**Анализ качества данных и рекомендации**

**по устранению ошибок**

Голубева Екатерина

Санкт-Петербург

2024

**Содержание**

[1 Таблица базы\_данных 3](#_Toc171533906)

[2 Таблица базы\_данных\_и\_уровень\_знаний\_сотру 4](#_Toc171533907)

[3 Таблица инструменты 6](#_Toc171533908)

[4 Таблица инструменты\_и\_уровень\_знаний\_сотр 7](#_Toc171533909)

[5 Таблица образование\_пользователей 9](#_Toc171533910)

[6 Таблица опыт\_сотрудника\_в\_отраслях 11](#_Toc171533911)

[7 Таблица опыт\_сотрудника\_в\_предметных\_обла 12](#_Toc171533912)

[8 Таблица отрасли 14](#_Toc171533913)

[9 Таблица платформы 15](#_Toc171533914)

[10 Таблица платформы\_и\_уровень\_знаний\_сотруд 16](#_Toc171533915)

[11 Таблица предметная\_область 18](#_Toc171533916)

[12 Таблица резюмедар 19](#_Toc171533917)

[13 Таблица сертификаты\_пользователей 20](#_Toc171533918)

[14 Таблица сотрудники\_дар 21](#_Toc171533919)

[15 Таблица среды\_разработки 22](#_Toc171533920)

[16 Таблица среды\_разработки\_и\_уровень\_знаний\_ 23](#_Toc171533921)

[18 Таблица технологии\_и\_уровень\_знаний\_сотру 25](#_Toc171533922)

[18 Таблица типы\_систем\_и\_уровень\_знаний\_сотру 26](#_Toc171533923)

[27 Таблица фреймворки\_и\_уровень\_знаний\_сотру 27](#_Toc171533924)

[28 Таблица языки\_пользователей 29](#_Toc171533925)

[30 Таблица языки\_программирования\_и\_уровень 30](#_Toc171533926)

[Рекомендации по улучшению качества данных 31](#_Toc171533927)

## **Таблица базы\_данных**

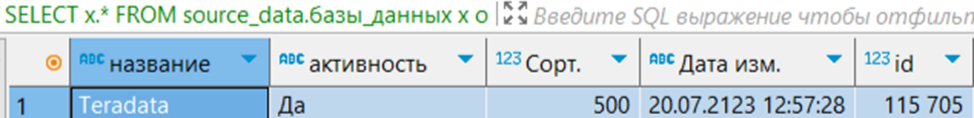
1. Столбец «название» должен быть переименован с заглавной буквы: «Название»
2. Столбец «Активность» должен быть переименован с заглавной буквы: «Активность»



1. Столбец «Сорт.» должен быть удален, так как не несет смысловой нагрузки
2. «Дата изм.» должен быть в формате DATE, время является избыточной информацией.
3. В поле Дата изм. значение имеет некорректные данные: 2123 г.

**SELECT** ***x***.\* **FROM** source\_data.базы\_данных x

**order** **by** **"Дата изм."** **ASC**



## **Таблица базы\_данных\_и\_уровень\_знаний\_сотру**

1. Столбец «название» следует переименовать в User ID и поставить тип данных int
2. Полные дубликаты строк найдены по запросу:

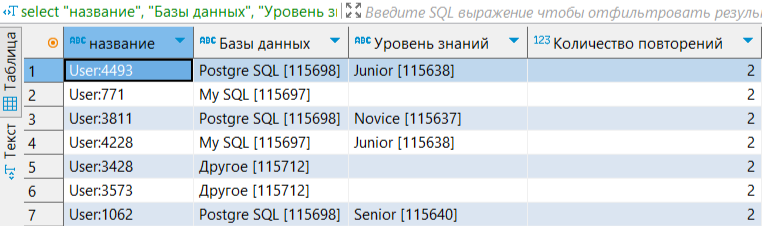
**select** **"название"**, **"Базы данных"**, **"Уровень знаний"**, **count**(\*)

**from** source\_data.базы\_данных\_и\_уровень\_знаний\_сотру x

**group** **by** "название", "Базы данных", "Уровень знаний"

**having** **count**(\*) >1

Результат:



1. Удалось выявить пустые значения в столбце «Уровень Знаний»:

**SELECT** ***x***.\* **FROM** source\_data.базы\_данных\_и\_уровень\_знаний\_сотру x

**WHERE** **"Уровень знаний"** = **''**

**order** **by** 8

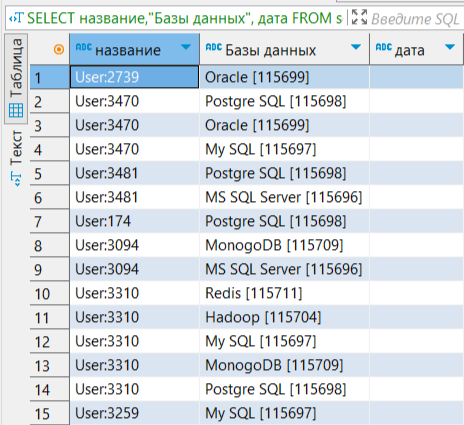
|  |  |
| --- | --- |
|  |  |

1. Отсутствуют некоторые данные в поле Дата, как NULL не обозначены:

**SELECT** ***x***.\* **FROM** source\_data.базы\_данных\_и\_уровень\_знаний\_сотру x

**WHERE** **"дата"** = **''**

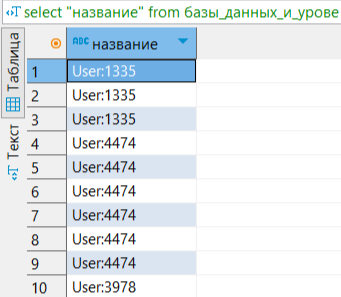
**order** **by** дата



1. User ID обозначен как «название», требуется переименовать поле в «User ID». Также в поле неверный тип данных, следует изменить формат на int.

**select** **"название"**

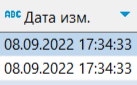
**from** базы\_данных\_и\_уровень\_знаний\_сотру



1. Поле «Сорт.» должно быть удалено, так как не несет смысловой нагрузки.

## **Таблица инструменты**

1. Неверный формат Даты изм. – text, следует изменить на date/



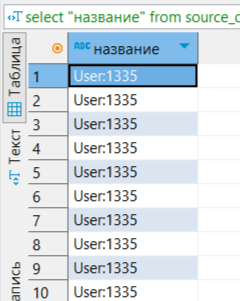
1. Столбец «Сорт.» должен быть удален, так как не несет смысловой нагрузки

## **Таблица инструменты\_и\_уровень\_знаний\_сотр**

1. User ID обозначен как «название», требуется переименовать поле в «User ID». Также в поле неверный тип данных, следует изменить формат на int.

