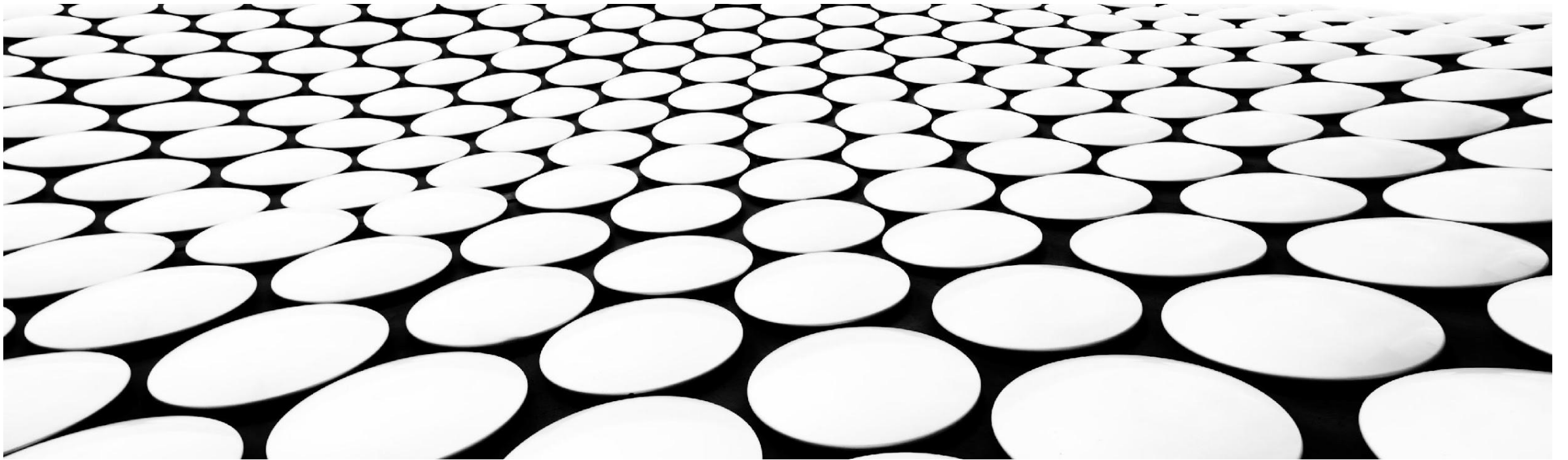


第3講 データの前処理とクリーニング グ



第3講 データの前処理とクリーニング

目次

1. はじめに
2. 前処理を行わなかった場合
3. 前処理の全体像
4. 異常値と外れ値
5. データ型の変換と整形
6. データ変換・カテゴリ変数
7. まとめ

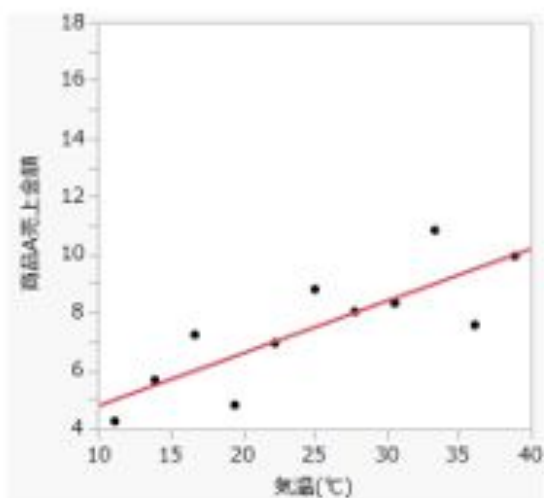
1. はじめに

- データサイエンスの最初のステップは「データの準備」
- 現実のデータには、抜けている値（欠損）、入力ミス（誤記）、同じデータの重複など、さまざまな問題が含まれている
- 前処理はデータを整えて「使える状態」にする作業
- 前処理はデータサイエンスの工程の約8割を示すと言われている

