</talentlabs>

CHAPTER 2

The Data Analyst Workflow

Learning Objectives

- Summarize the steps included in the data analytics workflow
- Setup and introduce Google Sheets
- Work through the entire data analytics workflow

</talentlabs>

</talentlabs>

AGENDA

- Data Analytics Workflow
- Step 1: Understanding the problem
- Step 2: Data collection
- Step 3: Data wrangling
- Step 4: Data Analysis
- Step 5: Documenting the Process
- Step 6: Sharing Insights
- Conclusion

Data Analytics Workflow

- Summarise the data analytics workflow
- Mention common tools used by analysts
- Setup a Google Sheets account

</talentlabs>

Data Analytics Workflow

1. **Understand the problem and the desired outcome**
2. **Data collection**
   - Acquiring data that best serves the problem
3. **Data wrangling**
   - Optimising data for the task
4. **Data analysis and visualisation**
   - Look for insights
5. **Document the process**
   - Ensure process can be repeated
6. **Communicate insights to stakeholders**

➢ In this course we'll go through the entire workflow

</talentlabs>

Software

- Spreadsheets
  - Easy to learn
  - Well documented
- Python
  - Open source with powerful libraries (NumPy, Pandas and Matplotlib)
- Tableau
  - Interactive data visualisation
  - Fast
  - Easy to use
  - Drag and drop features
  - Works well with huge datasets

How to pick a tool for analysis?

- Use case
- Infrastructure
- Team
- Ease of use

</talentlabs>

Google Sheets

Spreadsheet program that is free, web-based and developed by Google. It is similar to Microsoft Excel.

Google Sheets

How to use Google Sheets?

Go to: sheets.google.com

Login or create an account

Start a new spreadsheet +

 </talentlabs>

Step 1: Understanding the Problem

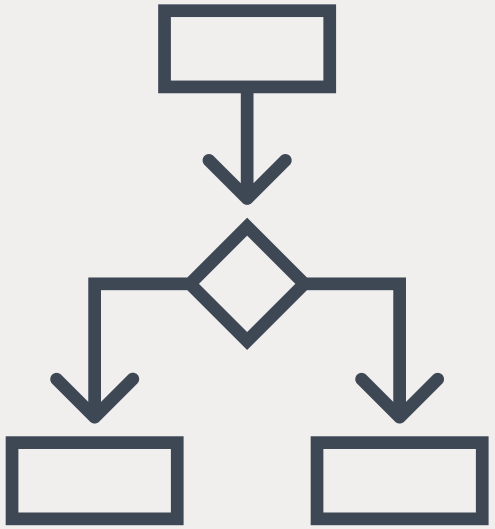
- How to go through the first step of the data analytics workflow
- Introduce the task for Chapter 2 analysis

</talentlabs>

Understanding the problem and the desired outcome

Where you are

Where you want to be



- Ask questions!

What questions to ask?

- What is the problem?
- What is the desired outcome?
- Who are the stakeholders?
  - Manager? Investors? Government Agency?
  - What is important to them?
  - What are their expectations?
  - How are they impacted by the problem?
  - What's their role in the business?
- Is there data available?

Start your analysis with a question

- Know what to solve
- Not just analysis for the sake of analysis

</talentlabs>

Your Task

- You are a freelancer hired to conduct an analysis on a collected dataset
- Dataset: Video Game Sales
  - Top 100 ever sold by volume
- Answer the following questions:
  1. What were the best and worst performing genres in total sales?
  2. Which game publisher was the most popular?
  3. What were the 5 most popular games in Japan?
  4. In which year were most of the games published?

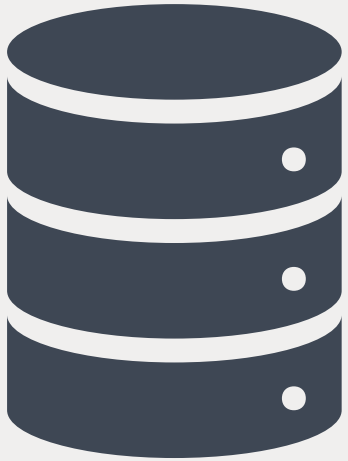
Step 2: Data Collection

- What does data collection involve?
- Types of data sources
- Loading our data into Google Sheets
- Setting up the file ready for data wrangling

</talentlabs>

Data Collection

- Gathering and storing data
- For larger projects: need to make a plan for the data collection process.
- Consider:
  - Time
  - Volume
  - Source