**select** **"название"**

**from** source\_data.инструменты\_и\_уровень\_знаний\_сотр x

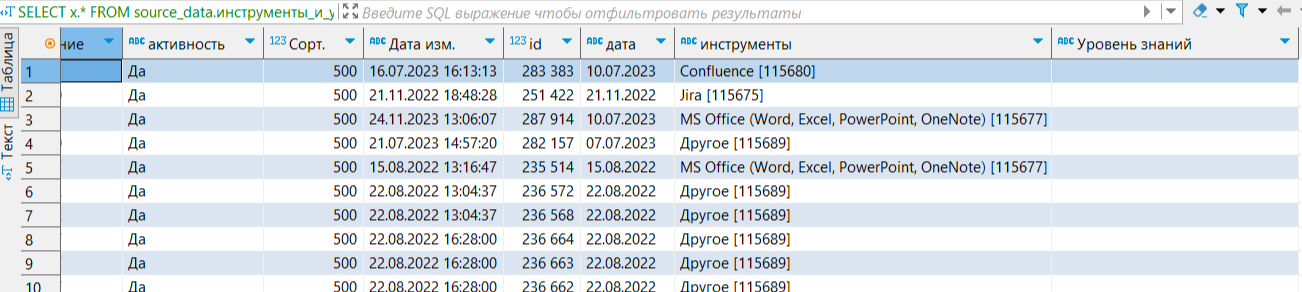


1. Отсутствуют уровни знаний сотрудников:

**SELECT** ***x***.\* **FROM** source\_data.инструменты\_и\_уровень\_знаний\_сотр x

**WHERE** **"Уровень знаний"** = **''**

**order** **by** 8

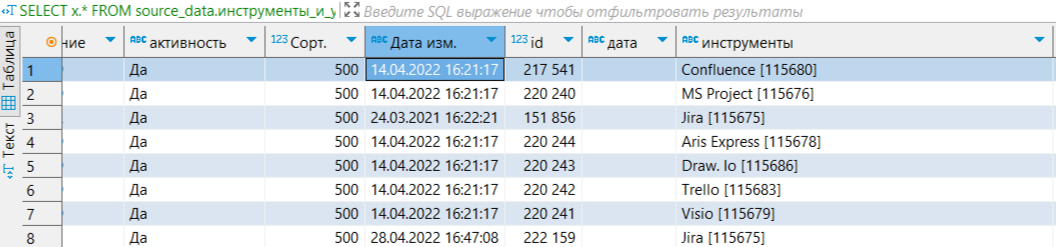


1. Отсутствуют даты

**SELECT** ***x***.\* **FROM** source\_data.инструменты\_и\_уровень\_знаний\_сотр x

**WHERE** **"дата"** = **''**

**order** **by** 6



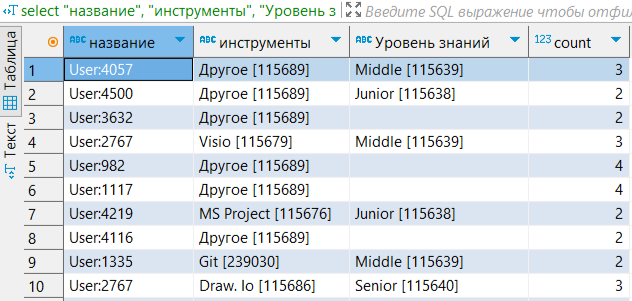
1. Присутствуют дубликаты строк, тем не менее «другое» имеет право на повторение, так как конкретное название не обозначено:

**select** **"название"**, **"инструменты"**, **"Уровень знаний"**, **count**(\*)

**from** source\_data.инструменты\_и\_уровень\_знаний\_сотр x

**group** **by** "название", "инструменты", "Уровень знаний"

**having** **count**(\*) >1



1. Поле «Сорт.» должно быть удалено, так как не несет смысловой нагрузки.

## **Таблица образование\_пользователей**

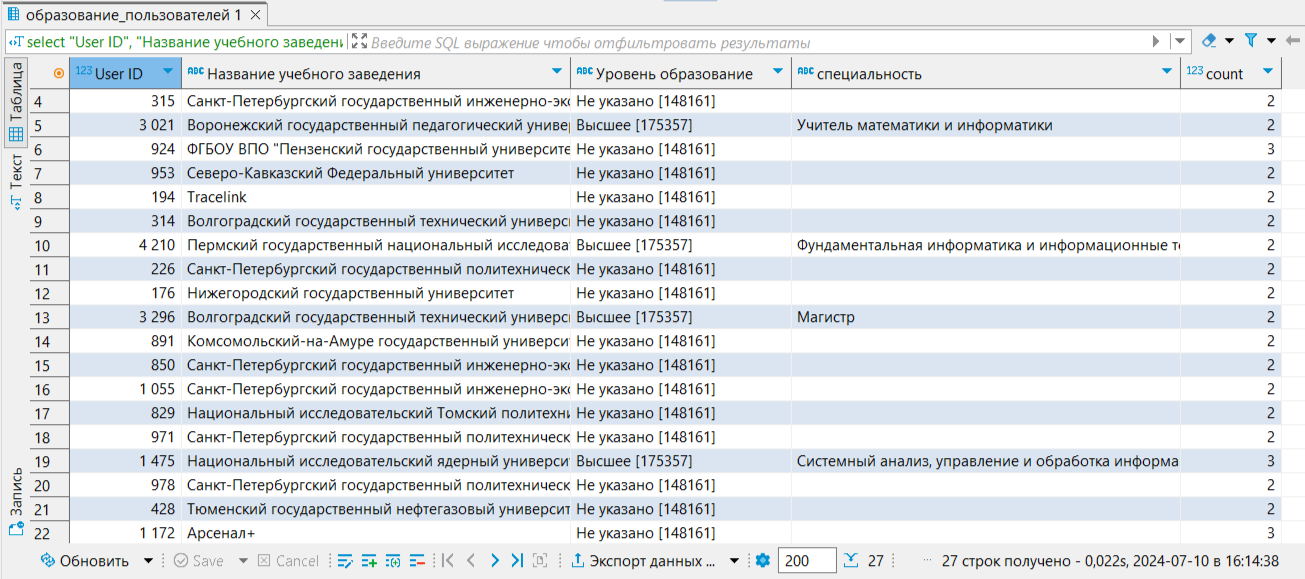
1. Дубликаты строк:

**select** **"User ID"**, **"Название учебного заведения"**,**"Уровень образование"**,специальность, **count** (\*)

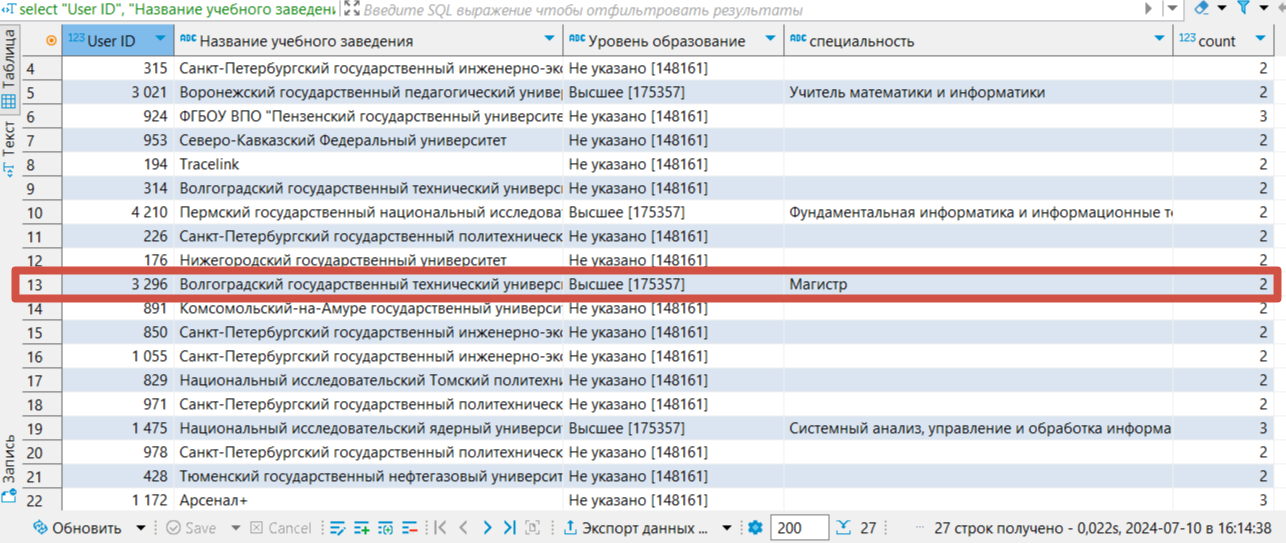
**from** source\_data.образование\_пользователей оп