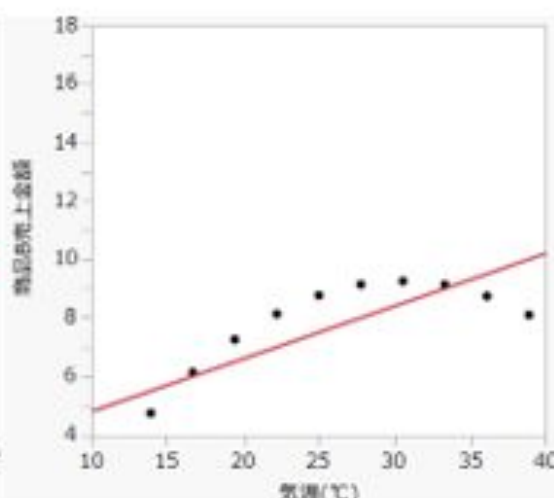
2. 前処理を行わなかった場合

- 下記の4つのグラフの回帰直線は、前処理しなかった場合、すべて同じ！！

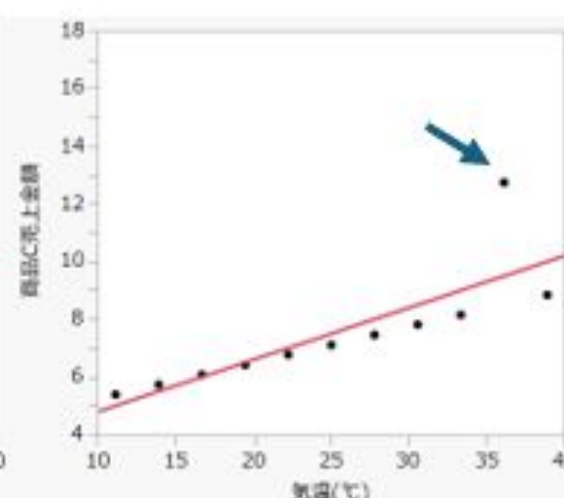
商品Aの売上金額（円）と気温（℃）



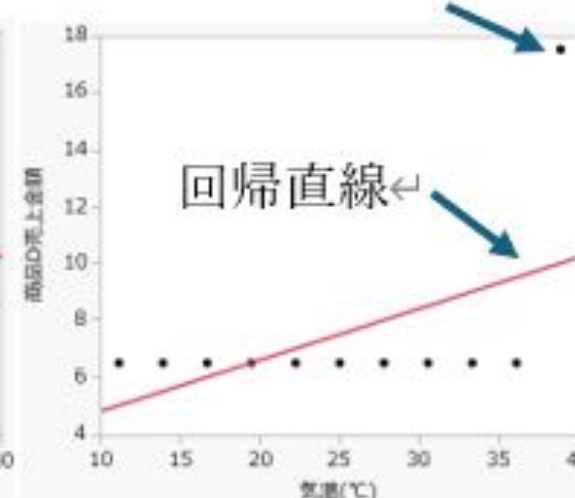
商品 A



商品 B



商品 C



商品 D

3. 前処理の全体像

- 欠損値の処理（データの一部が記録されていない状態）
- 異常値と外れ値の検出と対応
- データ型の変換
- 重複データの削除
- 変数のスケーリングや正規化
- カテゴリ変数のダミー化

3. 前処理の全体像 欠損値の対応法

対応法	どのような場合に対応できるのか、注意点など
欠損行の削除	■ データ量が十分な場合に対応できる。行自体を削除するので、データ総数が減ってしまう。(例えば、変数の数よりサンプルの数が3倍以上ある)
データ全体で平均、中央値、最頻値による補完	■ 欠測値を平均で埋めると、“だいたいの傾向”は見えるようになるが、“ちょっと変わった特徴”が見えにくくなることがある。
前後の値で補完	■ 時系列データ※などで欠測値の前後のデータがあり、前後の平均値で補完する。
モデルによる補完	■ 高度な手法であり、アプリケーションが必要。

