خط أنابيب تجهيز البيانات

(Data Preprocessing Pipeline)

دليل شامل للتعامل مع القيم المفقودة (Missing Values)، القيم المتطرفة (Outliers)، والمعيارية (Z-score).

محاور المحاضرة

1. القيم المفقودة

Missing Values

استراتيجيات التعامل مع البيانات الناقصة: الحذف أو التعويض (Imputation).

2. القيم المتطرفة

Outliers

كيفية اكتشاف القيم الشاذة وتأثيرها على النماذج الإحصائية.

3. المعيارية

Z-score

توحيد مقاييس البيانات لضمان كفاءة الخوارزميات (Algorithms).

القيم المفقودة

Missing Values

لماذا تعتبر القيم المفقودة مشكلة؟

معظم خوارزميات `sklearn` لا تقبل قيم `NaN` مثل:

- Logistic Regression
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)

⚠ تحذير

التعويض الخاطئ قد يؤدي إلى تحريف توزيع البيانات (Distribution Distortion) أو خلق تحيز (Bias) في النموذج.

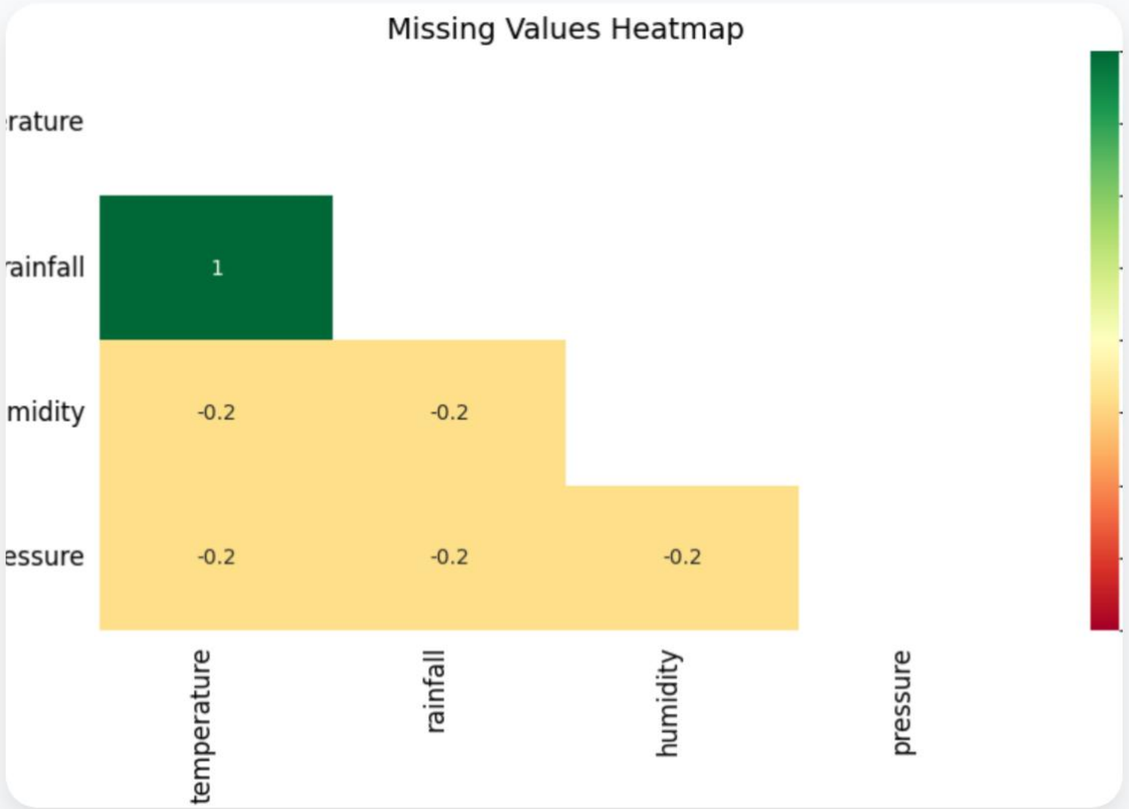
</> اكتشاف النسبة

نستخدم الكود التالي لحساب نسبة الفقد:

```
print("missing ratio:\n",
(df.isnull().sum() / len(df)) * 100)
```

- لو النسبة > 80%: **حذف (Drop)**
- لو النسبة 2-5%: **تعويض (Impute)**

تصوير البيانات المفقودة (Visualization)



خريطة حرارية (Heatmap)

استخدام مكتبة `seaborn` لرسم أماكن تواجد القيم المفقودة.