**group** **by** "User ID", "Название учебного заведения", "Уровень образование", специальность

**having** **count**(\*) >1



1. Ручным методом было выяснено, что перепутана специальность и квалификация у пользователя 3 296:

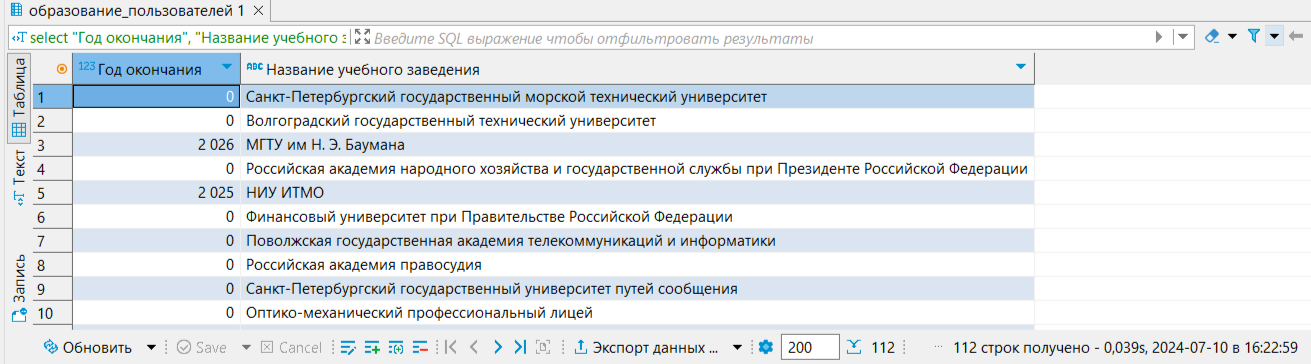


1. Не корректно указанный год окончания:

**select** **"Год окончания"**, **"Название учебного заведения"**

**from** source\_data.образование\_пользователей оп

**where** **"Год окончания"** **not** **between** 1920 **and** 2024



1. Поле «Сорт.» должно быть удалено, так как не несет смысловой нагрузки.
2. Название поля "Уровень образование" содержит ошибку – перименовать в «Уровень образования».

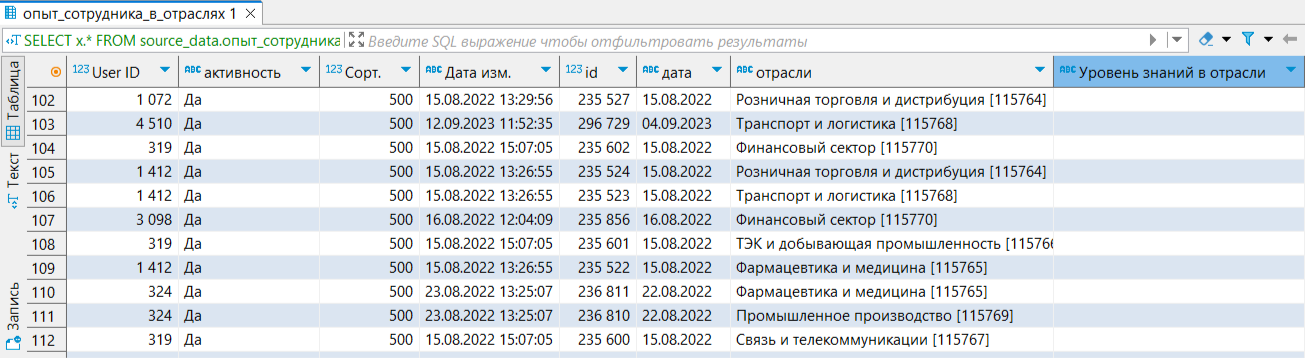
## **Таблица опыт\_сотрудника\_в\_отраслях**

1. Не указан опыт сотрудников (112 строк):

**SELECT** ***x***.\* **FROM** source\_data.опыт\_сотрудника\_в\_отраслях x

**WHERE** **"Уровень знаний в отрасли"** = **''**

**order** **by** **"Уровень знаний в отрасли"**



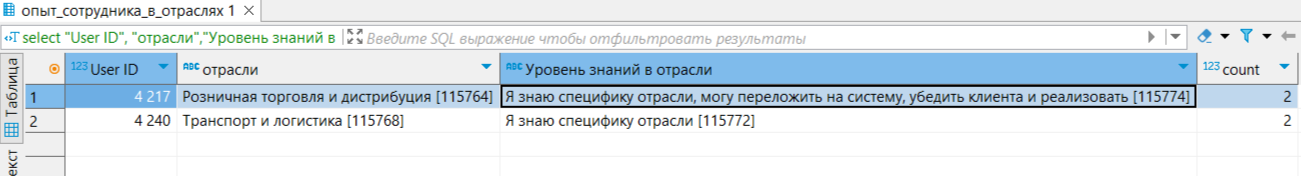
1. Дубликаты строк

**select** **"User ID"**, **"отрасли"**,**"Уровень знаний в отрасли"**, **count** (\*)

**from** source\_data.опыт\_сотрудника\_в\_отраслях x

**group** **by** "User ID", "отрасли", "Уровень знаний в отрасли"

**having** **count**(\*) >1

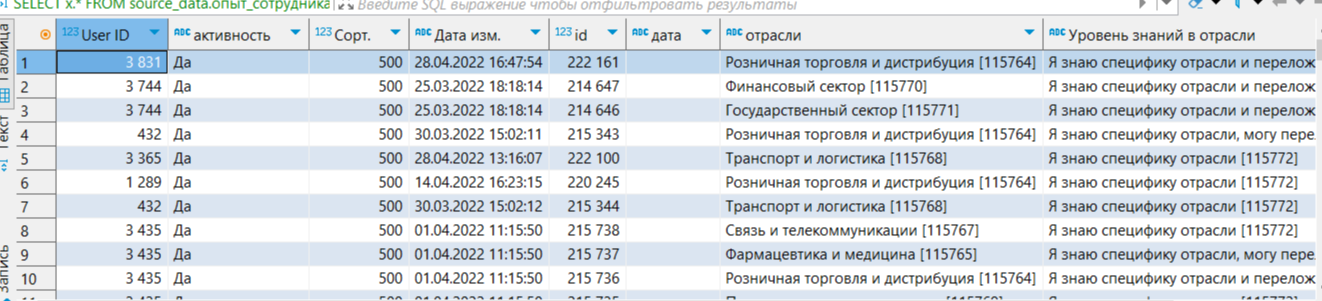


1. Отсутствие даты (56 стр):

**SELECT** ***x***.\* **FROM** source\_data.опыт\_сотрудника\_в\_отраслях x

**WHERE** **"дата"** = **''**

**order** **by** дата



1. Поле «Сорт.» должно быть удалено, так как не несет смысловой нагрузки.

