7. グループ化と集約

GROUP BY、SQL によるグループ化と集約、データの分析

（データベース演習）

URL: https://www.kkaneko.jp/de/de/index.html

金子邦彦

① SQLコマンドの習得

② グループ化と集約によるデータの把握

③ 実践的なデータ分析

7-1. イントロダクション

リレーショナルデータベースの仕組み

- データを**テーブル**と呼ばれる**表形式で保存**
- **テーブル間**は**関連**で結ばれる。複雑な構造を持ったデータを効率的に管理することを可能に。

ID	商品名	単価
1	みかん	50
2	りんご	100
3	メロン	500

ID	購入者	商品ID	数量
1	X	1	10
2	Y	2	5

集約

AVG, MAX, MIN, SUM: 平均、最大、最小、合計

COUNT: 行数

記録

名前	得点	居室
徳川家康	85	1階
源義経	78	2階
西郷隆盛	90	3階
豊臣秀吉	82	1階
織田信長	75	2階

SELECT AVG(得点) **FROM** 記録;

82

SELECT MAX(得点) **FROM** 記録;

90

SELECT MIN(得点) **FROM** 記録;

75

SELECT SUM(得点) **FROM** 記録;

410

グループ化

グループ化は、**同じ属性値を共有するデータ**を**集める**プロセス。

例：**科目の「国語」、「算数」、「理科」でグループ化**

科目	受講者	得点
国語	A	85
国語	B	90
算数	A	90
算数	B	96
理科	A	95

科目	受講者	得点
国語	A	85
国語	B	90

科目	受講者	得点
算数	A	90
算数	B	96

科目	受講者	得点
理科	A	95

例：**受講者の「A」、「B」でグループ化**

科目	受講者	得点
国語	A	85
国語	B	90
算数	A	90
算数	B	96
理科	A	95

科目	受講者	得点
国語	A	85
算数	A	90
理科	A	95

科目	受講者	得点
国語	B	90
算数	B	96

それぞれの値ごとにグループに分ける

グループ化と集約

① 成績表には科目、受講者、得点が記載されている

科目	受講者	得点
国語	A	85
国語	B	90
算数	A	90
算数	B	96
理科	A	95

②科目の「国語」のグループ

科目	受講者	得点
国語	A	85
国語	B	90

同様に、「算数」、「理科」のグループを形成

科目	受講者	得点
算数	A	90
算数	B	96

科目	受講者	得点
理科	A	95

③ 「国語」のグループを作ることにより国語の得点が見やすくなった

科目	受講者	得点
国語	A	85
国語	B	90
算数	A	90
算数	B	96
理科	A	95

元データ

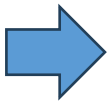
科目	受講者	得点
国語	A	85
国語	B	90

「国語」のグループ

④ グループで、集約（行数、平均、合計など）を実施

科目	受講者	得点
国語	A	85
国語	B	90

「国語」のグループ



行数：２　（受講者数を表す）
平均：８７．５
合計：１７５

グループ化と集約

グループ化

- **GROUP BY** は、特定の属性（「科目」、「受講者」）を基準として、**グループ化**を行う

集約を**GROUP BY**と**組み合わせる**ことで、**グループごとの集約結果**を得ることができる

GROUP BY の役割と書き方

- **SQL 問い合わせ「SELECT …」**の中で、**GROUP BY** を使用してデータをグループ化
- **１つ以上の属性を GROUP BY** に指定してグループ化の基準とする。

すべての科目ごとに、受講者の数を計算

SELECT 科目, COUNT(*) FROM 成績 GROUP BY 科目;

科目	COUNT(*)
国語	2
理科	1
算数	2

グループ化と分析の基本

グループ化と集約

- 概要（例：行数、平均、合計）の生成

データ分析

- グループ化と集約を用いて、カテゴリ別、時系列別などのデータ分析を実施

ビジネスインテリジェンス

- 売上のトレンド分析や顧客セグメント分析など、ビジネス意思決定に役立つ分析を実施

データの可視化

- グループ化と集約ののち、チャートやグラフで情報を視覚的に表現

7-2. 演習

いまから演習で行うこと、注意点

- 次のテーブルを作成

成績

科目	受講者	得点
国語	A	85
国語	B	90
算数	A	90
算数	B	96
理科	A	95

【Access での注意点】

- **SQLビューでは、SQL文を１つずつ実行**
  （複数まとめての一括実行ができない）
- **CREATE TABLE** では、「実行」の後、**画面が変化しない**が実行できている
- **INSERT INTO** では、「実行」の後、**確認表示**が出る。その後、**画面が変化しない**が実行できている

