



DATA ANALYSIS

Lecturer: Mark Gillan



DATA ANALYSIS DEFINED

Data Analysis: the process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, informing conclusions and supporting decision-making.

Data analysis involves the transformation of raw data into useful information or insights in a structured and organised way.



ANALYSIS STEPS

It is generally accepted that around 80% of any data analyst's time is spent cleaning and manipulating data. These activities are both important and time consuming. The other activities involve detailed understanding of the dataset before any manipulation and ensuring that conclusions are drawn, and actions are taken at the end of the process.

There is a structured approach to carrying out a data analysis, which, if followed, will minimise mistakes and maximise the validity of the conclusions or insights extracted from the data. This is what would be done within **PPDAC's analysis step**:



DATA UNDERSTANDING

All analysis activities should start with an understanding of the data being analysed.

This allows the analyst to get a feel for the data, understand its size and shape and identify any obvious issues with the data that may need to be addressed.



DATA UNDERSTANDING

Visual Inspection

By visually inspecting the data as a first step, it will give a feel for the size of the data, the frequency of missing values and the data types and formats.

Summary Statistics

Summary statistics for each variable should be created. This will be different depending on whether the variable is qualitative or quantitative.

Size and Shape

Information captured at this point should include the number of rows, the number of columns, the column names and the column formats.

sale	product	quantity	price
1	Apple	6	41
2	b@n@na	2	
3	kiwi	3	46
4	lemon	1	28
5			41

Initial observations:

- Mixture of numeric and character data
- Missing values
- Inconsistent format of product

shape	number
Number of rows	5
Number of columns	4

variable name	variable type
sale	integer
product	character
quantity	integer
price	float

var	mean	min	max	missing
sale	3	1	5	0
quantity	3	1	6	1
price	39	28	46	1

var	distinct values	missing
product	4	1



DATA DICTIONARIES



Data dictionaries are often a good place to start to understand the data being analysed.

Data dictionaries contain a lot of information about the dataset, not only explanations about the variables, but also formats, valid values and allowed ranges.

Data dictionaries can also provide valuable information about the required security classifications and retention periods.





TIDYING AND CLEANING DATA

Tidying, Cleaning, Fixing

EXAMPLE OF DATA ISSUES REQUIRING A FIX



Plain Unicode Text file named:

Datalssues.txt

Make a note of inconsistencies and mistakes requiring fixes or queries.



TIDYING DATA

The first step in preparing data is to tidy it up.

There are a number of activities that this could involve, including:

- **Removing metadata** at the top of a file, by either deleting or skipping rows to allow the dataset to be successfully read in.
- **Naming or renaming columns** so that the data is easily understood, and the names are informative.
- **Dropping unnecessary columns** so that only those required for the analysis remain. This saves disk space and speeds up processing.
- **Reformatting columns** so that numbers and dates/times are stored correctly and not as strings.
- **Fixing strings** so that the case (upper/lower) and spaces are consistent, allowing comparison.
- **Rescaling data** so that numeric values fit within a specific range.



FIXING STRINGS

An example of some string data that might require fixing is: “Star Wars”, “STARWARS”, “star_wars” and “Star wars”. A computer, being case-sensitive will not interpret these as the same.

Options for fixing strings include:

- **Upper case:** THE WHOLE STRING IS IN UPPER CASE
- **Lower case:** the whole string is in lower case
- **Title case:** The First Character of Each Word is Capitalised, Except Common Words
- **Capital case:** The First Character Of Each Word Is Capitalised
- **Sentence case:** Everything is lower case except the first character

Common options are upper and lower case, any can be chosen but must be consistent.



VARIABLE NAMING CONVENTIONS

A variable name should be descriptive and summarise what is contained in the column. A good variable name makes it easy for someone else to read through code and understand what is going on. Names such as “Var1” and “Var2” should be avoided.

The best practices for naming variables in datasets include:

- **Avoid whitespace in a name**: this is because a name such as “First Name” will always need to be referred to in code using quotes
- **Stick to lower case names**: most computer languages are case sensitive and by sticking to a consistent case ensures that errors are less likely to occur
- **Use a consistent style**: there are a number of common styles to choose from:
 - **Snake case (snake_case)**: this puts an underscore _ between each word e.g. first_name
 - **Camel case (camelCase)**: this capitalises the first letter of each word, except the first one e.g. firstName
 - **Pascal case (PascalCase)**: this capitalises the first letter of each word, including the first one e.g. FirstName
 - **Kebab case (kebab-case)**: this puts a dash - between each word e.g. first-name



MISSING VALUES AND OUTLIERS



Missing data includes empty cells, some zeroes, blank strings and sometimes system default values.



