

Базы данных

А5. Анализ данных на языке SQL



Московский государственный технический университет
имени Н.Э. Баумана

Факультет ИБМ

Мар 2025 года

Москва

Артемьев Валерий Иванович © 2025

Инструментальные навыки аналитика данных



- Запросы «на лету» (Ad hoc Query)
- Отчёты (Reporting)
- Интерактивный анализ данных (OLAP)
- Визуализация данных и панели мониторинга (Visualization & Dashboard)
- Профилирование данных (Data Profiling)
- Разведочный анализ данных (EDA)

Запросы «на лету» и отчёты



- MS Excel
 - Импорт таблиц БД
 - Умные таблицы (фильтр, сортировка, агрегаты)
 - QBE (форма ввода, расширенный фильтр)
 - Структуры с уровнями
 - SQL (подключения БД через Power Query)
- MS Access
 - QBE (конструктор запросов, формы поиска)
 - SQL (запросы, **агрегаты**, фильтры, сортировка)
 - Конструктор отчётов
- Генераторы отчётов

Интерактивный анализ данных



- MS Excel
 - Умные таблицы
 - Сводные таблицы
 - Power Pivot
- MS Access
 - Сводные таблицы
 - SQL PIVOT / UNPIVOT
- Средства бизнес-аналитики (BI)

Визуализация данных и панели мониторинга



- MS Excel
 - Умные таблицы
 - Сводные таблицы
 - Сводные диаграммы
 - Power Pivot
- MS Access
 - Сводные таблицы
 - Формы
 - Диаграммы
- Средства бизнес-аналитики (BI)

Визуализация данных и панели мониторинга



- MS Excel
 - Умные таблицы
 - Сводные таблицы
 - Сводные диаграммы
 - Power Pivot
- MS Access
 - Сводные таблицы
 - Формы
 - Диаграммы
- Средства бизнес-аналитики (BI)

Агрегирование данных GROUP BY

Элементы вывода запроса SELECT, содержащие выражения с *агрегатными функциями*, вычисляются и *группируются* по неаггегированным полям из списка GROUP BY.

Агрегатные функции SQL:

- Count(*выражение*) подсчёт значений без NULL
- Sum(*выражение*) сумма значений
- Avg(*выражение*) среднее значение
- Min, Max(*выражение*) минимум / максимум

SELECT DISTINCT *элементы вывода с агрегатами*
FROM *таблица*
WHERE *условия фильтрации*
GROUP BY *список полей для группировки*
HAVING *фильтр с агрегатами*
ORDER BY *список элементов сортировки*
LIMIT *количество записей*

Примеры агрегирования данных (1)

- Для анализа важно сформулировать *показатели или KPI*
- Назовите, какие показатели важны для управления кадрами
- Нужно понимать, *как считать эти метрики*
- Часто метрики определяются *путём агрегирования*
- Существуют *абсолютные и относительные метрики*
- *Одиночные показатели или распределения*
- Распределения *дискретные или непрерывные*
- Результаты в виде *обычных или сводных таблиц*
- Результаты в виде *диаграмм и графиков*

Примеры агрегирования данных (2)

Определение численности сотрудников:

- общей
- по полу
- по образованию

Минимальный, средний и максимальный:

- возраст сотрудника
- стаж работы

Доля сотрудников:

- женщин
- с высшим образованием
- пенсионеров

Распределение сотрудников:

- по возрасту
- по стажу
- по возрасту-полу
- по стажу-полу

Определение численности (1)

```
SELECT COUNT(*) AS [Всего сотрудников]  
FROM СОТРУДНИКИ;
```



31 Общее число сотрудников					
Всего сотрудников					
	11				

Определите, сколько мужчин на предприятии

Определение численности (2)



```
SELECT COUNT(*) AS [Всего сотрудников]  
FROM СОТРУДНИКИ;
```

31 Общее число сотрудников					
Всего сотрудников					
	11				

```
SELECT COUNT(Пол) AS [Всего мужчин]  
FROM СОТРУДНИКИ  
WHERE Пол="М";
```

32 Сколько мужчин					
Всего мужчин					
	7				

Расчёт численности в разрезе (1)

```
SELECT Пол, COUNT(Пол) AS [Кол-во]  
FROM СОТРУДНИКИ  
GROUP BY Пол;
```