演習１．Access の SQL ビューを用いたテーブル定義とデータの追加

【トピックス】

- **SQLビューを開く**
- **SQL文の編集**
- **create table**
- **insert into**
- **SQL文の実行**

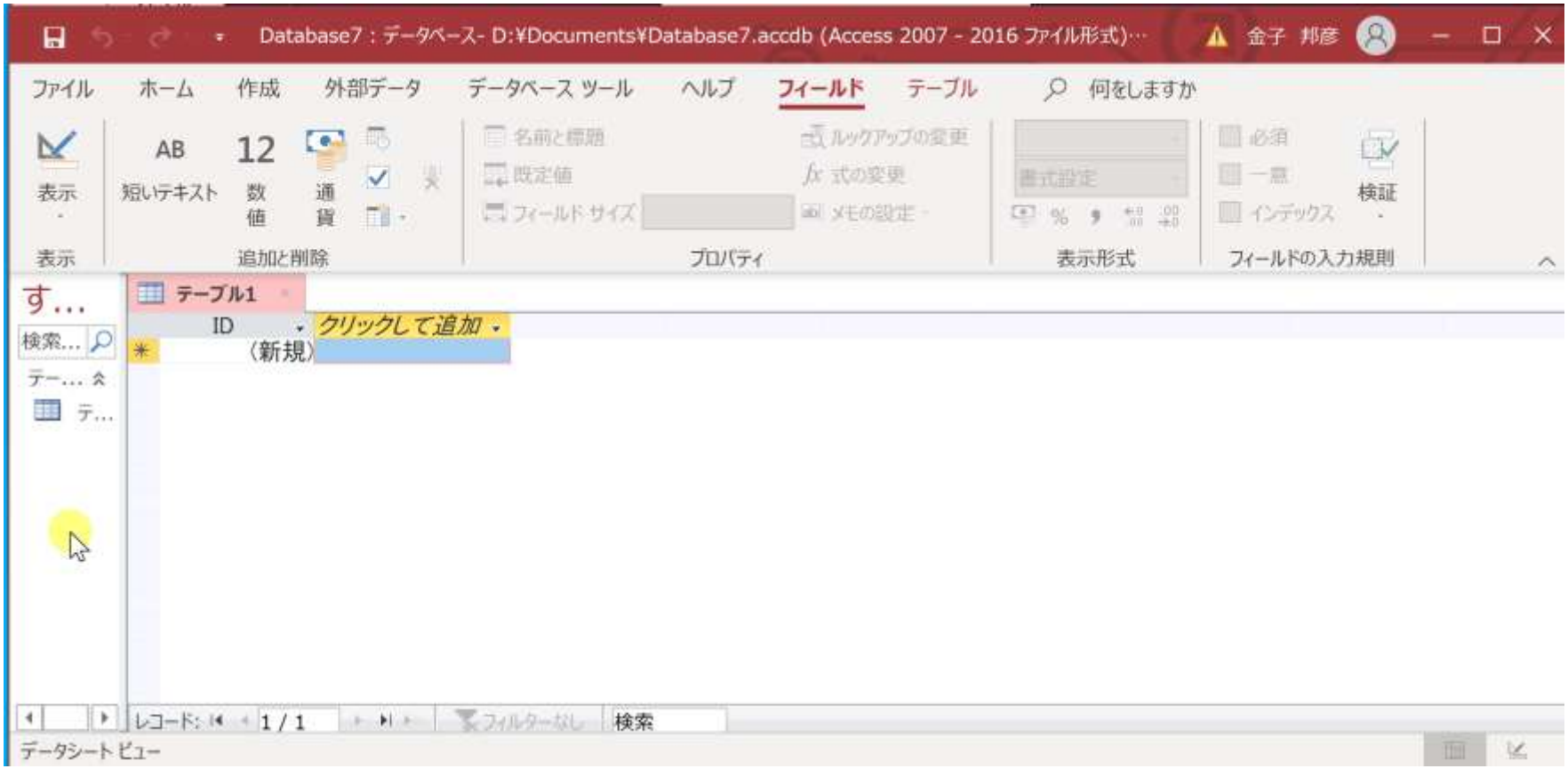
演習

1. パソコンを使用する
   前もって Access をインストールしておくこと

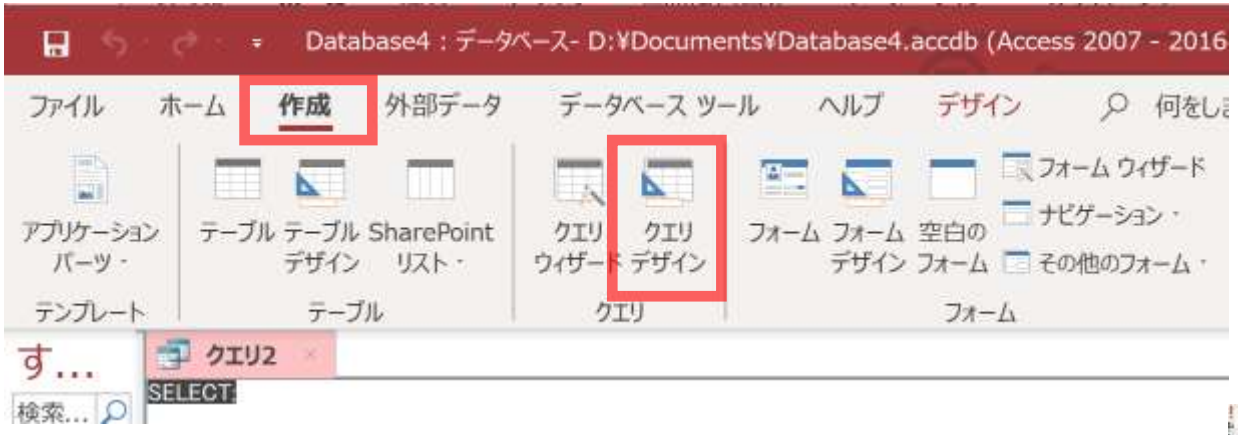
2. Access を起動する

3. Access で、「**空のデータベース**」を選び、「**作成**」をクリック.

4. **テーブルツール画面**が表示されることを確認

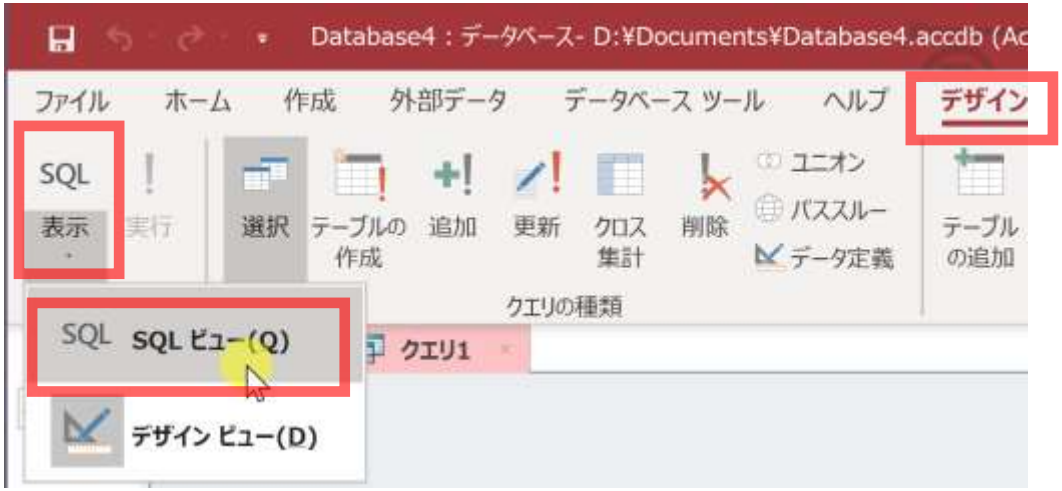


5. 次の手順で、**SQLビュー**を開く.



①「**作成**」タブで、「**クエリデザイン**」をクリック

