The Art and Science of Transportation Research in the AI Era

Data Validation With Pandas

Imen Meftah

AGENDA

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

#1 Why data quality matters

#2 What is data validation?

#3 Real-world example (Traffic Data)

#4 Live demo with Pandas

#5 Key takeaways

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

#1

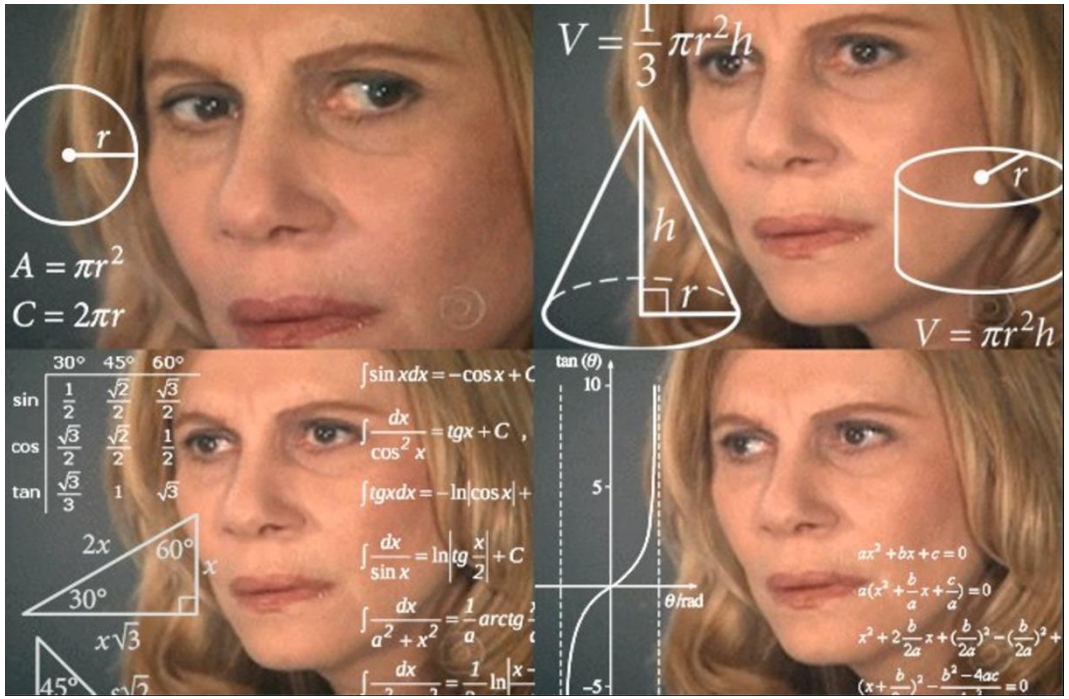
Why data quality matters

❌ Before Validation

- Negative values
- Missing data
- Unrealistic numbers
- Unreliable results

Can We Really Trust Our Data? 🤔

When you trust raw data without validation…
And your results make no sense 😅



✅ After Validation

- Clean values
- Logical ranges
- Consistent data
- Ready for analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

#2

What is Data Validation?

Simple Definition

Data validation = Checking whether data follows logical and physical rules before analysis

Examples:

- Speed cannot be negative
- Vehicle count cannot be negative
- Dates should not be in the future

Raw Data → Validation → Reliable Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

Why Validation is Important

Why Should We Care?

- Prevent wrong conclusions
- Improve decision-making quality
- Increase trust in reports and dashboards
- Essential for AI and automation systems

Good data quality = Good results

#3

Real Example: Traffic Sensor Data

Example Dataset (Raw Data)

id	*speed (km/h)*	*vehicle_count*
1	*50*	*120*
2	*-15*	*90*
3	*180*	*60*
4	*NaN*	*110*
5	*70*	-5

Problems:

- Negative speed
- Unrealistic speed
- Missing value
- Negative vehicle count

Validation Rules

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

Example Rules for Traffic Data

- Speed must be between 0 and 130 km/h
- Vehicle count must be ≥ 0
- Speed should not be missing

These rules protect data quality.