- كل خط ملون يمثل قيمة مفقودة.
- يساعد في اكتشاف الأنماط (Patterns) في البيانات الناقصة.

```
sns.heatmap(np.isnan(df), cbar=True,
            yticklabels=False, cmap='RdYlGn_r')
```

استراتيجيات التعويض (Imputation Strategies)

Mean Imputation

تعويض بالمتوسط الحسابي.

متى؟ عندما تكون البيانات موزعة بشكل طبيعي (Normal Distribution) وخالية من القيم الشاذة.

Median Imputation

تعويض بالوسيط.

متى؟ عندما تكون البيانات ملتوية (Skewed) أو تحتوي على قيم متطرفة (Outliers).

KNN Imputer

استخدام الجوار (Neighbors).

متى؟ للدقة العالية. يعتمد على قيم الصفوف المشابهة لتوقع القيمة المفقودة.

القيم المتطرفة

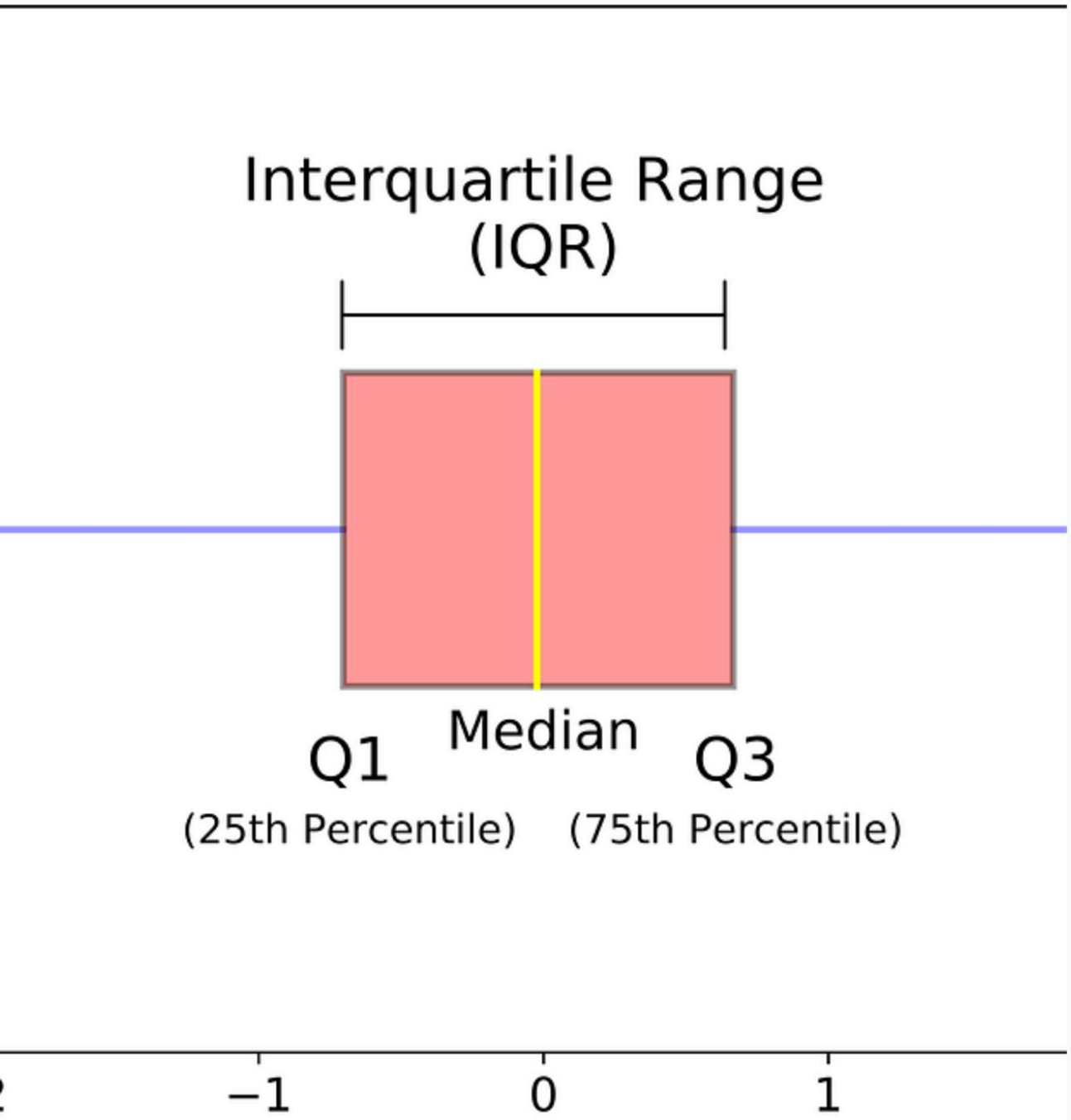
Outliers

اكتشاف القيم المتطرفة