There may be many reasons why data may be missing.



A question may not have been asked, or the system may not have captured the response.



Without knowing the real reason for missing data, it is difficult to decide the best approach to deal with missing values.



MISSING VALUES AND OUTLIERS (CONTINUED)

Similarly, there are several causes for outliers which can be both natural and due to system and human error:

- **Data entry errors:** entering a value in grams when the system expects kilograms will lead to an outlier
- **Measurement error:** the instrument has incorrectly measured the value
- **Experimental error:** incorrect data extraction or experimental approach
- **Intentional error:** misreporting of information
- **Data processing errors:** incorrect manipulation of data prior to capture and use
- **Sampling:** incorrect merging of datasets
- **A real outlying data point**



DEALING WITH MISSING VALUES AND OUTLIERS

There are four common approaches to deal with missing data and outliers.

Good data analysts will make active decisions around which method to use.

They will decide why it is suitable for the specific situation:

Whichever approach is chosen, the decision should be documented and justified.

Remove

- Remove the record

Treat

- Treat missing records separately

Input

- Input a suitable value e.g. average, median, best fit

Do

- Do Nothing

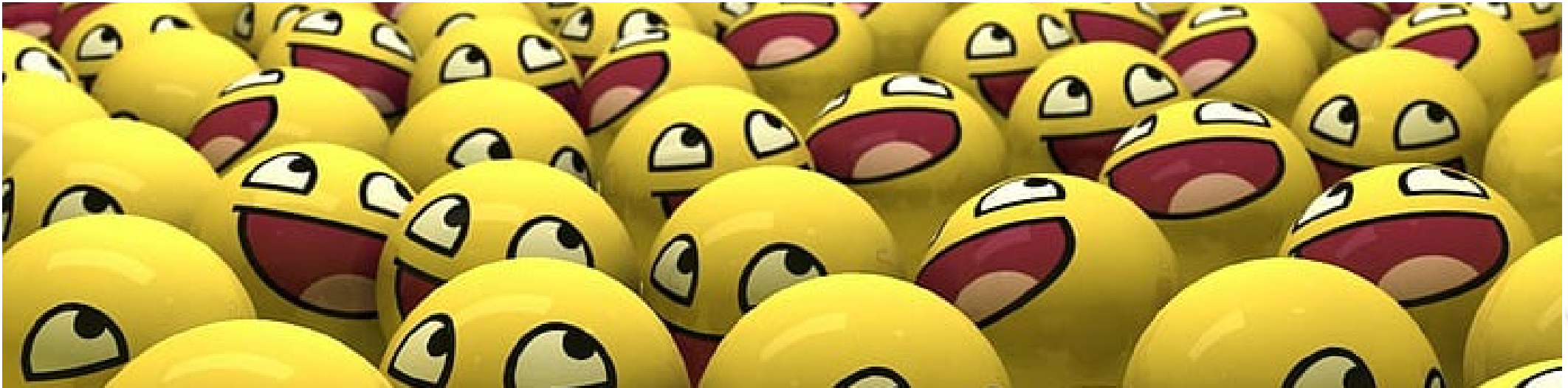


DUPLICATES

Duplicates can be exact copies of observations, that can be tidied up by deleting the additional copies.

However, duplicates can also be similar records, caused for example by an individual responding multiple times. In the case of partial duplicates, it is common for one version to be active and regularly updated and the remaining copies to remain unused.

In these cases, care should be taken when deduplicating to ensure the main active record is not deleted. The records should ideally be merged into a single active record prior to removal of the copies.



MANIPULATION



In addition to the common data manipulations and joins covered in the **Data Manipulation** section and the graphical display of data covered in the **Visualising Data** section it may also be useful to **cross tabulate** between variables to uncover hidden relationships.

It is normally carried out on categorical data.

This is the type of functionality that Excel's pivot table provides.



CROSS-TABULATION

Here is an example of the Titanic survival dataset (dataed.in/titanic)

		Survived		
		No	Yes	Total
Class	1st	122	203	325
	2nd	167	118	285
	3rd	528	178	706
	Crew	673	212	885
	Total	1490	711	2201

The outcome of all 2201 passengers has already been cross-tabulated by class, sex and age:

