目的に合わせたデータの利用

第5回
23j1-305
教科書　P108-P109

この時間の目標

- 目的に合わせたデータの利用

- S データの利用についてよく理解でき、活用しようと思った
- A データの利用についてよく理解できた
- B データの利用について理解できた
- C データの利用について理解できなかった

目的に合わせたデータの利用

1欠損値と外れ値

- 欠損値：数値が欠けている
- 外れ値：他の値から大きく外れている

- 取り除くべきか・近似値などで埋めるか検討が必要

教科書P108

2分析の目的とデータの関係

- 目的に合わないデータ
  - 間違った結果を導き出す

- 分析したい内容と関係が強いデータを集める

教科書P108

3データの解釈

- 因果関係
  - かき氷の売り上げが増えると熱中症が増える
- 擬似相関
  - 因果関係がないのに因果関係があるように見える

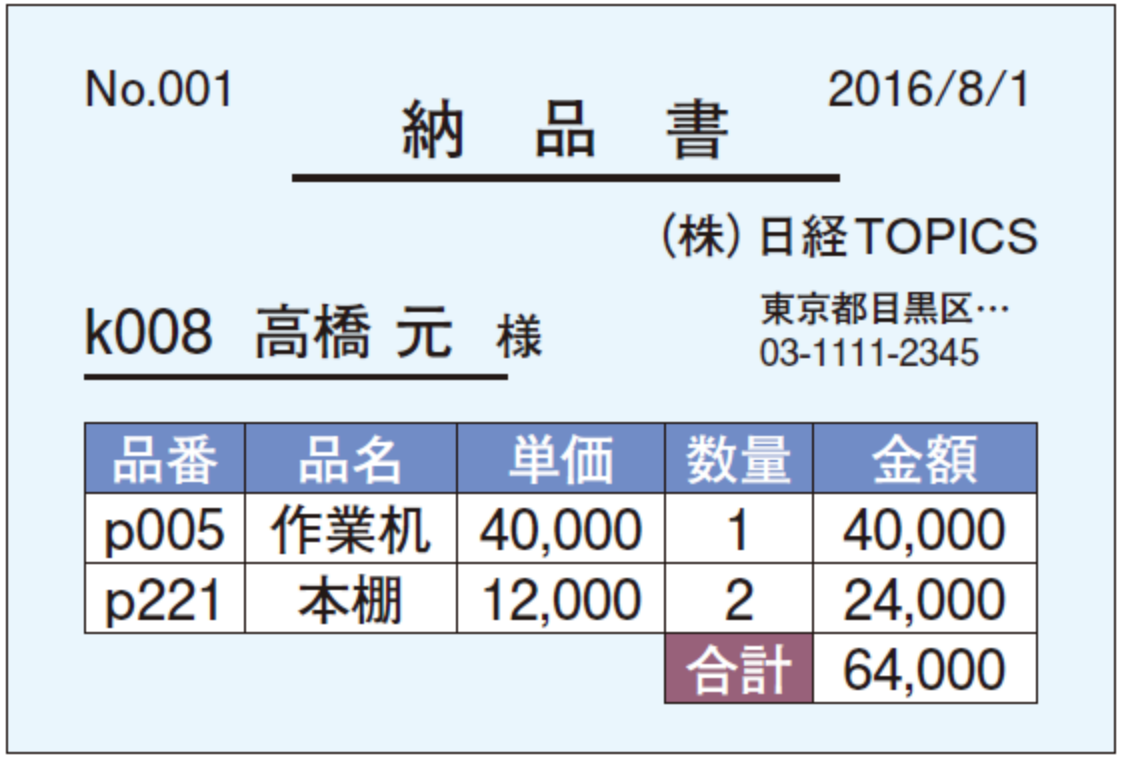
教科書P109

データの正規化

正規化

- データベースを設計するときの指標
- データの重複を少なくする
- トピック集P175

- 例:納品書
- 情報をデータベースに収めたい



No.001　　2016/8/1

納　品　書

(株) 日経TOPICS

k008　高橋 元　様

東京都目黒区…
03-1111-2345

品番	品名	単価	数量	金額
p005	作業机	40,000	1	40,000
p221	本棚	12,000	2	24,000
			合計	64,000

正規化-そのまま入力すると

伝票番号	日付	顧客番号	氏名	住所	電話番号	商品番号	商品名	単価	数量
001	2022.8.1	k008	高橋 元	東京都目黒区	03-1234-	（p005, p221）	（作業机, 本棚）	（40000, 12000）	（1, 2）