4. 異常値と外れ値

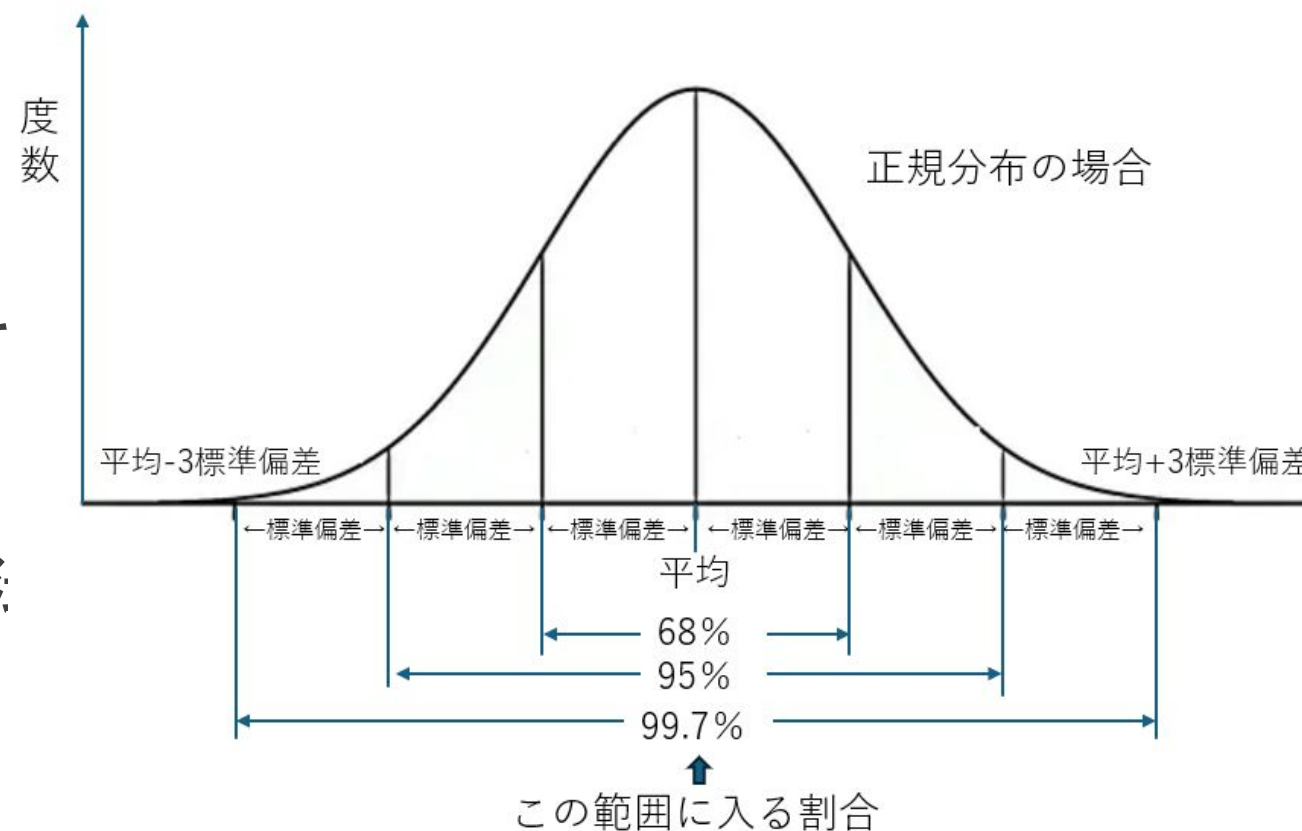
1. 異常値と外れ値の違い
2. 標準偏差を用いた統計的検出
3. ひげ図を用いた外れ値の判定
4. 外れ値の対応
5. 事例による説明

4. 1 異常値と外れ値の違い

- 異常値とは、通常とは異なる挙動やパターンで発生するデータ
- システム的・意味的におかしい値で、その原因は入力ミスや測定エラーなど様々
- 例えば、岐阜市の夏の気温のデータの中に、 -30°C や 0°C が含まれていた場合、実際にはありえない値
- -30°C は符号誤り、 0°C は誤って入力したと考えられる