Пол	Кол-во				
Ж	4				
М	7				

Рассчитайте численность в разрезе образования

Расчёт численности в разрезе (2)

```
SELECT Пол, COUNT(Пол) AS [Кол-во]  
FROM СОТРУДНИКИ  
GROUP BY Пол;
```



Пол	Кол-во
Ж	4
М	7

```
SELECT Образование, COUNT(Образование) AS [Кол-во]  
FROM СОТРУДНИКИ  
GROUP BY Образование;
```

Образование	Кол-во
высшее	4
среднее	5
среднее специальное	2

Расчёт численности в 2-х разрезах (1)

**SELECT Пол, Образование, COUNT(Образование) AS [Кол-во]
FROM СОТРУДНИКИ GROUP BY Пол, Образование;**



Пол	Образование	Кол-во
Ж	высшее	2
Ж	среднее	1
Ж	среднее специальное	1
М	высшее	2
М	среднее	4
М	среднее специальное	1

Постройте приведённую ниже сводную таблицу, используя суммирование SUM условий по полу с помощью функции ЕСЛИ, например IIF(Пол="М",1,0).

Образование	Мужчины	Женщины
высшее	2	2
среднее	4	1
среднее специальное	1	1

Расчёт численности в виде сводной таблицы для 2-х разрезов

**SELECT Пол, Образование, COUNT(Образование) AS [Кол-во]
FROM СОТРУДНИКИ GROUP BY Пол, Образование;**



Пол	Образование	Кол-во
Ж	высшее	2
Ж	среднее	1
Ж	среднее специальное	1
М	высшее	2
М	среднее	4
М	среднее специальное	1

**SELECT Образование, SUM(IIF(Пол="М",1,0)) AS Мужчины,
SUM(IIF(Пол="Ж",1,0)) AS Женщины
FROM СОТРУДНИКИ GROUP BY Образование;**

Образование	Мужчины	Женщины
высшее	2	2
среднее	4	1
среднее специальное	1	1

Трюк для расчёта численности в виде сводной таблицы (1)



Для подсчёта количества по какому либо признаку атрибута можно использовать суммирование условий равенства атрибута этому признаку. Нужно только учитывать, что в MS Access Истина равна -1.

```
SELECT Образование,  
SUM(-(Пол="М")) AS Мужчины,  
SUM(-(Пол="Ж")) AS Женщины  
FROM СОТРУДНИКИ  
GROUP BY Образование;
```

Образование	Мужчины	Женщины
высшее	2	2
среднее	4	1
среднее специальное	1	1

Добавьте столбец итога в сводную таблицу

Трюк для расчёта численности в виде сводной таблицы (2)



```
SELECT Образование, SUM(-(Пол="М")) AS Мужчины,  
SUM(-(Пол="Ж")) AS Женщины  
FROM СОТРУДНИКИ GROUP BY Образование;
```

Образование	Мужчины	Женщины
высшее	2	2
среднее	4	1
среднее специальное	1	1

```
SELECT Образование, SUM(-(Пол="М")) AS Мужчины,  
SUM(-(Пол="Ж")) AS Женщины, COUNT(Образование) AS Итого  
FROM СОТРУДНИКИ GROUP BY Образование;
```

Образование	Мужчины	Женщины	Итого
высшее	2	2	4
среднее	4	1	5
среднее специальное	1	1	2

Построение таблицы распределения для гистограммы с подзапросом (1)



Для распределения непрерывных величин (возраст, стаж) надо определять диапазоны, куда входят эти величины.

Для возрастных групп по 10 лет надо в MS Access вычислять:

Round(DateDiff("уууу", [Дата рождения], Now)/10)*10.

В других СУБД работает:

Round(DateDiff("уууу", [Дата рождения], Now)/10, -1)

Решая в лоб, получим:

SELECT Round(DateDiff("уууу", [Дата рождения], Now)/10)*10 &"-летние"

AS Возраст,

COUNT(Round(DateDiff("уууу", [Дата рождения], Now)/10)*10) AS [Кол-во]

FROM СОТРУДНИКИ

GROUP BY Round(DateDiff("уууу", [Дата рождения], Now)/10)*10;

Возраст	Кол-во
20-летние	1
30-летние	1
40-летние	3
50-летние	2
60-летние	4

Как устранить этот кошмар?

