

▼ Data Quality

▼ Rows and Columns

▼ Rows

Descriptions

What does the row mean? Is the row for a shipment, combined shipment, etc?

Count

How many rows? How many unique values?

▼ Columns

Descriptions

What does the column mean?

Count

how many columns?

▼ Types and Formats

▼ Data Types

what is each column consists of?

Types of data

Ordinal, Generative, etc

Is the type of the data correct

list, tuple, datetime, int, float, str, etc

...

▼ Data Formats

Are the dates loaded as dates?

Are the numbers loaded as numbers?

Are they strings?

Are the financial values correct?

str or numbers? EU format, US format?

▼ Missing Values

Are there missing values in each column

Percentage

Visualizations

e.g., missingno python package

▼ Different types of missing values

Standard missing values

nan, nat, None, na, null...

Represented with a specific value

-1, 0, MISSING, ...

▼ Duplications

Are there duplications of rows/columns?

Duplications of fields when the documentation says they are unique

▼ Distributions

▼ What is the generation process

Is it a histogram analysis of another row?

Is it a linear combination of other rows?

...

▼ Visualize the distribution of the values

Value count bar plot

For discrete data, list all possible values and counts

Histogram and KDE

for continuous data, use histograms or KDE.

Boxplot

Boxplot is easier to understand for business people

Scatter plot

Gut feeling of where the data points are located

Contour plot

▼ Numerical Summarization

Locations

Mean, median, quartiles, mode...

Spreads

range, variance, standard deviation, IQR

Skewness

asymmetries

Kurtosis

▼ Correlations, Similarities

Pairplot

▼ Correlations

Pearson

Kendall Tau Correlation

▼ Distances

Euclidean distance

Mahalanobis distance

Minkowski distance

Jaccard distance

▼ Size

How much space will the data take on our storage device?

In memory

Different Formats

▼ Combining Data Files

One dataset may come in different files

▼ Concat

The files should be concated

Validate overlap

Are there overlaps between the files?