Data Sources

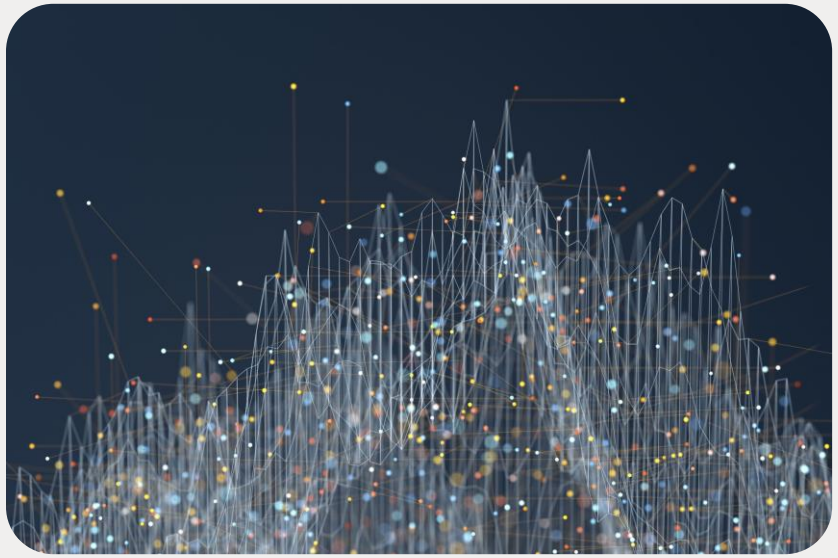
- Internal or external
  - Primary (original source)
  - Secondary (retrieved externally)
  - Tertiary (purchased)

Data Source Examples:

- The web
- Sensors
- Social Media

Security and Quality

- Data
- Source
- Collection method



</talentlabs>

Essentially:

- Collect data
- Get it ready for processing
- Preparation step

For our task:

- Dataset already collected
- Download + import + prepare

Our dataset

CSV file (comma-separated-values)

- Comes with Metadata

Metadata is data that provides information about other data

NA – North America

EU – Europe

Data Collection in Google Sheets

- Download the csv file
- Open Google Sheets
- Import the file
- Prepare the file

Google Sheets

 </talentlabs>

Loading a file

- Go to File
- Import
- Upload
- Select a file from your device

Google Sheets

 </talentlabs>

Prepare the file

- Name the sheet
- Move to suitable folder
- Delete whitespaces
- Resize columns
- Format headers
- Add borders

Google Sheets

 </talentlabs>

Prepare the file

Video Game Sales Raw Data

File Edit View Insert Format Data Tools Extensions Help Last edit was seconds ago

G5 | 3.73

	A	B	C	D	E	F	G	H	I
1	Rank	Name	Platform	Year	Genre and Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales
2	72	Donkey Kong Country	SNES	1994	Platform Nintendo	4.36	1.71	3	0.23
3	79	Wii Party	Wii	2010	Misc Nintendo	1.79	3.53	2.49	0.68
4	34	Call of Duty: Black Ops 3	PS4	2015		5.77	5.81	0.35	2.31
5	32	Call of Duty: Black Ops	X360	2010	Shooter Activision	9.67	3.73	0.11	1.13
6	81	Mario Party 8	Wii	2007	Misc Nintendo	3.81	2.3	1.58	0.73
7	36	Call of Duty: Black Ops II	X360	2012	Shooter Activision	8.25	4.3	0.07	1.12
8	54	Super Mario 3D Land	3DS	2011	Platform Nintendo	4.89	2.99	2.13	0.78
9	53	Gran Turismo	PS	1997	Racing Sony Computer Entertainment	4.02	3.87	2.54	0.52
10	39	Grand Theft Auto III	PS2	2001	Action Take-Two Interactive	6.99	4.51	0.3	1.3
11	63	Halo: Reach	X360	2010	Shooter Microsoft Game Studios	7.03	1.98	0.08	0.78
12	59	Pokemon FireRed/Pokemon LeafGreen	GBA	2004	Role-Playing Nintendo	4.34	2.65	3.15	0.35
13	18	Grand Theft Auto: San Andreas	PS2	2004	Action Take-Two Interactive	9.43	0.4	0.41	10.57
14	50	Pokemon Omega Ruby/Pokemon Alpha Sapphire	3DS	2014	Role-Playing Nintendo	4.23	3.37	3.08	0.65

