Reference card

Data quality checklist

Data quality

To ensure that the data are **accurate**, **complete**, **consistent**, and **reliable**, check for the following potential data issues:

☐ Missing data

Null - blank / no value at all
NaN - unrepresentable values

Watch out for informative missing values.

Date	Commodity	Unit	Kes_price	Usd_price
2/15/2022	Beans	KG	35.63	0.32
2/15/2022	Maize	KG	40.34	0.36
2/15/2022	Oil	L		
2/15/2022	Potatoes	KG	24.98	0.22
2/15/2022	Milk	L	110.00	0.97

Potential solutions:

- **Drop** observations/features with the missing values.
- **Fill** in with an appropriate estimate using imputation.
- **Flag** as missing by filling in with a uniform value.

☐ Duplicate observations

Repeated entries: **exact duplicates**, **near-duplicates**.

Watch out for seemingly duplicate values.

Date	Commodity	Unit	Kes_price	Usd_price
2/15/2022	Beans	KG	35.63	0.32
2/15/2022	Beans	KG	35.63	0.32
2/15/2022	Maize	KG	40.34	0.36
2/15/2022	Maize	50 KG	1480.00	13.10
2/15/2022	Oil	L	115.00	1.02
2/15/2022	Potatoes	KG	24.98	0.22
2/15/2022	Milk	L	110.00	0.97

Potential solutions:

- **Remove** the duplicate and retain only one occurrence.
- **Merge** the duplicate observations.

☐ Unwanted outliers

Values that are **significantly different** from the rest.

Watch out for useful outliers.

Date	Commodity	Unit	Kes_price	Usd_price
2/15/2022	Beans	KG	35.63	0.32
2/15/2022	Maize	50 KG	1480.00	13.10
2/15/2022	Oil	L	115.00	1.02
2/15/2022	Potatoes	KG	24.98	0.22
2/15/2022	Milk	L	110.00	0.97

Potential solutions:

- **Identify** the outlier by plotting the data, using statistical measures or domain knowledge.
- **Delete** the observation with the outlier.
- **Replace** the outlier with a more representative value.

☐ Irrelevant observations

Data that **do not contribute** to our analysis.

Watch out for interdependency.

Date	Commodity	Unit	Kes_price	Usd_price
#date	#item+name	#item+unit	#value	#value+usd
2/15/2022	Beans	KG	35.63	0.32
2/15/2022	Maize	50 KG	1480.00	13.10
2/15/2022	Oil	L	115.00	1.02
2/15/2022	Potatoes	KG	24.98	0.22
2/15/2022	Milk	L	110.00	0.97

Potential solutions:

- **Retain** to contribute to the story, but not to use for analysis.
- **Delete** the irrelevant values/rows/columns.

☐ Structural issues

Inconsistencies and **typing errors** within the data itself.

Date	Commodity	Unit	Kes_price	Usd_price
2/15/2022	Beans	KG	35.63	0.32
2022-02-15	Maize	50 KG	1480.00	13.10
2022-02-15	Oil	L	115.00	1.02
2/15/2022	Potatoes	KG	24.98	$0.22
2/15/2022	Milk	L	110.00	$0.97

Potential solutions:

- **Validate** data entries.
- **Standardise** the data: case, naming conventions, measurements, date formats, padding.
- **Convert** to the correct file format and data types.
- **Spell check** to correct typos.