Построение таблицы распределения для гистограммы с подзапросом (2)



Сначала построим подзапрос, который формирует таблицу с одним столбцом группы возраста:

```
SELECT Round(DateDiff("уууу", [Дата рождения], Now)/10)*10  
AS ВозрастГруппа FROM СОТРУДНИКИ
```

И уже из этой таблицы будем строить распределение:

```
SELECT ВозрастГруппа &"-летние" AS Возраст, COUNT(ВозрастГруппа) AS [Кол-во]  
FROM (SELECT Round(DateDiff("уууу", [Дата рождения], Now)/10)*10 AS  
ВозрастГруппа FROM СОТРУДНИКИ) AS Группы  
GROUP BY ВозрастГруппа;
```

Возраст	Кол-во
20-летние	1
30-летние	1
40-летние	3
50-летние	2
60-летние	4

Таблица распределения для гистограммы с несколькими разрезами (1)

Запрос для распределения пол – возраст с подзапросом:



```
SELECT Пол, ВозрастГруппа &"-летние" AS Возраст,  
COUNT(ВозрастГруппа) AS [Кол-во]  
FROM
```

```
(SELECT Пол, Round(DateDiff("уууу", [Дата рождения], Now)/10)*10 AS  
ВозрастГруппа FROM СОТРУДНИКИ) AS Пол_Возраст  
GROUP BY Пол, ВозрастГруппа;
```

38 Распределение пол-возраст			
	Пол	Возраст	Кол-во
	Ж	30-летние	1
	Ж	50-летние	1
	Ж	60-летние	2
	М	20-летние	1
	М	40-летние	3
	М	50-летние	1
	М	60-летние	2

Постройте распределение возраст - пол - образование

Таблица распределения для гистограммы с несколькими разрезами (2)



```
SELECT ВозрастГруппа &"-летние" AS Возраст, Пол, Образование,  
COUNT(ВозрастГруппа) AS [Кол-во]  
FROM
```

```
(SELECT Образование, Пол, Round(DateDiff("уууу", [Дата рождения], Now)/10)*10  
AS ВозрастГруппа FROM СОТРУДНИКИ) AS Группы  
GROUP BY ВозрастГруппа, Пол, Образование  
ORDER BY ВозрастГруппа, Пол, Образование;
```

Возраст	Пол	Образование	Кол-во
20-летние	М	среднее	1
30-летние	Ж	высшее	1
40-летние	М	высшее	1
40-летние	М	среднее	2
50-летние	Ж	высшее	1
50-летние	М	среднее специальное	1
60-летние	Ж	среднее	1
60-летние	Ж	среднее специальное	1
60-летние	М	высшее	1
60-летние	М	среднее	1

Постройте распределение по стажу с подзапросом

Таблица распределения численности по стажу работы



```
SELECT "от " & СтажГруппа & " лет" AS Стаж,  
COUNT(СтажГруппа) AS [Кол-во]  
FROM
```

```
(SELECT Round(DateDiff("уууу", [Дата приёма], Now)/10)*10 AS  
СтажГруппа FROM СОТРУДНИКИ) AS Группы  
GROUP BY СтажГруппа  
ORDER BY СтажГруппа;
```

40 Распределение по стажу работы			
Стаж	Кол-во		
от 10 лет	4		
от 20 лет	3		
от 30 лет	3		
от 40 лет	1		

Постройте распределение стаж - пол с подзапросом

Таблица распределения численности по стажу и полу



```
SELECT "от " & СтажГруппа & " лет" AS Стаж, Пол,  
COUNT(СтажГруппа) AS [Кол-во]  
FROM
```

```
    (SELECT Round(DateDiff("уууу", [Дата приёма], Now)/10)*10 AS  
    СтажГруппа, Пол FROM СОТРУДНИКИ)  
GROUP BY СтажГруппа, Пол  
ORDER BY СтажГруппа, Пол;
```

Стаж	Пол	Кол-во
от 10 лет	Ж	2
от 30 лет	Ж	1
от 40 лет	Ж	1
от 10 лет	М	2
от 20 лет	М	3
от 30 лет	М	2

Таблица распределения численности по стажу, полу и образованию



```
SELECT "от "& СтажГруппа & " лет" AS Стаж,  
Пол, Образование, COUNT(СтажГруппа) AS [Кол-во]  
FROM
```