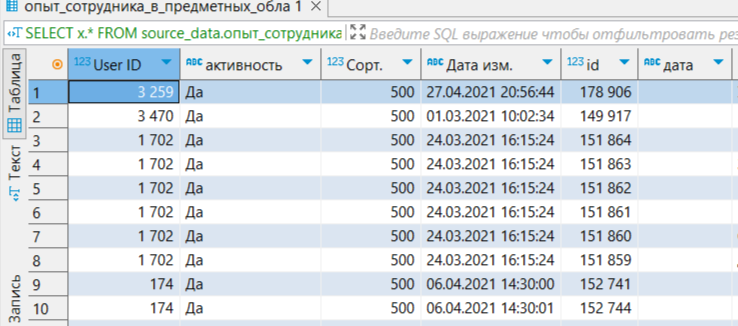
## **Таблица опыт\_сотрудника\_в\_предметных\_обла**

1. Отсутствует дата (66 стр):

**SELECT** ***x***.\* **FROM** source\_data.опыт\_сотрудника\_в\_предметных\_обла x

**WHERE** **"дата"** = **''**

**order** **by** дата



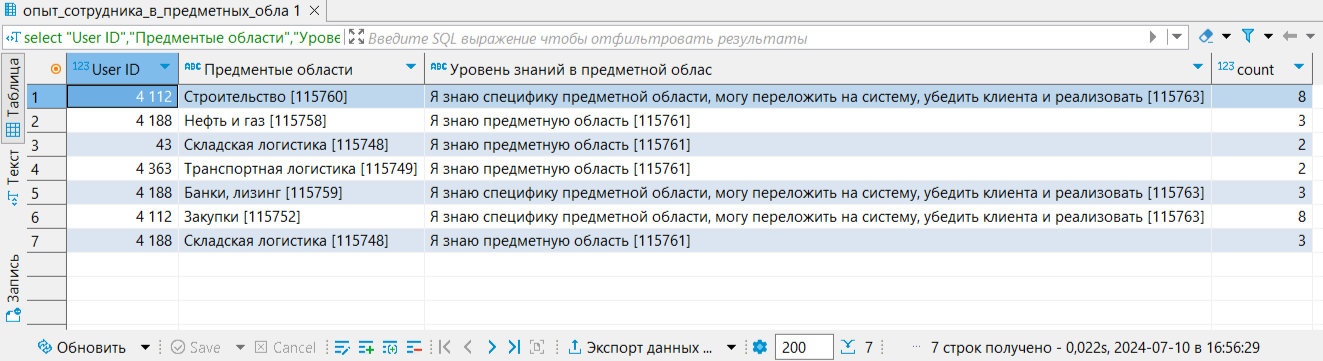
1. Дубликаты строк:

**select** **"User ID"**,**"Предментые области"**,**"Уровень знаний в предметной облас"**, **count** (\*)

**from** опыт\_сотрудника\_в\_предметных\_обла x

**group** **by** "User ID", "Предментые области","Уровень знаний в предметной облас"

**having** **count**(\*) >1

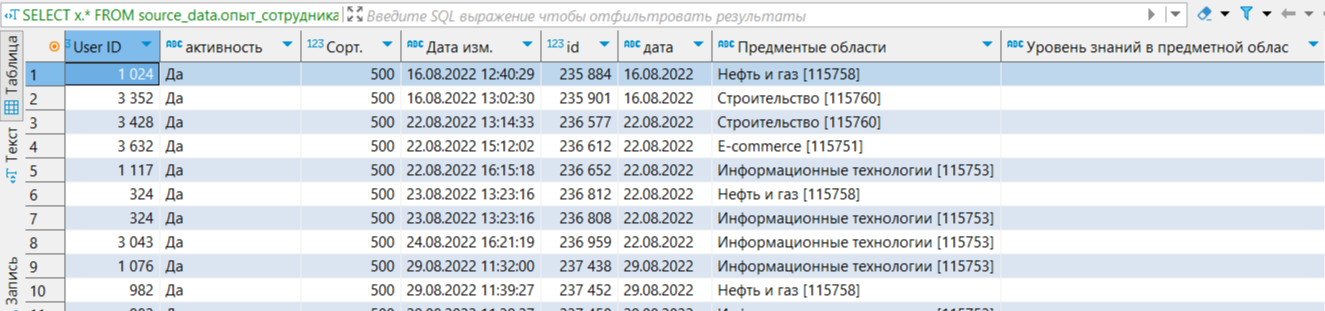


1. Поле «Сорт.» должно быть удалено, так как не несет смысловой нагрузки.
2. Отсутствуют уровни знаний (38 стр):

**SELECT** ***x***.\* **FROM** source\_data.опыт\_сотрудника\_в\_предметных\_обла x

**WHERE** **"Уровень знаний в предметной облас "** = **''**

**order** **by** **"Уровень знаний в предметной облас "**



## **Таблица отрасли**

1. Изменить формат даты на date
2. Поле «Сорт.» должно быть удалено, так как не несет смысловой нагрузки.

## **Таблица платформы**

1. Поле «Сорт.» должно быть удалено, так как не несет смысловой нагрузки.

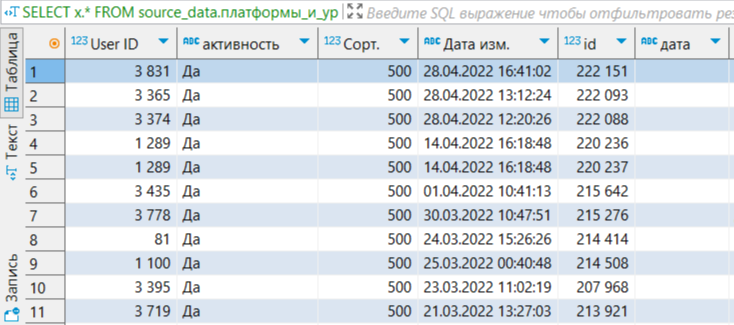
## **Таблица платформы\_и\_уровень\_знаний\_сотруд**

1. Отсутствующие даты (114 стр):

**SELECT** ***x***.\* **FROM** source\_data.платформы\_и\_уровень\_знаний\_сотруд x

**WHERE** **"дата"** = **''**

**ORDER** **BY** дата

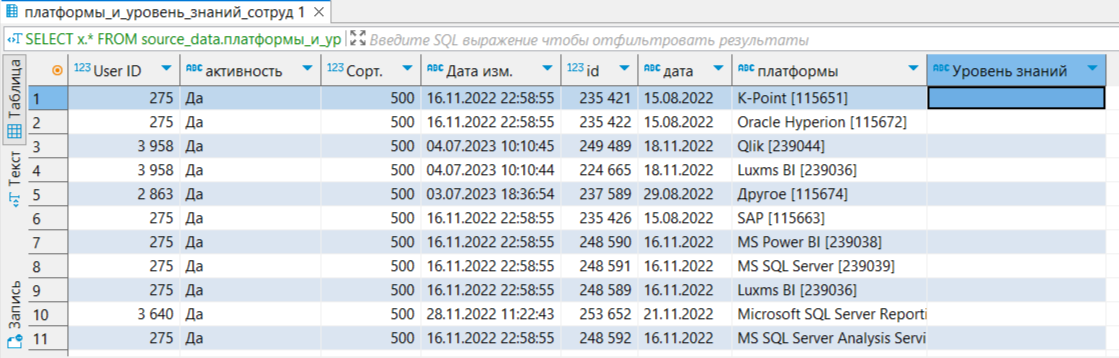


1. Отсутствующие уровни знания (207 стр):