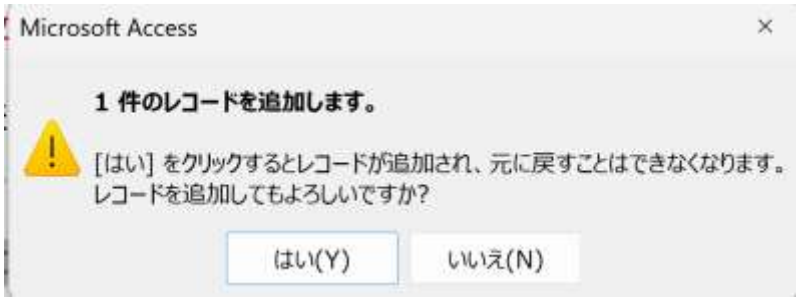
このような表示が出たときは「**閉じる**」をクリック



②「**デザイン**」タブで、「**表示**」を展開し「**SQLビュー**」を選ぶ

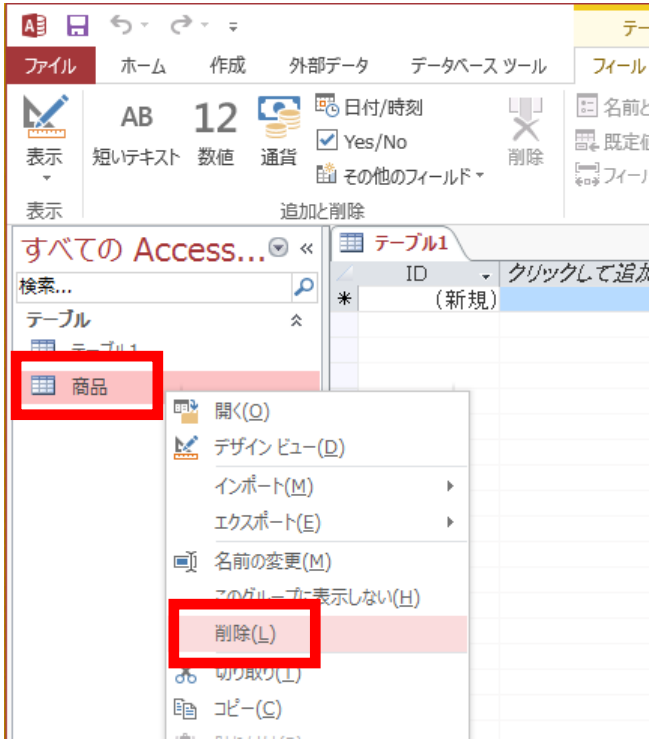
6. **SQL ビュー**に、次の SQL を１つずつ入れ、「**実行**」ボタンで、**SQL文**を実行．結果を確認

```
CREATE TABLE 成績 (
    科目 TEXT,
    受講者 TEXT,
    得点 INTEGER);

INSERT INTO 成績 VALUES('国語', 'A', 85);

INSERT INTO 成績 VALUES('国語', 'B', 90);

INSERT INTO 成績 VALUES('算数', 'A', 90);

INSERT INTO 成績 VALUES('算数', 'B', 96);

INSERT INTO 成績 VALUES('理科', 'A', 95);
```



