

Summary

Data cleaning is the process of data manipulating so that the data will be properly analyzed. Data cleaning involves handling duplicated data, handling missing data, and editing data.

Duplicates

Duplicated data makes our data unreliable and is often removed via `df.drop_duplicates()`.

PANDAS DROP_DUPLICATES DELETES DUPLICATE ROWS FROM A DATAFRAME

name	region	sales	expense
William	East	50000	42000
William	East	50000	42000
Emma	North	52000	43000
Emma	West	52000	43000
Anika	East	65000	44000
Anika	East	72000	53000

Irrelevant / unwanted data

There may be data that is not relevant or that you do not want to include in your analysis. Getting rid of unwanted data may be done using `df.drop(columns = [])` or selecting data that you want to drop by boolean indexing.

```
import pandas as pd

data = {
    "name": ["Sally", "Mary", "John"],
    "age": [50, 40, 30],
    "qualified": [True, False, False]
}

df = pd.DataFrame(data)
print(df)
print()

newdf = df.drop("age", axis='columns')
print(newdf)
print()

newdf2 = newdf.drop(newdf[newdf['qualified']== False].index)
print(newdf2)
```

Result Size: 729 x 553 [Get your own Python server](#)

	name	age	qualified
0	Sally	50	True
1	Mary	40	False
2	John	30	False

	name	qualified
0	Sally	True
1	Mary	False
2	John	False

	name	qualified
0	Sally	True

Missing data

There are multiple ways to deal with missing data. Getting rid of rows that contain missing data is an option to be considered when there isn't a lot of missing values. Filling in the missing values may be another option if reliable predictions could be made from the data that isn't missing. Dropping a column is also an option if most of the column consists of missing values. Some important methods are `fillna()` , `dropna()`, `ffill()`, `bfill()`, `replace()`

Editing data

Some data must have the same format or a specific dtype for analysis to be done properly. Some popular methods for doing that are: `apply()`, `replace()` , `astype()`,`str.strip()`,`str.split()`...