FOUNDATION LESSON

WEEK 5 · LESSON 2

Duplicates and safe numeric conversion

A duplicate is defined by a business key and a revision rule.

THE MENTAL MODEL

MENTAL MODEL

The mental model

A duplicate is defined by a business key and a revision rule.

EXACT DUPLICATES

CODE WALKTHROUGH W05-L02-S03

```
01  import pandas as pd
02  df = pd.DataFrame({"id": ["A", "A", "B"],
03                     "amount": [10, 10, 20]})
04  print(df.duplicated().tolist())
05  print(len(df.drop_duplicates()))
```

WHAT TO TRACK

Exact duplicates

PREDICT THE RESULT

Predict the output.
Explain each operation.
Pause before continuing.

EXACT DUPLICATES · RESULT

CODE WALKTHROUGH W05-L02-S04

```
01  import pandas as pd
02  df = pd.DataFrame({"id": ["A", "A", "B"],
03                     "amount": [10, 10, 20]})
04  print(df.duplicated().tolist())
05  print(len(df.drop_duplicates()))
```

WHAT TO TRACK

Exact duplicates

VALIDATE THE RESULT

```
[False, True, False]
2
```

LATEST REVISION BEFORE VALIDATION

CODE WALKTHROUGH W05-L02-S05

```
01 import pandas as pd
02 df = pd.DataFrame({"id": ["A", "A", "B"],
03  "revision": [1, 2, 1], "amount": ["10", "12", "oops"]})
04 latest = df.sort_values("revision").drop_duplicates(
05     "id", keep="last").copy()
06 latest["amount"] = pd.to_numeric(
07     latest["amount"], errors="coerce")
08 print(latest.sort_values("id").to_string(index=False))
09 print("Invalid:", latest["amount"].isna().sum())
```

WHAT TO TRACK

Latest revision before validation

PREDICT THE RESULT

Predict the output.
Explain each operation.
Pause before continuing.

LATEST REVISION BEFORE VALIDATION · RESULT

CODE WALKTHROUGH W05-L02-S06

```
01  import pandas as pd
02  df = pd.DataFrame({"id": ["A", "A", "B"],
03   "revision": [1, 2, 1], "amount": ["10", "12", "oops"]})
04  latest = df.sort_values("revision").drop_duplicates(
05      "id", keep="last").copy()
06  latest["amount"] = pd.to_numeric(
07      latest["amount"], errors="coerce")
08  print(latest.sort_values("id").to_string(index=False))
09  print("Invalid:", latest["amount"].isna().sum())
```

WHAT TO TRACK

Latest revision before validation

VALIDATE THE RESULT

```
id  revision  amount
 A         2    12.0
 B         1     NaN
Invalid: 1
```

CONFLICTING SAME-REVISION RECORDS

CODE WALKTHROUGH W05-L02-S07

```
01  import pandas as pd
02  df = pd.DataFrame({"id": ["A", "A"],
03   "revision": [2, 2], "amount": [12, 15]})
04  if df.duplicated(["id", "revision"]).any():
05      raise ValueError("Resolve repeated invoice revision")
```

WHAT TO TRACK

Conflicting same-revision records

PREDICT THE RESULT

Predict the output.
Explain each operation.
Pause before continuing.

CONFLICTING SAME-REVISION RECORDS · RESULT

CODE WALKTHROUGH W05-L02-S08

```
01  import pandas as pd
02  df = pd.DataFrame({"id": ["A", "A"],
03   "revision": [2, 2], "amount": [12, 15]})
04  if df.duplicated(["id", "revision"]).any():
05      raise ValueError("Resolve repeated invoice revision")
```

WHAT TO TRACK

Conflicting same-revision records

VALIDATE THE RESULT

ValueError: Resolve repeated invoice revision

GUIDED PRACTICE

LEARNER PRACTICE

Guided practice

Convert ["5", "bad", "0"] to numbers.
Print the invalid count and valid total.

PREDICT · MODIFY · RUN

01 Restate the supplied rules.

02 Predict the result.

03 Modify and run the code.

04 Compare actual output.

05 Explain any difference.

INDEPENDENT PRACTICE

LEARNER PRACTICE

Independent practice

For invoice X revisions 1 and 3 with amounts 9 and 11, select revision 3 and print 11.

PREDICT · MODIFY · RUN

01 Restate the supplied rules.

02 Predict the result.

03 Modify and run the code.

04 Compare actual output.

05 Explain any difference.

CHECK YOUR UNDERSTANDING

MENTAL MODEL

Check your understanding

Diagnostic: If the latest revision has invalid text, may we silently restore an older valid amount? No. That would introduce an unstated revision policy.

RECAP AND NEXT STEP

Recap and next step

Define identity and revision priority before deduplicating. Record conversion failures and validate business rules separately.

UNDERSTAND APPLY VALIDATE EXPLAIN