INSERT INTOでは、「実行」の後、確認表示が出る。その後、**画面が変化しない**が実行できている

間違ってしまったときは、テーブルの削除を行ってからやり直した方が早い場合がある



テーブルビューで、削除したいテーブルを**右クリック**して、「**削除**」

テーブルを削除するときは、
間違って必要な**テーブル**を削除しないように、十分に注意する！
（元に戻せない）

演習２．SQL によるグループ化と集約． Access の SQL ビューを使用.

【トピックス】

1. グループ化
2. 集約
3. GROUP BY
4. AVG
5. COUNT(*)
6. SUM

Access の SQL ビューを用いた問い合わせ

① Access の **SQLビュー**開く

② **SQL 文**の**編集。select, from, where** を使用

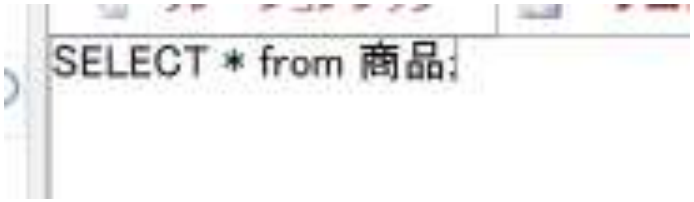
例: select * from テーブル名 where 列1 = 値1;

③ **SQL 文**の**実行**

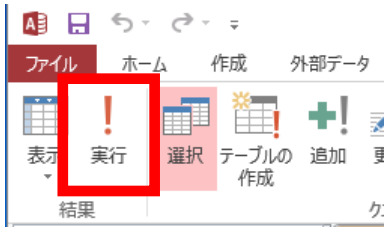
実行の結果、**データシートビュー**に画面が変わり、そこに**問い合わせの結果**が表示される

