

Basic EDA of two datasets -Report Structure for week 2

1. Dataset Overview

Item	Description
Dataset name	
Number of rows	
Number of columns	
Format file (.csv, .txt, etc)	
Authors of the dataset	
Source (name)	
Source (link)	

Short description (what is it about?)

2. Structure of the dataset

Column name	Data type	Non-null count	Unique values	Example values

3. Descriptive statistics

Numeric columns

	Column 1	Column 2	Column 3
Count			
Mean			
Standard deviation			
Min			
25%			
50%			
75%			
Max			

Categorical/object columns

	Column 1	Column 2	Column 3
Count			

Number of unique values			
Most frequent value			
Most frequent value (frequency)			
Least frequent value			
Least frequent value (frequency)			

3. Missing values and duplicates

Column name	Missing count	% Missing

Total missing values:

Percentage of dataset affected:

Duplicated rows found:

Percentage of rows in dataset affected:

4. Data consistency

Item	Description
Does the dataset contain unnecessary columns? Which?	
Do the data types correspond to the columns?	
Is the labelling of the columns appropriate?	
Are there mixed values in column (e.g., number and characters)?	
Are string column clean?	
Does the dataset look machine generated?	
Other	

5. Overall assessment

Is it worth it to further analyze the dataset?

What possible analysis can be performed?

6. Next steps

- Handling missing values? How?
- Removing duplicates?
- Cleaning the columns? Which?
- Creating a subset of the dataframe?