TECHNISCHE UNIVERSITÄT DARMSTADT

Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

#4

Live demo with Pandas

Live Demo : Load Data

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

Step 1: Load Example Data

```
import pandas as pd
data = {
 "speed": [50, -15, 180, None, 70],
 "vehicle_count": [120, 90, 60, 110, -5]
}
df = pd.DataFrame(data)

df
```

Live Demo: Detect Errors

TECHNISCHE UNIVERSITÄT DARMSTADT

Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

Step 2: Detect Invalid Values

Negative speeds:
df[df["speed"] < 0]

Unrealistic speeds:
df[df["speed"] > 130]

Missing values:
df.isnull().sum()
.

.isnull() : This checks where data is missing.
.sum() : Now Pandas **adds up the True values**.
True = 1
False = 0

Live Demo: Apply Validation Rules

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

Step 3: Apply Validation Rules

Keep only realistic speed values:

```
df_valid = df[(df["speed"] >= 0) & (df["speed"] <= 130)]
df_valid
```

Fix negative vehicle count:

```
df_valid["vehicle_count"] =
df_valid["vehicle_count"].clip(lower=0)
df_valid
```

clip() means: Limit values to a given range.

lower=0 , Any value **below 0** becomes **0**.

Before vs After

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

Raw Data vs Validated Data

Before:

- Errors
- Missing values
- Unrealistic numbers

After:

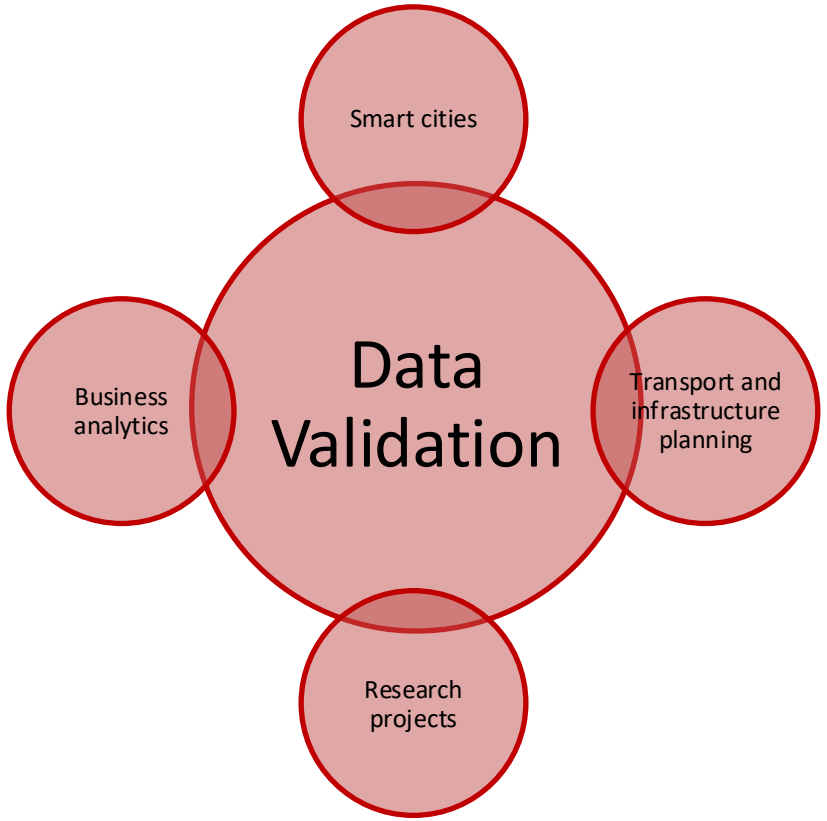
- Clean
- Logical
- Ready for analysis

Real Impact

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

Where Is This Used?



Validation improves reliability everywhere.

TECHNISCHE UNIVERSITÄT DARMSTADT

Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

#5

Key takeaways

Key Takeaways

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

What Should You Remember?

- Data validation is essential
- Pandas makes validation simple
- Better data = Better decisions

TECHNISCHE UNIVERSITÄT DARMSTADT

IVV Institut für Verkehrsplanung und Verkehrstechnik TU Darmstadt

Thank You!