④ さらにSQL 文の編集、実行を続ける場合には、**画面を SQL ビューに切り替える**

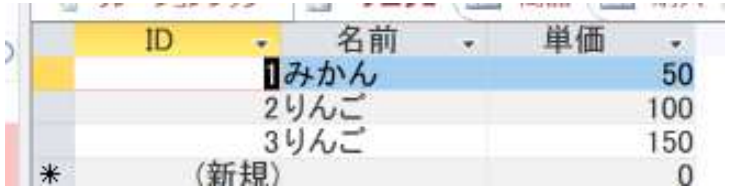
SQL 問い合わせ（クエリ）で使用する２つのビュー



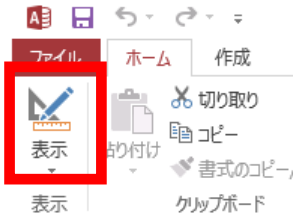
SQL ビュー
SQL 文の**作成**、**編集**



実行



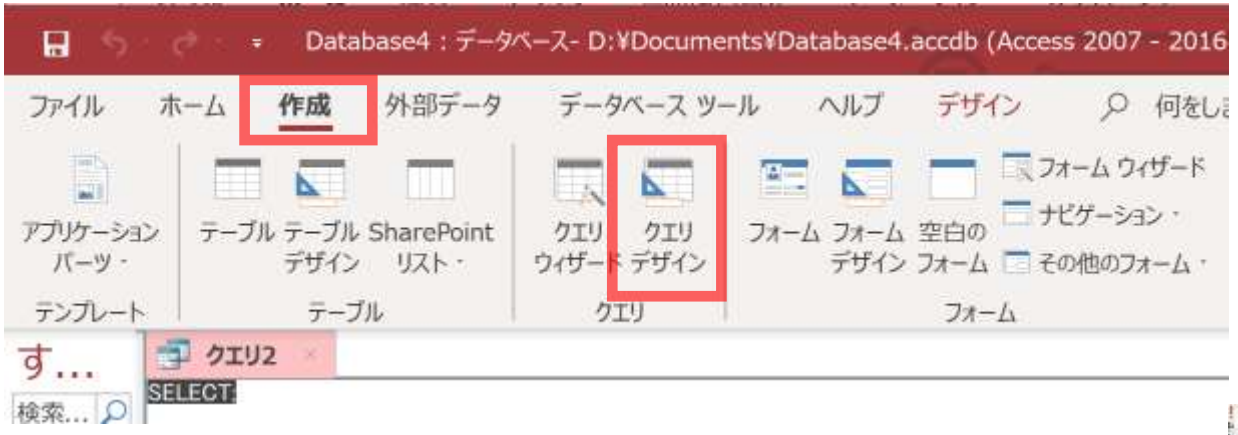
データシートビュー
問い合わせ（クエリ）の結果



表示＋SQL ビュー

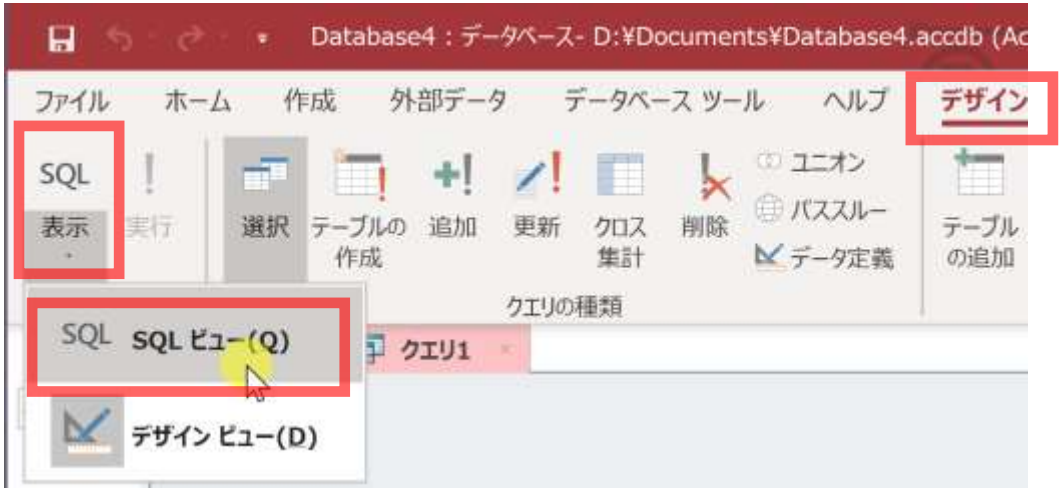
マウス操作でビューを切り替え

1. 次の手順で、**SQLビュー**を開く.



①「**作成**」タブで、「**クエリデザイン**」をクリック

このような表示が出たときは「**閉じる**」をクリック



②「**デザイン**」タブで、「**表示**」を展開し「**SQLビュー**」を選ぶ

2. **SQL ビュー**に、次の SQL を１つずつ入れ、「**実行**」ボタンで、**SQL文**を実行．結果を確認

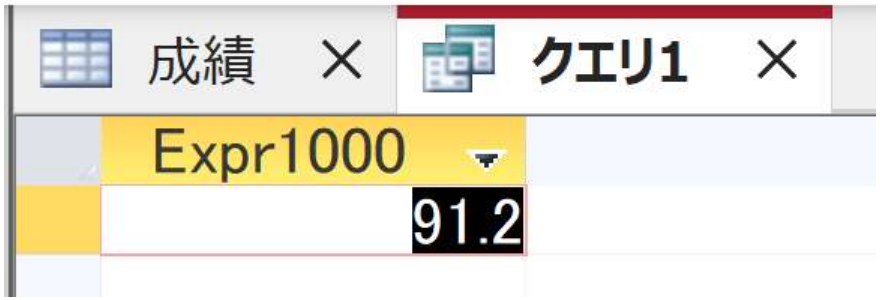
1. 単純な表示

SELECT * FROM 成績;

科目	受講者	得点
国語	A	85
国語	B	90
算数	A	90
算数	B	96
理科	A	95

2. 得点の平均

SELECT AVG(得点) FROM 成績;



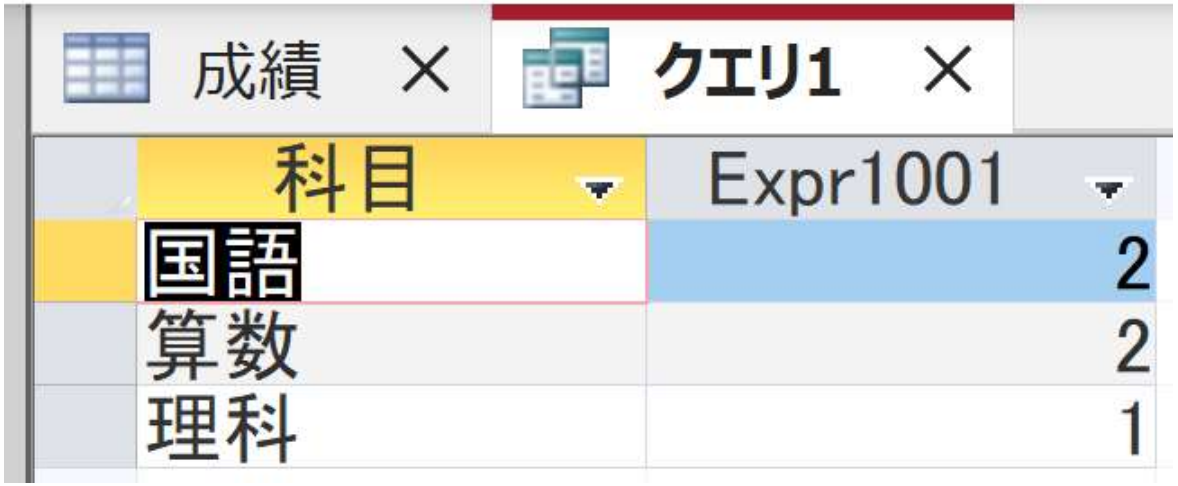
(続き)