**SELECT** ***x***.\* **FROM** source\_data.платформы\_и\_уровень\_знаний\_сотруд x

**WHERE** **"Уровень знаний"** = **''**

**order** **by** **"Уровень знаний"**



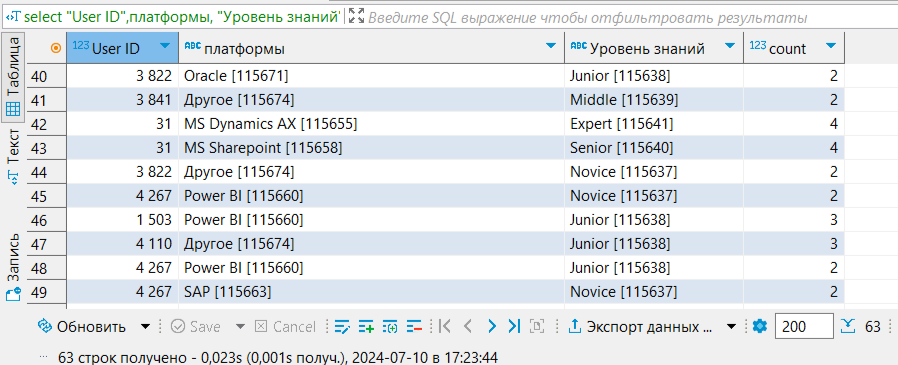
1. Дубликаты строк

**select** **"User ID"**,платформы, **"Уровень знаний"**, **count** (\*)

**from** source\_data.платформы\_и\_уровень\_знаний\_сотруд x

**group** **by** "User ID",платформы, "Уровень знаний"

**having** **count**(\*) >1



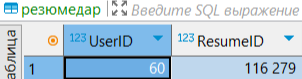
1. Поле «Сорт.» должно быть удалено.

## **Таблица предметная\_область**

1. Поле «Сорт.» должно быть удалено.

## **Таблица резюмедар**

1. Поле "UserID" следует переименовать в "User ID" (с пробелом), так же как и «ResumeID”



## **Таблица сертификаты\_пользователей**

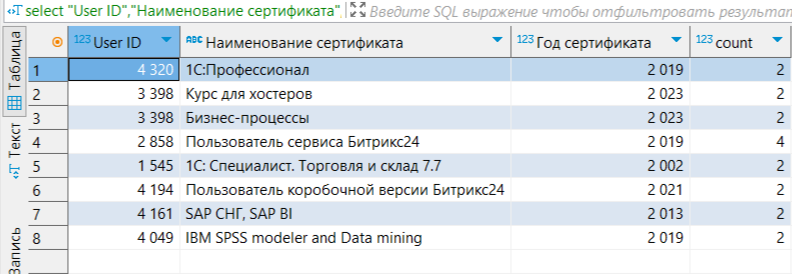
1. Дубликаты строк:

**select** **"User ID"**,**"Наименование сертификата"**, **"Год сертификата"**, **count** (\*)

**from** source\_data.сертификаты\_пользователей x

**group** **by** "User ID","Наименование сертификата", "Год сертификата"

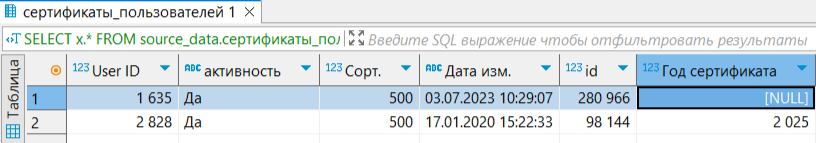
**having** **count**(\*) >1



1. Значение года, который еще не наступил и NULL:

**SELECT** ***x***.\* **FROM** source\_data.сертификаты\_пользователей x

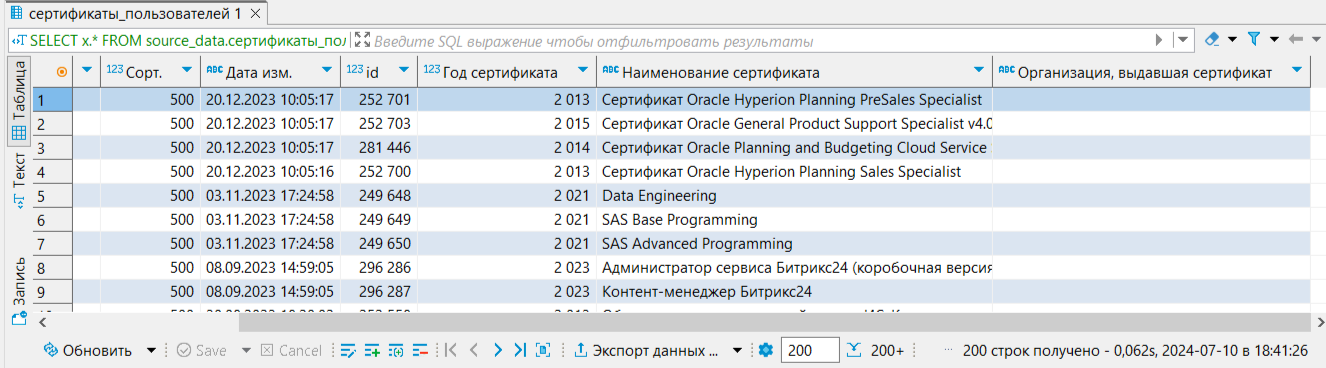
**where** **"Год сертификата"**>2024 **or** **"Год сертификата"** **is** **null**

****

1. Пустые значения в поле «Организация, выдавшая сертификат»

**SELECT** ***x***.\* **FROM** source\_data.сертификаты\_пользователей x

**where** **"Организация, выдавшая сертификат"** = **''**

****

## **Таблица сотрудники\_дар**

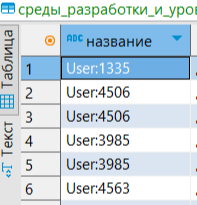
1. Удалить пустые поля.

## **Таблица среды\_разработки**

1. Поле «Сорт.» может быть удалено.

## **Таблица среды\_разработки\_и\_уровень\_знаний\_**

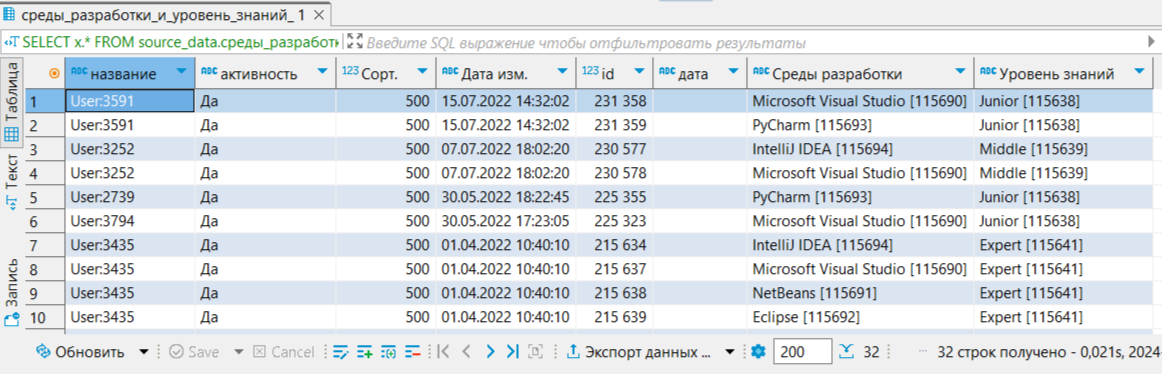
1. Поле «название» должно быть переименовано в «User ID», данные преобразованы в int.