- 一つの項目に複数のデータ
- 作業机を買った人を探したい
- →面倒なことに

No.001　　2022/8/1

納　品　書

（株）日経TOPICS

k008　高橋 元　様

東京都目黒区…
03-1111-2345

品番	品名	単価	数量	金額
p005	作業机	40,000	1	40,000
p221	本棚	12,000	2	24,000
			合計	64,000

トピック集P175

正規化-第1正規形

伝票番号	日付	顧客番号	氏名	住所	電話番号	商品番号	商品名	単価	数量
001	2022.8.1	k008	高橋 元	東京都目黒区	03-1234-	（p005, p221）	（作業机, 本棚）	（40000, 12000）	（1, 2）

- フィールドの内容を単純な値になるよう分割する

伝票番号	日付	顧客番号	氏名	住所	電話番号	商品番号	商品名	単価	数量
001	2022.8.1	k008	高橋 元	東京都目黒区	03-1234-	p005	作業机	40000	1
001	2022.8.1	k008	高橋 元	東京都目黒区	03-1234-	p221	本棚	12000	2

トピック集P175

正規化-第1正規形だけだと

伝票番号	日付	顧客番号	氏名	住所	電話番号	商品番号	商品名	単価	数量
001	2022.8.1	k008	高橋 元	東京都目黒区	03-1234-	p005	作業机	40000	1
001	2022.8.1	k008	高橋 元	東京都目黒区	03-1234-	p221	本棚	12000	2

- 同じ情報が複数のレコードに

- 高橋さん、住所変わったんだって！
- →直すのが大変

トピック集P175

正規化-第3正規形

伝票番号	日付	顧客番号	氏名	住所	電話番号	商品番号	商品名	単価	数量
001	2022.8.1	k008	高橋 元	東京都目黒区	03-1234-	p005	作業机	40000	1
001	2022.8.1	k008	高橋 元	東京都目黒区	03-1234-	p221	本棚	12000	2

- 氏名・住所・電話番号は顧客番号で特定→表を分割

伝票番号	日付	顧客番号	商品番号	商品名	単価	数量
001	2022.8.1	k008	p005	作業机	40000	1
001	2022.8.1	k008	p221	本棚	12000	2

顧客番号	氏名	住所	電話番号
k008	高橋 元	東京都目黒区	03-1234-
k005	山本 直	東京都新宿区	03-5678-

顧客テーブル

すべてのレコードがキーによって一意に特定できるが、キー以外やキーの一部分によって特定されることはない

トピック集P175

正規化-第3正規形　さらに分割

伝票番号	日付	顧客番号	商品番号	商品名	単価	数量
001	2022.8.1	k008	p005	作業机	40000	1
001	2022.8.1	k008	p221	本棚	12000	2

顧客番号	氏名	住所	電話番号
k008	高橋 元	東京都目黒区	03-1234-
k005	山本 直	東京都新宿区	03-5678-

- 商品名と単価は商品番号で特定→表を分割

伝票番号	日付	顧客番号	商品番号	数量
001	2022.8.1	k008	p005	1
001	2022.8.1	k008	p221	2

販売テーブル

商品番号	商品名	単価
p005	作業机	40000
p103	丸椅子	3500
p221	本棚	12000

商品テーブル

顧客番号	氏名	住所	電話番号
k008	高橋 元	東京都目黒区	03-1234-
k005	山本 直	東京都新宿区	03-5678-

顧客テーブル

トピック集P175

正規化-第3正規形　さらにさらに分割

伝票番号	日付	顧客番号	商品番号	数量
001	2022.8.1	k008	p005	1
001	2022.8.1	k008	p221	2

商品番号	商品名	単価
p005	作業机	40000
p103	丸椅子	3500
p221	本棚	12000

顧客番号	氏名	住所	電話番号
k008	高橋 元	東京都目黒区	03-1234-
k005	山本 直	東京都新宿区	03-5678-

- 購入した内容は伝票番号で特定→表を分割

伝票番号	日付	顧客番号
001	2022.8.1	k008
002	2022.8.3	k005

伝票テーブル

伝票番号	商品番号	数量
001	p005	1
001	p221	2
002	p221	1

明細テーブル

商品番号	商品名	単価
p005	作業机	40000
p103	丸椅子	3500
p221	本棚	12000

商品テーブル

顧客番号	氏名	住所	電話番号
k008	高橋 元	東京都目黒区	03-1234-
k005	山本 直	東京都新宿区	03-5678-

顧客テーブル

トピック集P175

みんなのデータ

データベースに収めるには

バーコードデータ

- バーコードデータ
  - データクレンジングしました
- データクレンジングの方法
  - 桁数の違うバーコードデータを削除(8桁・13桁のみ)
  - 2で始まるバーコードデータを削除(インストアコード)
  - 本のバーコードデータを削除
  - チェックデジット異常のデータを削除

- データベースにデータを格納するにはどうしたらよいか