3. 国語の得点の平均

SELECT AVG(得点) FROM 成績 WHERE 科目 = '国語';

4. それぞれの科目の受講者数

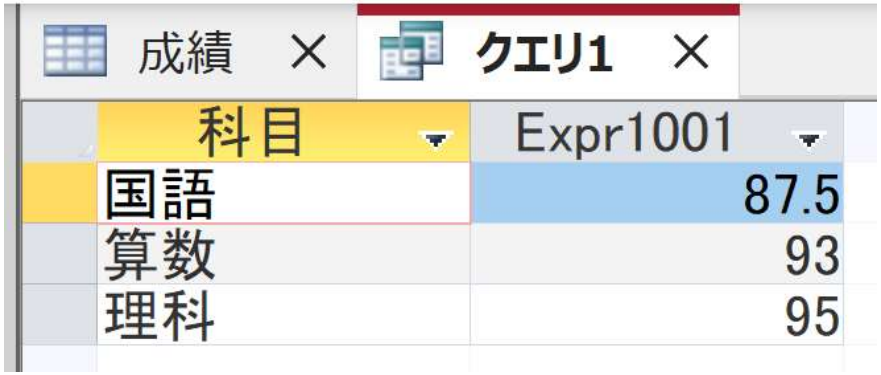
SELECT 科目, COUNT(*) FROM 成績 GROUP BY 科目;



(続き)

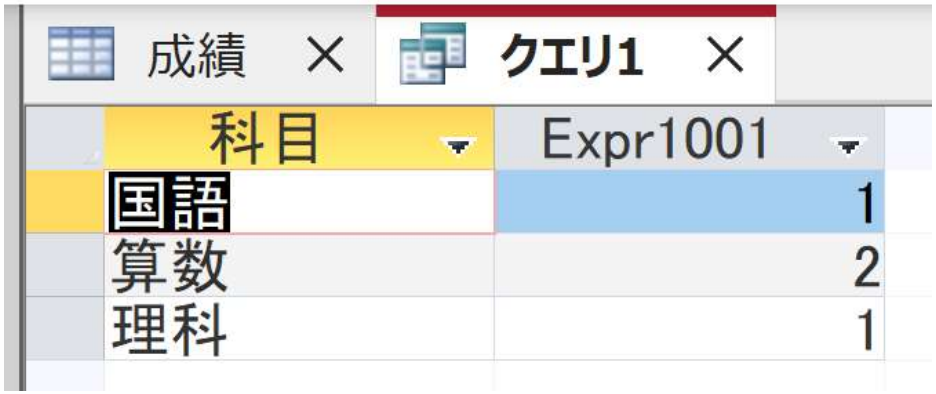
5.それぞれの科目の平均得点

SELECT 科目, AVG(得点) FROM 成績 GROUP BY 科目;



6. それぞれの科目について、得点が90点以上である受講者数

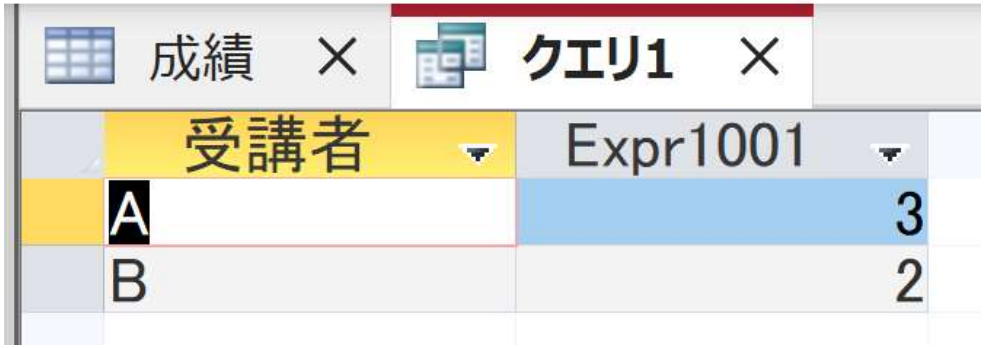
SELECT 科目, COUNT(*) FROM 成績 WHERE 得点 >= 90 GROUP BY 科目;



(続き)

7. それぞれの受講者が受講している科目数

SELECT 受講者, COUNT(*) FROM 成績 GROUP BY 受講者;



8.それぞれの受講者の得点合計

SELECT 受講者, SUM(得点) FROM 成績 GROUP BY 受講者;

(続き)

9. それぞれの受講者の得点平均

SELECT 受講者, AVG(得点) FROM 成績 GROUP BY 受講者;