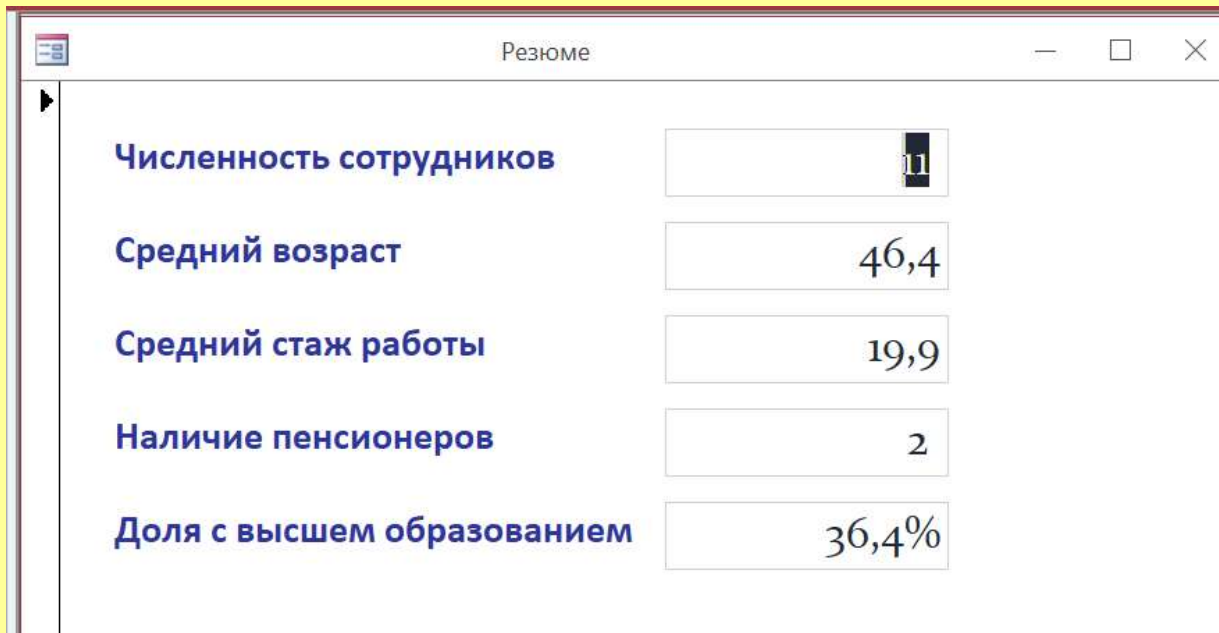
```
(SELECT Образование, Пол, Round(DateDiff("уууу", [Дата приёма],  
Now)/10)*10 AS СтажГруппа FROM СОТРУДНИКИ) AS Группы  
GROUP BY СтажГруппа, Пол, Образование  
ORDER BY СтажГруппа, Пол, Образование;
```

Стаж	Пол	Образование	Кол-во
от 10 лет	Ж	высшее	1
от 10 лет	Ж	среднее специальное	1
от 10 лет	М	среднее	2
от 20 лет	М	высшее	1
от 20 лет	М	среднее	2
от 30 лет	Ж	высшее	1
от 30 лет	М	высшее	1
от 30 лет	М	среднее специальное	1
от 40 лет	Ж	среднее	1