Google Sheets

 </talentlabs>

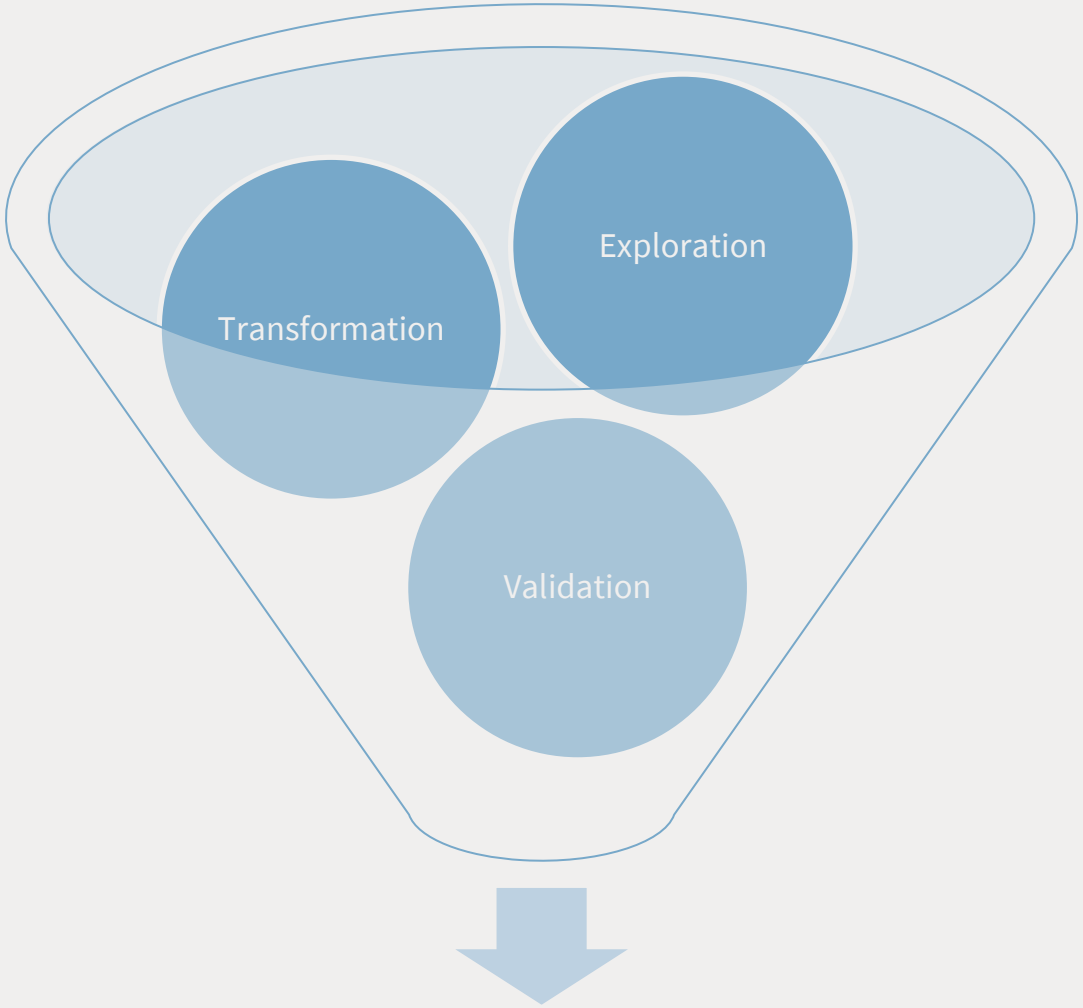
Step 3: Data Wrangling

- What does data wrangling involve?
- Preparing data for analysis in Google Sheets

</talentlabs>

What does Data Wrangling involve?

- Iterative process



</talentlabs>

Data exploration involves understanding what data you have and examining that data

</talentlabs>

Understanding the data

- Looking at each column of data

Date	Coffee	Price	Quantity
1/1/2022	Latte	2.5	3
2/1/2022	Cappuccino	3	2
3/1/2022	Latte	2.5	4

- Numeric (871, 3.02)
- Categorical (Type of coffee, car brands, country)

Examining the data

Explore data to see what transformations are required

Look out for:

- Missing values
- Errors (typos, inconsistent data entry, white spaces)
- Duplicates
- Data type
- Outliers
- Measurement Units

Transformations

- Bulk of the process
- Data cleaning

Validation

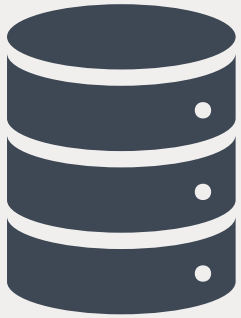
Assessing the quality of the data

Examining the data (iterative process)

