Data Cleaning Checklist

datacamp

Data cleaning takes up 80% of the data science workflow. Use this checklist to identify and resolve any quality issues with your data

Checklist	Examples in Action	Potential Solutions

Data Constraints Problems

Data type constraints

Ensuring that different columns have the correct data type before beginning analysis.

Example A `revenue_usd` column that is a string, and not a numeric data type.

revenue_usd	revenue_usd
$1,000	1,000
$3,210	3,210

- ✓ Convert to the correct data type

Data range constraints

Ensuring that different columns have the correct range. This is especially the case for columns that have limits.

Example A `gpa` column should be constrained to [0.0, 4.0]

gpa
3.1
4.0
5.1

- ✓ Check for typos, like a decimal point in the wrong place.
- ✓ Drop rows where data points break range constraints
- ✓ Set the data point that breaks range constraints to the maximum, or minimum
- ✓ Treat the data point that breaks range constraints to missing, and impute it

Uniqueness Constraints

Ensuring that there are no exact or almost exact duplicates within your rows.

Example A duplicate row where the name and `phone_number` columns are identical, but not the `height_cm` column

name	height_cm	phone_number
Carl Rosseel	177	(555) 200-5598
Carl Rosseel	178	(555) 200-5598

- ✓ Keep only one of the exact duplicate rows
- ✓ Merge rows that have non-exact duplicate rows

Text and Categorical Data Problems

Membership constraints for categorical data

Ensuring that categorical columns have correct and consistent categories

Example Two different entries for “New York” in the `city` column

name	city
Carl Rosseel	New York
Sara Billen	New York City

- ✓ Drop rows that are affected by inconsistent categories
- ✓ Remap inconsistent categories to the correct category name
- ✓ Infer categories based on other data points if it’s not clear how it should be remapped

Length violation for text data

Ensuring that text columns that follow a specific standard have the same string length

Example A `phone_number` column that is 9 characters instead of 14

name	height_cm	phone_number
Carl Rosseel	177	(555) 200-5598
Carl Rosseel	178	(555) 200-5598

- ✓ Drop rows that are affected by length violation
- ✓ Set affected observations to missing

Text data inconsistent formatting

Ensuring that text columns that follow a specific standard have the same string formatting

Example A `phone_number` number column that contains different phone number formats

name	height_cm	phone_number
Carl Rosseel	177	(555) 200-5598
Carl Rosseel	178	(555) 200-5598

- ✓ Standardize formatting for affected observations
- ✓ Drop rows that are affected by the inconsistency

Data Uniformity Problems

Unit uniformity for numeric columns

Ensuring that numeric columns have the same units (Temperature being Celsius, or Fahrenheit across all observations. This is especially relevant when joining datasets from different countries or sources.)

Example A `temperature` column in celsius that has absurdly high or low temperature values

date	city	temperature
05-18-2022	New York	27
05-18-2022	New York	80.6

** please note that absurdly high or low temperature values can be caused by other data quality issues, such as sensor malfunctions*

- ✓ Dropping rows where no context on units appears and don’t pass a sanity check
- ✓ Standardize the units where possible

Unit uniformity for date columns

Ensuring that date columns have the same datetime format

date	birthday
05-18-2022	Carl Rosseel
05-19-2022	Sara Billen
20-05-2022	Isabella Leslie-Miller

- ✓ Standardizing datetime formats where possible
- ✓ Dropping rows where no context on datetime format appears and don’t pass a sanity check

Crossfield validation for numeric columns

Crossfield validation is when we use multiple fields in a dataset to ensure the validity of another. For example, ensuring that part to whole columns add to a relevant total (Flight bookings per class add up to the total recorded bookings)

date	economy	first class	total
05-18-2022	250	50	300
05-19-2022	200	50	200

- ✓ Dropping rows where sanity checks fail
- ✓ Apply rules from domain knowledge based on knowing the data

Crossfield validation for date columns

Ensuring that date and temporal columns pass sanity checks (for example, ensuring that webinar registration dates always precede webinar attendance dates)

name	date of birth	age
John Doe	02-07-1994	27
Jane Doe	10-12-2000	34

- ✓ Dropping rows where sanity checks fail
- ✓ Apply rules from domain knowledge based on knowing the data

Missing Data Problems

Missing Completely at Random Data

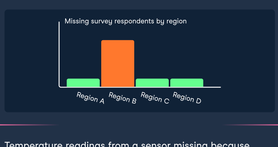
When there is no systematic relationship between missing values and other values within the dataset

There is no observed relationship between missing data and other values within the dataset

Missing at Random Data

When there is a systematic relationship between missing data and other observed values

Missing census data from a specific region, because the postal service doesn’t have full coverage in that region



- ✓ Drop missing rows
- ✓ Impute missing rows with measures of centrality such as median or mean
- ✓ Impute missing rows with algorithmic, machine-learning based approaches
- ✓ Collect new data points and features

Missing Not at Random Data

When there is a systematic relationship between missing data and other unobserved values

Temperature readings from a sensor missing because temperate was too low, or high