University of Lancashire

Data Cleaning

Karen O’Shea

Where opportunity creates success

Starter Activity

- Why is data cleaning important in data analysis?
- Can you list common data quality issues?

Learning Objectives

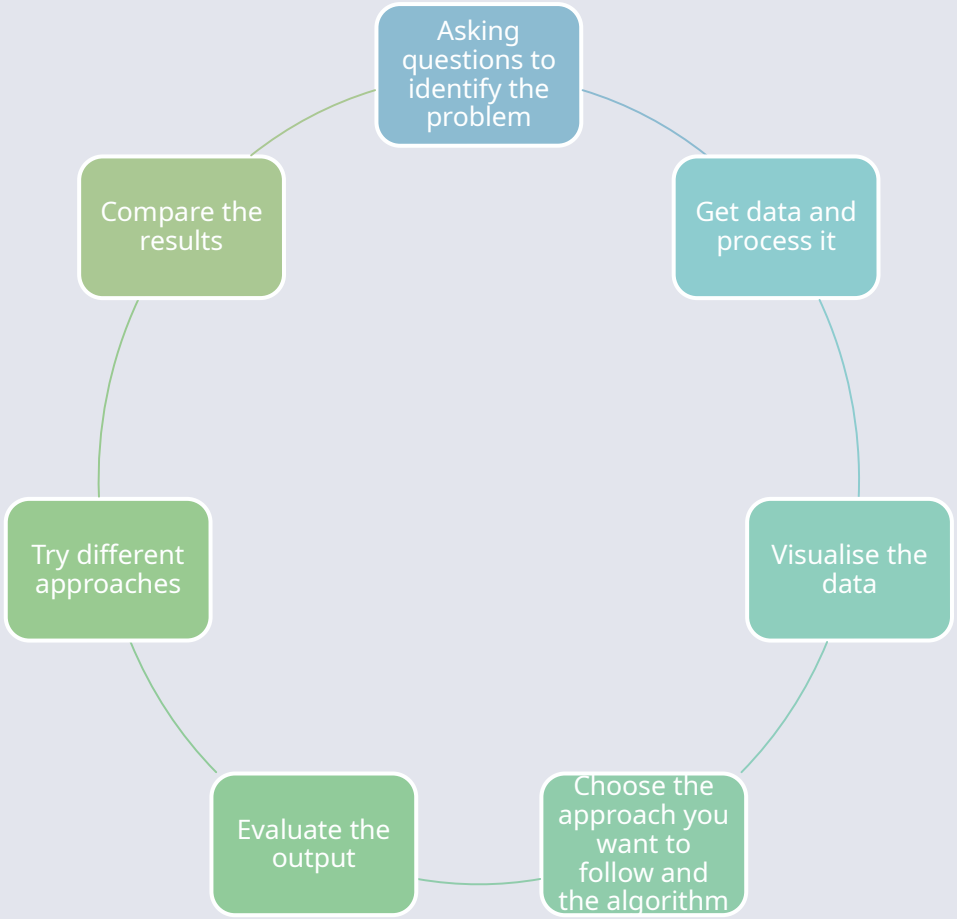
University of Lancashire

Discuss the process for ‘Data Cleaning’

Evaluate a dataset for quality control

University of Lancashire

Recap: Data Science: Workflow Model



Recap: Data Science Workflow

Recap: Import Dataset and Libraries

University of Lancashire

```
import pandas as pd
import matplotlib as plt
import numpy as np
```

```
df=pd.read_csv("train.csv")
```

*Reads dataset into DataFrame ready for viewing

Recap: Data Exploration

- df.head(15)

  Will display 15 rows

- df.describe()

  Will provide you with mean, quartile, count, min, max and standard deviation as well as their outputs.

- df['Property_Area'].value_counts()

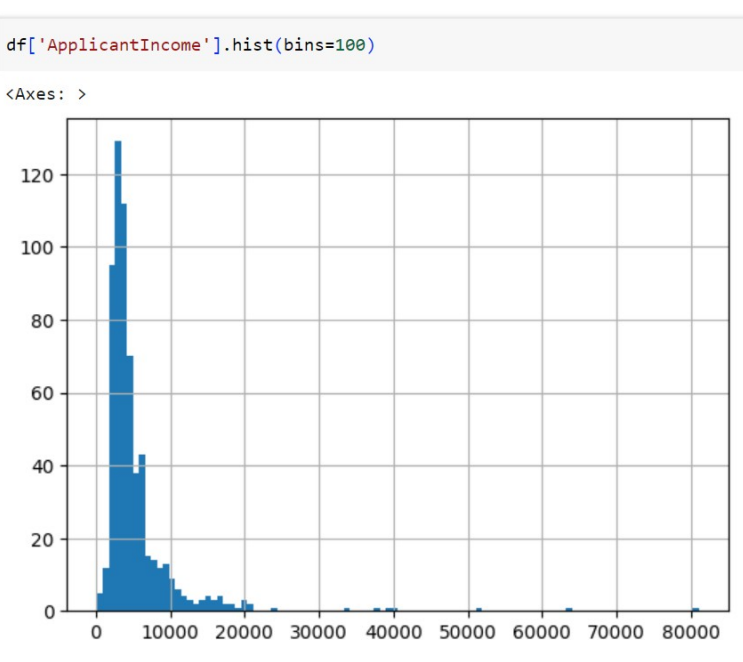
  Accessing a particular column in the data frame