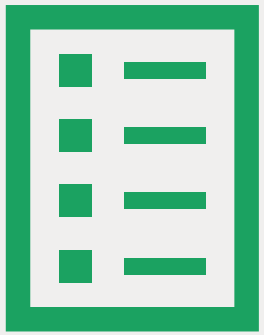
Output

Saved

Exported



Backed up



Documented

Data Wrangling in Google Sheets

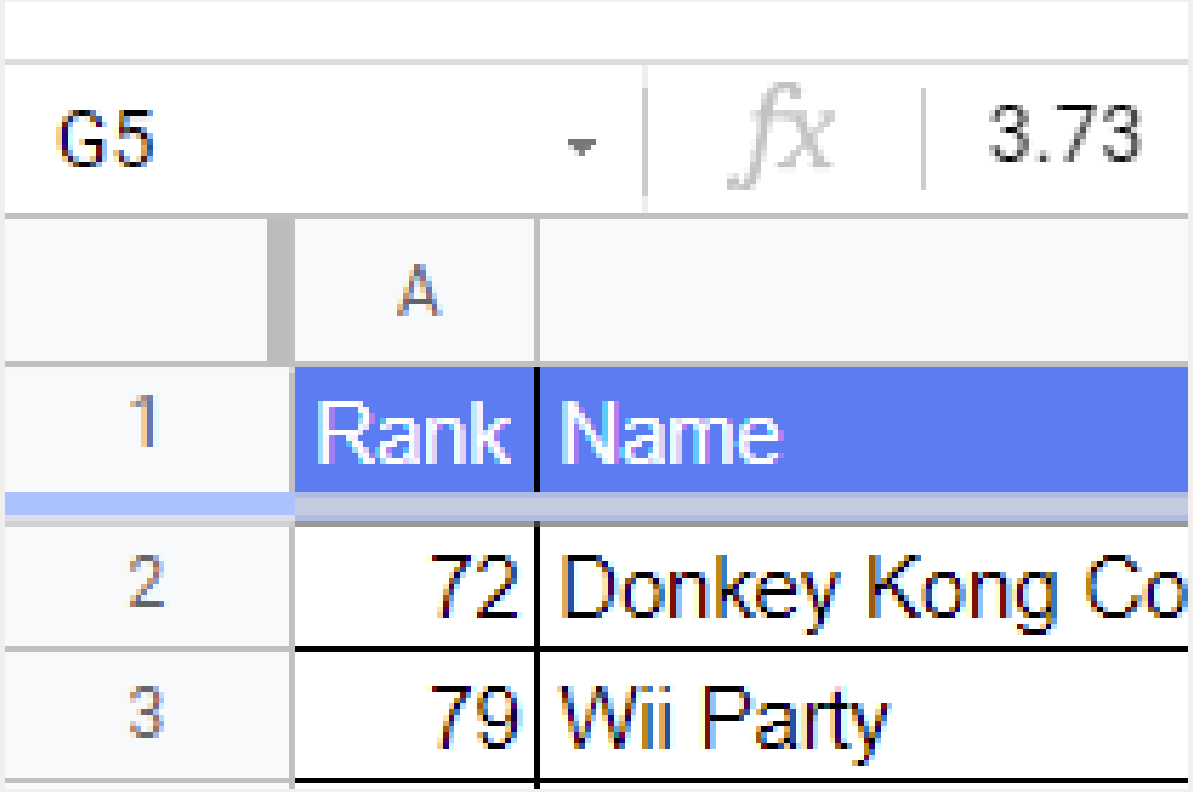
- Examine every column
- Transform the data

Google Sheets

 </talentlabs>

Data Wrangling in Google Sheets

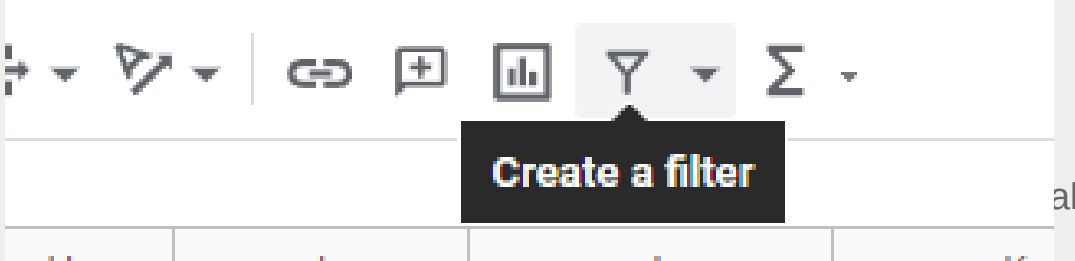
- Make header row (drag slider)



Google Sheets

Data Wrangling in Google Sheets

- Sort a column:
  - Right click selected column and choose ‘Sort Sheet’
- Remove duplicates:
  - Data > Data clean-up > Remove duplicates
- Filter values
  - Select row(s) to filter > press filter shortcut



Google Sheets

Data Wrangling in Google Sheets

- Split one column in two
  - Select column > Data > Split text to columns
  - Use LEFT and RIGHT functions
- Conditional formatting
  - Highlight cells to format > Format > Conditional Formatting