Вычислите и выведите важные показатели управления кадрами

Вычисление резюме с важными показателями управления кадрами

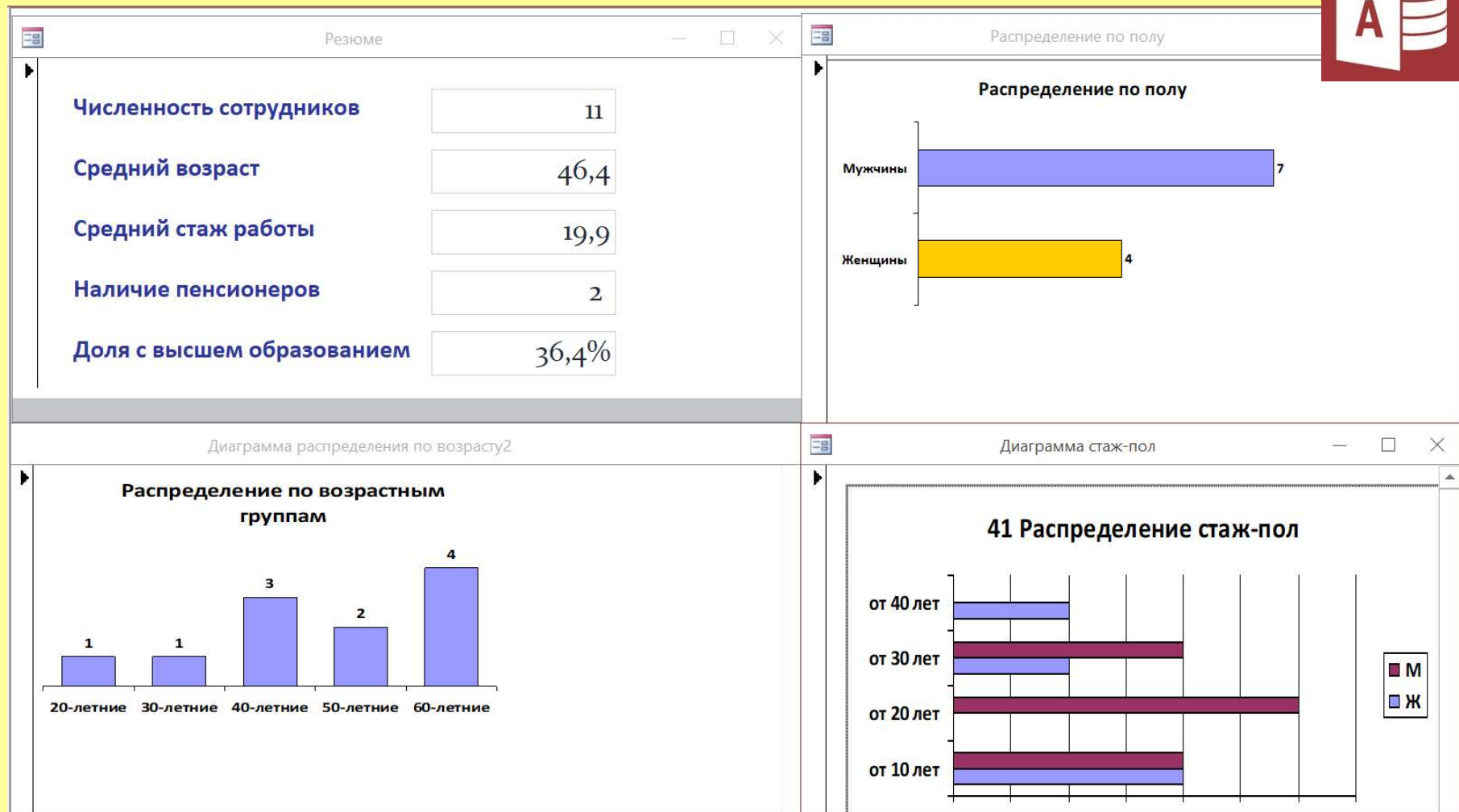
```
SELECT Count([Сотрудник_ид]) AS Численность,  
       Round(Year(Now)-AVG(Year([Дата рождения])), 1) AS [Средний возраст],  
       Round(Year(Now)-AVG(Year([Дата приёма])), 1) AS [Средний стаж  
работы],  
       -SUM(Пенсионер) AS [Наличие  
пенсионеров],  
       -Round(SUM(Образование="высшее")/COUNT(*), 3) AS [Доля с высшем  
образованием]  
FROM СОТРУДНИКИ;
```



The screenshot shows a Microsoft Access window titled "Резюме". It displays a list of five calculated metrics with their corresponding values in input fields:

Показатель	Значение
Численность сотрудников	11
Средний возраст	46,4
Средний стаж работы	19,9
Наличие пенсионеров	2
Доля с высшем образованием	36,4%

Пример вывода ключевых данных по управлению кадрами



**Терпения и удачи всем, кто
связан с базами данных**

Спасибо за внимание!

Валерий Иванович Артемьев

МГТУ имени Н.Э. Баумана, кафедра ИУ-5

Банк России

Департамент данных, проектов и процессов

Тел.: +7(495) 753-96-25

e-mail: viart@bmstu.ru