7-3. 実データを用いた演習

演習の目的と形式

- 目的：実データを使い、グループ化と集約の有用性を確認する。SQLのスキルアップも行う

- 形式：自習形式（**資料を見ながら各自実施してください**）

演習の内容

- **SQL を用いたグループ化と集約**、そのバリエーションと有用性を知る

- **米国成人調査データ**を利用

調査に協力した人たちの年齢分布は？

米国成人調査データ　クエリ1

年齢	Expr1001
17	395
18	550
19	712
20	753
21	720
22	765
23	877
24	798
25	841
26	785
27	835
28	867

教育と年収の関係を見る

米国成人調査データ　クエリ1

教育	年収5万ドル	Expr1002
10th	<=50K	871
10th	>50K	62
11th	<=50K	1115
11th	>50K	60
12th	<=50K	400
12th	>50K	33
1st-4th	<=50K	162
1st-4th	>50K	6
4年制大学	<=50K	3134
4年制大学	>50K	2221
5th-6th	<=50K	317
5th-6th	>50K	16
7th-8th	<=50K	606
7th-8th	>50K	40
9th	<=50K	487
9th	>50K	27
Preschool	<=50K	51
何らかの大学	<=50K	5904
何らかの大学	>50K	1387
高校	<=50K	8826
高校	>50K	1675
職業技術訓練校	<=50K	1021
職業技術訓練校	>50K	361
専門職大学院	<=50K	153
専門職大学院	>50K	423
大学院修士	<=50K	764
大学院修士	>50K	959
大学院博士	<=50K	107
大学院博士	>50K	306
短大、コミュニティカレッジ	<=50K	802
短大、コミュニティカレッジ	>50K	265

実演・実習で使うデータベース

米国成人調査データ

（１９９４年、米国における統計調査データのうち 32561 人分）

米国成人調査データ

ID	年齢	職業の分類	教育	教育年数	職業	性別	週当たり労働時間	母国	年収5万ドル
1	39	州政府	4年制大学	13	管理、事務	男性	40	米国	<=50K
2	50	法人でない自営業	4年制大学	13	執行、経営	男性	13	米国	<=50K
3	38	民間	高校	9	各種取扱者、清掃	男性	40	米国	<=50K
4	53	民間	11th	7	各種取扱者、清掃	男性	40	米国	<=50K
5	28	民間	4年制大学	13	専門職	女性	40	キューバ	<=50K
6	37	民間	大学院修士	14	執行、経営	女性	40	米国	<=50K
7	49	民間	9th	5	その他のサービス	女性	16	ジャマイカ	<=50K
8	52	法人でない自営業	高校	9	執行、経営	男性	45	米国	>50K
9	31	民間	大学院修士	14	専門職	女性	50	米国	>50K
10	42	民間	4年制大学	13	執行、経営	男性	40	米国	>50K
11	37	民間	何らかの大学	10	執行、経営	男性	80	米国	>50K
12	30	州政府	4年制大学	13	専門職	男性	40	インド	>50K
13	23	民間	4年制大学	13	管理、事務	女性	30	米国	<=50K
14	32	民間	短大、コミュニティカレッジ	12	販売	男性	50	米国	<=50K

※　このデータを使います

（演習では、特定の職業、学歴、性別、母国を差別的に見ないようにしてください）

データの出典：Lichman, M. (2013).

UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].

Irvine, CA: University of California, School of Information and Computer Science （米国）

演習用のデータベースファイル

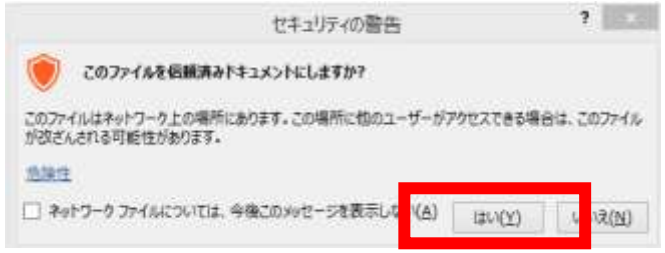
- **演習用の Access データベースファイル**

セレッソの利用者は，セレッソからもダウンロード可能

ファイル名: **db4-4.accdb**

- 「**コンテンツの有効化**」のメッセージが出たときは、確認のうえ、次にすすむ

- つぎのような表示が出たときは、確認のうえ、「**はい**」



米国成人調査データ

ID	年齢	職業の分類	教育	教育年数	職業	性別	週当たり労働時間
4	53	民間	11th	7	各種取扱者、清掃	男性	
5	28	民間	4年制大学	13	専門職	女性	
6	37	民間	大学院修士	14	執行、経営	女性	
7	49	民間	9th	5	その他のサービス	女性	
8	52	法人でない自営業	高校	9	執行、経営	男性	
9	31	民間	大学院修士	14	専門職	女性	
10	42	民間	4年制大学	13	執行、経営	男性	
11	37	民間	何らかの大学	10	執行、経営	男性	
12	30	州政府	4年制大学	13	専門職	男性	
13	23	民間	4年制大学	13	管理、事務	女性	
14	32	民間	短大、コミュニティカレッジ	12	販売	男性	
15	40	民間	職業技術訓練校	11	工作、修理	男性	
16	34	民間	7th-8th	4	運輸、交通	男性	
17	25	法人でない自営業	高校	9	農業、漁業	男性	
18	32	民間	高校	9	機器操作、診断	男性	
19	38	民間	11th	7	販売	男性	
20	43	法人でない自営業	大学院修士	14	執行、経営	女性	
21	40	民間	大学院博士	16	専門職	男性	
22	54	民間	高校	9	その他のサービス	女性	

