logs, sensors.
- **External Sources:** APIs, web scraping, public datasets (e.g., Kaggle, Open Data).

Types of Data:

- **Structured:** SQL tables
- **Unstructured:** text, images
- **Semi-structured:** JSON, XML



Exploring Your Data

Techniques:

- Summary **statistics** (mean, median, std dev).
- **Visualizations** (histograms, scatter plots, heatmaps).
- **Correlation** analysis.

Tools:

- Python (Pandas, Matplotlib, Seaborn)
- R (dplyr, ggplot2)
- SQL (for querying databases)

Red Flags in Data Quality

Issue	Example	Impact	Solution ***
Missing Values	30% of customer ages are null	Biased analysis	Imputation or removal
Outliers	A salary of \$1M in a dataset of \$50K	Skewed statistics	Investigate or cap outliers
Inconsistent Formats	Dates as "01/02/2023" and "Feb-1-23"	Errors in time-series analysis	Standardize formats
Duplicates	Same customer ID appears twice	Overcounting	Remove duplicates
Bias	Dataset lacks diversity	Unfair or inaccurate models	Collect more representative data

Does Your Data Answer the Business Question?

Checklist:

- Does the data cover the required time period?
- Are all relevant variables included?
- Is the data granular enough (e.g., daily vs. yearly)?
- Are there legal/ethical concerns (e.g., GDPR compliance)?

Example:

- Business Goal: Predict customer churn.
- Data Needed: Customer transactions, demographics, support tickets.

Validation is about ensuring the data can actually solve the problem.

What if your data lacks a key variable like customer satisfaction scores?

Workshop – Hands-On Data Exploration

Scenario: You're given a dataset of online retail transactions. Your goal is to understand customer purchasing behavior.

Tasks:

- Identify the **variables** and their types.
- Calculate **summary statistics** for numerical columns.
- **Plot** the distribution of one categorical variable (e.g., product category).
- Flag any **data quality issues**.

Tools: Use Python (Pandas/Seaborn) or Excel.

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom

Workshop Debrief

Variables: CustomerID (categorical), Quantity (numeric (categorical), invoiceDate (datetime).

Summary Statistics:

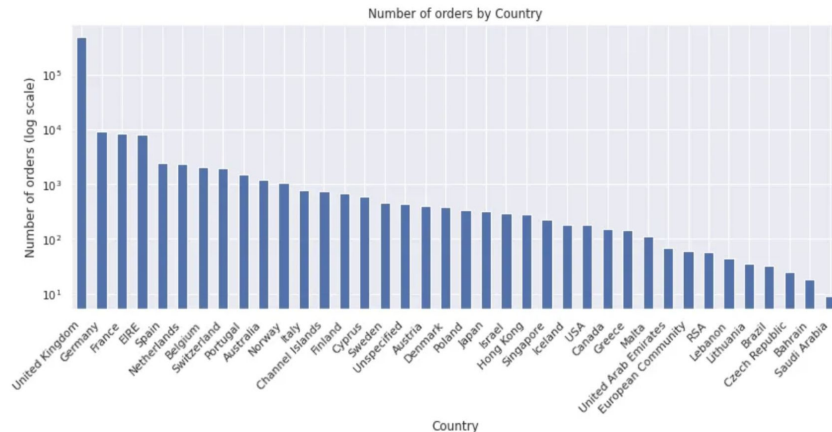
- Mean Quantity:
- Most frequent category: Electronics

Data Quality Issues:

- Missing values in CustomerID (134374 transactions)
- Negative value in Quantity (9719 transactions).
- Outliers...

Visualization: Histogram of Quantity shows right skew.

	Quantity	UnitPrice	CustomerID
count	534532.000000	534532.000000	401528.000000
mean	9.701713	4.644063	15281.185711
std	218.789309	97.424304	1714.034187



Tools to Master Data Understanding

Python Libraries:

- Pandas: data manipulation 
- NumPy: numerical operations 
- Matplotlib/Seaborn: visualization  

Other Tools:

- SQL (querying)
- OpenRefine (data cleaning)
- Tableau/Power BI (visualization)

Datasets for Practice:

- Kaggle: <https://www.kaggle.com/datasets>
- UCI Machine Learning Repository: <https://archive.ics.uci.edu/>

What to Remember

- Data Understanding is **iterative** – You'll revisit it as you learn more.
- Quality vs. Quantity
 - **clean, relevant** data beats **big, messy** data.
- Visualize Early, Visualize Often
 - Plots reveal insights that numbers can't.

