```
from google.colab import files

uploaded = files.upload()
```

```
Choose files sales_data_sample.csv
sales_data_sample.csv(text/csv) - 527958 bytes, last modified: 01/02/2026 - 100% done
Saving sales_data_sample.csv to sales_data_sample.csv
```

Handling missing values

```
import pandas as pd
```

```
df = pd.read_csv('sales_data_sample.csv', encoding='latin1')
```

```
print(df.isnull().sum())
```

```
ORDERNUMBER           0
QUANTITYORDERED       0
PRICEEACH             0
ORDERLINENUMBER       0
SALES                 0
ORDERDATE             0
STATUS                0
QTR_ID                0
MONTH_ID              0
YEAR_ID               0
PRODUCTLINE           0
MSRP                  0
PRODUCTCODE           0
CUSTOMERNAME          0
PHONE                 0
ADDRESSLINE1          0
ADDRESSLINE2       2521
CITY                  0
STATE              1486
POSTALCODE           76
COUNTRY               0
TERRITORY          1074
CONTACTLASTNAME       0
CONTACTFIRSTNAME      0
DEALSIZE              0
dtype: int64
```

```
df_cleaned = df.dropna()
```

```
df['ADDRESSLINE2'] = df['ADDRESSLINE2'].fillna('')
```

Removing duplicates

```
df = df.drop_duplicates()
```

Standardizing texts and column headers

```
df.columns = [col.lower().replace(' ', '_') for col in df.columns]
```

```
df['country'] = df['country'].str.strip().str.title()
```

Consistent date formats

```
df['orderdate'] = pd.to_datetime(df['orderdate'])
```

```
df['formatted_date'] = df['orderdate'].dt.strftime('%d-%m-%Y')
```

Rename column headers

```
df.columns = df.columns.str.lower().str.replace(' ', '_')
```

Fixing data types

```
print(df.dtypes)
```

```
ordernumber                int64
quantityordered            int64
priceeach                float64
orderlinenumber            int64
sales                    float64
orderdate          datetime64[ns]
status                    object
qtr_id                     int64
month_id                   int64
year_id                    int64
productline               object
msrp                       int64
productcode               object
customername              object
phone                     object
addressline1              object
addressline2              object
city                      object
state                     object
postalcode                object
country                   object
territory                 object
contactlastname           object
contactfirstname          object
dealsize                  object
formatted_date            object
dtype: object
```

```
df['quantityordered'] = df['quantityordered'].astype(int)
# df['sales'] = df['sales'].astype(float)
```

Check for outliers

```
print(df.describe())
```

```
       ordernumber  quantityordered    priceeach  orderlinenumber  \
count   2823.000000      2823.000000  2823.000000      2823.000000
mean   10258.725115        35.092809    83.658544         6.466171
min    10100.000000         6.000000    26.880000         1.000000
25%    10180.000000        27.000000    68.860000         3.000000
50%    10262.000000        35.000000    95.700000         6.000000
75%    10333.500000        43.000000   100.000000         9.000000
max    10425.000000        97.000000   100.000000        18.000000
std       92.085478         9.741443    20.174277         4.225841

              sales                      orderdate       qtr_id     month_id
count   2823.000000                           2823  2823.000000  2823.000000
mean    3553.889072  2004-05-11 00:16:49.989373056     2.717676     7.092455
min      482.130000            2003-01-06 00:00:00     1.000000     1.000000
25%     2203.430000            2003-11-06 12:00:00     2.000000     4.000000
50%     3184.800000            2004-06-15 00:00:00     3.000000     8.000000
75%     4508.000000            2004-11-17 12:00:00     4.000000    11.000000
max    14082.800000            2005-05-31 00:00:00     4.000000    12.000000
std     1841.865106                            NaN     1.203878     3.656633

         year_id         msrp     quantity
count  2823.00000  2823.000000  2823.000000
mean   2003.81509   100.715551    35.092809
min    2003.00000    33.000000     6.000000
25%    2003.00000    68.000000    27.000000
50%    2004.00000    99.000000    35.000000
75%    2004.00000   124.000000    43.000000
max    2005.00000   214.000000    97.000000
std       0.69967    40.187912     9.741443
```