ما هي الـ Outlier؟ نقطة بيانات تختلف بشكل كبير عن باقي الملاحظات. قد تكون خطأ في الإدخال أو حالة نادرة.

أدوات الرسم (Visual Tools)

- **Boxplot:** النقاط التي تقع خارج "الشاربين" (Whiskers) تعتبر شاذة.
- **Histogram/KDE:** لمراقبة شكل التوزيع والذيول الطويلة (Long Tails).
- **Scatter Plot:** رؤية النقاط البعيدة عن السرب.



طرق التعامل مع القيم المتطرفة

التحويل اللوغاريتمي (Log Transform)

يستخدم عندما تكون البيانات ملتوية لليمين (Right Skewed).

يضغط القيم الكبيرة ويجعل التوزيع أقرب لـ Normal.

```
np.log(df[col] + 1)
```

القص (Capping / Winsorizing)

بدل الحذف، نقوم بتحديد سقف للقيم.

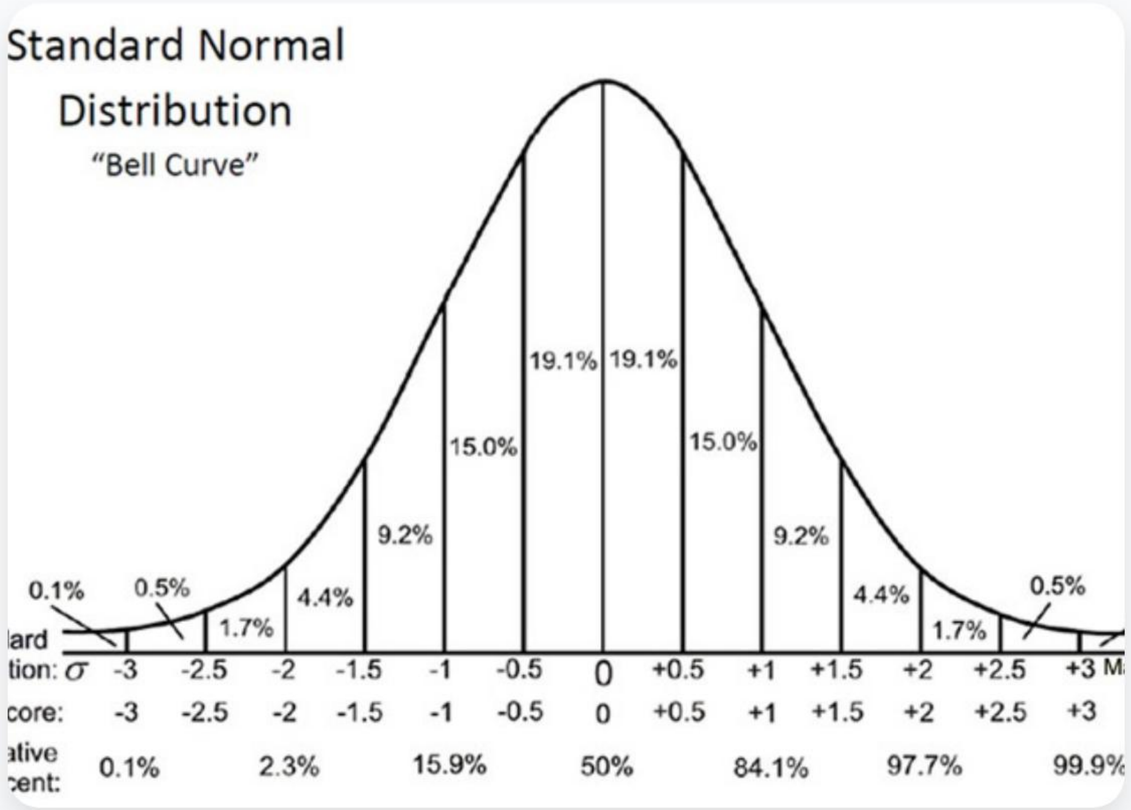
مثلاً: أي قيمة أعلى من 99% يتم استبدالها بقيمة الـ 99th percentile.

```
df[col].clip(lower, upper)
```