1. Поле «Сорт.» может быть удалено.
2. Отсутствуют значения даты

**SELECT** ***x***.\* **FROM** source\_data.среды\_разработки\_и\_уровень\_знаний\_ x

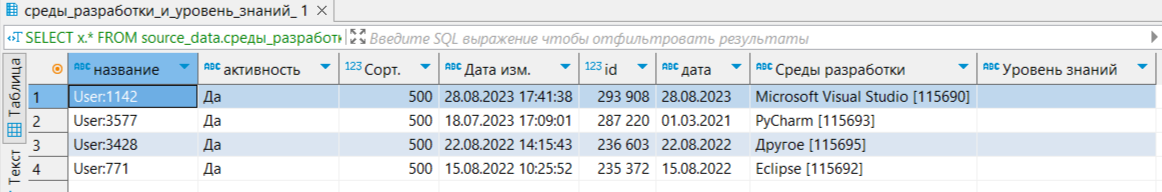
**where** **"дата"** = **''**



1. Отсутствует значение уровня знаний

**SELECT** ***x***.\* **FROM** source\_data.среды\_разработки\_и\_уровень\_знаний\_ x

**where** **"Уровень знаний"** = **''**



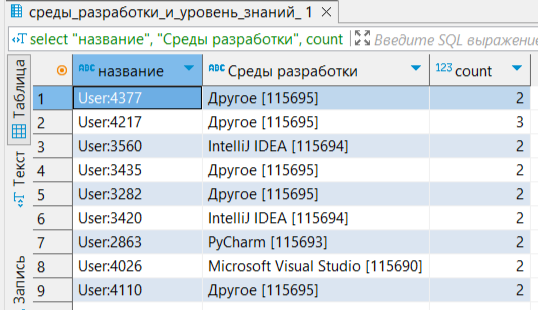
1. Дубликаты:

**select** **"название"**, **"Среды разработки"**, **count** (\*)

**from** source\_data.среды\_разработки\_и\_уровень\_знаний\_ x

**group** **by** "название", "Среды разработки"

**having** **count**(\*) >1



## **18 Таблица технологии\_и\_уровень\_знаний\_сотру**

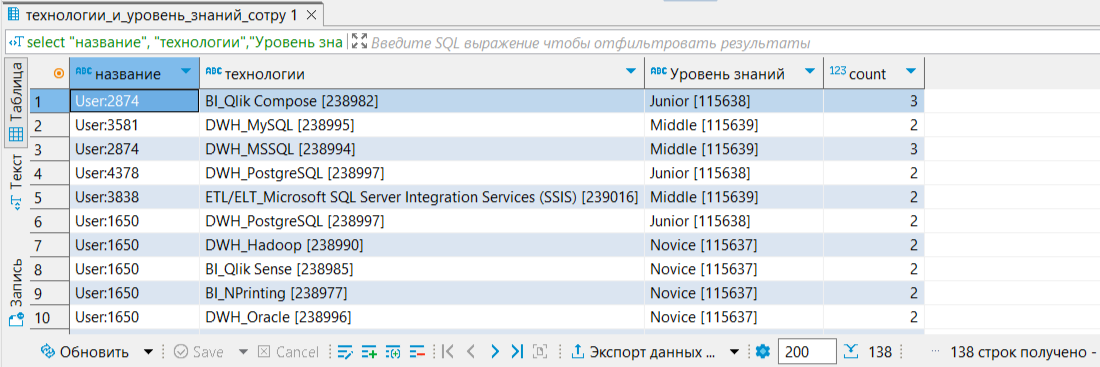
1. Дубликаты:

**select** **"название"**, **"технологии"**,**"Уровень знаний"**, **count** (\*)

**from** source\_data.технологии\_и\_уровень\_знаний\_сотру x

**group** **by** "название", "технологии","Уровень знаний"

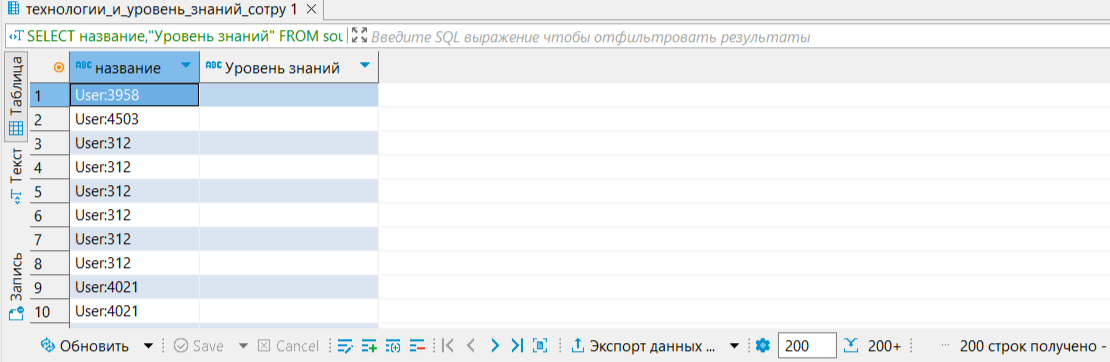
**having** **count**(\*) >1



1. Отсутствующий уровень знаний:

SELECT название,"Уровень знаний" FROM source\_data.технологии\_и\_уровень\_знаний\_сотру x

where "Уровень знаний" = ''

****

## **18 Таблица типы\_систем\_и\_уровень\_знаний\_сотру**

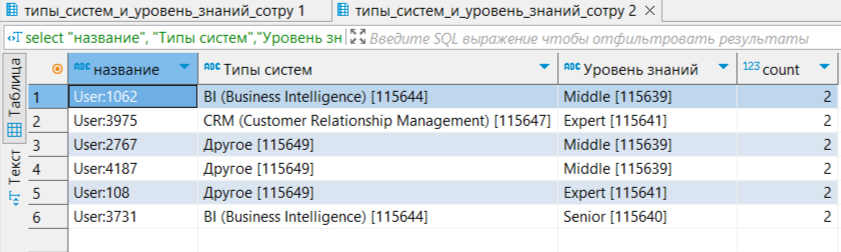
1. Дубликаты:

**select** **"название"**, **"Типы систем"**,**"Уровень знаний"**, **count** (\*)

**from** source\_data.типы\_систем\_и\_уровень\_знаний\_сотру x

**group** **by** "название","Типы систем","Уровень знаний"

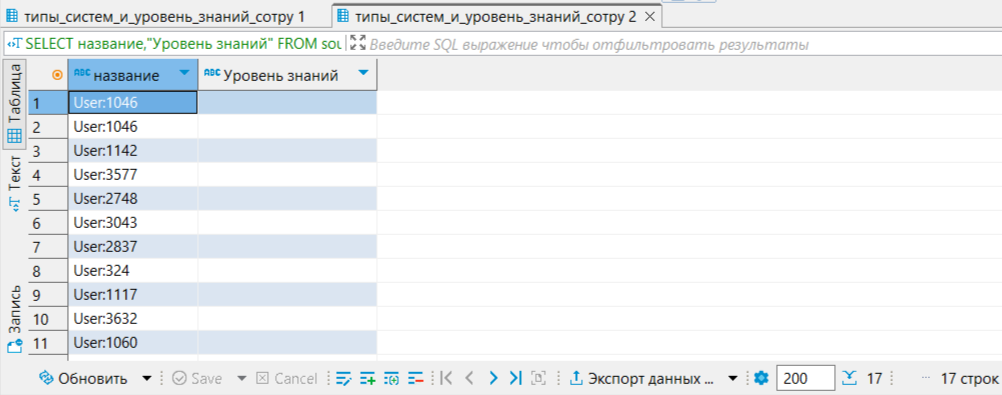
**having** **count**(\*) >1



1. Отсутствие уровня знаний

**SELECT** название,**"Уровень знаний"** **FROM** source\_data.типы\_систем\_и\_уровень\_знаний\_сотру x

**where** **"Уровень знаний"** = **''**

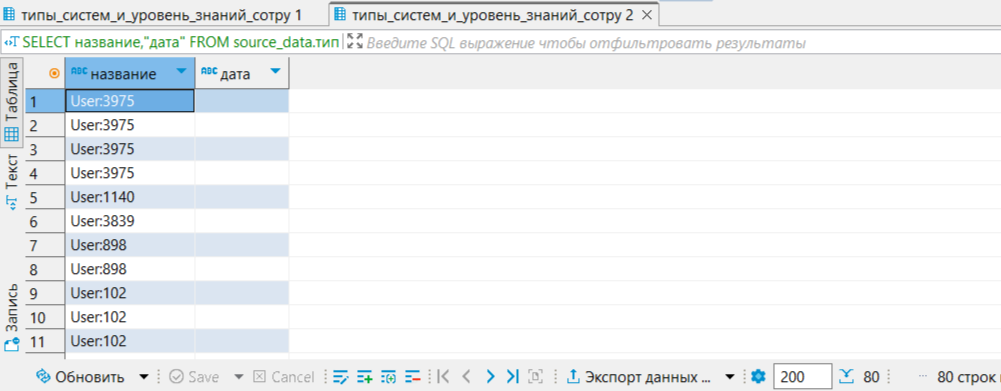


1. Даты

**SELECT** название,**"дата"**

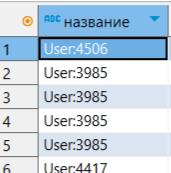
**FROM** source\_data.типы\_систем\_и\_уровень\_знаний\_сотру x

**where** **"дата"** = **''**

****

## **Таблица фреймворки\_и\_уровень\_знаний\_сотру**

1. Переименовать поле «название» в «User ID» типа данных int



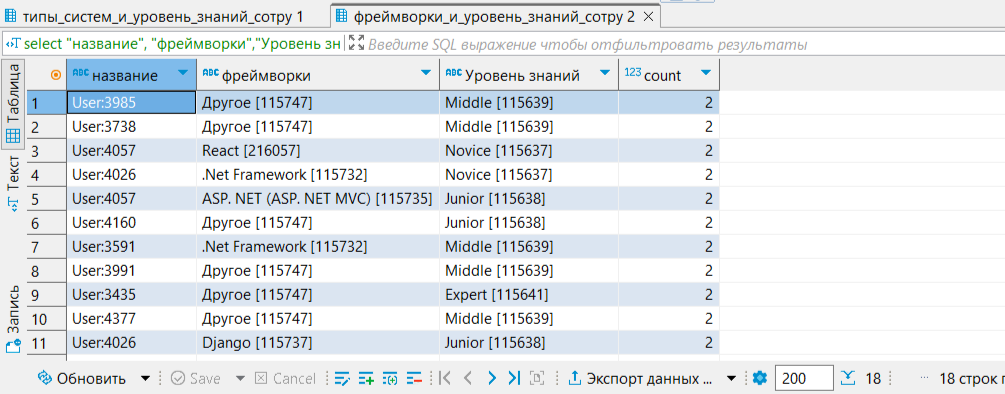
1. Дубликаты:

**select** **"название"**, **"фреймворки"**,**"Уровень знаний"**, **count** (\*)

**from** source\_data.фреймворки\_и\_уровень\_знаний\_сотру x

**group** **by** "название","фреймворки","Уровень знаний"

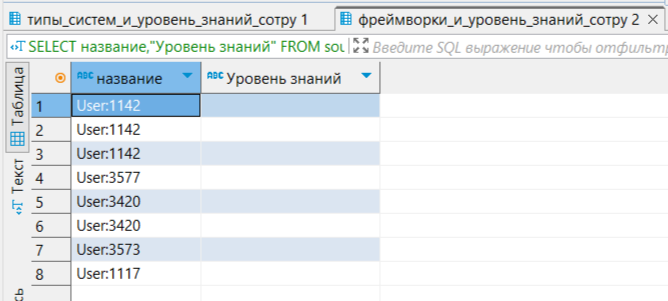
**having** **count**(\*) >1



1. Отсутствующие уровни:

**SELECT** название,**"Уровень знаний"** **FROM** source\_data.фреймворки\_и\_уровень\_знаний\_сотру x

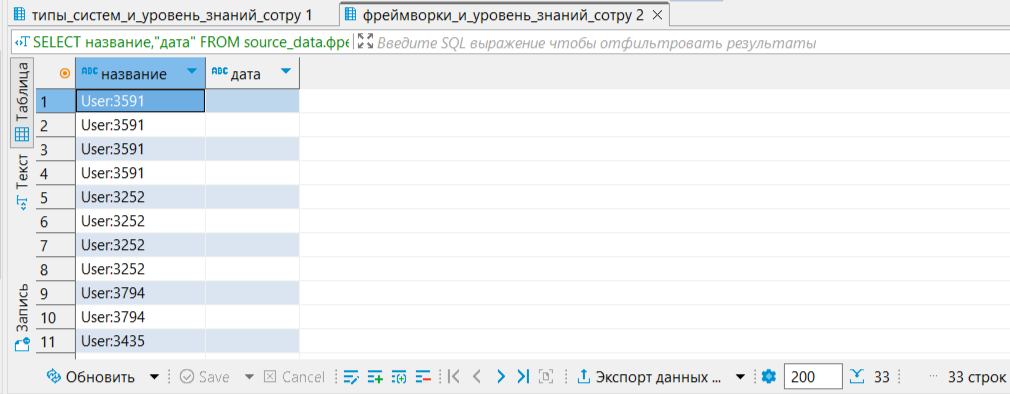
**where** **"Уровень знаний"** = **''**



1. Отсутствующие даты

**SELECT** название,**"дата"** **FROM** source\_data.фреймворки\_и\_уровень\_знаний\_сотру x

**where** **"дата"** = **''**



## **Таблица языки\_пользователей**

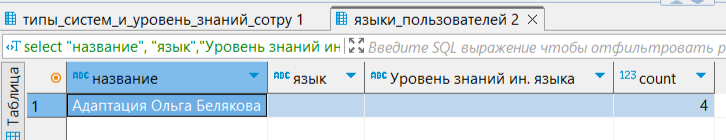
1. Дубликат и имя вместо id

**select** **"название"**, **"язык"**,**"Уровень знаний ин. языка"**, **count** (\*)

**from** source\_data.языки\_пользователей x

**group** **by** "название","язык","Уровень знаний ин. языка"

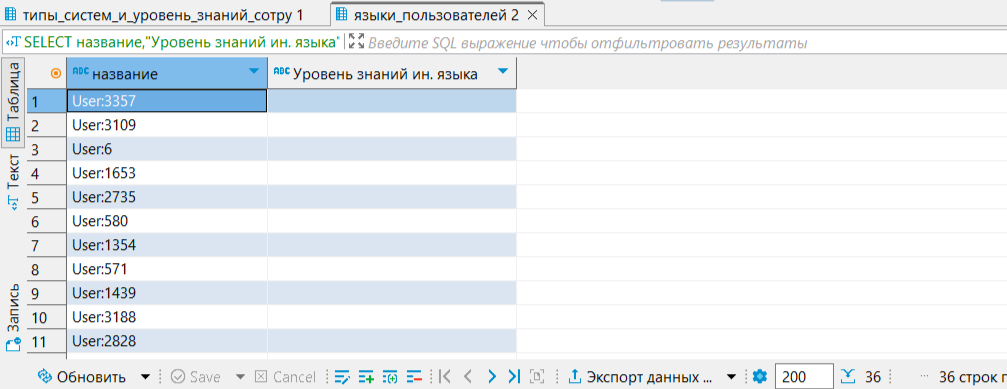
**having** **count**(\*) >1



1. Пустое значение уровня знаний

**SELECT** название,**"Уровень знаний ин. языка"** **FROM** source\_data.языки\_пользователей x

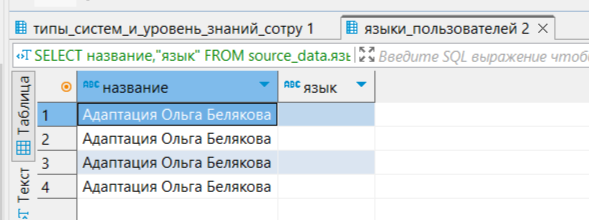
**where** **"Уровень знаний ин. языка"** = **''**



3)Отсутствие данных о языке сотрудника

**SELECT** название,**"язык"** **FROM** source\_data.языки\_пользователей x

**where** **"язык"** = **''**



## **Таблица языки\_программирования\_и\_уровень**

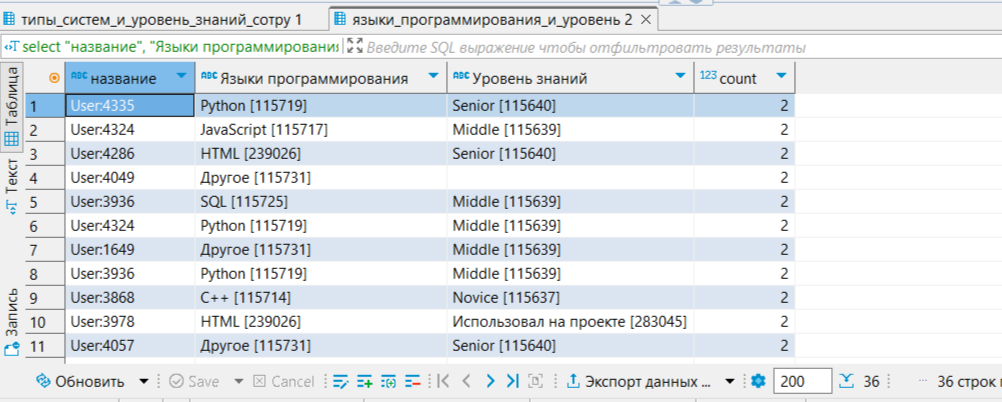
1. Дубликаты

**select** **"название"**, **"Языки программирования"**,**"Уровень знаний"**, **count** (\*)

**from** source\_data.языки\_программирования\_и\_уровень x

**group** **by** "название", "Языки программирования","Уровень знаний"

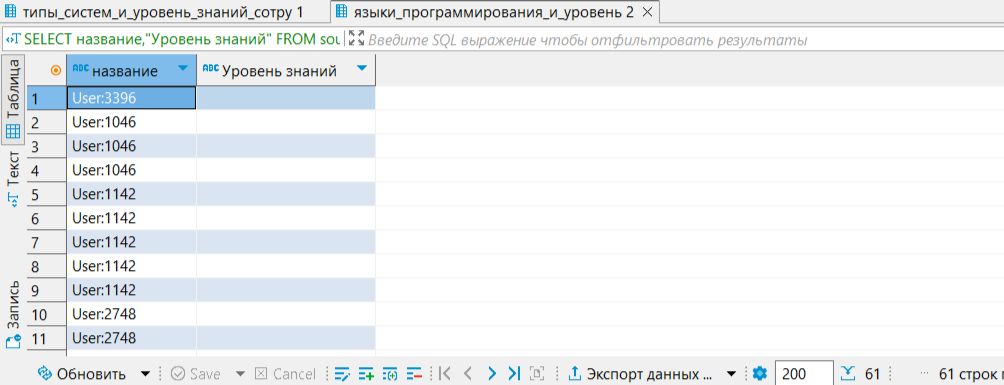
**having** **count**(\*) >1



1. Отсутствие уровней знаний

**SELECT** название,**"Уровень знаний"** **FROM** source\_data.языки\_программирования\_и\_уровень x

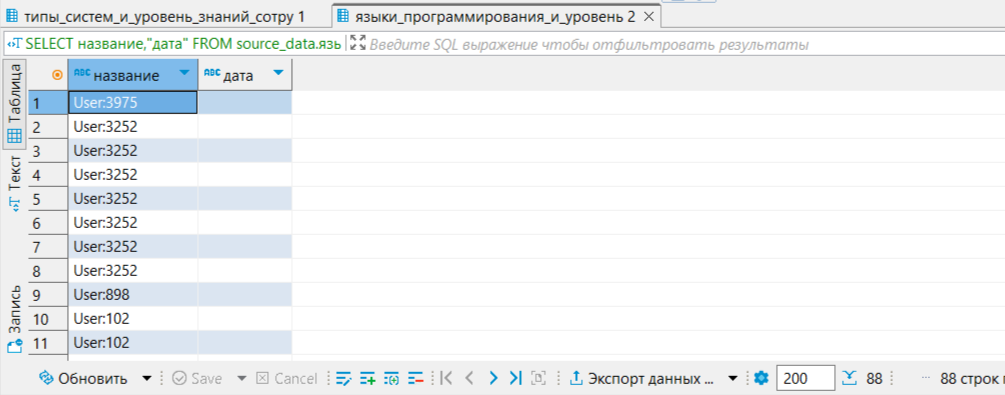
**where** **"Уровень знаний"** = **''**



1. Отсутствие даты

**SELECT** название,**"дата"** **FROM** source\_data.языки\_программирования\_и\_уровень x

**where** **"дата"** = **''**



## **Рекомендации по улучшению качества данных**

1. Много значений отсутствует при этом не обозначены как NULL, их нужно оформить как полагается или удалить при невозмоджности восстановления данных.
2. Сделать селективными все варианты в которых ограничены названия (уровень, навык, платформа) из списка, чтобы пустых значений в последствие не добавлялось
3. Убрать текст из ID
4. Столбцы с полем “название”, где под п=ним понимается id - переименовать в User ID
5. Заменить формат ячеек на соответствующий, где это возможно, а не text (id, date)
6. Даты не раньше условно 1920х и не позже настоящего момента.
7. Маппинг резюмедар и других таблиц
8. Столбцы то со строчной то с заглавной буквы, раз большинство с начинаются с заглавной, то лучше сделать такими все
9. Столбец Сорт. удалить
10. Изменить формат времени на data