Google Sheets

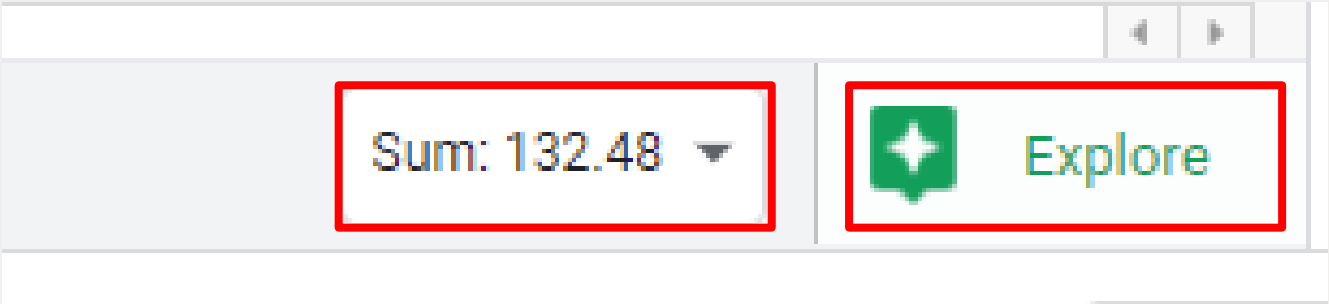
Step 4: Data Analysis and Visualisation

- Useful analysis features in Google Sheets
- Analyse our prepared dataset

</talentlabs>

Data Analysis in Google Sheets

- Quick insights and machine learning generated suggestions bottom right of Sheets



- Search Sheet shortcut: Ctrl + F

  (or Command + F on Mac)

Google Sheets

 </talentlabs>

Data Analysis in Google Sheets

Pivot Tables:

- Select whole dataset
  - Insert > Pivot table

Charts:

- Select data > Insert chart

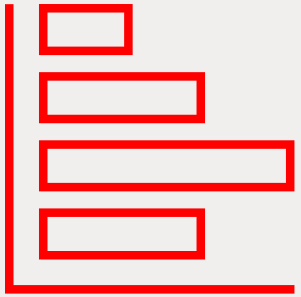
Google Sheets

 </talentlabs>

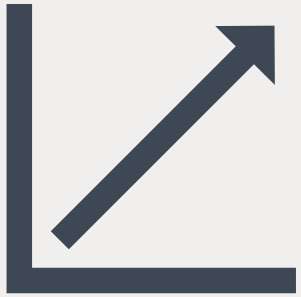
Common Types of Charts

Column

Pie



Bar



Line

1. What were the best and worst performing genres in total sales?
2. Which game publisher was the most popular?
3. What were the 5 most popular games in Japan?
4. In which year were most of the games published?

Step 5: Documenting the Process

- Why is documenting important?
- What should be documented?

</talentlabs>

Why is it important?

- Future reference
- Ensure process repeatable

</talentlabs>

What to document?

- What is the problem?
- How was the data source collected?
  - Metadata
- How did you transform the data for analysis?
- What analysis was carried out?
- What were the conclusions?

- What were the considerations behind each decision?