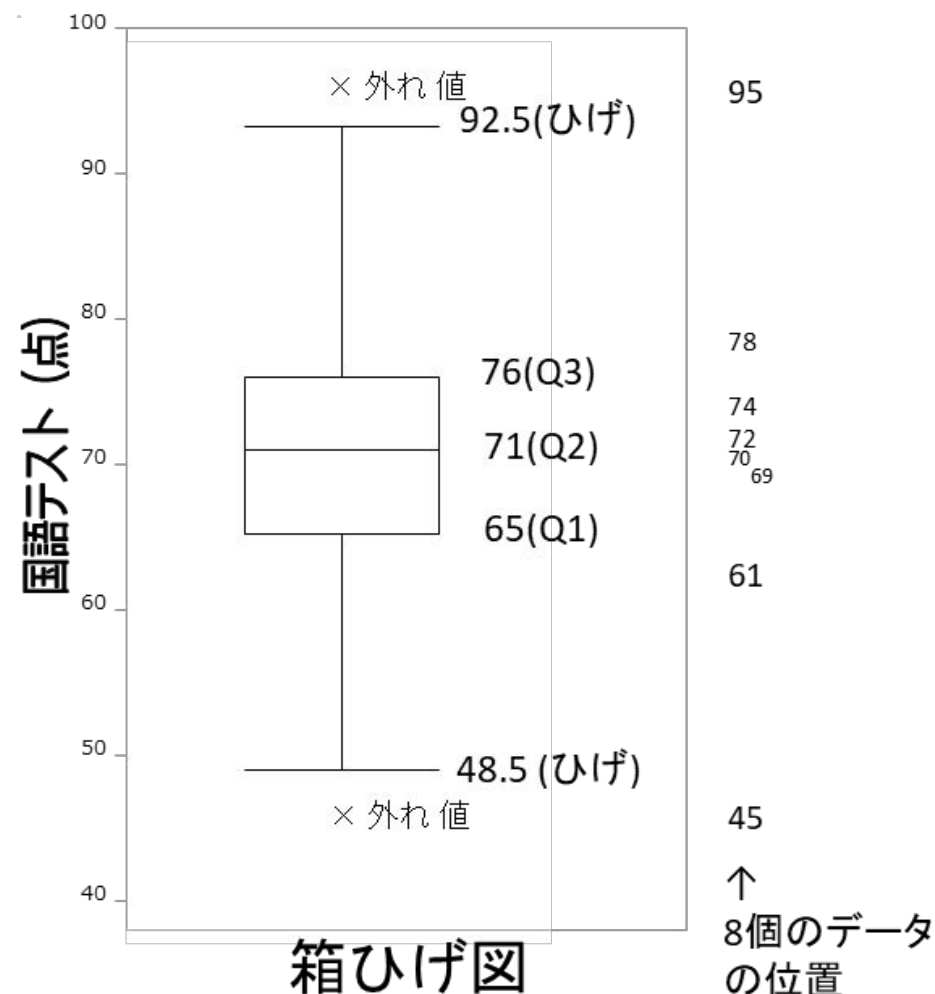
4. 2 標準偏差を用いた統計的検出

- 標準偏差の値から3倍以上離れているかどうかを基準とし、3倍より外れたものを、外れ値と見なす方法



4. 3 ひげ図を用いた外れ値の判定

- 8人の国語テストの結果があった場合
- まずデータを小さいものから順番に並べて4等分する
- 45 61 69 70 72 74 78 95
- 最小値: 45
- 第1四分位数(Q1): 65 (61と69の平均)
- 中央値(Q2): 71 (70と72の平均)
- 第3四分位数(Q3): 76 (74と78の平均)
- 最大値: 95
- 四分位範囲(IQR) = 第3四分位数(Q3) - 第1四分位数(Q1) = 76 - 65 = 11
- ひげの上端: 第3四分位数(Q3) + 1.5 × IQRより小さい最大値 = 76 + 1.5 × 11 = 92.5
- ひげの下端: 第1四分位数(Q1) - 1.5 × IQRより大きい最小値 = 65 - 1.5 × 11 = 48.5
- ひげの範囲から外れた値が外れ値であり、95と45が該当



4. 4 外れ値の対応

対処法は大きく3通り

- 正しい値に修正
データの入力ミスやシステムのエラーなど、外れ値が生じた要因が判明している場合は正しい値に修正
- 行ごと除外する
要因が不明な場合や、要因が分かってもあまりに外れ具合が大きいデータは行ごと除外することを検討する
- そのまま使用する
外れ値の要因が判明しており、かつ分析結果に大きな影響がなさそうな場合は、そのまま使用することがある

4. 5 事例による説明

- ある生徒の通学時間(分)に関するデータを100件取得し、スプレッドシートに右記の通り入力
- 通学時間は通常、約60分かかりが、ばらつきがある
- このデータの中で、2025/4/7の-58分は、マイナス？
- 2025/4/8のゼロ？ 同様に時間がゼロということはない
- 2025/4/11は病欠となっており、通学していないので、欠損値として処理が必要

C10 ✕ ✓ fx 病欠			
	A	B	C
1	Data No	年月日	通学時間 (分)
2	1	2025/4/1	65
3	2	2025/4/2	58
4	3	2025/4/3	62
5	4	2025/4/4	101
6	5	2025/4/7	-58
7	6	2025/4/8	0
8	7	2025/4/9	60
9	8	2025/4/10	61
10	9	2025/4/11	病欠
11	10	2025/4/14	62
12	11	2025/4/15	75
13	12	2025/4/16	67

: 省略↵

100	99	2025/8/22	59
101	100	2025/8/25	60

5. データ型の変換と整形

同じ列(変数)の中で、数値・文字列・日付などの形式がバラバラだと処理ができないので、整形が必要

行 \ 列	名前	誕生日	クラス番号
1	A	平成15年5月15日	3
2	B	2003年8月1日	1
3	C	31/12/2003	01



誕生日	クラス番号
2003/05/15	3
2003/08/01	1
2003/12/31	1

日付型

YYYY/MM/DDに統一する

数値で

統一する

6. データ変換・カテゴリ変数

- 「性別」「地域」などのカテゴリ変数は、機械学習モデルで扱うためにデータ変換、数値化が必要
- 代表的な方法は「ダミー変数化」で、例えば「性別」が「男性」「女性」の場合、 $\text{male} = 1$ 、 $\text{female} = 0$ のように変換

7. まとめ

- 前処理は単なる技術ではなく、「データの意味を理解する力」を養う学び
- 生徒の成績データやアンケート結果など、身近なデータを題材にすることで、実践的な理解が深まる
- 前処理を体験することで、データ活用の可能性を広げる第一歩となる