SQLビューで次を実行し、結果を確認

```
SELECT 年齢, count(*)
FROM 米国成人調査データ
GROUP BY 年齢;
```

調査に協力した人たちの<u>年齢分布</u>は？

米国成人調査データ　クエリ1

年齢	Expr1001
17	395
18	550
19	712
20	753
21	720
22	765
23	877
24	798
25	841
26	785
27	835
28	867

SQLビューで次を実行し、結果を確認

```
SELECT 教育, count(*)
FROM 米国成人調査データ
GROUP BY 教育;
```

調査に協力した人たちの教育の分布は？

米国成人調査データ　クエリ1

教育	Expr1001
10th	933
11th	1175
12th	433
1st-4th	168
4年制大学	5355
5th-6th	333
7th-8th	646
9th	514
Preschool	51
何らかの大学	7291
高校	10501
職業技術訓練校	1382
専門職大学院	576
大学院修士	1723
大学院博士	413
短大、コミュニティカレッジ	1067

SQLビューで次を実行し、結果を確認

```
SELECT 週当たり労働時間, count(*)
FROM 米国成人調査データ
GROUP BY 週当たり労働時間;
```

調査に協力した人たちの週当たり労働時間の分布は？

米国成人調査データ　クエリ1

週当たり労働時間	Expr1001
1	20
2	32
3	39
4	54
5	60
6	64
7	26
8	145
9	18
10	278
11	11
12	173
13	23
14	34
15	404
16	205
17	29

SQLビューで次を実行し、結果を確認

```
SELECT 年収５万ドル以上か, count(*)
FROM 米国成人調査データ
GROUP BY 年収５万ドル以上か;
```

年収５万ドル以上の人とそうでない人の人数

SQLビューで次を実行し、結果を確認

```
SELECT 教育, 年収５万ドル以上か, count(*)
FROM 米国成人調査データ
GROUP BY 教育, 年収５万ドル以上か;
```

教育と年収の関係を見る

米国成人調査データ　クエリ1

教育	年収5万ドル	Expr1002
10th	<=50K	871
10th	>50K	62
11th	<=50K	1115
11th	>50K	60
12th	<=50K	400
12th	>50K	33
1st-4th	<=50K	162
1st-4th	>50K	6
4年制大学	<=50K	3134
4年制大学	>50K	2221
5th-6th	<=50K	317
5th-6th	>50K	16
7th-8th	<=50K	606
7th-8th	>50K	40
9th	<=50K	487
9th	>50K	27
Preschool	<=50K	51
何らかの大学	<=50K	5904
何らかの大学	>50K	1387
高校	<=50K	8826
高校	>50K	1675
職業技術訓練校	<=50K	1021
職業技術訓練校	>50K	361
専門職大学院	<=50K	153
専門職大学院	>50K	423
大学院修士	<=50K	764
大学院修士	>50K	959
大学院博士	<=50K	107
大学院博士	>50K	306
短大、コミュニティカレッジ	<=50K	802
短大、コミュニティカレッジ	>50K	265

① SQLコマンドの習得

SQLを学ぶことで、データベース内の情報を整理し、必要なデータを効率的に抽出する能力が身につく。

② グループ化と集約によるデータの把握

同じ属性を持つデータをグループ化し、AVG, SUM, COUNTなどの集約を用いてそれぞれのグループの平均値、合計、行数などを求める。そのことで、データ全体の傾向を理解し、意味のある洞察を得る。

③ 実践的なデータ分析

集約されたデータを用いて、実際のデータ分析を行い、ビジネスインテリジェンスの観点から売上のトレンドや顧客セグメント分析を実施することが可能になる。このスキルは、リレーショナルデータベースを使用した実践的な問題解決に活用できる。

自習１．テーマ: 科目別の平均得点の計算

目的: GROUP BY を使用して、科目ごとに平均得点を計算する方法を学ぶ。

成績テーブルから、科目ごとに平均得点を計算するSQ文を書いてください。

ヒント: AVG と GROUP BY を組み合わせて使用し、科目 でグループ化します。

自習２．テーマ:　得点が90点以上の受講者数の計算

特定の得点基準を満たす受講者の数を科目ごとに調べる。

各科目について、得点が90点以上である受講者数をカウントするSQL文を書いてください。

ヒント: WHERE を使って得点が90点以上のものを選択。
GROUP BY で科目ごとにグループ化します。

- 自習１の正解例

SELECT 科目, AVG(得点) FROM 成績 GROUP BY 科目;

- 自習２の正解例

SELECT 科目, COUNT(*) FROM 成績 WHERE 得点 >= 90 GROUP BY 科目;