Recap: Distribution Analysis

- df[‘ApplicantIncome’].hist(bins=100)

  Using a histogram to observe extreme values

- Consider data cleaning:
  - Resolving problems with the dataset (ie. missing values, extreme values)
  - How?



Reflect: Data Cleaning…

University of Lancashire

What is data cleaning? Why bother?

Key terms:

Pre-processing
Encoding
Filtering

Can you propose a series of steps to prepare for data cleaning….?

Reflect:
Data Cleaning…

Data Analysis?

- Only as good as your data
- Garage data in, garbage data analysis out

- Quality data decision-making!

So What is Data Cleaning?

- The process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset
- If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct
- There is no one absolute way to prescribe the exact steps in the data cleaning process
- Establish a method for good practice

How to Clean Data – Basic Template

University of Lancashire

Remove Unwanted/Duplicate Values

University of Lancashire

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations.

- Duplicate observations will happen most often during data collection.
- May occur when you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments.

Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyse.

- For example, if you want to analyse data regarding millennial customers, but your dataset includes different generations, you might remove those irrelevant observations.

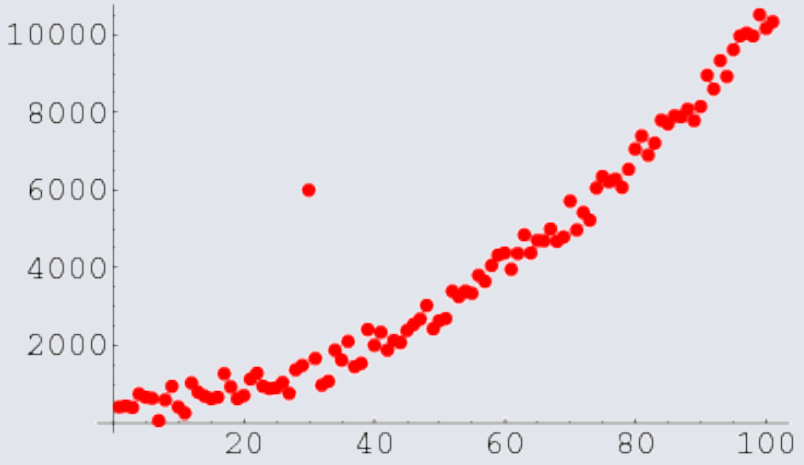
Fix Structural Errors

- Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalisation. These inconsistencies can cause mislabelled categories or classes:
  - For example, you may find “N/A” and “Not Applicable” both appear, but they should be analysed as the same category.

Filter Unwanted Outliers

- Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analysing.
  - If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with.

- Sometimes it is the appearance of an outlier that will prove a theory you are working on.
  - Remember: just because an outlier exists, doesn't mean it is incorrect. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.



Handle Missing Data

University of Lancashire

You can’t ignore missing data because many algorithms will not accept missing values.

- **As a first option**: you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
- **As a second option**: you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
- **As a third option**: you might alter the way the data is used to effectively navigate null values.

Quality Control

Validate:

- Does the data make sense?
- Does the data follow the appropriate rules for its field?
- Does it prove or disprove your working theory, or bring any insight to light?
- Can you find trends in the data to help you form your next theory?
- If not, is that because of a data quality issue?

False conclusions because of incorrect or “dirty” data can inform poor business strategy and decision-making. Does your data stand up to scrutiny.

* What does data quality mean to you?

Components of Quality Data

University of Lancashire

Validity: The degree to which your data conforms to defined business rules or constraints.

Accuracy: Ensure your data is close to the true values.

Completeness: The degree to which all required data is known.

Consistency: Ensure your data is consistent within the same dataset and/or across multiple datasets.

Uniformity: The degree to which the data is specified using the same unit of measure.

Summative Assessment Reflection

- What is your understanding of the assignment brief?
- What are the key sections that you think are important at this stage?
- Can you propose some sample headings and format for structuring report?

Learning Objectives

University of Lancashire

Discuss the process for ‘Data Cleaning’

Evaluate a dataset for quality control