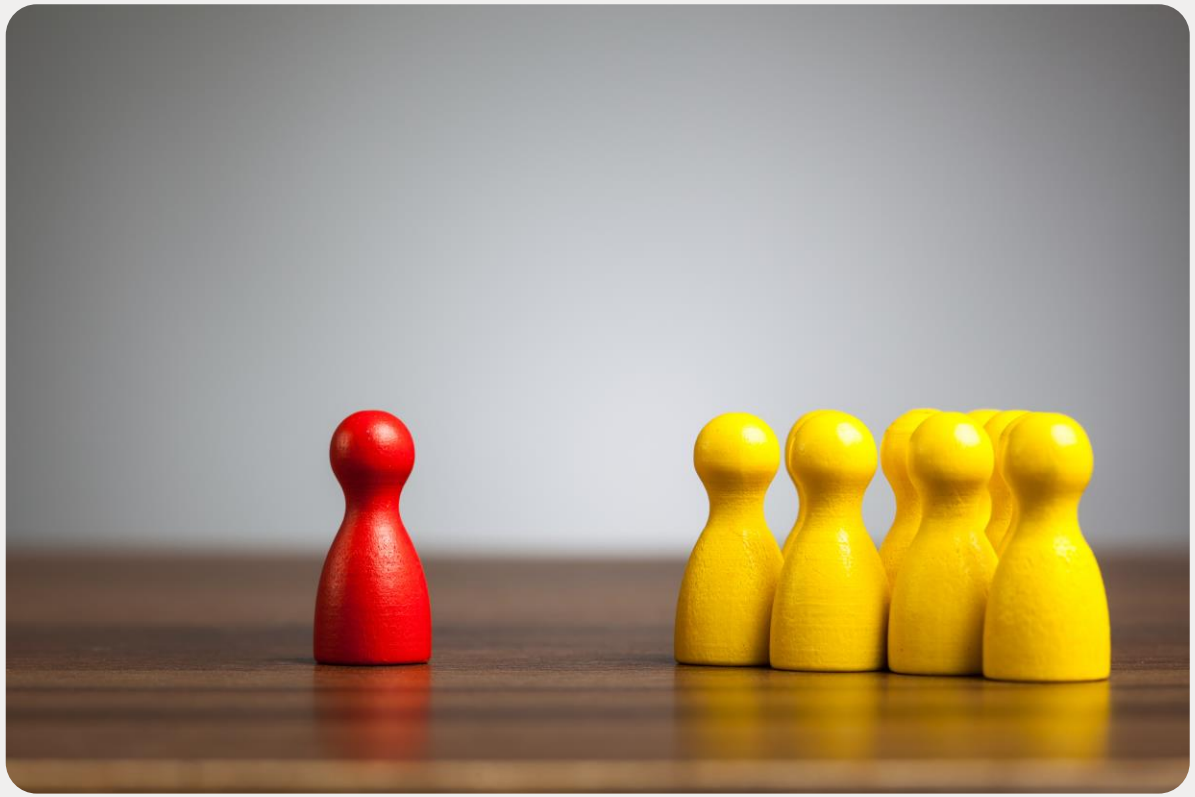
Step 6: Sharing Insights

- Why reporting your analysis is important to get right
- How to share your insights

</talentlabs>

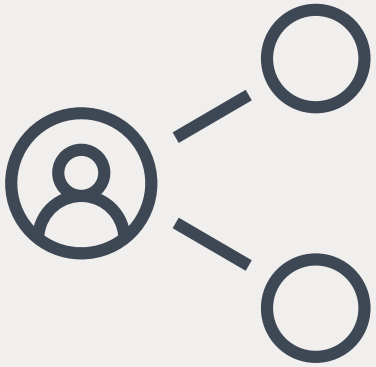
Why reporting your analysis is important to get right

- Stakeholders make **decisions**
- Informs decision making



How to share your insights

- What are the expectations?
  - Visualisation? Dashboard? Detailed report?
- Presentation is not a data dump
  - Story: Present the process
  - Include relevant information only
  - Visualisation – effective communication tool
- Use domain language of the industry



</talentlabs>

Conclusion

</talentlabs>

Data Analytics Workflow

1. **Understand the problem and the desired outcome**
2. **Data collection**
   - Acquiring data that best serves the problem
3. **Data wrangling**
   - Optimising data for the task
4. **Data analysis and visualisation**
   - Look for insights
5. **Document the process**
   - Ensure process can be repeated
6. **Communicate insights to stakeholders**

➢ In this course we'll go through the entire workflow

</talentlabs>

Software

- Spreadsheets
  - Easy to learn
  - Well documented
- Python
  - Open source with powerful libraries (NumPy, Pandas and Matplotlib)
- Tableau
  - Interactive data visualisation
  - Fast
  - Easy to use
  - Drag and drop features
  - Works well with huge datasets

How to use Google Sheets?

Go to: sheets.google.com

Login or create an account

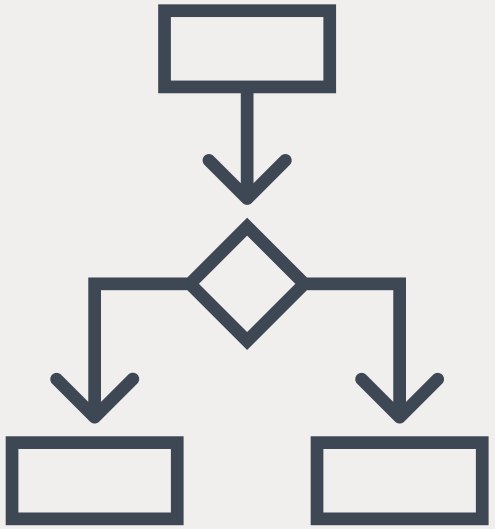
Start a new spreadsheet +

 </talentlabs>

Understanding the problem and the desired outcome

Where you are

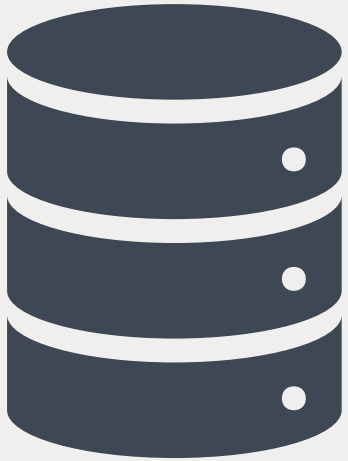
Where you want to be



- Ask questions!

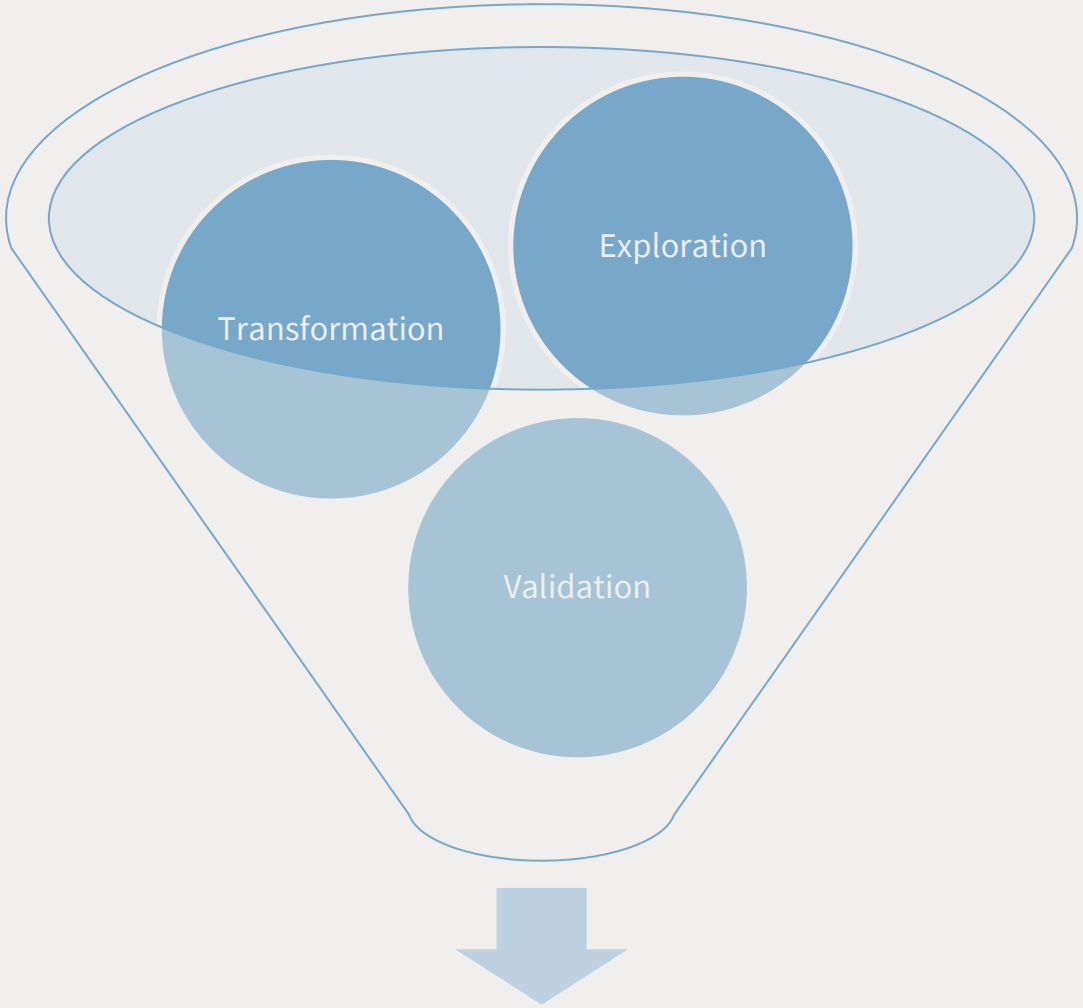
Data Collection

- Gathering and storing data
- For larger projects: need to make a plan for the data collection process.
- Consider:
  - Time
  - Volume
  - Source



What does Data Wrangling involve?

- Iterative process



</talentlabs>

Data Wrangling in Google Sheets

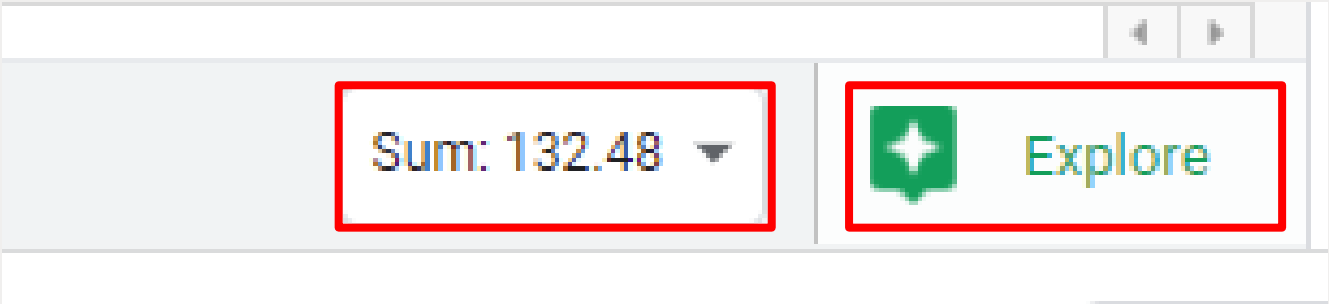
- Examine every column
- Transform the data

Google Sheets

 </talentlabs>

Data Analysis in Google Sheets

- Quick insights and machine learning generated suggestions bottom right of Sheets



- Search Sheet shortcut: Ctrl + F

  (or Command + F on Mac)

Google Sheets

 </talentlabs>

Data Analysis in Google Sheets

Pivot Tables:

- Select whole dataset
  - Insert > Pivot table

Charts:

- Select data > Insert chart

Google Sheets

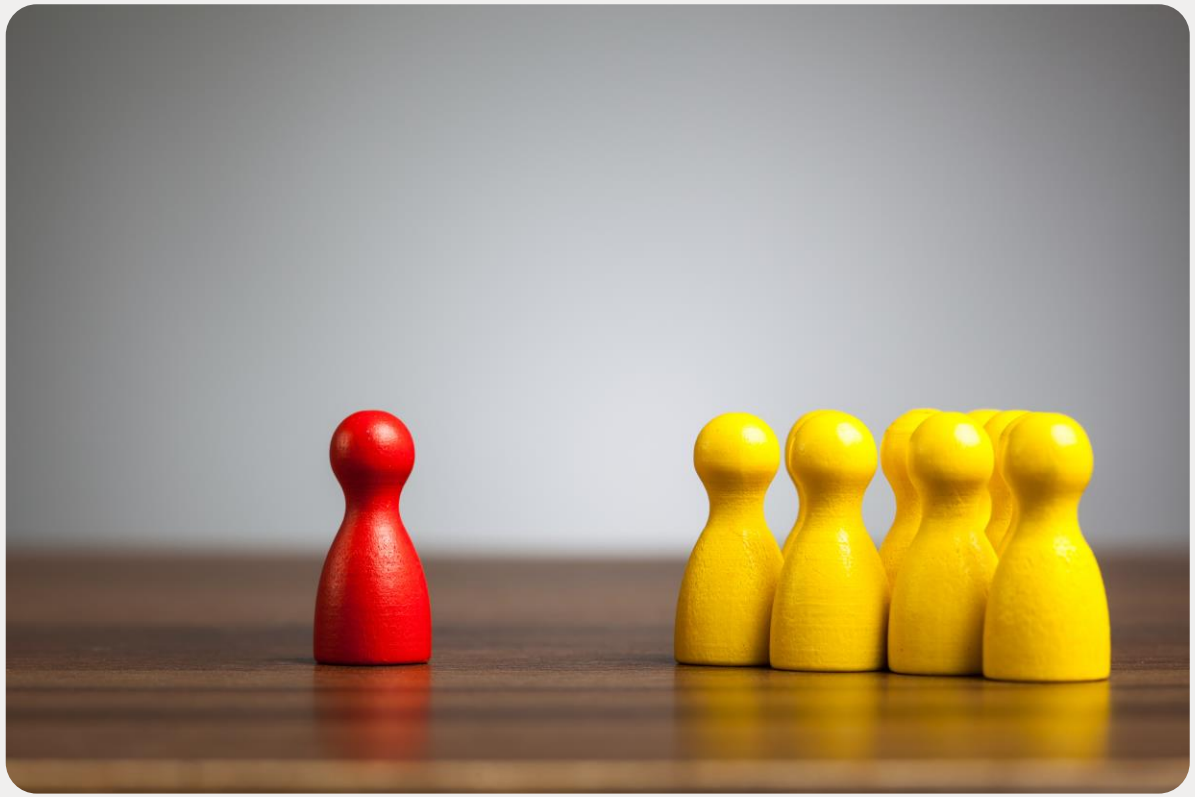
 </talentlabs>

Documentation

- Future reference
- Ensure process repeatable

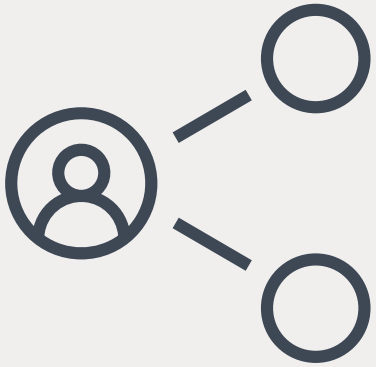
Why reporting your analysis is important to get right

- Stakeholders make **decisions**
- Informs decision making



How to share your insights

- What are the expectations?
  - Visualisation? Dashboard? Detailed report?
- Presentation is **not a data dump**
  - Story: Present the process
  - Include relevant information only
  - Visualisation – effective communication tool
- Use domain language of the industry



</talentlabs>

Next Chapter

Google Analytics

Assignment

- First **five** questions are **theory** based.
- Next **five** questions are a continuation of the **analysis** in Google Sheets of the Video Game Sales dataset that we cleaned in this chapter.