المعيارية

Z-score Standardization

مفهوم الـ Z-score



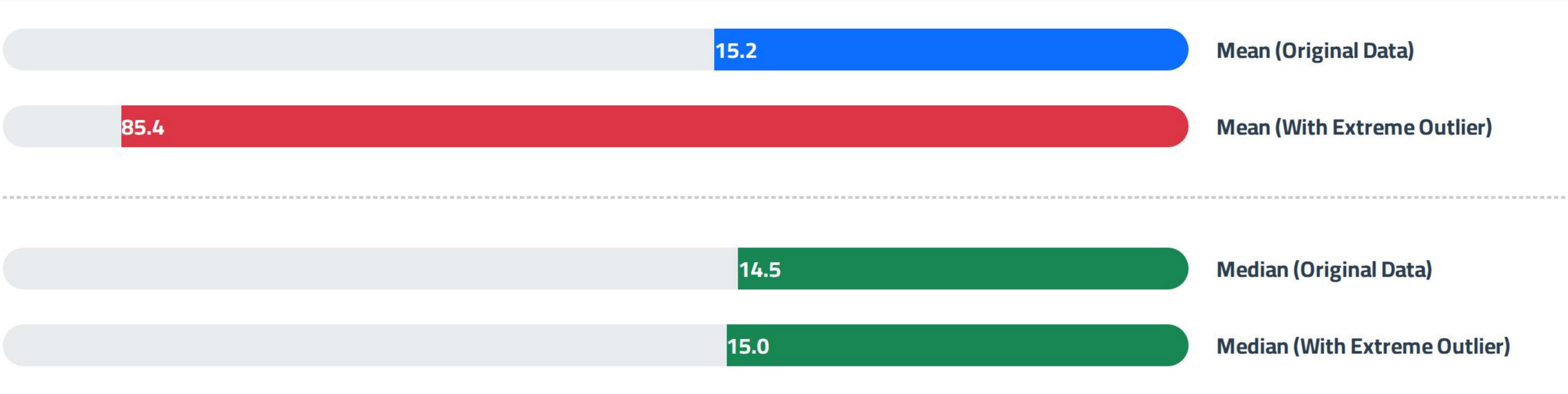
يحول البيانات إلى توزيع طبيعي قياسي (Standard Normal Distribution) بمتوسط 0 وانحراف معياري 1.

$$z = \frac{x - \mu}{\sigma}$$

- **μ (Mean):** المتوسط الحسابي.
- **σ (Std Dev):** الانحراف المعياري.
- إذا كان $|z| > 3$ تعتبر القيمة غالباً Outlier.

لماذا الـ Median أفضل عند وجود Outliers؟

مقارنة توضح حساسية الـ Mean للقيم الشاذة مقابل ثبات الـ Median.



لاحظ كيف قفز الـ Mean (بالأحمر) بشكل جنوني بسبب الـ Outlier، بينما ظل الـ Median (بالأخضر) مستقراً.

أسئلة؟

Q & A

شكراً لاستماعكم

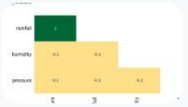
Image Sources

https://media.istockphoto.com/id/2079936911/vector/big-data-analytics-ai-technology-neural-network-and-artificial-intelligence-analyzing.jpg?s=612x612&w=0&k=20&c=K1yaUbNX=P-tohQfMoh7EH3_l-yhAAbYX_pLabDTPxoA

Source: www.istockphoto.com

---

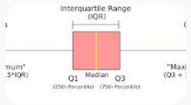
https://miro.medium.com/1*8-UKYPpVrhdQOewnTzwesw.png



Source: medium.com

---

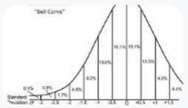
https://miro.medium.com/v2/resize:fit:1400/1*0MPDTLn8KoLApoFvl0P2vQ.png



Source: medium.com

---

https://mathbitsnotebook.com/Algebra2/Statistics/normalstandard.jpg



Source: mathbitsnotebook.com